Pre-existing immunity drives the response to neoadjuvant chemotherapy in esophageal adenocarcinoma

SUPPLEMENTARY MATERIALS AND METHODS

WES data acquisition and analysis

WES was performed on EAC biopsies at an average coverage of 120X. Germline controls were established using autologous PBMCs at an average coverage of 80X. Libraries were generated from 30 ng of input DNA using the SureSelect QXT Human All Exon V7 kit (Agilent Technologies), and sequencing was performed on the NextSeq 550 platform (Illumina) with 2x150 bp read length. BCL files were converted to FastQ files using bcl2fastq2 software (Illumina) after demultiplexing.

To ensure the reproducibility, scalability, and portability of the code, 3 paired-end sequencing batches were analyzed independently using Snakemake (55), a workflow management and development system for data analysis. The analyzed sequencing batches were as follows: Batch1 (samples 8, 10, 11, 12, 15, 17, 18), Batch2 (samples 20, 24, 25, 26, 27, 29, 30, 31, 33, 34), and Batch3 (samples 35, 37, 39, 40, 41, 43, 45, 48, 54, 55, 57). Version control was maintained using Conda (https://docs.conda.io/en/latest/). Quality checks on raw sequencing data were performed with FastQC v0.11.9 (56). Trimming was performed using BBMap bbduck v38.90 [BBMap - Bushnell B. - sourceforge.net/projects/bbmap/] (with parameters k=23 mink=11 rcomp=t ktrim=f kmask=N qtrim=rl trimq=5 forcetrimleft=5 forcetrimright2=0 overwrite=true) to remove adapter sequences and low-quality bases from the reads. The FastQ files from Batch2 samples were cleaned using BBMap filterbytile.sh v38.90 (with parameters ud=0.75 qd=1 ed=1 ua=.5 qa=.5 ea=.5) to eliminate the loss in quality associated with some tiles of the flowcell. FastQC was then re-applied to verify the quality of the reads after the pre-filtering processes outlined above. Clean reads in FastQ format were aligned to the reference human genome (GATK_bundle Hg38, v0) using Burrows-Wheeler Aligner (bwa-mem2 v2.1) (57). Duplicate reads were identified using picard MarkDuplicates v2.24.2 [http://picard.sourceforge.net/] after sorting and indexing of bam alignment files, which was performed using samtools v1.11 (58). To correct for the effect of overlapping read pairs on SNV coverage, BAM files were processed with BamUtil ClipOverlap

v1.0.15 (59) to address the risk of potential PCR amplification/sequencing errors that could otherwise carry errors to both paired reads, causing a single-base error to appear as two independent mismatches and resulting in false positive variant calls. Mapping metrics were generated using Qualimap v.2.2.2-dev (60) and picard CollectHsMetrics v2.24.2. Sample quality metrics were collected in an html report by MultiQC v1.9 (61).

The clean sequencing data were analyzed to identify somatic single nucleotide variations (SNVs) and small indels using GATK MuTect2 v4-4.2.2.0 (62), following the Somatic short variant discovery Best Practices. To correct for systematic bias that affects the assignment of base quality scores by the sequencer, Base Quality Score Recalibration (BQSR) steps were applied to the tumor, matching normal BAM files. First, the BQSR recalibration tables were generated from BAM files using GATK BaseRecalibrator v4-4.2.2.0. Then, the tables were provided with the BAM files as input to GATK ApplyBQSR to recalibrate the base qualities of the input reads. A panel of normals (PoN) containing germline and artifactual sites present in normal samples was then created with GATK GenomicsDBImport v4-4.2.2.0 and GATK CreateSomaticPanelOfNormals v4-4.2.2.0.

To call somatic variants only, GATK MuTect2 v4-4.2.2.0 was used in tumors with matched normal mode, using the tumor match normal, the panel of normals, the Agilent probe interval list, and the gnomad Hg38 germline-resource. Soft-clipped bases were excluded from the call using the optional parameter --dont-use-soft-clipped-bases true. The raw output of Mutect2 was filtered using GATK FilterMutectCalls v4-4.2.2.0. Gene-level and COSMIC v95 (63) information was added to each variant using GATK Funcotator v4-4.2.2.0 and SnpSift annotate v4.3t (64), respectively. Ensembl Variant Effector Predictor (VEP) (65) was used to predict the effect of the detected variants. The resulting tumor somatic Variant Call Files (VCFs) were converted to Mutation Annotation Format (MAF) files using vcf2maf.pl v2.0 (66). Maftools v2.6.5 (67) and bcftools stats v1.8 were used to generate summary plots and statistics, while the VCFs MPOS field was checked to exclude the presence of calling biases along the read length.

Sequenza v3.0.0 tool (68) was used to estimate tumor cellularity and ploidy, as well as to calculate allele-specific copy number profiles, using the recalibrated BAM files from the 28 samples. Optional parameters low_ploidy=1, up_ploidy=7, cellularity=seq(low_cell, up_cell, 0.01) were utilized, where low_cell corresponded to the sample cellularity estimated by the pathologist -20% and up_cell corresponded to the sample cellularity estimated by the pathology +20%.

The code for WES analysis has been uploaded to GitHub https://github.com/auroramaurizio/WES_ESOCA and is available in Zenodo at the following hyperlink https://zenodo.org/badge/latestdoi/575453310.

Oncoplots were designed considering mutations derived from the gene list of the Reactome Class I antigen processing and presentation, downloadable from the Molecular Signature database (69), as well as a gene list of known EAC drivers described elsewhere (70).

RNA sequencing data analysis

RNA-seq libraries were prepared using 50 ng of total RNA with an RNA integrity index (RIN) of ≥7, using the TruSeq Stranded mRNA library preparation kit (Illumina) in accordance with low-throughput protocol. After PCR enrichment (15 cycles) and purification of adapter-ligated fragments, the concentration and length of DNA fragments were measured using the D1000 Screen Tape System (Agilent), resulting in a median insert size of 311 nucleotides. The RNA-seq libraries were then sequenced using the Illumina NovaSeq platform with 1x100 bp read length, generating on average 100 million single reads per sample.

To quantify gene expression levels, read tags were pseudo-aligned to the GENCODE 38 transcriptome (71) using Kallisto v.0.44.0 (72) (parameters: “-t –single --rf-stranded -l 200 -s 20”). Two samples (39 and 40) were excluded from downstream analysis as they failed to pass the quality check test (i.e., pseudo-aligned reads on the coding sequence/total reads of >50%). Abundances for genes were summarized using the TXImport package (73) and analyzed using edgeR (74). A volcano plot illustrating the enriched genes in CRs versus NRs was generated,

considering the log2 fold change of gene expression and the p-value ($p<0.05$ for significantly enriched genes; red dots representing genes differentially expressed, $p<0.01$). Genes highlighted in the volcano plot were selected among the top 500 differentially enriched in CRs versus NRs ($p<0.01$).

Gene set enrichment analysis (GSEA) was performed using hallmark gene sets from the MolecularSignature database (MsigDB) (65), comparing RNA-seq data from CR, PR, and NR samples. The frequency of immune cell populations was predicted in silico with digital cytometry using CIBERSORTx (75). The analysis utilized a transcripts per million (TPM) kallisto-counts mixture matrix and LM22 signature, run in absolute mode with 100 permutations.

HLA allotypes, HLA loss of heterozygosity, and neoantigen load prediction

We used Optitype v.1.3.2 (76) to predict HLA class I allotypes for all samples by providing normal sample WES FastQ files as input, with the parameters unpaired_weight=0 and use_discordant=false.

To predict the loss of heterozygosity in HLA class I genes, we utilized LOHHLA (28) but modified the original *LOHHLAscript.R* script to ensure compatibility with the human genome release GRCh38. Specifically, we modified the boundaries of the regions covering the HLA-A, -B, and -C genes and updated the names of alternate chr6 contigs that included HLA contigs. Input to LOHHLA included the following for each sample: i) purity and ploidy estimated by Sequenza (64), ii) normal HLA allotypes predicted by Optitype, and iii) BAM files from the WES of both normal and tumor samples. LOHHLA was run using the following parameters: --mappingStep TRUE --fishingStep TRUE --minCoverageFilter 5.

For each sample, we generated a list of missense mutations and in-frame indels from the VCF files containing all the somatic mutations predicted by Mutect2 and annotated using Funcotator (59). We excluded variants with an allele frequency (VAF) of ≤0.05 and those occurring on genes featuring one or more frameshift indels, as most of the resulting transcripts are believed to undergo

nonsense-mediated decay (77). Next, using an in-house Python script, we generated all possible peptides 8-11 amino acids in length (referred to as neoepitopes from now on) and spanning each of the annotated mutations. For variants found on the same transcript, we considered phasing information from Mutect2 when building our mutant peptides. If variants were predicted to be in *cis*, we built the mutant peptides resulting from all the variants combined (this was relevant for mutations positioned within 11 amino acids away from each other); if variants were predicted to be in *trans*, we built all mutant peptides generated by the individual variants separately. If multiple variants were present on the same transcript, but no phasing information was available from Mutect2, we always considered them in *cis*. For missense mutations, we additionally generated a list of wild-type peptides associated with the mutant ones.

We predicted the peptide-HLA binding affinity and elution likelihood of mutant peptides using *netMHCpan* (version 4.1b) (78). For each sample, we ran *netMHCpan* against the full list of mutant peptides separately for each HLA class I allotype associated with the sample. To calculate the neoantigen load of a given sample, we considered neoepitopes that met all of the following criteria: they originated from a variant found on an expressed gene (i.e., a gene with a TPM >0, as determined from the sample's bulk RNA-seq data); their elution likelihood % rank was ≤0.5%; and their binding affinity was ≤500nM with respect to at least one of the HLA allotypes associated with the sample.

Supplementary References

55. Köster J, Rahmann S. Snakemake—a scalable bioinformatics workflow engine. Bioinformatics 28.19 (2012): 2520-2522.

56. Andrews S. (2010). FastQC: A Quality Control Tool for High Throughput Sequence Data. Online: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

57. Li, Heng. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv preprint arXiv:1303.3997 (2013).

58. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. The sequence alignment/map format and SAMtools. Bioinformatics 25.16 (2009): 2078-2079.

59. Jun G, Wing MK, Abecasis GR, Kang HM. An efficient and scalable analysis framework for variant extraction and refinement from population-scale DNA sequence data. Genome research 25.6 (2015): 918-925.

60. García-Alcalde F, Okonechnikov K, Carbonell J, Cruz LM, Götz S, Tarazona S, Dopazo J, Meyer TF, Conesa A. Qualimap: evaluating next-generation sequencing alignment data. Bioinformatics 28.20 (2012): 2678-2679.

61. Ewels P, Magnusson M, Lundin S, Käller M. MultiQC: summarize analysis results for multiple tools and samples in a single report. Bioinformatics 32.19 (2016): 3047-3048.

62. Benjamin D, Sato T, Cibulskis K, Getz G, Stewart C, Lichtenstein L. Calling somatic SNVs and indels with Mutect2. BioRxiv (2019): 861054.

63. Tate JG, Bamford S, Jubb HC, Sondka Z, Beare DM, Bindal N, Boutselakis H, Cole CG, Creatore C, Dawson E, Fish P, Harsha B, Hathaway C, Jupe SC, Kok CY, Noble K, Ponting L, Ramshaw CC, Rye CE, Speedy HE, Stefancsik R, Thompson SL, Wang S, Ward S, Campbell PJ, Forbes SA. COSMIC: the catalogue of somatic mutations in cancer." Nucleic acids research 47.D1 (2019): D941-D947.

64. Cingolani P, Patel VM, Coon M, Nguyen T, Land SJ, Ruden DM, Lu X. Using Drosophila melanogaster as a Model for Genotoxic Chemical Mutational Studies with a New Program, SnpSift. Front Genet. 2012 Mar 15;3:35.

65. McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GR, Thormann A, Flicek P, Cunningham F. The ensembl variant effect predictor." Genome biology 17.1 (2016): 1-14.

66. Cyriac Kandoth. mskcc/vcf2maf: vcf2maf v1.6.19. (2020). doi:10.5281/zenodo.593251.

67. Mayakonda A, Lin DC, Assenov Y, Plass C, Koeffler HP. Maftools: efficient and comprehensive analysis of somatic variants in cancer." Genome research 28.11 (2018): 1747-1756.

68. Favero F, Joshi T, Marquard AM, Birkbak NJ, Krzystanek M, Li Q, Szallasi Z, Eklund AC "Sequenza: allele-specific copy number and mutation profiles from tumor sequencing data." Ann Oncol. 2015 Jan;26(1):64-70.

69. Liberzon, A. Birger C, Thorvaldsdóttir H, Ghandi M, Mesirov JP, Tamayo P. The Molecular Signatures Database Hallmark Gene Set Collection. Cell Syst. 1, 417–425 (2015).

70. Secrier M, Li X, de Silva N, Eldridge MD, Contino G, Bornschein J, MacRae S, Grehan N, O'Donovan M, Miremadi A, Yang TP, Bower L, Chettouh H, Crawte J, Galeano-Dalmau N, Grabowska A, Saunders J, Underwood T, Waddell N, Barbour AP, Nutzinger B, Achilleos A, Edwards PA, Lynch AG, Tavaré S, Fitzgerald RC; Oesophageal Cancer Clinical and Molecular Stratification (OCCAMS) Consortium. Mutational signatures in esophageal adenocarcinoma define etiologically distinct subgroups with therapeutic relevance. Nat Genet. 2016 Oct;48(10):1131-41.

71. Harrow, J. Frankish A, Gonzalez JM, Tapanari E, Diekhans M, Kokocinski F, Aken BL, Barrell D, Zadissa A, Searle S, Barnes I, Bignell A, Boychenko V, Hunt T, Kay M, Mukherjee G, Rajan J, Despacio-Reyes G, Saunders G, Steward C, Harte R, Lin M, Howald C, Tanzer A, Derrien T, Chrast J, Walters N, Balasubramanian S, Pei B, Tress M, Rodriguez JM, Ezkurdia I, van Baren J, Brent M, Haussler D, Kellis M, Valencia A, Reymond A, Gerstein M, Guigó R, Hubbard TJ. GENCODE: the reference human genome annotation for The ENCODE Project. Genome Res. 22, 1760–74 (2012).

72. Bray, N. L., Pimentel, H., Melsted, P. & Pachter, L. Near-optimal probabilistic RNA-seq quantification. Nat. Biotechnol. 34, 525–527 (2016).

73. Soneson, C., Love, M. I. & Robinson, M. D. Differential analyses for RNA-seq: Transcript-level estimates improve gene-level inferences. F1000Research 4, 1–23 (2016).

74. Anders, S. McCarthy DJ, Chen Y, Okoniewski M, Smyth GK, Huber W, Robinson MD. Count-based differential expression analysis of RNA sequencing data using R and Bioconductor. Nature protocols 8, 1765–86 (2013).

75. Newman AM, Steen CB, Liu CL, Gentles AJ, Chaudhuri AA, Scherer F, Khodadoust MS, Esfahani MS, Luca BA, Steiner D, Diehn M, Alizadeh AA. Determining cell type abundance and expression from bulk tissues with digital cytometry. Nat Biotechnol. 2019 Jul;37(7):773-782.

76. Szolek A, Schubert B, Mohr C, Sturm M, Feldhahn M, Kohlbacher O. OptiType: precision HLA typing from next-generation sequencing data. Bioinformatics. 2014 Dec 1;30(23):3310-6.

77. Litchfield K, Reading JL, Lim EL, Xu H, Liu P, Al-Bakir M, Wong YNS, Rowan A, Funt SA, Merghoub T, Perkins D, Lauss M, Svane IM, Jönsson G, Herrero J, Larkin J, Quezada SA, Hellmann MD, Turajlic S, Swanton C. Escape from nonsense-mediated decay associates with anti-tumor immunogenicity. Nat Commun. 2020 Jul 30;11(1):3800.

78. Reynisson B, Alvarez B, Paul S, Peters B, Nielsen M. NetMHCpan-4.1 and NetMHCIIpan-4.0: improved predictions of MHC antigen presentation by concurrent motif deconvolution and integration of MS MHC eluted ligand data. Nucleic Acids Res. 2020 Jul 2;48(W1):W449-W454.