Supplementary Material

Supplementary Methods

Supplementary information is available online at https://github.com/tkarn/G9-TMB.

Whole Exome sequencing

Whole exome sequencing was conducted on fresh-frozen pre-therapeutic core biopsies and patient-matched blood samples. Tumor purity and cellularity can have impact on the detection of mutations from whole exome sequencing and thereby potentially confound TMB estimation (see below). Therefore, selection of tumor tissue was performed by experienced pathologists to obtain samples with high cellularity (>50%). For each sample 10x 5µm cryosections were used for whole exome sequencing with *Illumina HiSeq 4000* (Illumina Inc., San Diego, CA, USA) for 2 x 100 paired end reads. High quality data was obtained for 149 (85.6%) of the 174 patients from the trial (Supplementary Figure S1). The sequence reads were mapped to human reference genome hg19 using *BWA* (Burrows-Wheeler aligner) [1]. *Samtools* and *Bedtools* were used to calculate mapping rate and coverage for quality control (QC) [2, 3]. *Mutect2* [4] was used to make somatic variant calls, including single-nucleotide-variant (SNV) and small-insertion-deletion (indel). Gene and functional annotation of the called SNVs and indels was carried out using *ANNOVAR* [5]. We used COSMIC v86 and dbSNP v152 variants as white and black lists for *MuTect2* respectively. This resulted in 34,082 mutations in 10,969 genes. 96 genes from a black list and 16,471 SNVs flagged in germline by *ANNOVAR* were excluded. Of the remaining 15,682 mutations 3,368 synonymous SNV were filtered resulting in a final list of 12,314 mutations. For somatic copy number alterations *pure CN* [6] was used.

Tumor mutational burden and mutational signatures

For TMB calculation we used the final list of 12,314 non-synonymous SNVs and indels together with an effective DNA coverage of 46 Mb to determine mutations per Mb for each sample. Mutational signatures were identified as described by Alexandrov et al [7]. R package *SomaticSignatures* [8] was employed to estimate the proportion of each sample's mutations that have been assigned to each of the 21 mutational signatures.

Based on our previous analyses of TNBC from TCGA we were concerned that high immune infiltration cells could bias towards a lower TMB, because of reduced tumor cell content. In that previous study we found a negative correlation between TMB and immune cell infiltration (despite ≥50% tumor cell content mandatory for TCGA samples) [9]. However, in the GeparNuevo dataset we did not detect such negative correlation between immune cell gene expression and TMB (Fig. 2A). Moreover, in the TCGA dataset we had also found a positive correlation between TMB and tumor cellularity (as measured by ASCAT) [9]. Again, in the GeparNuevo data we did not observe such a correlation between TMB and purity (as measured by pureCN, Spearman's rho=0.006). Thus, despite no systematic data on histological assessment of tumor content was available, these observations argue against a strong bias of TMB by tumor cellularity in the GeparNuevo dataset.

RNA sequencing

RNA sequencing was performed on formalin-fixed paraffin-embedded (FFPE) tissue using a *HTG EdgeSeq* instrument (HTG Molecular Inc, Tucson, AZ, USA) with the *HTG EdgeSeq Oncology Biomarker Panel* (2549 genes) according to the manufacturer's instructions. Tumor area was marked on an H&E stained slide and the area of invasive breast cancer recorded. From a corresponding unstained slide 15 mm^2 tissue was scraped and used for library preparation. The method is based on an RNA-extraction-free chemistry and a nuclease protection assay [10]. Libraries were quantified, pooled and sequenced on an *Ion Torrent S5* instrument (Thermo Fisher Scientific, Waltham, MA, USA). Count tables were generated using the *HTG parsing tool*. For quality control, we transformed the reads to counts-per-million and calculated the mean of five negative and four positive internal controls for each sample. We repeated processing for a sample if the mean of its positive controls was below two standard deviations (SDs) of the grand mean across all samples or if the mean of its negative controls were above two SDs from the grand mean (Supplementary Figure S6). The data was median normalized within a sample and across the experiment by calculating a scaling factor for each sample as the median gene expression value for each sample-gene count adjusted by the geometric mean over all genes. RNA sequencing data from pre-therapeutic cores were available for 159 of the 174 patients (Supplementary Figure S1).

Molecular subtyping by AIMS from RNA-Seq

Molecular subtyping from RNA-Seq was performed using the *Absolute Intrinsic Molecular Subtyping* (AIMS) method by applying the R package *AIMS*. Of the 136 patients with both WES and RNA-Seq data, 82 (60.3%) were classified as basal-like, 50 (36.7%) as HER2-enriched, 3 (2.2%) as normal-like, and 1 (0.7%) as luminal A. The *HTG-EdgeSeq Oncology Biomarker Panel* encompasses 2549 genes, which includes 41 of the 100 gene pairs proposed as rules for the AIMS algorithm. The respective list of available gene pairs is given in Supplementary Table S5 with 6-10 rules per subtype. In the original AIMS publication [11] the agreement between AIMS and PAM50 classification dropped below 10 rules per subtype. Therefore, the lower number of rules available in our study may be one explanation for the relatively high frequency of HER2-enriched cases (36.7%) among TNBC in our cohort in contrast to PAM50 data from the literature e.g. 7.8% [12].

Immune gene expression profiles (GEP) from RNA-Seq

We evaluated a predefined immune gene expression profile (GEP) predictive for neoadjuvant response that was created from a list of genes we previously identified in the GeparSixto study ("GeparSixto immune signature": CXCL9, CCL5, CD8A, CD80, CXCL13, IDO1, PDCD1, CD274, CTLA4, FOXP3) [13]. The genes CD21 and IGKC were omitted because they were not covered by the HTG-EdgeSeq Oncology Biomarker Panel. The immune GEP was calculated as mean of the expression of the genes from the signature. We also verified substantial agreement of this pre-specified signature with various other gene signatures for lymphocyte infiltration (Pearson corrlation values 0.74-0.97, Supplementary Figure S7).

Aggregation of WES and HTG-RNA-Seq data

The analyses of the pseudonymized genomic datasets were performed fully blinded to any clinical or pathological sample information. The final blinded WES and HTG-RNA-Seq datasets were transferred to GBG-headquarters. WES and RNA-Seq were available for 149 and 159 patients, respectively, with both data available for 136 patients (Supplementary Figure S1). A comparison of the complete trial cohort, the WES cohort, and the WES+RNA-Seq cohort is provided in Supplementary Table S1. An integrated view of all meaures is presented in Supplementary Figure S8.

Supplementary References

1. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics (Oxford, England) 2009;25(14):1754–60.

2. Li H, Handsaker B, Wysoker A et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics (Oxford, England) 2009;25(16):2078–79.

3. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics (Oxford, England) 2010;26(6):841–42.

4. Cibulskis K, Lawrence MS, Carter SL et al. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. Nature biotechnology 2013;31(3):213–19.

5. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic acids research 2010;38(16):e164.

6. Riester M, Singh AP, Brannon AR et al. PureCN: copy number calling and SNV classification using targeted short read sequencing. Source code for biology and medicine 2016;11:13.

7. Alexandrov LB, Nik-Zainal S, Wedge DC et al. Signatures of mutational processes in human cancer. Nature 2013;500(7463):415–21.

8. Gehring JS, Fischer B, Lawrence M, Huber W. SomaticSignatures: inferring mutational signatures from single-nucleotide variants. Bioinformatics (Oxford, England) 2015;31(22):3673–75.

9. Karn T, Jiang T, Hatzis C et al. Association Between Genomic Metrics and Immune Infiltration in Triple-Negative Breast Cancer. JAMA oncology 2017;3(12):1707–11.

10. Martel RR, Botros IW, Rounseville MP et al. Multiplexed screening assay for mRNA combining nuclease protection with luminescent array detection. Assay and drug development technologies 2002;1(1 Pt 1):61–71.

11. Paquet ER, Hallett MT. Absolute assignment of breast cancer intrinsic molecular subtype. J. Natl. Cancer Inst. 2015;107(1):357.

12. Prat A, Adamo B, Cheang MCU et al. Molecular characterization of basal-like and non-basal-like triple-negative breast cancer. The oncologist 2013;18(2):123–33.

13. Denkert C, Minckwitz G von, Brase JC et al. Tumor-infiltrating lymphocytes and response to neoadjuvant chemotherapy with or without carboplatin in human epidermal growth factor receptor 2-positive and triple-negative primary breast cancers. Journal of clinical oncology : official journal of the American Society of Clinical Oncology 2015;33(9):983–91.

14. Rooney MS, Shukla SA, Wu CJ et al. Molecular and genetic properties of tumors associated with local immune cytolytic activity. Cell 2015;160(1-2):48–61.

15. Ayers M, Lunceford J, Nebozhyn M et al. IFN-γ-related mRNA profile predicts clinical response to PD-1 blockade. The Journal of clinical investigation 2017;127(8):2930–40.

16. Karn T, Meissner T, Weber KE et al. A Small Hypoxia Signature Predicted pCR Response to Bevacizumab in the Neoadjuvant GeparQuinto Breast Cancer Trial. Clinical cancer research : an official journal of the American Association for Cancer Research 2020.

17. Rody A, Karn T, Liedtke C et al. A clinically relevant gene signature in triple negative and basal-like breast cancer. Breast cancer research : BCR 2011;13(5):R97.