SCISPACE
formerly Typeset

Open access • Posted Content • DOI:10.1101/2020.10.28.359240

Whole-genome analysis of Nigerian patients with breast cancer reveals ethnic-driven somatic evolution and distinct genomic subtypes — Source link

Naser Ansari-Pour, Yonglan Zheng, Jason J. Pitt, Stefan C. Dentro ...+22 more authors

Institutions: University of Oxford, University of Chicago, National University of Singapore, Wellcome Trust Sanger Institute ...+5 more institutions

Published on: 28 Oct 2020 - bioRxiv (Cold Spring Harbor Laboratory)

Topics: Somatic evolution in cancer, Population and Breast cancer

Related papers:

- Identification of Distinct Heterogenic Subtypes and Molecular Signatures Associated with African Ancestry in Triple Negative Breast Cancer Using Quantified Genetic Ancestry Models in Admixed Race Populations.
- Admixture Mapping of African-American Women in the AMBER Consortium Identifies New Loci for Breast Cancer and Estrogen-Receptor Subtypes.
- A Genome-Wide Breast Cancer Scan in African Americans
- Abstract 1599: Determining patient ancestry based on targeted tumor comprehensive genomic profiling
- Fine mapping of breast cancer genome-wide association studies loci in women of African ancestry identifies novel susceptibility markers

Share this paper:

View more about this paper here: https://typeset.io/papers/whole-genome-analysis-of-nigerian-patients-with-breast-5b45z1mbbh

Whole-genome analysis of Nigerian patients with breast cancer reveals ethnic-driven somatic evolution and distinct genomic subtypes

Naser Ansari-Pour[1,#], Yonglan Zheng[2,#], Jason J. Pitt[3], Stefan Dentro[4,5], Toshio F. Yoshimatsu[2], Ayodele Sanni[6], Mustapha Ajani[7], Anna Woodard[8], Padma Sheila Rajagopal[2], Dominic Fitzgerald[9], Andreas J. Gruber[1,10], Abayomi Odetunde[11], Abiodun Popoola[12], Adeyinka G. Falusi[11], Chinedum Peace Babalola[13], Temidayo Ogundiran[14], John Obafunwa[6], Oladosu Ojengbede[15], Nasiru Ibrahim[16], Jordi Barretina[17], Peter Van Loo[18,19], Mengjie Chen[2,20], Kevin P. White[21], Dezheng Huo[22], David C. Wedge[1,10,*], Olufunmilayo I. Olopade[2,*]

[1]Big Data Institute, Nuffield Department of Medicine, University of Oxford, Oxford, OX3 7LF, United Kingdom

[2]Center for Clinical Cancer Genetics and Global Health, Department of Medicine, The University of Chicago, Chicago, IL 60637, United States

[3]Cancer Science Institute of Singapore, National University of Singapore, Singapore, 117599, Singapore

[4]European Molecular Biology Laboratory, European Bioinformatics Institute, Cambridge, CB10 1SD, United Kingdom

[5]Wellcome Trust Sanger Institute, Hinxton, Cambridge, CB10 1SA, United Kingdom

[6]Department of Pathology and Forensic Medicine, Lagos State University Teaching Hospital, Ikeja, Lagos, Nigeria

[7]Department of Pathology, University of Ibadan, Ibadan, Oyo, Nigeria

[8]Department of Computer Science, The University of Chicago, Chicago, IL 60637, United States

[9]Institute for Genomics and Systems Biology, University of Chicago, Chicago, IL 60637, United States

[10]Manchester Cancer Research Centre, University of Manchester, Manchester, M20 4GJ, UK

[11]Institute for Advanced Medical Research and Training, College of Medicine, University of Ibadan, Ibadan, Oyo, Nigeria

[12]Oncology Unit, Department of Radiology, Lagos State University, Ikeja, Lagos, Nigeria

[13]Department of Pharmaceutical Chemistry, Faculty of Pharmacy, University of Ibadan, Ibadan, Oyo, Nigeria

[14]Department of Surgery, University College Hospital, Ibadan, Oyo, Nigeria

[15]Centre for Population and Reproductive Health, College of Medicine, University of Ibadan, Ibadan, Oyo, Nigeria

[16]Department of Surgery, Lagos State University Teaching Hospital, Ikeja, Lagos, Nigeria

[17]Girona Biomedical Research Institute (IDIBGI), Hospital Universitari de Girona Dr Josep Trueta, Girona, Spain.

[18]The Francis Crick Institute, London, NW1 1AT, United Kingdom

[19]Department of Human Genetics, University of Leuven, 3000 Leuven, Belgium

[20]Department of Human Genetics, The University of Chicago, Chicago, IL 60637, United States

[21]Tempus Labs Inc., Chicago, IL 60654, United States

[22]Department of Public Health Sciences, The University of Chicago, Chicago, IL 60637, United States

[#]These authors contributed equally

[*]Corresponding authors

Abstract

Black women of African ancestry experience more aggressive breast cancer with higher mortality rates than White women of European ancestry. Although inter-ethnic germline variation is known, differential somatic evolution has not been investigated in detail. Analysis of deep whole genomes of 97 breast tumors, with RNA-seq in a subset, from indigenous African patients in Nigeria in comparison to The Cancer Genome Atlas (n=76) revealed a higher rate of genomic instability and increased intra-tumoral heterogeneity as well as a unique genomic subtype defined by early clonal *GATA3* mutations and a 10.5-year younger age at diagnosis. We also found evidence for non-coding mutations in two novel drivers (*ZNF217* and *SYPL1*) and a novel INDEL signature strongly associated with African ancestry proportion. This comprehensive analysis of an understudied population underscores the need to incorporate diversity of genomes as a key parameter in fundamental research with potential to tailor clinical intervention and promote equity in precision oncology care.

Introduction

Black women of African ancestry worldwide face breast cancer at younger ages, present with more advanced disease at diagnosis, experience more clinically aggressive disease and suffer higher mortality relative to women of other ancestries[1,2]. While socioeconomic and structural barriers explain some of this disparity, women of African ancestry also experience higher rates of estrogen receptor-negative (ER-) and progesterone receptor-negative (PR-) [hormone receptor-negative, HR-] or human epidermal growth factor receptor 2 (*ERBB2*)-amplified [HER2+] subtypes of breast cancer[3-5]. At least 40 percent of this subtype distribution is estimated to be genetic in origin[6].

Studies of breast cancer genomes reveal population-specific differences in germline predisposition mutation frequency, somatic mutation landscapes and mutational signatures, mirroring population differences in molecular subtypes[7-10]. Breast cancer patients of African ancestry demonstrate more *TP53* alterations and fewer *PIK3CA* alterations[6,8], and Nigerian HR+/HER2- tumors are characterized by increased homologous recombination deficiency (HRD) signature[8]. Tumors from patients of African ancestry have also previously been shown to demonstrate increased intra-tumor heterogeneity (ITH)[11].

Whole genome sequencing (WGS) with paired germline tissue can be used to reconstruct the evolutionary “life history” of breast tumors, providing a detailed roadmap for early or late clonal and subclonal genomic events that help prioritize therapeutic targets[12-14]. To date, however, the evolutionary and clonal structure of breast cancers have only been derived using tumors predominantly of non-African ancestry. We hypothesized that studying the evolutionary trajectory of tumors from indigenous African women would provide insight into population-specific genomic features relevant to the breast cancer burden in previously understudied populations.

High-depth WGS was performed on 100 breast tumors (90x depth; of which 49 had complementary RNA-seq) and paired normal tissue (30x depth) from women with breast cancer from Nigeria as previously described[8]. Key events in the somatic evolution of these tumors were identified, and compared with similar analysis of WGS from 76 breast cancer cases from The Cancer Genome Atlas (TCGA). This study is the first to perform life history analysis on an indigenous Black African population, underscoring the critical paucity of genomic data previously available from breast cancer patients of African ancestry[15,16].

Results

Three samples from Nigeria were excluded due to low purity estimates (<10%) resulting in a final set of 173 samples comprising Nigerian Black (Nigerian for short, n=97), White TCGA (White for short, n=46) and Black TCGA (Black for short, n=30) groups (Supplementary Table S1). The ancestry of breast cancer patients from TCGA was estimated as previously described[6].

Nigerian-specific molecular features

Somatic mutation and drivers

We observed a higher insertion and deletion (indel) burden in the Nigerian group compared with the White and Black groups ($P=6.5\times10^{-5}$ and $P=2\times10^{-4}$ respectively), which remained significant after adjusting for clinical subtype. However, the single nucleotide variant (SNV) rate did not significantly differ between races/ethnicities. Somatic coding drivers were identified with cDriver[17] (recurrence≥2%, false discovery rate [FDR]<0.01) and MutSigCV[18] (FDR<0.05) independently. In total, thirteen driver genes were identified (Table 1). Using the 20/20 principle[19], driver genes were classified into oncogenes (ONC) and tumor-suppressor genes (TSG). *GATA3* showed the strongest TSG signal and, of the five novel genes detected, three showed a TSG signal (Supplementary Fig. S1). Most drivers occurred clonally, i.e. in all tumor cells (Supplementary Fig. S2). However, *BCLAF1*, a transcription regulator involved in DNA damage response[20] which also displayed a strong TSG score, was found to occur predominantly subclonally. Based on the union set of previously identified breast cancer drivers[8,21] and those identified here, 30 were identified in our samples, and 93.6% of all samples were mutated in at least one known driver (Fig. 1).

Driver enrichment analysis identified *GATA3* as the only driver significantly enriched in the Nigerian group (FDR=0.038, odds ratio [OR]=6.3, 95% confidence interval [CI] 1.8-34.3). Subtype stratification identified *LAMB3* enriched in HER2+ tumors (OR=13.7), although lacking significance following multiple testing correction (FDR=0.15). Interestingly, *LAMB3* occurred only in Nigerian HR-/HER2+ patients. *TP53* was enriched in ER- patients (FDR=0.0021, OR=3.8, 95% CI 1.9-7.8). Interestingly, although *GATA3* has been reported to be strongly enriched and unique to ER+ tumors[22], we did not observe such an enrichment (OR=2.7, FDR=0.17) in the Nigerian group, which included ten *GATA3*-mutant, ER- tumors.

We identified hotspots for non-coding mutations by comparing the Nigerian, Black and White groups. Two regions across the genome showed significant differences (FDR<0.1) in mutation rates between the Nigerian and White groups (Fig. 2), both over-represented in the Nigerian group. No significant differences were identified between the Nigerian and Black groups. The strongest signal (42.3% versus 4.3%, FDR=0.037) was found at 20q13.2 where mutations clustered immediately upstream of *ZNF217*, a gene encoding a transcription factor which is a key regulator of tumorigenesis[23] and previously associated with clinical outcomes in breast cancer[24]. The second hotspot (28.9% versus 0%, FDR=0.097) was found at 7q22.3 within and flanking *SYPL1* (Synaptophysin-like 1). Although there is no evidence of its association with breast cancer, this gene has been previously associated with clinical outcomes in hepatocellular carcinoma[25] and pancreatic ductal adenocarcinoma[26]. Interestingly, we saw a significantly positive cline in the prevalence of mutations in both genes from White to Black to Nigerian groups (proportion trend test; $P=3.4\times10^{-6}$ for *ZNF217* and 3.3×10^{-4} for *SYPL1*), suggesting an association with African ancestry.

Mutational Signatures

Somatic mutational signatures may provide an etiological explanation for both exogenous and endogenous risk factors of breast cancer. Mutational signature analysis identified 13 single-base substitution (SBS) COSMIC signatures (Fig. 3A-B). Those observed in >5% of samples have been previously reported with a similar order of prevalence[22], with the exception of SBS39 which is a recently detected signature[27]. The rare SBS signatures (SBS17a/b, SBS7b, SBS15 and SBS28) were observed primarily in Nigerian and Black groups with mean prevalence of these signatures at 2.8% and 3.3% respectively. The HRD signature SBS3 was observed in all groups. However, compared with the White group, the Black (OR=2.74, P=0.048) and Nigerian (OR=1.87, P=0.13) groups had slightly higher activity. Nine double-base substitution (DBS) signatures were also identified, of which five were novel (Supplementary Fig. S3). Although DBS-B was observed in similar frequencies across the groups (OR~1.3, P>0.58), DBS-B showed higher activity in Nigerian than in the White (1.55-fold, P=0.0035) and Black (1.42-fold, P=0.018) groups.

Twelve INDEL signatures were detected, of which ID8 and INDEL-B were the most frequent. Of note, novel signature INDEL-B was not only significantly depleted in the White group compared with the Nigerian ($P=1.1\times10^{-18}$) and the Black groups ($P=1.2\times10^{-4}$), but it also showed a clear positive cline from White to Black to Nigerian groups both in prevalence (proportion trend test $P=3.9\times10^{-18}$; Fig. 3C) and activity (proportion trend test $P=2.6\times10^{-6}$; Fig. 3D). Moreover, unlike common INDEL signatures such as ID6 which are driven by short deletions, this signature comprises short insertions (Supplementary Fig. S4). Although the etiology of this signature remains to be elucidated, the data suggest a strong association with recent African ancestry. This association was not observed for any other INDEL signature. Notably, the indel burden in Nigerians was bimodally distributed (Supplementary Fig. S5), suggesting greater activity in a subset of patients. Assessment of the high burden samples identified ID2, ID4 and ID6 as the dominant signatures, all of which showed at least 2-fold higher mean activity than in low burden samples.

A comparison of hormone subtypes (Supplementary Fig. S6) revealed that SBS3, ID6 and ID8 were more prominent in the HR-/HER2- subtype while INDEL-B and DBS11 had highest activity in the HER2+ subtype.

The HRD signature (SBS3) was detected in 4/7 and 7/8 germline and somatic *BRCA*-positive tumors respectively. While four samples with either germline or somatic *BRCA* variants lacked SBS3, all four samples exhibited high activity of INDEL signatures ID6 and ID8 (Supplementary Table S2 and Supplementary Fig. S7), both of which are associated with DSB repair by non-homologous end-joining[27]. It seems likely that, due to the high similarity of ‘flat’ SBS signatures

(Supplementary Fig. S8), SBS3 activity may have been mis-assigned mainly to SBS39 during signature deconvolution, an interpretation supported by the mutual exclusivity of these two signatures in these samples (Supplementary Fig. S7).

Copy Number Aberrations (CNA)

The CNA landscape of the Nigerian group (Fig. 4A) was very similar to that of the Black group with all enriched CNA in the Nigerian group also observed in the Black group. We therefore compared the CNA landscape only with the White group. The key enriched CNA events unique to the Nigerian group were 5p15.33-13.3 Gain, 7p22.1-14.2 Gain, 17p13.3 Gain and 14q LOH. We further analyzed these enriched CNA at the clinical subtype level and found that clonal 14q LOH was highly enriched in the Nigerians in the HR-/HER2+ subtype (0.58 versus 0.07, 8.6-fold, $P=7\times10^{-4}$; Fig. 4b) even though the proportion of this subtype was comparable between the two groups (43.8% and 33.3% respectively, P=0.358). 14q LOH enrichment in the Nigerian group is an interesting observation since it is known to be associated with aggressive breast cancer progression[28,29]. Its effect may be particularly exacerbated in Nigerian patients as HR-/HER2+ has been reported to be enriched within younger Nigerian patients for reasons that are poorly understood[8]. 14q LOH has been reported in *BRCA2* mutation carriers using array-CGH[30,31]. We also observed 14q LOH in all *BRCA2* carriers, however the presence of 14q LOH in *BRCA2*-negative tumors suggests that this CNA event is present more widely in breast cancer but has been understudied.

Genomic Instability (GI)

GI is a known hallmark of cancer which manifests as WGD, chromosomal instability (CIN) and kataegis. We observed a 3-fold higher rate of WGD in Nigerians compared with the White group (FDR=0.02) but no significant difference was observed between either group and the Black group. Interestingly, we observed a significant positive trend in WGD rate from White to Black to Nigerian groups (proportion trend test, P=0.004)

The proportion of the genome altered (PGA) by CNAs was calculated for all samples as a measure of CIN. We observed a 2.2-fold higher PGA in WGD tumors compared with non-WGD tumors ($P=3.7\times10^{-17}$), which was consistent across all groups and subtypes (FDR<0.1; Supplementary Fig. S9). PGA was significantly correlated with mutation burden (R=0.47, $P=9.3\times10^{-11}$). A modified metric, PGAn that took account of both the length and number of CNA segments was found to correlate more strongly with mutation burden (R=0.55, 7×10^{-15}; Supplementary Fig. S10), suggesting that PGAn may be a superior measure of CIN than PGA.

We observed kataegis in 64.2% (111/173) of samples (Supplementary Table S3), with 3.6% of these, all Nigerian, harboring more than ten kataegis events (Supplementary Fig. S11). The majority of SNVs at kataegis foci were C>T and C>G mutations, associated with APOBEC mutational signatures SBS2 and SBS13. At the group level, the Nigerian group exhibited a higher number of foci (Supplementary Fig. S12) compared with the White (2.1-fold, $P=6.4\times10^{-4}$) and Black groups (2.8-fold, P=0.002), but was significantly higher after adjusting for subtype and multiple-testing only than the Black group. No association was observed between the number of foci and subtype or age at diagnosis.

Chronological ordering of genomic aberrations

The Plackett-Luce probabilistic framework was used to reconstruct the most likely chronological order of genomic aberrations across all tumors. Genomic aberrations included in the analysis were enriched CNAs, WGD and common mutational drivers, all of which were ordered based on clonality. Fig. 5 depicts the relative timing of genomic events for the Nigerian dataset. Mutations in both *GATA3* and *TP53* were found to be early drivers. In addition to known early events such as 8p LOH and 17p LOH, 9q34.2 LOH, 14q LOH, 15q14-q21.3 LOH and 19p13.3 LOH were among the early drivers. In contrast, in the White group, 19p13.3 (*STK11*) LOH did not occur pre-WGD (Supplementary Fig. S13), but 8p11.21 gain did occur pre-WGD. This gain event and 19p13.3 LOH encompass *ANK1* and *STK11* genes respectively, both of which have been implicated in tumorigenesis[14,32], and copy number loss of *STK11* has been reported in metastatic breast cancer[33]. Ordering just the HR-/HER2+ subtype, 13q LOH, 14q LOH and 8p11.21 gain occurred as early as the known early drivers (8p LOH and 17p LOH), of which 14q LOH is virtually absent in the White HR-/HER2+ group.

ITH analysis

ITH was assessed using weighted cancer cell fraction (wCCF), a metric that incorporates both the number and CCF of subclonal mutations. Significantly higher ITH was observed in cancers from the Nigerian group than the White (Generalised linear model, P=0.005; 3.4% increase) and Black ($P=1.7\times10^{-4}$; 5.7% increase) groups after adjusting for the higher sequencing depth of Nigerian samples. No significant difference was observed between White and Black groups (P=0.13). The five samples with the highest subclonality (wCCF range 0.66-0.73) were all Nigerian. One of these samples had subclonal mutations in known driver genes *MAP2K4* and *RB1*[21], and a second sample in the novel gene *F5*. The remaining three samples carried possible driver mutations in COSMIC tier 1 genes *ATR*, *ATRX* and *KMT2D* respectively. Further, mutations in *PRDM14*, an epigenetic regulator associated with increase in ITH[34], were observed in two of the five samples (Supplementary Fig. S14).

Genomic subtypes

We analyzed the mutational and CNA drivers for potential pairwise interactions and identified the majority to be co-occurrences of CNA events (Fig. 6). However, *GATA3* and *TP53* mutations were found to be almost mutually exclusive ($P=2.46\times10^{-7}$, OR=0.065, 95% CI 0.012–0.24). *TP53* and *GATA3* mutations had similar CCF distributions (Kolmogorov-Smirnov test, P=0.24) and were predominantly clonal (95% and 96% respectively). In addition, the timing model identified both genes as pre-WGD drivers. To assess likely gene function, we combined the CNA and mutation data to assess the rate of double-hits (bi-allelic inactivation) of these two genes. In total, 74 tumors had both clonal LOH and mutation at *TP53* and thus were predicted to have no *TP53* activity. As expected, WGD was significantly enriched in this double-hit subtype ($FDR=5.08\times10^{-5}$; Supplementary Fig. S15), but not the single-hit group. In contrast, of the 23 samples that carried clonal LOH at the *GATA3* locus, none had a *GATA3* mutation (Fisher exact test, P=0.047), suggesting that either biallelic inactivation of *GATA3* is lethal for cells and therefore selected against or that loss of one *GATA3* copy causes haploinsufficiency. *GATA3* gene dosage was not associated with WGD.

The early, clonal, near mutually exclusive occurrence of *TP53* and *GATA3* suggests that they define distinct genomic subtypes of breast cancer, at least in Nigerian patients with breast cancer. We therefore proceeded to further characterize these subtypes in the Nigerian cohort. Interestingly, patients in the *GATA3* mutant subtype were diagnosed an average of 10.5 years younger (42.9 versus 53.4 years, $P=4.8\times10^{-4}$). From chronological ordering of CNA events in each subtype (Supplementary Fig. S16), 5q35.1 gain was observed as an early event only in the *TP53* subtype. In contrast, HD at 9p21.3-p11.2 was an early event in the *GATA3* subtype, albeit with a high variance due to the small number of samples with this event. The latter is an interesting observation since HDs generally appear late in tumor evolution[35,36]. We observed no overall difference in the wCCF distribution of the two subtypes (Kolmogorov-Smirnov test, P=0.8), suggesting no significant difference in ITH patterns. With respect to mutational signatures, we observed statistically significant increases in SBS1, SBS18, ID5 and the novel INDEL-B in the *GATA3* subtype, while SBS8, SBS39, novel DBS-D, ID8 and ID9 signatures were significantly over-represented in the *TP53* subtype (Supplementary Fig. S17). A higher prevalence of kataegis (OR=6.79, P=0.057) was observed in the *GATA3* subtype, which also affected more foci (1.53-fold, P=0.033). In contrast, the *TP53* subtype showed greater GI in general with significantly higher mutation burden (1.84-fold, Wilcoxon test, P=0.007) and WGD (OR=7.28, P=0.004), consistent with previous studies[37,38]. Mean PGA and PGAn were also both higher in the *TP53* subtype ($P=7.6\times10^{-4}$ and P=0.012).

In the Nigerian cohort, 20.6% of tumors (n=20) were not mutated at either *TP53* or *GATA3* (Fig. 7). These samples had lower mutation burden ($P=5\times10^{-4}$), WGD rate (P=0.007), PGA (P=0.0013) and PGAn (P=0.0014) than the *TP53* subtype, and a lower rate (P=0.007) and prevalence (P=0.02) of kataegis foci compared with the *GATA3* subtype. Common clonal coding drivers were *PIK3CA* (25%; 28% and 5% in *TP53* and *GATA3* subtypes respectively) and *RB1* (10%; 3% and 0% in *TP53* and *GATA3* subtypes respectively) while 35% of these tumors had neither clonal mutations in known mutational drivers nor noncoding variants in either *ZNF217* or *SYPL1*. We therefore further explored this subset of 'quiet' genomes. We observed no difference in tumor purity between this subtype and the *GATA3* and *TP53* subtypes (P>0.24; 0.42 versus 0.44 and 0.48 respectively), suggesting that the observation of lower GI in this subtype is not due to lower tumor content. Similar to the *GATA3* subtype, we observed an enrichment of INDEL-B compared with the *TP53* subtype (FDR=0.011) and a lower rate of enriched CNA events (Fig. 7). However, HD of 17p11.2 (*MAP2K3*) was an early event in the quiet tumors present in 15% (3/20) of samples, while absent in the *GATA3* and late occurring in one tumor in the *TP53* subtypes respectively (Supplementary Fig. S16). Comparison of the whole-transcriptomes of a subset (n=49) of samples from the three subtypes did not show a distinct cluster for the quiet genomes (Supplementary Fig. S18). However, this subtype demonstrated significant overexpression of genes (FDR<0.05; Supplementary Fig. S19) previously associated with breast cancers including casein (*CSN1S1*, logFC=7.0, $P_{adj}=0.007$)[39] and Nectin-4 (*PRR4*, logFC=4.1, $P_{adj}=2.74\times10^{-8}$) as well as genes associated with epithelial development (*SPRR2G*/*SPRR2E*, logFC≥5, $P_{adj}<7.34\times10^{-5}$) , mucin production (*MUC7*, logFC=4.7, $P_{adj}<0.0001$)[40], worse chemotherapeutic response (*TUBA3D*, logFC=2.6, $P_{adj}=0.0004$)[41], and metastatic potential (*LOXL4*, logFC=2.6, $P_{adj}=2.43\times10^{-6}$, and *SERPINE2*, logFC=2.4, $P_{adj}=0.005$)[42,43].

Discussion

Previous genomic landscape studies have sketched out evolutionary trajectories of cancers including breast cancer, primarily focused on White patients of European ancestry ascertained in the US, Canada and Europe[12,14]. Similar studies in other racial/ethnic groups and geographic locations have lagged far behind, despite the reported higher mortality rates of Black patients with breast cancer in the US and sub-Saharan Africa. Here, we used deep WGS to characterize the genomic landscape of somatic events and reconstruct the chronological ordering of events in breast tumors from 97 indigenous Nigerian women and compared the findings with tumors from White and Black patients in TCGA.

We observed key differences in the somatic events occurring during the evolution of breast cancer in our Nigerian cohort, but the clinical implications are impossible to address due to inadequate access to quality cancer care in Nigeria. Nonetheless, we were able to gain improved understanding of breast cancer heterogeneity across populations. First, genomic instability, a hallmark of cancer[44,45], was observed at a higher rate in the Nigerians in the form of WGD, PGA and kataegis, all of which may provide raw material for aggressive tumor behavior. Second, a higher level of ITH was observed in the Nigerian group. Given that ITH can serve as an indicator of tumor fitness for evolutionary adaptation and impinge upon the efficacy of therapeutic treatments[46], this may in part explain the biologically aggressive behavior of these tumors and poor clinical outcome in an underscreened population. Third, key somatic events were enriched in the HR-/HER2+ subtype, which may point to etiology but could also potentially impact response to HER2-targeted therapies. Of note, 14q LOH, which encompasses breast cancer genes such as *SERPINA1* and *DICER1*, was highly enriched in HR-/HER2+ Nigerian women. Loss of the former has been shown to be associated with poor outcome in this subtype[29] and that of the latter is associated with tumor progression and recurrence in this subtype[28]. Other enriched events include *LAMB3* (regulator of the PI3K/Akt signaling pathway in multiple cancers[47] and the novel INDEL-B (unknown etiology).

Epidemiological studies have shown a younger age of onset in women of African ancestry[48]. The significant enrichment of the early clonal driver *GATA3* in the Nigerian group and a positive trend in its recurrence with African ancestry (proportion trend test, P=0.0035) along with a significantly lower age at diagnosis in patients with tumors carrying *GATA3* mutations is likely to be an underlying genomic event associated with young onset breast cancer. Furthermore, non-coding mutation hotspots at *ZNF217* and *SYPL1*, which are both associated with poor outcomes, and the novel INDEL-B showed a strong positive trend with African ancestry, suggesting that these genomic features may also be associated with different evolutionary patterns of breast cancer in Nigeria.

Substantial progress has been made in unravelling the genomic complexity of breast cancer[12,22] one key improvement being the development of genomic classification of breast cancer[49,50]. In our cohort, we identified three genomic subtypes of which the *GATA3* subtype was strongly enriched in the Nigerian group. These genomic subtypes presented distinct mutational properties. In the *TP53* subtype, all of which were double-hit *TP53* tumors, mutation burden was higher and WGD was significantly enriched. A recent study has shown that the loss of both copies of *TP53* drives the poor outcome of patients with myelodysplastic syndromes, and different evolutionary trajectories were evident between single-hit and double-hit tumors[51]. On the contrary, kataegis was more frequent in the *GATA3* subgroup. The quiet genome subtype displayed low genomic instability and demonstrated associations to genes previously incorporated across breast cancer subtyping, prediction and prognostication approaches to date without a clearly consistent, previously-described pattern. That a large proportion of the tumors had no known clonal driver and that a number of breast cancer related genes were highly upregulated, suggests that tumor evolution in this subtype is complex and remains to be elucidated.

We acknowledge that the current sample size was modest, however, to the best of our knowledge, this is the largest breast cancer life history study on non-White patients to date. While starting to redress the imbalance with larger European cohorts, it will also supplement existing studies of tumor evolution in breast cancer. In addition, the Southwest Nigerian breast cancer patients included in the present study cannot fully represent all the Nigerian ethnic groups that are known to vary in environmental, cultural, and socioeconomic indices. Nigerian and coastal West African populations contributed a significantly large proportion of genetic makeup of African Americans and African Caribbeans[52]. Thus, future investigations are warranted to integrate germline and somatic genetics, as well as epigenetic and environmental factors, to extend our understanding of the dynamic nature of breast tumor evolutionary trajectories in diverse populations of African ancestry. Also, exploring the complexity of genomic classifications and neoplastic progression in terms of treatment using matched primary and metastatic tumors from individual patients will shed light on rational therapeutic strategies to close the mortality gap. The inherent challenge in treating breast cancer effectively among Black women across the African Diaspora is the pervasive institutional racism that has led to inequality of access to quality cancer care and under-representation of Black patients in genomic studies[53,54]. Our findings underscore the need to expand access to innovative and life-saving biomarker-informed clinical trials in Nigeria and across sub-Saharan Africa to accelerate progress in precision oncology and promote equity[55]. For instance, enhanced by modern genomics and international partnerships, the ARETTA trial (ClinicalTrials.gov identifier: NCT03879577) provides a blueprint for implementing evidence-based cancer treatment in low-resource settings[56].

Methods

Patient cohort, ethics, and pathological assessment. This study was embedded within the Nigerian Breast Cancer Study (NBCS) and approved by the Institutional Review Board of all participating institutions[1,7,9,57]. A grand total of 493 subjects were recruited from University College Hospital, Ibadan (UCH; n=284) and Lagos State University Teaching Hospital (LASUTH; n=209) between February 2013 and September 2015. Each patient gave written informed consent before participation in the study. Six biopsy cores and peripheral blood were collected from each patient. Two biopsy cores were used for routine formalin fixation for clinical diagnosis and the remaining four cores were preserved in PAXgene Tissue containers (Qiagen, CA) for subsequent genomic material extraction. In addition, 27 mastectomy tissues were preserved in RNAlater. Complete pathology assessment was performed centrally by study pathologists. Tumor burden was assessed based on cellularity, histology type, and morphological quality of tissue using TCGA best practices[49]. IHC on ER, PR, and HER2 were performed centrally in Nigeria and further reviewed in the United States. Cases with discordant results were again reviewed and resolved by the study pathologists. IHC scoring variables for Allred scoring algorithm were captured according to the 2013 ASCO/CAP standard reporting guidelines. Briefly, for ER and PR testing, immunoreactive tumor cells <1% was recorded as negative and those with ≥1% were reported positive. All the positive ER and PR cases were graded in percentages of stained cells and further scored in line with the Allred scoring system. Percentage of tumor staining for HER2 test were also reported along with a score of 0 and 1+ as negative, 2+ as equivocal, and 3+ as positive case. HER2 equivocal cases were further confirmed with genomic copy number calls.

Sample selection and genomic material preparation. Tumor samples containing >60% tumor cellularity were selected for DNA extraction using PAXgene Tissue DNA kit (Qiagen). Gentra Puregene Blood Kit (Qiagen) was used to extract genomic DNA from blood. Extracted DNA was quality controlled for its purity, quantity, and integrity. Identity of each extracted DNA sample was tested using AmpFlSTR Identifiler PCR Amplification Kit (Thermo Fisher Scientific). Samples that match >80% of the short tandem repeat profiles between tumor and germline DNA were considered authentic. RNA was extracted from PAXgene fixed tissues using the PAXgene Tissue RNA kit (Qiagen). RNA integrity (RIN) was determined for all samples by the RIN score given by the TapeStation (Agilent) read out. RNA samples that had RIN scores of 4 and above were included in downstream sequencing analysis.

Next-generation sequencing data generation. A total of 100 WGS were performed at the University of Chicago High-throughput Genome Analysis Core (HGAC) and at the New York Genome Center (NYGC). Libraries were prepared using the Illumina Truseq DNA PCR-free Library Preparation Kit and were sequenced on an Illumina HiSeq 2000 sequencer at HGAC using 2×100bp paired-end format and HiSeq× sequencer (v2.5 chemistry) at NYGC using 2×150 bp cycles. Mean coverage depth tumor was at 103.2× and normal was at 35.1×. A total of 103 RNA-seq were carried out at the Novartis Next Generation Diagnostics facility. Average number of mapped reads per sample was 97 million. Seven samples failed QC and were excluded. Among the remaining 96 samples, 49 have WGS data available from the same patients. Total RNA were constructed into poly-A selected Illumina-compatible cDNA libraries using the Illumina TruSeq RNA Sample Prep kit. Passing cDNA libraries were combined in equimolar pools with other libraries of compatible adapter barcodes and later sequenced on the Illumina HiSeq 2500 sequencer.

Alignment of DNA sequence to reference genome. WGS reads were aligned to GRCh37 from GATK data bundle (v2.8; https://software.broadinstitute.org/gatk/) using BWA-MEM (v0.7.12; http://bio-bwa.sourceforge.net/). Duplicate reads were removed using PicardTools MarkDuplicates (v1.119; https://broadinstitute.github.io/picard/).

Calling germline SNVs and indels. Reads from WGS were aligned to GRCh37 from GATK data bundle (v2.8; https://software.broadinstitute.org/gatk/). Duplicate reads were removed using PicardTools MarkDuplicates (v1.119; https://broadinstitute.github.io/picard/). Both SNVs and indels were called using Platypus (v0.7.9.1; https://github.com/andyrimmer/Platypus) in single-sample mode. Only variants passing the Platypus 'PASS' filter were considered for downstream analysis.

Calling somatic SNVs and indels. SNVs were called using both MuTect (v1.1.7; https://software.broadinstitute.org/cancer/cga/mutect) and Strelka (v1.0.13; ftp://strelka:@ftp.illumina.com/v1-branch/v1.0.13/) with default parameters using the SwiftSeq workflow (https://github.com/PittGenomics/SwiftSeq) as previously described[8]. Variants were called on the entirety of the genome in order to detect and retain any high-quality off-target calls. Any variant call that did not meet 'PASS' criteria for either algorithm was discarded. For a given tumor-normal pair, only SNVs called by both MuTect and Strelka were retained. Furthermore, we constructed a panel of 1088 Nigerian and TCGA normal samples[8]. For a given normal sample, a site needed to be covered by a minimum of ten reads to be included. Any SNV that was supported by 5% or more of reads (MAPQ [MAPping Quality] ≥20; Base quality ≥20) in two or more samples was removed. SNVs were later annotated with Oncotator (v1.5.3.0; https://software.broadinstitute.org/cancer/cga/oncotator) and those that met the required criteria ("COSMIC_n_overlapping_mutation >1" AND "1000gp3_AF ≤0.005" AND "ExAC_AF ≤0.005") were considered likely to be somatic and were retained. Small indels were called using cgpPindel (v3.0.1) within cgpWGS container (v2.0.1; https://dockstore.org/containers/quay.io/wtsicgp/dockstore-cgpwgs:2.0.1?tab=info) with default filters implemented. In addition, any indel calls found in the 1000 Genomes Project Phase 3 release (http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/ALL.wgs.phase3_shapeit2_mvncall_integrated_v5b.20130502.sites.vcf.gz) or the dbSNP (b151; ftp://ftp.ncbi.nih.gov/snp/organisms/human_9606_b151_GRCh37p13/VCF/All_20180423.vcf) were removed, unless they were found in the Catalogue of Somatic Mutations in Cancer (COSMIC v91; https://cancer.sanger.ac.uk/cosmic/download/CosmicCodingMuts.vcf.gz).

Variant annotation. SNVs and indels at both germline and somatic levels were annotated by ANNOVAR[58] (version May2018; http://annovar.openbioinformatics.org/) for functional consequence. In addition, variants were identified based on dbSNP (v150) and population frequency of variants were reported based on the Exome Aggregation Consortium dataset (ExAC v0.3; http://exac.broadinstitute.org/) and the Genome Aggregation Database (gnomAD v2.1.1; https://gnomad.broadinstitute.org/).

Clonality of somatic variants. To measure clonality, we calculated cancer cell fraction (CCF) of each variant by adjusting the variant allele frequency (VAF) for copy number aberration (CNA) status, tumor purity and multiplicity of the variant[59]. Given that VAF of indels are reference-biased in standard variant calling algorithms, we used vafCorrect (https://github.com/cancerit/vafCorrect) to obtain accurate VAF from BAM files directly by leveraging unmapped reads[60]. These re-estimated VAFs were used to calculate accurate CCF for coding indels. To assign coding mutations as clonal or subclonal, the CCF of all SNV and indel coding mutations were statistically assessed for clonal status. Briefly, the observed VAF was modelled using a binomial distribution and values representing the 95% interval were used to generate the 95% confidence interval (CI) of the observed CCF. Any variant with an upper CI above 1 was considered to not deviate from a clonal state and, in turn, was assigned a CCF of 1. Otherwise, variants were considered subclonal and the original CCF value was retained. This allowed us to assess the clonality of coding mutations without introducing an arbitrary CCF cut-off.

Somatic drivers. cDriver[17] (https://github.com/hanasusak/cDriver) was used to identify cancer drivers by not only taking into account the recurrence against the background mutation rate and functional impact (CADD score[61]; https://cadd.gs.washington.edu/), but also the CCF of each variant. In addition, MutSigCV[18] (v1.3; https://software.broadinstitute.org/cancer/cga/mutsig) was used independently to identify drivers based on recurrence given background mutation processes. The 20/20 principle[19] was applied to all detected drivers to classify, based on mutation patterns in the dataset, which are tumor suppressor gene (TSG), oncogene (ONC) or both. Enrichment analysis was undertaken for all mutational drivers (detected and previously known;

n=30) across ethnic groups, clinical subtypes and ER status using Fisher's exact test to identify differential prevalence of drivers. The mutational landscape plot was generated using Maftools[62] (https://www.bioconductor.org/packages/release/bioc/html/maftools.html).

Non-coding hotspots. We partitioned each chromosome in the genome into discrete bins of 100kb and undertook a genome-wide screening of variant recurrence in each of the non-overlapping bins. Similar to a genome-wide association study construct, we compared the Nigerian group with both the White and Black groups using pairwise Fisher's exact test followed by multiple testing correction for differential prevalence to detect potential non-coding mutation hotspots in or near coding genes enriched in the Nigerians. This analysis was based only on SNVs since overall rate of SNV was not significantly different between Nigerian and the other two groups. Therefore, no overall bias is present in the rate of SNVs and local over-representation signals are likely to be genuine ethnicity-specific hotspot signals.

Mutational signatures. De-novo extraction and decomposition to known cosmic mutation signatures in single base substitution (SBS), double base substitution (DBS), as well as small insertion and deletion (ID) formats were implemented based on a non-negative matrix factorization (NNMF) framework using SigProfilerExtractor[27] (v0.0.5.77; https://github.com/AlexandrovLab/SigProfilerExtractor). Because NNMF is more accurate with a larger number of samples[63], we increased our sample set by adding 128 additional breast cancer WGS samples from the Pan-cancer Analysis of Whole Genomes (PCAWG) study[64] and eight TCGA (Asian and unassigned ancestries) WGS samples[8]. Our final input dataset for SigProfilerExtractor thus included a total of 309 samples. Signatures identified as singletons in the entire dataset were removed from analysis.

Calling somatic CNA. Genome-wide copy number profiles of all samples in the entire dataset were obtained by Battenberg (v2.2.8; https://github.com/Wedge-lab/battenberg) which has been described in detail previously[22,36]. In addition to calling clonal and subclonal allele-specific CNAs, it was also used to estimate purity and average ploidy of each tumor. As part of quality control analyses, in the entire dataset, we detected three samples in the Nigerian group which had very low purity estimates based on copy number analysis (<10%). However, one sample did not show a consistent low mutation burden (n_{SNV}=4,285, n_{SNV} of other two samples <100). An SNV-centric VAF-based purity estimation analysis was undertaken to independently evaluate the purity estimate for this sample. Briefly, all SNVs within diploid regions were identified and VAFs were calculated. In the VAF density distribution, the local peak with maximum VAF was considered as the clonal peak. The purity, calculated as 2*clonalVAF, was consistent with that based on the CNA analysis. These samples were thus removed from subsequent analyses in the Nigerian cohort (final n=97). To call recurrent CNA events, first, CNA of each type (i.e. Gain, LOH and HD) were aggregated across all samples along the chromosomes to obtain the frequency landscape of each CNA type based on all observed breakpoints. Next, a permutation test (n=1,000) followed by multiple testing correction was undertaken to identify regions that were significantly enriched above the random background copy change rate. The enriched regions that encompassed the HLA region (6p21), or specific to telomeric ends or present as a singleton were excluded.

Genomic instability analysis. Whole-genome duplication (WGD) was called in samples where the proportion of the genome with balanced 2:2 copy number status was larger than that with 1:1

diploid copy number. Samples were also manually inspected to see WGD features such as multiple copy losses post-WGD (3:1 copy number status) and LOH events with 2:0 status. For the reconstruction of the chronological ordering of somatic events, WGD, as an event, was called in samples that had an average ploidy greater than three[36]. Proportion of genome altered (PGA), which is the proportion of genome bases encompassed by CNAs, was calculated for all samples based on the Battenberg output. In WGD samples, PGA was calculated as the proportion of genome that did not have a balanced tetraploid copy number state (i.e. 2:2). PGA does not take into account the number of CNA events. A modified metric (PGAn) was calculated as the geometric mean of PGA and number of breakpoints to not only take into account length of CNA segments, but also the number of CNA segments to allow for genomes with focal or global shattering. We followed previous studies in defining kataegis events[22,63]. KataegisPCF (https://github.com/nansari-pour/KataegisPCF) was used to detect the kataegis loci and visualize the kataegis events based on SNVs. A minimum of six consecutive SNVs with mean distance ≤1kb were required for kataegis events, which were identified systematically by applying piecewise constant fitting (PCF)[65] on inter-variant distance of all SNVs across the genome.

Timing model of ordering events. To reconstruct the chronological ordering of somatic events, we developed a timing model to order the occurrence of mutational drivers and enriched CNAs based on the clonality of the events. Briefly, for CNAs, Battenberg copy number calls were used to assign clonality (whether CCF=1 or <1) and describe their type (i.e. gain, LOH and HD). CCF of each variant was estimated by adjusting VAF according to the CNA status of the locus and purity of the tumor sample as previously described[59]. Variants were then classified as clonal

(CCF=1) and subclonal (CCF<1). All events were combined per sample and ordered based on CCF. Where more than one tree could be inferred based on subclonal events, all possible trees were generated and randomly chosen in each iteration of ordering events. To time the events based on the entire dataset, events were ordered based on clonality (randomized clonal events followed by a sampled tree of subclonal events) in each sample. To classify events with respect to WGD, we used major/minor copy number status and the estimated number of chromosomes bearing the mutation (NCBM) to call pre-WGD and post-WGD CNA and mutations respectively by using logical rules described previously for CNA[36] and extended them here for mutations. For instance, in a tumor with WGD, a clonal coding mutation with NCBM≥2 was considered as a pre-WGD event while that with NCBM=1 was defined as post-WGD. The Plackett-Luce model[66,67] for ordering partial rankings was implemented using the PlackettLuce package in R (https://github.com/hturner/PlackettLuce) based on the ordering matrix of the entire dataset to infer the order of events at the population level while allowing for unobserved events in individual tumors. This analysis was undertaken for 1,000 iterations to obtain the 95% CI of the timing estimate of each event. In this implementation of the Plackett-Luce model, the clonality level of an event across the population dictates the overall ranking. However, its frequency affects the variance of the timing estimate, such as rarer events show higher 95% CI. We repeated this analysis within each clinical and genomic subtype.

ITH analysis. To infer subclonal architecture of each tumor, a Bayesian Dirichlet process algorithm was implemented (DPClust v2.2.2; https://github.com/Wedge-lab/dpclust), to cluster somatic SNVs based on CCF as previously described[59,68]. Mutation clusters were identified as local peaks in the posterior mutation density obtained from DPClust. In addition to the clonal cluster, the number of subclonal clusters and their respective mutation burden were also estimated. To quantify subclonality, we calculated weighted CCF (wCCF) which is defined by the mean of the CCF of mutation cluster peaks adjusted by the mutation burden of clusters. The ability to detect subclones depends, not on the number of detected SNVs, but on the number of reads per chromosome copy (NRPCC), as previously described[69]. This metric takes tumor purity, ploidy and sequencing coverage into account. We control for this effect by including only tumors with NRPCC≥10. In these tumors, we should be sufficiently powered to detect subclones with CCF>0.3.

Somatic interactions. To test for somatic interactions, we undertook mutual exclusivity and co-occurrence analysis by using pairwise Fisher's exact test to detect significant pairing within and among mutational driver and CNA events. Negative associations with an odds ratio (OR) between 0 and 1 (exclusive) were considered mutually exclusive and positive associations with OR>1 were considered co-occurring with the magnitude of OR being inversely and directly proportional to the strength of the association respectively.

RNA-seq and differential gene expression analysis. RNA-seq genomic material extraction and sequencing data generation with library preparation has been previously described[8]. Read alignment of RNA-seq to GRCh37 (hg19) as reference genome and GENCODE (v19; https://www.gencodegenes.org/human/release_19.html) for gene annotation was performed using STAR (v2.4.2a; https://github.com/alexdobin/STAR) and HTSeq (v0.6.1p1; https://github.com/htseq/htseq). Quality control metrics were calculated using RNA-SeQC (v1.1.8; https://software.broadinstitute.org/cancer/cga/rna-seqc), featureCounts (v1.5.1;

http://subread.sourceforge.net/), PicardTools (v1.128; https://broadinstitute.github.io/picard/), and Samtools (v1.3.1; http://www.htslib.org/). Differential expression analysis of raw read counts of protein-coding genes from HTSeq was then performed using DESeq2 (v1.24.0; https://bioconductor.org/packages/release/bioc/html/DESeq2.html), with subtype information based on immunohistochemistry to maintain consistency with genomic data. Analysis was performed with ancestral populations only to avoid batch effect artifacts.

Statistical methods. All statistical calculations were implemented in R (v3.4.3; https://www.r-project.org/). For categorical data, we used Fisher's exact test (*fisher.test*) and for continuous data, we used wilcoxon rank test (*wilcox.test*) or Student's t-test (*t.test*) wherever appropriate. Where applicable, P-values were adjusted for multiple testing (*p.adjust*) based on the false discovery rate (FDR) proposed by Benjamini and Hochberg[70] with FDR<0.05 considered significant, unless stated otherwise. This was done to not only reduce type I error, but to also minimize type II error[71].

Given the difference in coverage between Nigerian and TCGA WGS samples, to detect true differences in ITH, a generalized linear model (*glm*) was used to model the association of the ITH metric (wCCF) with ethnicity while adjusting for the confounding effect of the covariable NRPCC. The Cochran-Armitage trend test (*prop.trend.test*) was used to assess whether proportions of a variable across the three groups were monotonic with the ordered variable (i.e. increasing African ancestry proportion). The two-sample Kolmogorov-Smirnov test (*ks.test*) was implemented to detect significant differences in the distribution of a variable across different groups.

Data availability

The raw sequencing data from Nigerian cases is available through dbGaP under Study Accession phs001687.v1.p1.

Code availability

The tools used to analyze the study data are provided in the Methods and are publicly available.

Acknowledgements

We are greatly indebted to all the patients who agreed to participate in this study and graciously donated their biological materials. This study was supported by National Institutes of Health (U01 CA161032, R01 MD013452), Susan G. Komen for the Cure (SAC110026) and Breast Cancer Research Foundation (BCRF-19-120). A.J.G. was funded by a postdoctoral research fellowship (grant number: P2BSP3_178591). The computational aspects of this research were funded from the NIHR Oxford BRC with additional support from the Wellcome Trust Core Award (203141/Z/16/Z). We want to thank the New York Genome Center and Novartis Institutes for BioMedical Research (NIBR) for the quality of the sequencing services provided. We want to thank Markus Riester and his colleagues at NIBR, and West Africa Breast Cancer Study investigators for their supports and contributions to the project. We also want to thank Qiagen for their generous donation of PAXgene Tissue Containers and DNA Extraction Kits for this study. The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR or the Department of Health.

Author contributions

O.I.O., D.C.W. and D.H. conceived the study. N.A-P., Y.Z., J.J.P., S.D. and T.F.Y. designed the experimental approach. N.A-P. led the computational analyses, undertook statistical analyses and interpreted results. Y.Z., J.J.P., S.D., T.F.Y., A.W., P.S.R., and A.J.G. conducted bioinformatics analyses and interpreted results. D.F. provided computational support. A.P., A.F., C.P.B., T.O., and N.I. recruited the patients and collected specimens from patients. A.S. and M.A. performed pathological assessment of patient specimens. A.O. procured patient specimens and prepared DNA/RNA. O.O. served as Site-PI and provided overall supervision of the study at UCH. J.O. served as Site-PI and provided overall supervision of the study at LASUTH. J.B. served as Site-PI and provided overall supervision of the study at NIBR. M.C. and D.H. provided statistical support and discussion. P.V.L. and K.P.W. provided supervisory support and discussion. N.A-P., Y.Z. and P.S.R. drafted the manuscript. All authors reviewed and edited the manuscript. O.I.O. and D.C.W. equally supervised this work.

Competing interests

K.P.W. is a Scientific Advisor and Fellow at Tempus. O.I.O is co-founder at CancerIQ and serves as Scientific Advisor at Tempus. All other authors declare no competing interest.

Figure legends

Figure 1: Landscape of driver genes in breast cancer across different ethnic groups.

Genes were identified using two different detection methods (cDriver and MutSigCV; n=13). Breast cancer drivers not detected due to insufficient statistical power (n=173 independent tumors) but frequent in this dataset (≥2%) were also added to the overall list of drivers of breast cancer (n=30) to visualize their distribution in the Nigerian, Black and White groups. Multi-hit: more than one non-silent variant detected in a gene in one tumor.

Figure 2: Manhattan plot for a genome-wide non-coding variant enrichment analysis in the Nigerian group.

The dotted horizontal line represents the genome-wide significance threshold (Benjamini-Hochberg FDR<0.1) with two bins showing significant enrichment of non-coding variants in two novel breast cancer drivers.

Figure 3: Mutational signatures in breast cancer tumors across different ethnic groups.

(A,B) Single base substitution (SBS) signatures in all groups; A) from top to bottom: number of tumors with SBS signatures across the entire dataset (dotted line represents total sample size, n=173) with signatures sorted left to right by descending frequency, number of mutations per sample (color representing groups) in respective signatures and proportion of samples carrying each signature in each group, B) proportion of mutations assigned to each SBS signature across the three groups. (C,D) Identical plots as in (A,B) respectively for insertion/deletion (INDEL) signatures identified in the three groups. INDEL-B is a novel signature characterized mainly by 5+bp insertions.

Figure 4: Copy number landscape of Nigerian breast tumors.

A) Genome-wide landscape of gain, loss of heterozygosity (LOH) and homozygous deletion (HD) events in the Nigerian group. The y-axis represents fraction of tumors with a particular event. LOH and HD are shown in opposite direction for better visualization. B) Differential landscape (Nigerian versus White) of LOH events in HER2+ tumors. Events in the positive direction are more frequent in the Nigerian group. 14q LOH was virtually exclusive to the Nigerian HER2+ subtype.

Figure 5: Chronological ordering of genomic events in Nigerian breast tumors.

Clonality-based ordering of significantly enriched copy number events (FDR<0.05), whole-genome duplication (WGD) and key frequent mutational drivers (*TP53* and *GATA3*). A Plackett-Luce model was used to order the events by sampling from all possible tumor phylogenies across the entire dataset (1,000 iterations). Violins represent the 95% confidence interval of the relative timing estimate for each event. LOH: loss of heterozygosity, HD: homozygous deletion, Mut: mutational driver.

Figure 6: Somatic interaction analysis in breast cancer tumors.

Pairwise associations within and between mutational drivers and significantly enriched copy number aberrations were assessed. Significant associations after correction for multiple testing (FDR<0.05) are shown with positive associations (co-occurrence; OR>1) having a positive sign (red) and negative associations (mutual exclusivity; 0<OR<1) having a negative sign (blue).

Figure 7: Copy number burden across genomic subtypes identified in the Nigerian group.

Distribution of significantly enriched copy number aberrations in the form of gain, loss of heterozygosity (LOH) and homozygous deletion (HD) were assessed for all Nigerian patients. *GATA3*-positive tumors displayed a lower copy number burden than *TP53*-positive tumors but were similar to tumors negative for both *TP53* and *GATA3*.

Table 1: Characteristics of the somatic mutational drivers detected in breast tumors. TSG: tumor-suppressor genes, ONC: oncogenes, COSMIC: Catalogue of Somatic Mutations in Cancer.

Supplementary Table S1: Characteristics of study participants.

Supplementary Table S2: Details of germline and somatic BRCA-positive patients.

Supplementary Table S3: Number of kataegis foci in all breast cancer tumors.

Supplementary Fig. S1: Oncogene and tumor suppressor gene categorization scores for detected mutational drivers. The dashed orange vertical lines represent the threshold scores according to the 20/20 rule.

Supplementary Fig. S2: Cancer cell fraction (CCF) distribution of mutations in detected mutational drivers.

Supplementary Fig. S3: Double base substitution (DBS) signatures in all groups. A) from top to bottom: number of tumors with DBS signatures across the entire dataset (dotted line represents total sample size, n=173) with signatures sorted left to right by descending frequency, number of mutations per sample (color representing groups) in respective signatures and proportion of samples carrying each signature in each group, B) proportion of mutations assigned to each DBS signature across the three groups.

Supplementary Fig. S4: Mutational profile of the novel INDEL-B signature.

Supplementary Fig. S5: Bimodal distribution of indel variant count in Nigerian breast cancer tumors.

Supplementary Fig. S6: Mean activity of SBS, DBS and INDEL mutational signatures across clinical subtypes in breast cancer tumors.

Supplementary Fig. S7: Activity levels of NHEJ, HRD and SBS39 signatures in BRCA-positive tumors. Flat signatures SBS39 and SBS3 (HRD) are observed mutually exclusively.

Supplementary Fig. S8: Mutational profile of signatures SBS3 and SBS39.

Supplementary Fig. S9: Comparison of PGA between WGD and non-WGD samples across groups and clinical subtypes.

Supplementary Fig. S10: Correlation of PGAn and mutation burden.

Supplementary Fig. S11: Kataegis profile of breast cancer tumors with high (n>10) foci rate.

Supplementary Fig. S12: Comparison of kataegic foci rate across the three groups.

Supplementary Fig. S13: Chronological ordering of genomic events in White breast cancer tumors. Clonality-based ordering of significantly enriched copy number events (FDR<0.05), whole-genome duplication (WGD) and key frequent mutational drivers (*TP53* and *GATA3*) based on a Plackett-Luce model. LOH: loss of heterozygosity, HD: homozygous deletion, Mut: mutational driver.

Supplementary Fig. S14: Cancer cell fraction density distribution in tumors with the highest intra-tumoral heterogeneity. The cancer cell fraction of most likely clonal and subclonal drivers in each tumor is shown.

Supplementary Fig. S15: Association of *TP53* status with whole-genome duplication rate.

Supplementary Fig. S16: Chronological ordering of events within genomic subtypes of the Nigerian group. Clonality-based ordering of significantly enriched copy number events (FDR<0.05) and whole-genome duplication (WGD) based on a Plackett-Luce model. LOH: loss of heterozygosity, HD: homozygous deletion.

Supplementary Fig. S17: Differential activity of SBS, DBS and INDEL signatures between *GATA3* and *TP53* genomic subtypes. Asterisks represent significant differences after multiple testing correction (FDR<0.05). Signatures in the positive direction are more frequent in the *GATA3* subtype.

Supplementary Fig. S18: Principal components analysis of 49 Nigerian whole-transcriptomes representing the three genomic subtypes. Dual represents samples positive for both *GATA3* and *TP53*.

Supplementary Fig. S19: Volcano plot of differential expression analysis between quiet-genome tumors and *GATA3*/*TP53* positive tumors in the Nigerian group.

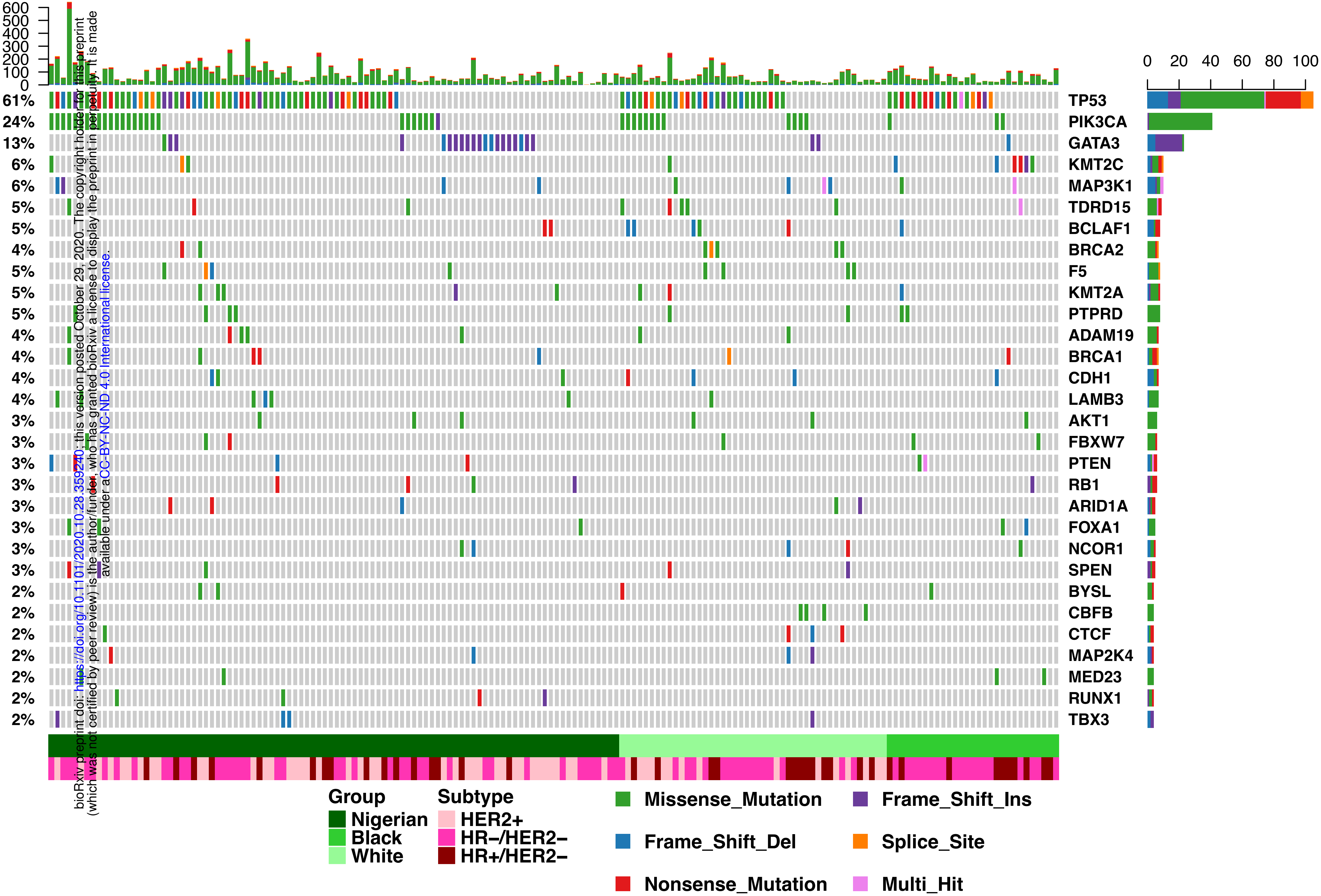
References

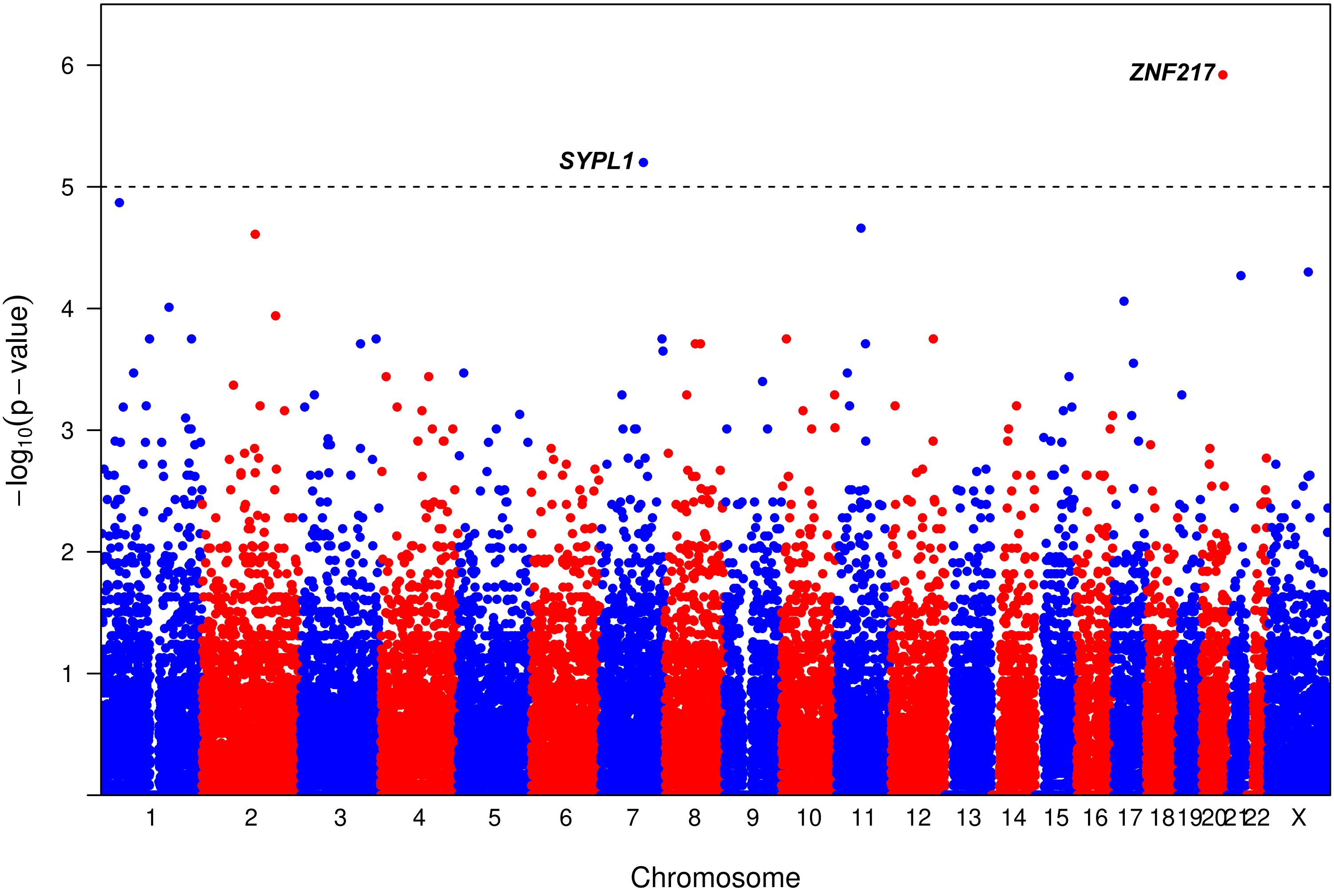
1. Huo, D. *et al.* Population differences in breast cancer: survey in indigenous African women reveals over-representation of triple-negative breast cancer. *J Clin Oncol* **27**, 4515-21 (2009).
2. Wright, N. *et al.* Distinctions in Breast Tumor Recurrence Patterns Post-Therapy among Racially Distinct Populations. *PLoS One* **12**, e0170095 (2017).
3. Kamangar, F., Dores, G.M. & Anderson, W.F. Patterns of cancer incidence, mortality, and prevalence across five continents: defining priorities to reduce cancer disparities in different geographic regions of the world. *J Clin Oncol* **24**, 2137-50 (2006).
4. Daly, B. & Olopade, O.I. A perfect storm: How tumor biology, genomics, and health care delivery patterns collide to create a racial survival disparity in breast cancer and proposed interventions for change. *CA Cancer J Clin* **65**, 221-38 (2015).
5. DeSantis, C.E. *et al.* Cancer statistics for African Americans, 2016: Progress and opportunities in reducing racial disparities. *CA Cancer J Clin* **66**, 290-308 (2016).
6. Huo, D. *et al.* Comparison of Breast Cancer Molecular Features and Survival by African and European Ancestry in The Cancer Genome Atlas. *JAMA Oncol* **3**, 1654-1662 (2017).
7. Zheng, Y. *et al.* Inherited Breast Cancer in Nigerian Women. *J Clin Oncol* **36**, 2820-2825 (2018).
8. Pitt, J.J. *et al.* Characterization of Nigerian breast cancer reveals prevalent homologous recombination deficiency and aggressive molecular features. *Nat Commun* **9**, 4181 (2018).
9. Wang, S. *et al.* Germline variants and somatic mutation signatures of breast cancer across populations of African and European ancestry in the US and Nigeria. *Int J Cancer* **145**, 3321-3333 (2019).
10. Yuan, J. *et al.* Integrated Analysis of Genetic Ancestry and Genomic Alterations across Cancers. *Cancer Cell* **34**, 549-560 e9 (2018).
11. Keenan, T. *et al.* Comparison of the Genomic Landscape Between Primary Breast Cancer in African American Versus White Women and the Association of Racial Differences With Tumor Recurrence. *J Clin Oncol* **33**, 3621-7 (2015).
12. Nik-Zainal, S. *et al.* The life history of 21 breast cancers. *Cell* **149**, 994-1007 (2012).
13. Shah, S.P. *et al.* The clonal and mutational evolution spectrum of primary triple-negative breast cancers. *Nature* **486**, 395-9 (2012).
14. Gerstung, M. *et al.* The evolutionary history of 2,658 cancers. *Nature* **578**, 122-128 (2020).
15. Popejoy, A.B. & Fullerton, S.M. Genomics is failing on diversity. *Nature* **538**, 161-164 (2016).
16. Spratt, D.E. *et al.* Racial/Ethnic Disparities in Genomic Sequencing. *JAMA Oncol* **2**, 1070-4 (2016).
17. Zapata, L. *et al.* Signatures of positive selection reveal a universal role of chromatin modifiers as cancer driver genes. *Sci Rep* **7**, 13124 (2017).
18. Lawrence, M.S. *et al.* Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* **499**, 214-218 (2013).
19. Vogelstein, B. *et al.* Cancer genome landscapes. *Science* **339**, 1546-58 (2013).

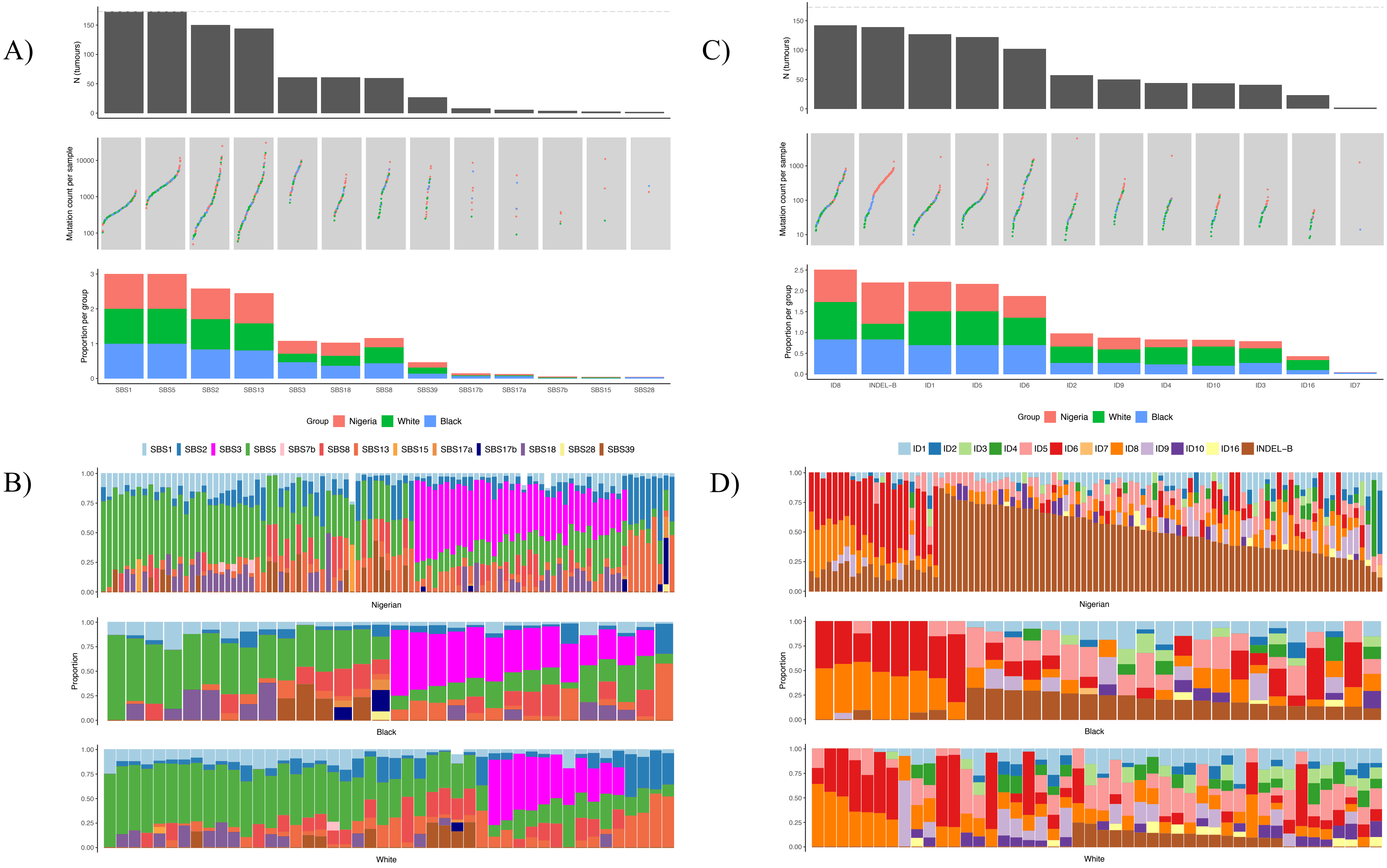
20. Vohhodina, J. *et al.* The RNA processing factors THRAP3 and BCLAF1 promote the DNA damage response through selective mRNA splicing and nuclear export. *Nucleic Acids Res* **45**, 12816-12833 (2017).
21. Bailey, M.H. *et al.* Comprehensive Characterization of Cancer Driver Genes and Mutations. *Cell* **173**, 371-385 e18 (2018).
22. Nik-Zainal, S. *et al.* Landscape of somatic mutations in 560 breast cancer whole-genome sequences. *Nature* **534**, 47-54 (2016).
23. Cohen, P.A., Donini, C.F., Nguyen, N.T., Lincet, H. & Vendrell, J.A. The dark side of ZNF217, a key regulator of tumorigenesis with powerful biomarker value. *Oncotarget* **6**, 41566-81 (2015).
24. Vendrell, J.A. *et al.* ZNF217 is a marker of poor prognosis in breast cancer that drives epithelial-mesenchymal transition and invasion. *Cancer Res* **72**, 3593-606 (2012).
25. Chen, D.H., Wu, Q.W., Li, X.D., Wang, S.J. & Zhang, Z.M. SYPL1 overexpression predicts poor prognosis of hepatocellular carcinoma and associates with epithelial-mesenchymal transition. *Oncol Rep* **38**, 1533-1542 (2017).
26. Song, Y. *et al.* SYPL1 Inhibits Apoptosis in Pancreatic Ductal Adenocarcinoma via Suppression of ROS-Induced ERK Activation. *Front Oncol* **10**, 1482 (2020).
27. Alexandrov, L.B. *et al.* The repertoire of mutational signatures in human cancer. *Nature* **578**, 94-101 (2020).
28. Khoshnaw, S.M. *et al.* Loss of Dicer expression is associated with breast cancer progression and recurrence. *Breast Cancer Res Treat* **135**, 403-13 (2012).
29. Chan, H.J. *et al.* SERPINA1 is a direct estrogen receptor target gene and a predictor of survival in breast cancer patients. *Oncotarget* **6**, 25815-27 (2015).
30. Joosse, S.A. *et al.* Prediction of BRCA2-association in hereditary breast carcinomas using array-CGH. *Breast Cancer Res Treat* **132**, 379-89 (2012).
31. Rouault, A. *et al.* Deletion of chromosomes 13q and 14q is a common feature of tumors with BRCA2 mutations. *PLoS One* **7**, e52079 (2012).
32. Hall, A.E. *et al.* The cytoskeleton adaptor protein ankyrin-1 is upregulated by p53 following DNA damage and alters cell migration. *Cell Death Dis* **7**, e2184 (2016).
33. Paul, M.R. *et al.* Genomic landscape of metastatic breast cancer identifies preferentially dysregulated pathways and targets. *J Clin Invest* **130**, 4252-4265 (2020).
34. Tracey, L.J. & Justice, M.J. Off to a Bad Start: Cancer Initiation by Pluripotency Regulator PRDM14. *Trends Genet* **35**, 489-500 (2019).
35. Williams, J.L., Greer, P.A. & Squire, J.A. Recurrent copy number alterations in prostate cancer: an in silico meta-analysis of publicly available genomic data. *Cancer Genet* **207**, 474-88 (2014).
36. Wedge, D.C. *et al.* Sequencing of prostate cancers identifies new cancer genes, routes of progression and drug targets. *Nat Genet* **50**, 682-692 (2018).
37. Aylon, Y. & Oren, M. p53: guardian of ploidy. *Mol Oncol* **5**, 315-23 (2011).
38. Bielski, C.M. *et al.* Genome doubling shapes the evolution and prognosis of advanced cancers. *Nat Genet* **50**, 1189-1195 (2018).
39. Bach, K. *et al.* Differentiation dynamics of mammary epithelial cells revealed by single-cell RNA sequencing. *Nat Commun* **8**, 2128 (2017).
40. Mukhopadhyay, P. *et al.* Mucins in the pathogenesis of breast cancer: implications in diagnosis, prognosis and therapy. *Biochim Biophys Acta* **1815**, 224-40 (2011).

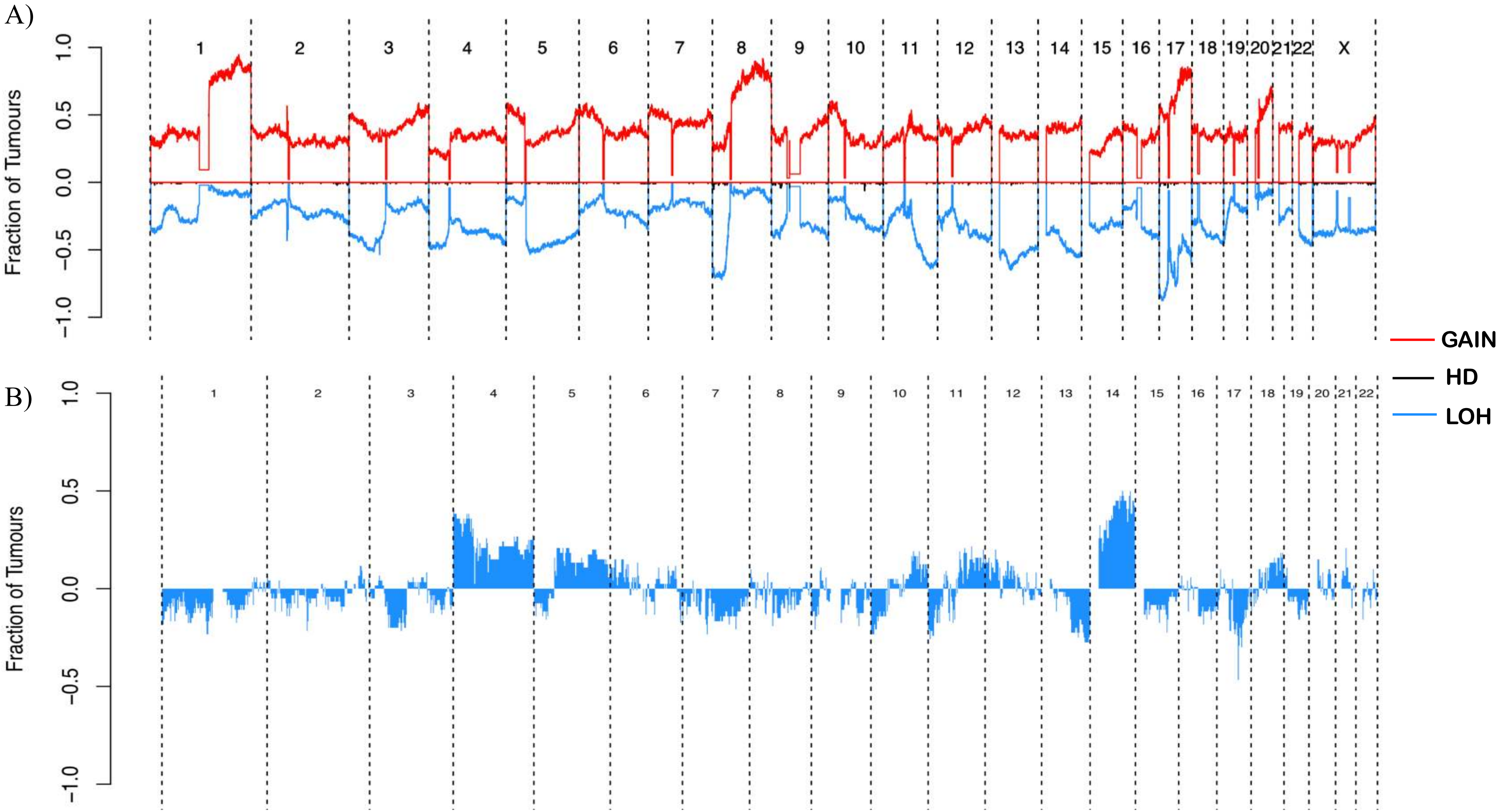
41. Nami, B. & Wang, Z. Genetics and Expression Profile of the Tubulin Gene Superfamily in Breast Cancer Subtypes and Its Relation to Taxane Resistance. *Cancers (Basel)* **10**(2018).
42. Choi, S.K., Kim, H.S., Jin, T. & Moon, W.K. LOXL4 knockdown enhances tumor growth and lung metastasis through collagen-dependent extracellular matrix changes in triple-negative breast cancer. *Oncotarget* **8**, 11977-11989 (2017).
43. Smirnova, T. *et al.* Serpin E2 promotes breast cancer metastasis by remodeling the tumor matrix and polarizing tumor associated macrophages. *Oncotarget* **7**, 82289-82304 (2016).
44. Hanahan, D. & Weinberg, R.A. Hallmarks of cancer: the next generation. *Cell* **144**, 646-74 (2011).
45. McGranahan, N. & Swanton, C. Biological and therapeutic impact of intratumor heterogeneity in cancer evolution. *Cancer Cell* **27**, 15-26 (2015).
46. Shibata, D. Cancer. Heterogeneity and tumor history. *Science* **336**, 304-5 (2012).
47. Zhang, H. *et al.* LAMB3 mediates apoptotic, proliferative, invasive, and metastatic behaviors in pancreatic cancer by regulating the PI3K/Akt signaling pathway. *Cell Death Dis* **10**, 230 (2019).
48. Bowen, R.L., Stebbing, J. & Jones, L.J. A review of the ethnic differences in breast cancer. *Pharmacogenomics* **7**, 935-42 (2006).
49. Cancer Genome Atlas, N. Comprehensive molecular portraits of human breast tumours. *Nature* **490**, 61-70 (2012).
50. Dawson, S.J., Rueda, O.M., Aparicio, S. & Caldas, C. A new genome-driven integrated classification of breast cancer and its implications. *EMBO J* **32**, 617-28 (2013).
51. Bernard, E. *et al.* Implications of TP53 allelic state for genome stability, clinical presentation and outcomes in myelodysplastic syndromes. *Nat Med* (2020).
52. Micheletti, S.J. *et al.* Genetic Consequences of the Transatlantic Slave Trade in the Americas. *Am J Hum Genet* **107**, 265-277 (2020).
53. Oluwasanu, M. & Olopade, O.I. Global disparities in breast cancer outcomes: new perspectives, widening inequities, unanswered questions. *Lancet Glob Health* **8**, e978-e979 (2020).
54. Rajagopal, P.S. & Olopade, O.I. Black Lives Matter Worldwide: Retooling Precision Oncology for True Equity of Cancer Care. *Cell Rep Med* **1**, 100079 (2020).
55. Ntekim, A. *et al.* Implementing oncology clinical trials in Nigeria: a model for capacity building. *BMC Health Serv Res* **20**, 713 (2020).
56. Ntekim, A.I. *et al.* ARETTA: Assessing Response to Neoadjuvant Taxotere and Subcutaneous Trastuzumab in Nigerian Women With HER2-Positive Breast Cancer: A Study Protocol. *JCO Glob Oncol* **6**, 983-990 (2020).
57. Huo, D. *et al.* Parity and breastfeeding are protective against breast cancer in Nigerian women. *Br J Cancer* **98**, 992-6 (2008).
58. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res* **38**, e164 (2010).
59. Bolli, N. *et al.* Heterogeneity of genomic evolution and mutational profiles in multiple myeloma. *Nat Commun* **5**, 2997 (2014).
60. Yates, L.R. *et al.* Genomic Evolution of Breast Cancer Metastasis and Relapse. *Cancer Cell* **32**, 169-184 e7 (2017).
61. Kircher, M. *et al.* A general framework for estimating the relative pathogenicity of human genetic variants. *Nat Genet* **46**, 310-5 (2014).

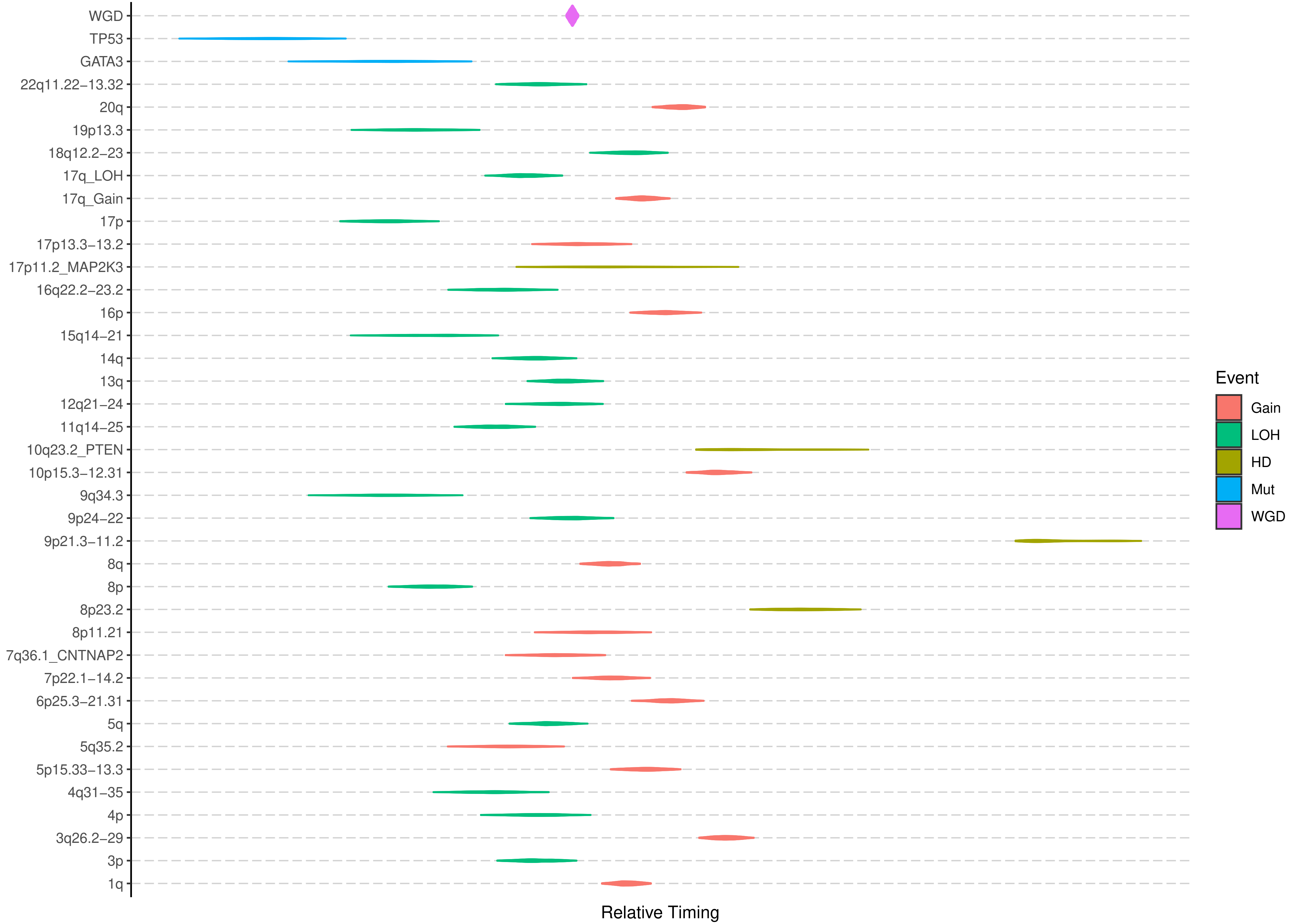
62. Mayakonda, A., Lin, D.C., Assenov, Y., Plass, C. & Koeffler, H.P. Maftools: efficient and comprehensive analysis of somatic variants in cancer. *Genome Res* **28**, 1747-1756 (2018).
63. Alexandrov, L.B., Nik-Zainal, S., Wedge, D.C., Campbell, P.J. & Stratton, M.R. Deciphering signatures of mutational processes operative in human cancer. *Cell Rep* **3**, 246-59 (2013).
64. Consortium, I.T.P.-C.A.o.W.G. Pan-cancer analysis of whole genomes. *Nature* **578**, 82-93 (2020).
65. Nilsen, G. *et al.* Copynumber: Efficient algorithms for single- and multi-track copy number segmentation. *BMC Genomics* **13**, 591 (2012).
66. Duncan, L.R. Individual choice behavior: A theoretical analysis. (New York: Wiley, 1959).
67. Plackett, R.L. The analysis of permutations. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **24**, 193-202 (1975).
68. Gundem, G. *et al.* The evolutionary history of lethal metastatic prostate cancer. *Nature* **520**, 353-357 (2015).
69. Cmero, M. *et al.* Inferring structural variant cancer cell fraction. *Nat Commun* **11**, 730 (2020).
70. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)* **57**, 289-300 (1995).
71. Jafari, M. & Ansari-Pour, N. Why, When and How to Adjust Your P Values? *Cell J* **20**, 604-607 (2019).

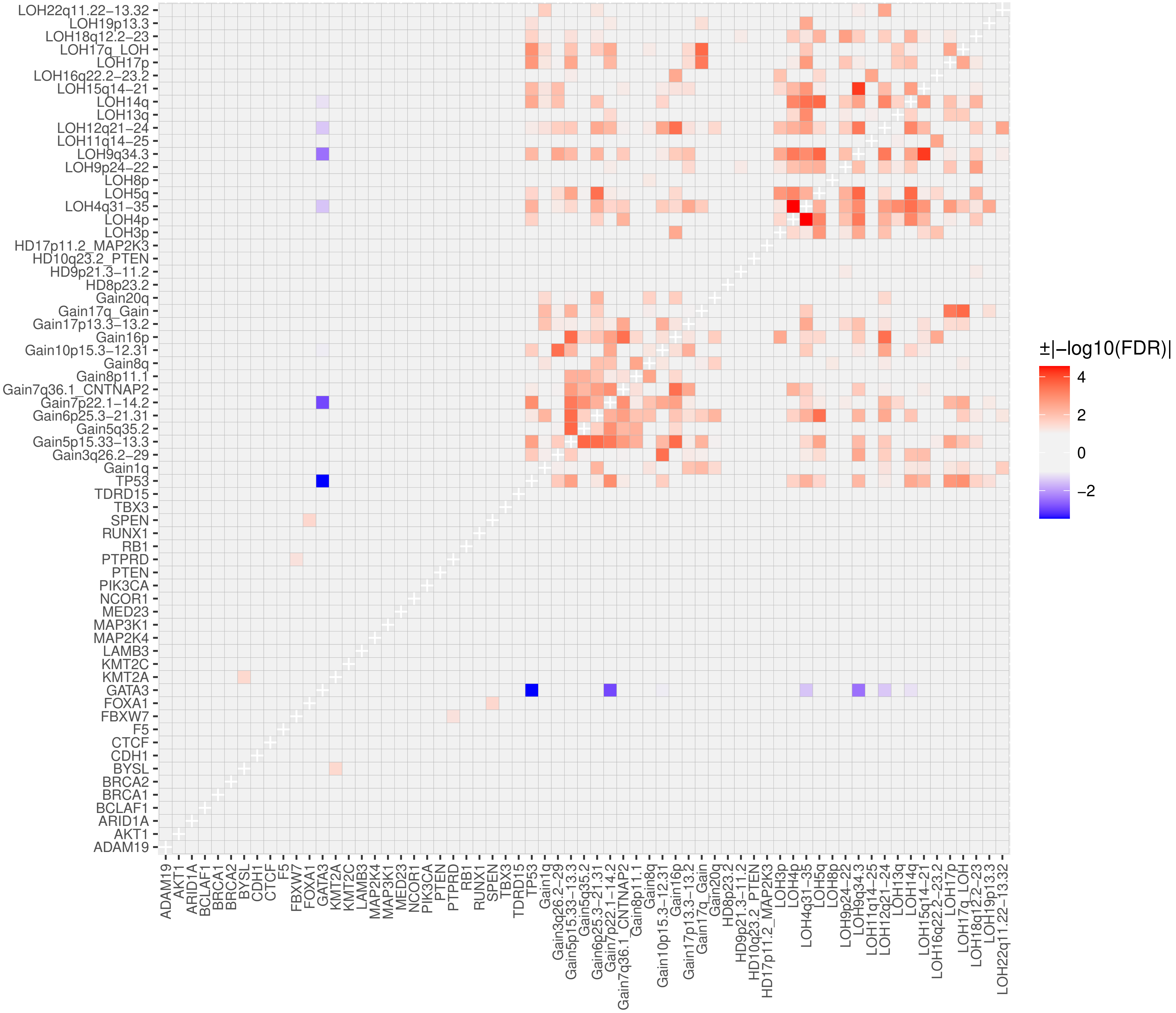

600
500
400
300
200
100
0
0
20
40
60
80
100
61% TP53
24% PIK3CA
13% GATA3
6% KMT2C
6% MAP3K1
5% TDRD15
5% BCLAF1
4% BRCA2
5% F5
5% KMT2A
5% PTPRD
4% ADAM19
4% BRCA1
4% CDH1
4% LAMB3
3% AKT1
3% FBXW7
3% PTEN
3% RB1
3% ARID1A
3% FOXA1
3% NCOR1
3% SPEN
2% BYSL
2% CBFB
2% CTCF
2% MAP2K4
2% MED23
2% RUNX1
2% TBX3
Group
Nigerian
Black
White
Subtype
HER2+
HR−/HER2−
HR+/HER2−
Missense_Mutation
Frame_Shift_Ins
Frame_Shift_Del
Splice_Site
Nonsense_Mutation
Multi_Hit


SYPL1
ZNF217
−log10(p − value)
1
2
3
4
5
6
Chromosome
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
X


A)
N (tumours)
Mutation count per sample
Proportion per group
SBS1
SBS5
SBS2
SBS13
SBS3
SBS18
SBS8
SBS39
SBS17b
SBS17a
SBS7b
SBS15
SBS28
Group
Nigeria
White
Black
B)
SBS1
SBS2
SBS3
SBS5
SBS7b
SBS8
SBS13
SBS15
SBS17a
SBS17b
SBS18
SBS28
SBS39
Nigerian
Proportion
Black
White
C)
N (tumours)
Mutation count per sample
Proportion per group
ID8
INDEL−B
ID1
ID5
ID6
ID2
ID9
ID4
ID10
ID3
ID16
ID7
Group
Nigeria
White
Black
D)
ID1
ID2
ID3
ID4
ID5
ID6
ID7
ID8
ID9
ID10
ID16
INDEL−B
Nigerian
Proportion
Black
White


A)
Fraction of Tumours
1.0
0.5
0.0
−0.5
−1.0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
X
GAIN
HD
LOH
B)
Fraction of Tumours
1.0
0.5
0.0
−0.5
−1.0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22


WGD
TP53
GATA3
22q11.22–13.32
20q
19p13.3
18q12.2–23
17q_LOH
17q_Gain
17p
17p13.3–13.2
17p11.2_MAP2K3
16q22.2–23.2
16p
15q14–21
14q
13q
12q21–24
11q14–25
10q23.2_PTEN
10p15.3–12.31
9q34.3
9p24–22
9p21.3–11.2
8q
8p
8p23.2
8p11.21
7q36.1_CNTNAP2
7p22.1–14.2
6p25.3–21.31
5q
5q35.2
5p15.33–13.3
4q31–35
4p
3q26.2–29
3p
1q
Relative Timing
Event
Gain
LOH
HD
Mut
WGD


LOH22q11.22–13.32
LOH19p13.3
LOH18q12.2–23
LOH17q_LOH
LOH17p
LOH16q22.2–23.2
LOH15q14–21
LOH14q
LOH13q
LOH12q21–24
LOH11q14–25
LOH9q34.3
LOH9p24–22
LOH8p
LOH5q
LOH4q31–35
LOH4p
LOH3p
HD17p11.2_MAP2K3
HD10q23.2_PTEN
HD9p21.3–11.2
HD8p23.2
Gain20q
Gain17q_Gain
Gain17p13.3–13.2
Gain16p
Gain10p15.3–12.31
Gain8q
Gain8p11.1
Gain7q36.1_CNTNAP2
Gain7p22.1–14.2
Gain6p25.3–21.31
Gain5q35.2
Gain5p15.33–13.3
Gain3q26.2–29
Gain1q
TP53
TDRD15
TBX3
SPEN
RUNX1
RB1
PTPRD
PTEN
PIK3CA
NCOR1
MED23
MAP3K1
MAP2K4
LAMB3
KMT2C
KMT2A
GATA3
FOXA1
FBXW7
F5
CTCF
CDH1
BYSL
BRCA2
BRCA1
BCLAF1
ARID1A
AKT1
ADAM19
ADAM19
AKT1
ARID1A
BCLAF1
BRCA1
BRCA2
BYSL
CDH1
CTCF
F5
FBXW7
FOXA1
GATA3
KMT2A
KMT2C
LAMB3
MAP2K4
MAP3K1
MED23
NCOR1
PIK3CA
PTEN
PTPRD
RB1
RUNX1
SPEN
TBX3
TDRD15
TP53
Gain1q
Gain3q26.2–29
Gain5p15.33–13.3
Gain5q35.2
Gain6p25.3–21.31
Gain7p22.1–14.2
Gain7q36.1_CNTNAP2
Gain8p11.1
Gain8q
Gain10p15.3–12.31
Gain16p
Gain17p13.3–13.2
Gain17q_Gain
Gain20q
HD8p23.2
HD9p21.3–11.2
HD10q23.2_PTEN
HD17p11.2_MAP2K3
LOH3p
LOH4p
LOH4q31–35
LOH5q
LOH8p
LOH9p24–22
LOH9q34.3
LOH11q14–25
LOH12q21–24
LOH13q
LOH14q
LOH15q14–21
LOH16q22.2–23.2
LOH17p
LOH17q_LOH
LOH18q12.2–23
LOH19p13.3
LOH22q11.22–13.32
±|−log10(FDR)|
4
2
0
−2

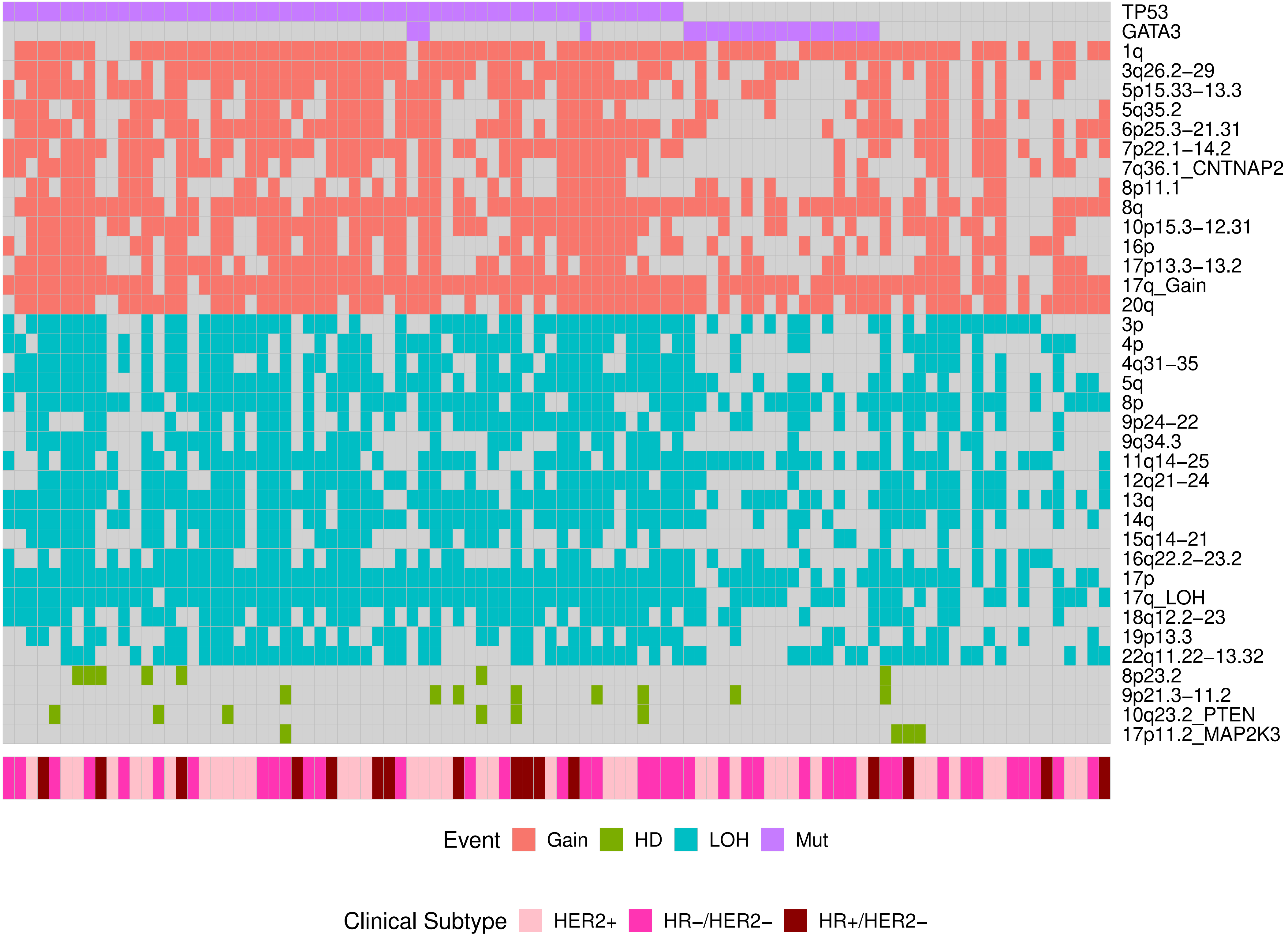

TP53
GATA3
1q
3q26.2−29
5p15.33−13.3
5q35.2
6p25.3−21.31
7p22.1−14.2
7q36.1_CNTNAP2
8p11.1
8q
10p15.3−12.31
16p
17p13.3−13.2
17q_Gain
20q
3p
4p
4q31−35
5q
8p
9p24−22
9q34.3
11q14−25
12q21−24
13q
14q
15q14−21
16q22.2−23.2
17p
17q_LOH
18q12.2−23
19p13.3
22q11.22−13.32
8p23.2
9p21.3−11.2
10q23.2_PTEN
17p11.2_MAP2K3
Event
Gain
HD
LOH
Mut
Clinical Subtype
HER2+
HR−/HER2−
HR+/HER2−

Table 1. Characteristics of the somatic mutational drivers detected in breast tumors.

Driver	**Detection Method**	**Mutational pattern**	**Cancer Gene Status**
ADAM19	cDriver	Neither	Novel
BCLAF1	MutSigCV	TSG	COSMIC
BYSL	cDriver	TSG	Novel
CDH1	cDriver/MutSigCV	TSG	COSMIC
F5	cDriver	TSG	Novel
GATA3	cDriver/MutSigCV	TSG	COSMIC
LAMB3	cDriver	Neither	Novel
MAP3K1	cDriver/MutSigCV	TSG	COSMIC
PIK3CA	cDriver/MutSigCV	ONC	COSMIC
PTEN	cDriver/MutSigCV	TSG	COSMIC
RB1	cDriver	TSG	COSMIC
TDRD15	cDriver	TSG	Novel
TP53	cDriver/MutSigCV	TSG/ONC	COSMIC