NAR Genomics and Bioinformatics, 2025, lqaf026
https://doi.org/10.1093/nargab/lqaf026
Computational Methods for Transcriptome Analysis

OXFORD

Efficient and effective identification of cancer neoantigens from tumor only RNA-seq

Danilo Tatoni[1,2,*], Mattia Dalsass[3], Giulia Brunelli[1,2], Mario Chiariello[4], Guido Grandi[3], Romina D'Aurizio[1,*]

[1]Institute of Informatics and Telematics (IIT), National Research Council (CNR), 56124 Pisa, Italy
[2]Department of Medical Biotechnologies, University of Siena, 53100 Siena, Italy
[3]Department of Cellular, Computational and Integrative Biology (CIBIO), University of Trento, 38131 Trento, Italy
[4]Institute of Clinical Physiology (IFC), National Research Council (CNR) and Core Research Laboratory, Istituto per lo Studio, la prevenzione e la Rete Oncologica (ISPRO), 53100 Siena, Italy

[*]To whom correspondence should be addressed. Email: romina.daurizio@iit.cnr.it
Correspondence may also be addressed to Danilo Tatoni. Email: danilo.tatoni@gmail.com

Abstract

The growing accessibility of sequencing experiments has significantly accelerated the development of personalized immunotherapies based on the identification of cancer neoantigens. However, predicting neoantigens involves lengthy and inefficient protocols, which require simultaneous analysis of sequencing data from paired tumor/normal exomes and tumor transcriptome, often resulting in a low success rate. To date, the feasibility of adopting a more efficient strategy has not been fully evaluated. To this end, we developed ENEO, a computational approach to detect cancer neoantigens using solely the tumor RNA-seq data while addressing the lack of matched control through a Bayesian probabilistic model. ENEO was assessed on the TESLA benchmark dataset, reporting efficient identification of DNA-alterations derived neoantigens and compelling results against state-of-art exome-based methods. We further validated the method on two independent cohorts, encompassing different tumor types and experimental procedures. Our work demonstrates that a tumor-only RNA-based approach, such as the one implemented in ENEO, maintains accuracy in identifying mutated peptides resulting from expressed genomic alterations while also broadening the pool of potential neoantigens with RNA-specific mutations in a faster and cost-effective way. ENEO is freely available at the URL: https://github.com/ctglab/ENEO

Introduction

Somatic mutations responsible for the change of the aminoacidic sequence of a given protein are referred to as non-synonymous, and could lead to the generation of tumor-specific neoantigens [1]. The absence of these mutations from the healthy human genome makes these peptides not subjected to self-tolerance, increasing the attractiveness towards their use for the development of targeted immunotherapies [2].

Owing to the growing accessibility of high-throughput sequencing experiments, the development of individualized immunotherapies based on patient-specific neoantigens has been significantly sped up, showing promising results in multiple cancer types, such as melanoma [3, 4], gastrointestinal cancers [5, 6], and even particularly hard-to-target brain tumors as glioblastoma [7]. The engineering of these therapies requires the identification of cancer neoantigens with high binding affinity towards the patient-specific human leukocyte antigen (HLA) [1, 8]. The identification of patient-specific neoantigens is a critical step that typically starts by calling somatic variants using genomic data, from either the whole exome (WES) or whole genome (WGS) sequencing, obtained from the patient's tumor and healthy tissue [8]. The expression of the resulting aberrant transcripts must be then confirmed, requiring the concurrent profiling of tumor RNA-seq data [9]. The obtained mutations are then used to generate candidate peptides whose binding affinity towards major histocompatibility complex (MHC) is predicted using computational tools [8].

Although widely adopted, this approach is not exempt of limitations. Even though the genomic profiling of cancer mutations could lead to a notable set of predicted candidate neoantigens, only a minimal fraction of them may result in the effective stimulation of an immunogenic response in the patient [5, 6, 10]. Indeed, while WES opens to the high-throughput profiling of nearly all the open reading frames of the tumor genome, DNA variants located in untranscribed regions are not predicted to generate neoantigens [1]. Moreover, with the increasing adoption and optimization of protocols for profiling the MHC-I restricted immunopeptidome in cancer [11], it has become clear that DNA variants alone failed to fully explain the variety of detected peptides [12–14]. Conversely, the profiling of RNA-alterations using RNA-seq from the same cell lines [12, 14] or from the same tumor biopsy [14–16] empowered the spectra resolution, demonstrating how alterations detectable from tumor RNA offer a consistently better representation of the MHC-presented peptides space. These evidences were supported and enforced by the

Received: September 4, 2024. **Revised:** January 21, 2025. **Editorial Decision:** February 21, 2025. **Accepted:** March 11, 2025

recent work of Tretter and colleagues [17], which demonstrated, using a wide cohort of patients spanning multiple tumor types, that only a small fraction of immunogenic neoantigen detected by MS-based immunopeptidomics could be explained using DNA variants.

Detecting somatic alterations from the tumor RNA-seq alone has been proven to represent a feasible approach [18], even though not exempt of challenges [19]. Due to the nature of the RNA-seq experiment, selecting an appropriate matched control remains complex and not fully resolved [20]. Additionally, RNA-seq reads are more error-prone due to reverse transcriptase artifacts and to the splicing mechanism, which may result in alignment errors nearby splicing junctions [21]. Despite these obstacles, advancements in variant calling algorithms have driven the exploration of tumor-only RNA-seq methods for capturing tumor mutational burden and identifying driver mutations across various cancer types [18].

To date, the feasibility of adopting solely the tumor RNA-seq for the effective prediction of immunogenic neoantigens in a personalized immunotherapy scenario has not been fully evaluated. To this end, we developed ENEO, an easy-to-use computational workflow encompassing all necessary analytical steps, from variant calling to peptide-MHC (pMHC) binding affinity, in a reproducible and scalable manner. ENEO addresses the absence of a matched control sample by utilizing a Bayesian probabilistic model that takes advantage of genetic population databases. We demonstrate that ENEO effectively detects tumor neoantigens across various tumor types and experimental setups using publicly available data. These analyses open to the use of tumor RNA-seq as a rapid and cost-effective method for detecting tumor neoepitopes, enhancing the role of deep transcriptional profiling in precision oncology.

Materials and methods

Data used in this study

We downloaded matched WES data (tumor/normal) and tumor RNA data of five patients generated by the TESLA consortium [22], three of which with melanoma (namely TESLA_1, TESLA_2, and TESLA_3), and with non-small cell lung cancer (TESLA_12 and TESLA_16) from Synapse (syn21048999, access granted). Tested neoepitopes and their relative validation response were collected from the supplementary materials attached to the related publication [22]. RNA-seq data of the three melanoma samples from [23] were downloaded from SRA (SRP064661) and eight patients with gastric cancer from [5] and [6] from SRA (SRP278662). Within the last cohort, the selected patients are those with a single sequencing experiment carried out from the resected metastasis. Peptide validation results for these two cohorts, together with the detected somatic variants, were obtained from the retrospective dataset built by Gartner *et al.* [24].

Identification of germline and somatic variants from WES

WES data were preprocessed using *fastp* [25] (v.0.23.2) to remove sequencing adapter and low quality reads. Reads mapping was performed using *bwa* [26] (v.0.7.17) with default parameters and the GRCh38 human genome assembly as reference. Resulting alignment files were then processed following the GATK Best Practices [27]. Somatic variants were called from paired tumor/matched control with *Mutect2* [28] (v.4.2.0) and subsequently filtered as previously described [27]. Germline variants were called from control WES using *DeepVariant* [29] (v.1.6.0) using default parameters. To exclude variants falling in known challenging regions of the human genome [30, 31], we downloaded target lists for repetitive and homopolymeric regions from the genome in a bottle archive (https://ftp-trace.ncbi.nlm.nih.gov/ReferenceSamples/giab/release/genome-stratifications/v3.4/GRCh38@all/) and removed overlapping variants using *bedtools* [32] (v.2.30.0). Variant effect prediction annotation was performed on both the call sets using *vep* [33] with the Ensembl v.105.

Identification of candidate neoantigens from tumor RNA-seq with ENEO

We implemented a modular and reproducible computational pipeline within the Snakemake framework [34], named ENEO, to identify putative neoantigens using solely tumor RNA-seq data (depicted in Fig. 1). The overall workflow encompasses three main step which are described in full detail below.

Reads alignment and file processing

First, tumor paired end RNA-seq data are preprocessed to trim adapter sequences and low-quality reads using *fastp* [25] (v.0.23.2). To increase the read confidence, the base correction method of *fastp* is adopted. It corrects poor-quality bases (PHRED < 14) in a read if the other sequencing pair has a discordant base with excellent sequencing quality (PHRED > 30), given an overlapping region with a minimum length of 30 bases. Reads are then aligned with STAR [35], using the GRCh38 assembly genome and the Ensembl 105 annotation, with the "*2Pass-mode*," to increase specificity by generating sample-specific exon-splicing junctions and to enhance the overall mapping confidence. Aligned reads are further processed according to the GATK Best Practices [27] for short variants discovery from RNA-seq. In short, duplicate reads are flagged using *MarkDuplicates* (v.4.2.0) to discard PCR and optical duplicates; aligned reads are then passed through *SplitNCigar* (v.4.2.0), which is responsible for the splitting of those containing "N" in their CIGAR string and for the hard-clipping of overhanging regions containing mismatches (a common scenario across exon junctions); the base quality is then recalibrated using Base Quality Score Recalibrator (v.4.2.0) to adapt for the subsequent variant calling step.

Variant calling, filtering, and annotation

As described in [17] the variant calling is performed with *strelka2* [36] using the mutation calling procedure specific to RNA (specified with the argument "*rna*"). Despite being designed originally for germline variant calling, *strelka2*'s relaxed modeling of genotypes enables the inclusion of the plethora of different measured allele frequencies coming from somatic alterations of different clonality from RNA-seq as showed by Tretter *et al.* [17]. Variants are called over a provided set of protein-coding exons, as obtained from the Ensembl v.105, after the exclusion of the genes belonging to the antigen processing and presentation. The list of excluded genes and their relative chromosomal regions is reported in the Supplementary Table S1.

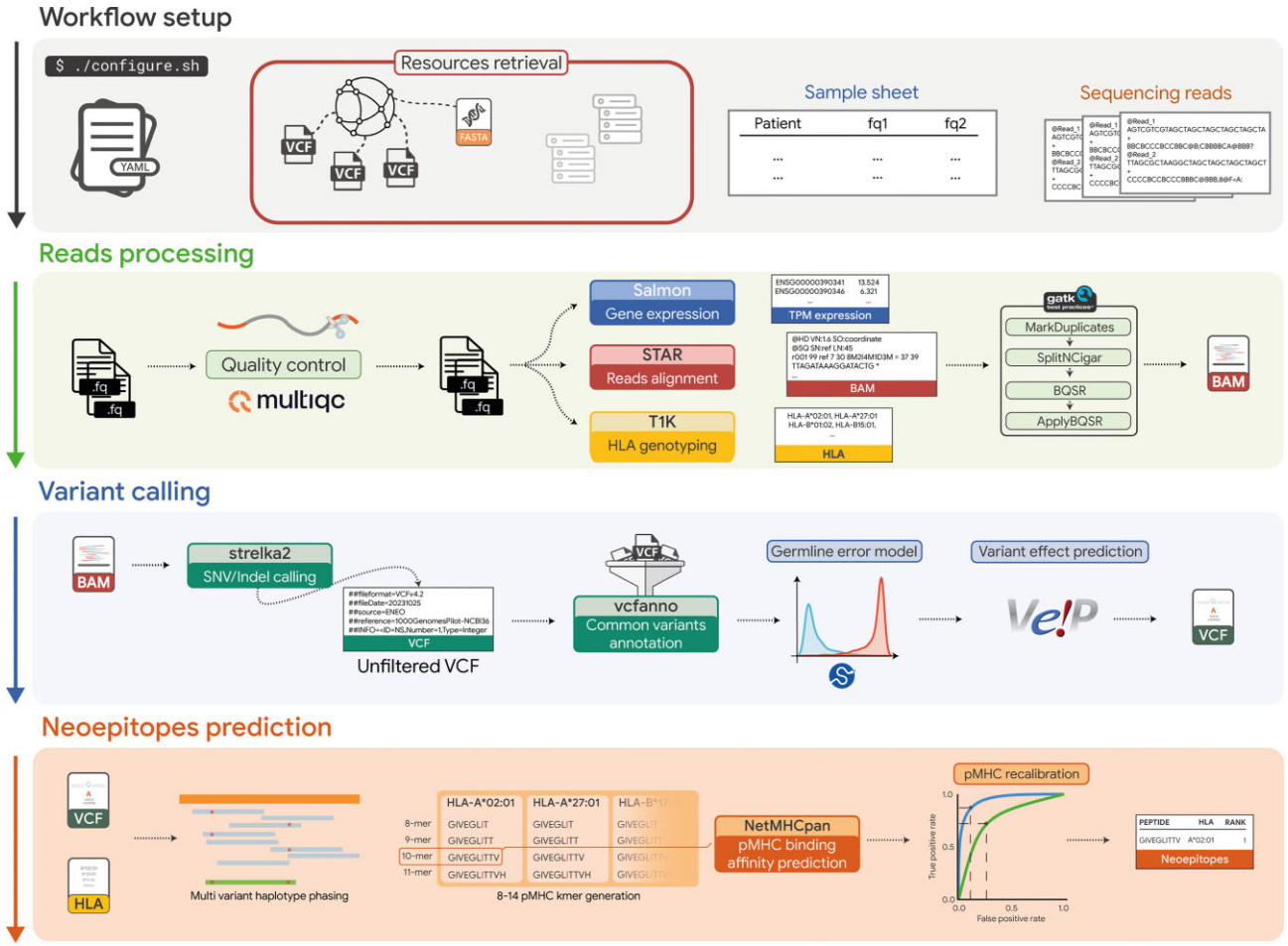

Figure 1. Overview of the ENEO computational workflow. The pipeline is composed of three main phases: reads processing (first step), variant calling (second step), and neoepitopes prediction (third step). Input files required by the workflow are indicated on the top gray box, together with the employed public resources whose first download and setup is guided. Intermediate post-processed output files (i.e. alignment (BAM), variant calling file (VCF), HLA genotyping) used throughout the pipeline are saved and made available for further user inspection.

We added multiple filtering criteria to account for intrinsic technical noise arising from the tumor-only RNA-seq assay and discarded:

- variants in region with recurrent technical artifacts, as determined by the Panel of Normals built on the 1000 Genomes project, retrieved from the Broad Institute cloud bucket (gs://gatk-best-practices/somatic-hg38/1000g_pon.hg38.vcf.gz);
- variants potentially affected by sequencing leakage error, defined as variants whose alternative allele matches at least three bases within a genomic window of 7 bp centered on the variant;
- variants not covered by at least five reads and with fewer than three reads spanning the alternative allele;
- variants falling in pseudogenes and non-coding regions.

Furthermore, as in the WES case, variants falling in known error-prone regions (as homopolymer and repetitive regions) were filtered out from the calling set. To report common germline events, candidate variants are annotated with *vcfanno* [37] (v.0.3.5) using multiple population databases: exAC [38](r.1), gnomAD [39] (v.3.1), and the ALFA project [40]. The obtained allele counts and frequencies are weighted using geometric average, to account for the different sample size in each dataset. Variants' consequence annotation is performed with VEP (v.105), using additionally the plugins *Wildtype* and *Frameshift* to report the protein sequence within the VCF file.

Definition of the Bayesian probabilistic germline error model

In the current restricted scenario without a matched-normal sample, germline and somatic variants are not easily distinguishable. To overcome this limitation, we built and tested a Bayesian probabilistic model that leverages, when available, the measured population allele frequency retrieved from germline resources (i.e. ExAC, gnomAD, and the ALFA project), here denoted as f, to derive the likelihood of a called variant being germline in the diploid scenario. In contrast, whenever f is not known, we adopted the same approximation introduced by Benjamin and colleagues in the Mutect2 paper [28]. Briefly, the number of variant alleles n in a germline resource composed by K patients, with $N = 2K$ chromosomes, could be modeled as coming from a binomial distribution as $n \sim Binom(N, f)$. Supposing a prior for the frequency f to be $f \sim Beta(\alpha, \beta)$, centered on the average human heterozygosity $\mathbb{E}(f) \approx 10^{-3}$, the probability of sampling a site not found in

the germline resource is the result of a beta-binomial process as $P(n = 0) = BetaBinom(0|\alpha, \beta, N)$. This could be approximated to the measurable number of exonic sites not mutated, resulting in $P(n = 0) \approx 7/8$. Updating the Beta prior on f, it results in a posterior $f \sim Beta(\alpha, \beta + N)$ whose mean value is approximated to α/N, given that $N \gg \beta$. Using the gnomAD 3.1 reference ($N = 71{,}702$) we obtained a mean value approximal to $7e^{-8}$, which is used as f for variants not found in the germline resource. Considering the diploid scenario, the unnormalized probability of three main events is defined as follow:

$$\begin{aligned} P((0|1)) &= L_{(0|1)} f(1-f) \\ P((1|1)) &= L_{(1|1)} f^2 \\ P_{\text{som}} &= \pi \end{aligned} \tag{1}$$

Where:

- $L_{(0|1)}$ and $L_{(1|1)}$ represents the likelihood emitted by *strelka2* for the given variant, respectively, for the heterozygous and homozygous genotype.
- f is the population allele frequency derived from the population databases as previously described.
- π is the somatic occurrence ratio.

The parameter π inside the probability estimation represents the likelihood of observing a somatic event in a position within an exon of a protein-coding gene, without accounting for the given clonality and/or ploidy. We modeled it as the results of two subsequent Bernoulli processes, where the first represents the successful event of finding an exon E that contains at least a variant, and the second as the successful event of finding a mutated base B out of all the bases n of that exon. Considering each exon E as independent each other, finding a mutated exon E_{MUT} is dependent upon the probability p_E and could be defined as:

$$E_{\text{MUT}} \sim \text{Bernoulli}(p_{\text{E}})$$

Within an altered exon E_{MUT}, considering each of the composing bases as independent each other, the probability p_B of picking the position borrowing the variant is defined as the conditioned probability

$$B_{\text{MUT}} = P(B|E_{\text{MUT}}) \sim \text{Bernoulli}(p_{\text{B}})$$

This entangles the model as dependent on two distinct probabilities which, to encompass the different mutation susceptibility observable throughout the human genome, could not be coerced to fixed values. We then suppose an independent prior Beta distribution for each, defined as:

$$p_{\text{E}} \sim \text{Beta}(\alpha_{\text{E}}, \beta_{\text{E}})$$

$$p_{\text{B}} \sim \text{Beta}(\alpha_{\text{B}}, \beta_{\text{B}})$$

To estimate the parameters of these two Beta distributions, we collected somatic variants of 2,683 cancer patients from the PCAWG consortium through cBioPortal (cbioportal.org). We retained only those alterations falling within the regions used by ENEO for variant calling to ensure consistency. To speed-up model convergence, we excluded patients with <50 somatic calls, resulting in a final set of 1,523 patient across 26 tumor types. From each patient we collected the observed probability of a mutated exon $\theta_{\text{E}} = E_{\text{MUT}}/E_{\text{TOT}}$ and the observed probability of a mutated base $\theta_B = B_{(\text{MUT}|E_{\text{MUT}})}/B_{(\text{TOT}|E_{\text{TOT}})}$ for all the exons. Given that both the parameters of a Beta distribution must fall in $\mathbb{R}^+$, we assigned non-informative Gamma distribution as priors

$$\alpha \sim \text{Gamma}(0.1, 0.1)$$

$$\beta \sim \text{Gamma}(0.1, 0.1)$$

Non-informative priors were chosen to do not enforce beliefs about the parameter values, allowing the data to inform the posterior estimates. We used the Metropolis–Hastings algorithm [41] to approximate the posterior distributions of α and β based on the collected data. The algorithm was run for 5,000 iterations with a burn-in period of 1,000 iterations to ensure convergence and reduce the influence of initial values. The highest probability density (HPD) estimated for each parameter was then used to update the respective parameters of both the Beta distributions. This approach allows us to identify the most likely parameter values given the observed data. Model inference was conducted using the *scipy* [42] library (v.1.12) in Python (v.3.10).

Power analysis for variant detection assessment

To assess the performance of the proposed approach in calling somatic variants, in terms of both precision and recall, we reciprocally compared the variant sets coming from the two assays (DNA or RNA) using the TESLA cohort. To account for specific constrains and make the two assays comparable, we conducted a power analysis as in [18] to identify the assay-specific truthset (named "powered set"). Given a variant in one assay (DNA or RNA), whose alternative read count is x and reference read count is y, we computed the power to detect the variant in the other assay (RNA or DNA) given a coverage of N. To compute the probability of observing at least k reads, we employed the survival function S of a Beta-binomial model, defined as

$$S(k \mid x, y, N) = \sum_{i=k+1}^{N} P(i \mid x, y, N)$$

where P is the probability mass function of the Beta-binomial, defined as

$$P(i \mid x, y, N) = \binom{N}{i} \frac{B(i+x+1, N-i+y+1)}{B(x+1, y+1)}$$

where B is the Beta function. To obtain the minimum number of reads k, we took into account the error rate r at a given genomic site, obtained as the maximal allele fraction of the three possible alternative alleles, applying a Laplace correction of 1. The minimum number of reads k is then defined as the number of alternate reads that have a probability <1% of being generated by noise, as the result of a Binomial model employing the computed error rate r.

$$k = \min \left\{ x \in \mathbb{N} : \sum_{i=0}^{x} \binom{N}{i} r^i (1-r)^{N-i} < 0.01 \right\}$$

We then proceed considering a variant from one assay (RNA or DNA) to be powered in the other assay (DNA or RNA) if the probability P was greater than or equal to 0.8, given a number of reads N. The nucleotide coverage across all the variant regions was collected using *perbase* [43].

Gene expression quantification and HLA typing

Transcript expression quantification is conducted using *salmon* [44] (v.1.9.0), and subsequently summarized at gene-level transcript per million (TPM) using the R package *tximport* [45](v.1.30). HLA genotyping is performed with T1K [46] (v.1.0.1) using raw reads (FASTQ) as input. Only the genotypes for the HLA-(A|B|C) loci with the highest coverage and likelihood are kept as input for the latter stage.

Prediction of neoantigens binding affinity

Annotated variants and patient HLA genotypes are fed for the last step of ENEO, which is responsible for the generation of candidate neoantigens and the prediction of their binding affinity. To increase the specificity of predictions, we used the haplotype phasing information to obtain multivariant phased representations of the altered protein, as reported previously [47]. From the resulting mutated aminoacidic sequence, all k-mer peptides (from 8 to 14aa), which include the mutated position, are extracted by moving a sliding window through the sequence. These peptides, along with each of the patient's HLA alleles, are then passed to the NetMHCpan 4.1 [48] algorithm, which predicts the binding affinity of the peptide to the HLA molecules. This results in a list of pMHC complexes, with the binding likelihood expressed as a percentile rank against a set of random peptides. To obtain an high confidence value specific for a given allele, instead of adopting a fixed acceptance boundary for the percentile rank, we resembled the approach adopted by Reardon and colleagues [49]. After obtaining pMHC data from the IEDB [50] v3 (iedb.org) and training data of NetMHCpan4.1 from the supplementary materials of the publication [48], we generated the negative background peptide dataset by randomly slicing k-mer peptides within the length range of 8–14 from the human RefSeq proteome (GCF_000001405.40). Then, after removing pMHC pairs already present in the training data, we kept positive and negative pairs until reaching a proportion of positive:negative of 1:9. To ensure minimum confidence, we proceeded using only alleles with at least 100 experimentally validated unique peptides.

For each allele, we used the peptides' predicted percentile ranks by NetMHCpan4.1 and the corresponding labels to obtain the optimal percentile threshold, defined as the value that maximizes the difference between the true positive rate (sensitivity) and false positive rate (1—specificity). This resulted in the identification of HLA-specific optimal values for 114 different HLA alleles, which are used to filter the set of candidate pMHCs arising from the detected variants. To reduce biases due to the choice of a fixed threshold, we repeated the analysis using increasing positive-to-negative ratios ranging from 1 to 20 and computed the optimal percentile for each scenario. We then assessed the statistical significance of the differences using a one-tailed *t*-test followed by Bonferroni correction. We found that, for all but one of the alleles (HLA-C*03:04, adj-p = 0.01), the obtained threshold for the percentile was not influenced by the choice of the positive-to-negative ratio. The resulting percentile values are reported in Supplementary Table S5. Natural non-malignant HLA ligands, potentially resulting from healthy tissues and thus subject to self-tolerance, are removed using the data collected from the HLA Ligand Atlas [51] (release 2020.12). Resulting neoantigens are then ranked according to their predicted percentile rank and reported in the ENEO output along with the expression of the mutated gene, allowing for optional additional ranking strategies.

Results

Comprehensive identification of cancer neoantigens from tumor-only RNA-seq with ENEO

To identify and prioritize putative candidate neoepitopes using solely the tumor transcriptome, we developed ENEO, a scalable, reproducible, and extensible computational workflow implemented using the Snakemake [34] framework. ENEO's workflow encompasses three main steps (i.e. "reads processing," "variant calling," and "neoepitopes prediction"), whose graphical representation is depicted in Fig. 1 and detailed in the "Materials and methods" section. The first step is responsible for the reads pre-processing, mapping, and alignment post-processing, accomplishing three main tasks: the gene expression quantification, the delivery of a processed alignment file compliant with GATK Best Practices, and the HLA genotyping. The second step implements Strelka2 for variant calling, the ENEO's germline error model for somatic variants likelihood estimation and the functional annotation with VEP tool.

In the absence of a matched normal sample, discriminating between germline and somatic variants is crucial. Here, we present a Bayesian probabilistic model that derives the likelihood of observing a somatic event using either the calculated or estimated germline population frequency and the genotype likelihood derived from sequencing data. In our model, we assumed that the likelihood of identifying a somatic variant within a specific coding exon is contingent upon two Bernoulli trials. The first involves sampling a mutated exon out of the pool of expressed protein-coding exons, while the second regards sampling the mutated nucleotide. We assumed that the probabilities of both events follow a Beta distribution, characterized by two shape parameters (i.e. α and β). Using data from 1,523 patients across 26 tumor types from the ICGC (see the "Materials and methods" section), we approximated the posterior distributions for these parameters using the Metropolis–Hastings algorithm. For the exon mutation probability, the posterior mode for α and β were found to be 0.87 ([0.78, 0.96] HPD interval) and 612.23 ([532.92, 702.62] HPD interval), respectively. Conversely, we obtained 0.95 ([0.94, 0.96] HPD interval) and 199.58 ([197.98, 203.61] HPD interval) as the posterior mode for α and β of the probability of sampling the mutated nucleotide. The posterior density plot confirms that our model proposes for both a highly skewed Beta distribution, that fits accurately the observed data distribution (Supplementary Fig. S1).

Probabilities sampled from the target posterior distributions are used, together with the population allele frequency and the variant genotypes' likelihood (methods), to annotate variants in the resulting VCF file with the estimated probabilities. The output of the second step is a VCF file with the full set of variants along with their variant effect prediction, used to determine the resulting mutated protein sequences. Lastly, in the third step, all variants are then used for generating candidate peptides and, coupling with the HLA genotype, the pMHC binding affinity is then predicted using NetMHCpan 4.1. The final output of ENEO is a ranked list of patient's specific candidate neoepitopes arising from

RNA mutations. ENEO is freely available atthe URL https://github.com/ctglab/ENEO.

Efficient disjoin of germline variants from the tumor-only RNA callset through ENEO germline error model

We evaluated the ability of our model to identify most of the germline variants present in tumor cells, given the absence of the genetic profile of matched non-tumor cells, by comparison with the standard calling approach which exploits both the DNA sequencing of tumor and paired non-tumor samples (T/N WES). To accomplish that, we used data provided by the TESLA consortium [22]. These data, indeed, include exome and transcriptome sequencing of cancer biopsies collected from five patients with melanoma and non-small cell lung cancer, along with exome sequencing from PBMCs as healthy controls (methods).

At first glance, we evaluated the sensitivity of the implemented germline error model in correctly distinguishing the origin of reported variants, as either germline or somatic from tumor-only RNA-seq. To do so, we called germline alterations from control WES using DeepVariant (Methods), somatic alterations from T/N WES using Mutect2 (Methods), and finally annotated them to keep only those with a predicted effect on downstream protein. Despite the different algorithms employed, no overlap was observed between WES-derived germline and somatic calls across all patients (Supplementary Fig. S2). We identified a total of 3,850 non-synonymous somatic variants across all patients (ranging from a minimum of 120 to a maximum of 2,182) with melanoma patients showing an overall higher number of variants (Supplementary Fig. S2), in line with the expected burden of the tumor types under investigation [52]. Within the intersection between WES-variants and RNA-variants from ENEO, we assessed the sensitivity of the germline error model. Starting from a set of 15,508 RNA-variants detected also from germline WES analysis, the adoption of a probability cutoff of 0.5 resulted in the successful labeling of 15,143 (∼97.64%) non-synonymous germline variants, showing consistent performance for all the patient under investigation (Fig. 2A). After filtering, we retained a total of 2,959 somatic variants with a predicted impact on the downstream protein sequence across all patients. These variants were predominantly transitions (C>T and G>A for melanoma, A>G and T>C for NSCLC, Fig. 2B), aligning with the expected alterations' frequency of these tumor types [52] (Fig. 2B).

As the reported calls are expected to include somatic DNA alterations, detectable also from the T/N WES analysis, and RNA-specific alterations [17], undetectable from genomics data, we quantified and described the overlap between the two call sets. As expected, and in accordance with published studies [17–19, 53], significant discrepancies were found at the sample level (Supplementary Fig. S2). Indeed, in both assays, the variant calling is strongly influenced by intrinsic features and technical limitations, which makes the comparison a non-trivial task. The main reason for the missing detection of a DNA alteration in the RNA is the insufficient sequencing coverage resulting from low expression levels. Conversely, in WES, the covered regions are limited by the enrichment probe kit in use, and even within them, the sequencing depth may be extremely variable and insufficient to drive somatic variant calling [54]. To gain more insights into these discrepancies, we employed a power analysis with a binomial model (methods) to determine whether, given an assay (RNA or DNA), an alteration could be detectable.

Out of the total 1,633 non-synonymous WES variants detectable from RNA-seq (Fig. 2C, "Powered in RNA"), ENEO successfully identified 1,377 of them, resulting in an average recall of 0.84% (±0.034, Fig. 2E). The remaining alterations not identified by our pipeline (Supplementary Fig. S3A) include missed calls or the rejected from the overly stringent filtering operated by Strelka2 (217, 13.28% of the total), mislabelled germlines called by the error model of ENEO (26, 1.59% of the total) and variants rejected by the RNA-leakage filter (13, 0.79% of the total) (Supplementary Fig. S3B).

Conversely, to further investigate the specificity of the RNA variants identified by ENEO, we focused on those not already identified as editing events in the REDIportal (resulting in a total of 2,452). Also in this case, we performed a power analysis in DNA (methods) to identify sites with sufficient coverage to be detectable (Fig. 2D, "powered in DNA"). They represented the 74.4% (1,824) of the total RNA variants, and 75.5% of them (1,377) were indeed detected in the WES analysis (Supplementary Fig. S4A), achieving a mean precision of 0.62 (±0.18, Fig. 2E). In the remaining set of RNA-exclusive alterations reported by ENEO, we could confidently report as wrong calls (i.e. false positives) those that were identified as germline in the WES analysis (251, 10.2% of the total). As the detected somatic alterations from WES were not subjected to orthogonal validation with other assays, we examined the depth of the alternative allele in DNA to determine how many RNA alterations were not called in WES-powered regions due to technical limitations of the DNA method. Interestingly, we found that for the remaining 218 alterations, there is sufficient evidence of the alternative allele in DNA (depth ≥4, mean = 115), and a significant fraction of them (38%, 83) are reported in COSMIC as known somatic alterations in tumors (Supplementary Fig. S4B). Furthermore, for these alterations, the empirical allele fraction computed in DNA showed a significant correlation with the VAF obtained from RNA ($r^2 = 0.477$, $P = 3.014 \times 10^{-32}$), further supporting the plausible existence of these alterations in cancer cells (Supplementary Fig. S4B). To rule out sequencing coverage as the main cause of the missing calls [55], we compared both WES variants and RNA variants by the corresponding gene expression levels. This analysis revealed significantly lower expression levels for genes harboring variants undetected by RNA-seq, compared to those detected by both methods (Fig. 2F, Kruskal–Wallis test, $P < 1e-5$).

The tumor-only RNA-seq approach favors the identification of immunogenic neoantigens

The simultaneous detection of germline and somatic events enables the deciphering of the alterations at the transcription level, leveraging both those occurring in healthy and tumor cells. To this end, ENEO exploits the haplotype information coming from the variant calling process to phase multiple variants across the same haplotype and to generate mutated proteins (methods). Using the previously described RNA-variants from TESLA patients, we generated all the possible 8–14mer mutated peptides and, coupling with patient HLA genotype, we predicted the binding likelihood as the elution rank percentile [48] using NetMHCpan 4.1 (methods). As demonstrated by Reardon and colleagues [49], adopting the same

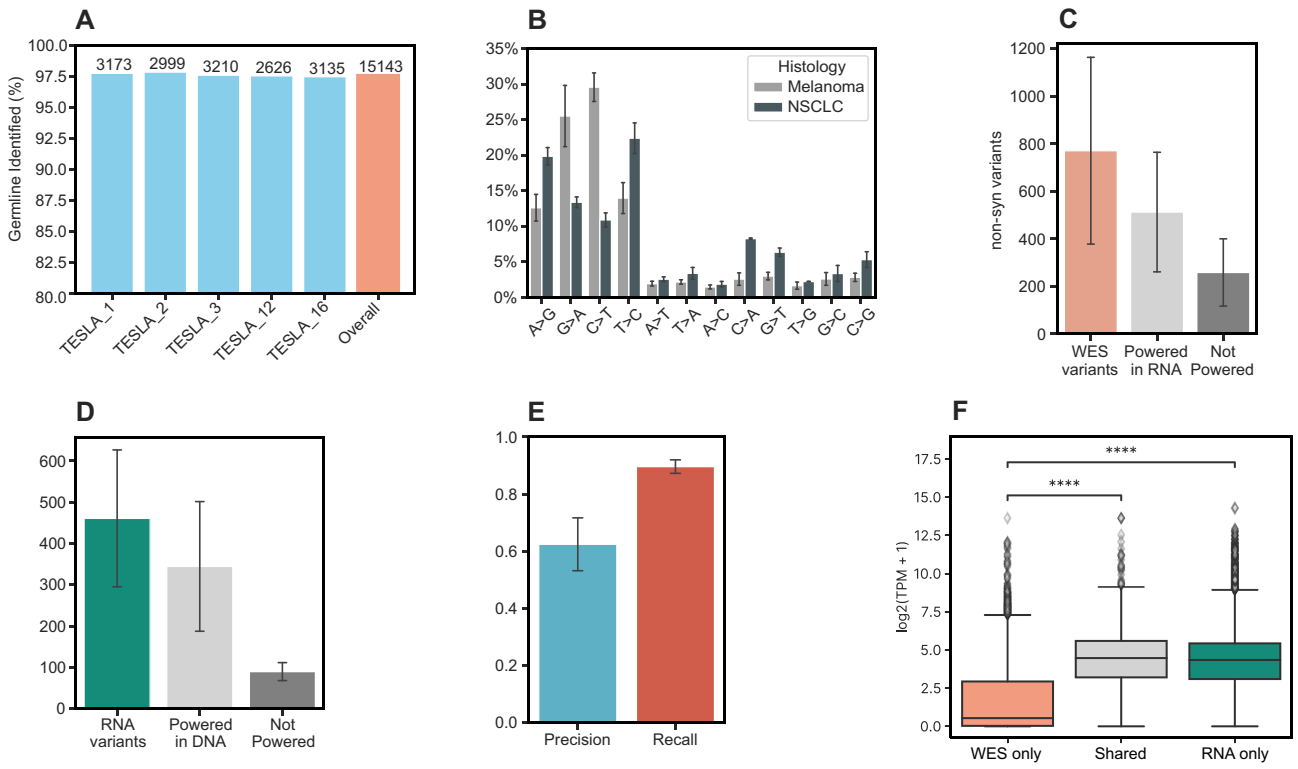

Figure 2. Characterization of somatic variants identified by ENEO using tumor-only RNA-seq in comparison with variants identified from WES in the TESLA cohort. (**A**) Percentage of non-synonymous germline variants in patients of the TESLA cohort correctly identified by the germline error model of ENEO. (**B**) Distribution of ENEO non-synonymous single nucleotide variants (SNVs) within each single nucleotide variant type. Percentage data from each histology under investigation are aggregated into the same bar. (**C**) Barplot showing the average number of non-synonymous variants detected from WES samples (*WES variants*), the average number of those detectable in RNA (*powered in RNA*), and those not detectable from RNA (*not powered*). Error bars represent the standard error of the mean. (**D**) Barplot showing the average number of non-synonymous variants detected in RNA by ENEO (*RNA variants*), the average number of those detectable in tumor WES (*powered in DNA*), and those not detectable in tumor WES samples (*not powered*). Error bars represent the standard error of the mean. (**E**) Calculated average precision and recall of ENEO in detecting powered non-synonymous variants from tumor-only RNA-seq. (**F**) Expression level of genes reporting WES-specific variants (*WES only*), RNA-specific (*RNA only*), or common calls in both assays (*shared*). Expression reported as TPM. Kruskal–Wallis test. $****P \leq .0001$

threshold for elution percentile over multiple HLA to delimit the list of candidate neoantigens may result in loss of specificity or sensitivity. Thus, we computed HLA-specific optimal elution percentile thresholds over >100 HLA alleles using IEDB data (methods) and used them to select the set of candidates pMHCs for each patient. This led to the generation of a total of 5,172 pMHCs across all the cohort, resulting in 4,883 candidate peptides (Supplementary Table S2).

We measured the sensitivity of our approach in identifying TESLA tested immunogenic peptides. Notably, applying ENEO to the tumor RNA-seq data of patients resulted in the successful identification of 26 out of the 34 positive tested immunogenic peptides, with variable performances between patients and respective histology (Fig. 3A). In melanoma cases (TESLA_1, TESLA_2, and TESLA_3), ENEO showed a higher recall, reporting 23 out of the 26 immunogenic peptides (Fig. 3A). Conversely, we observed a lower resolution for the two NSCLC patients under investigation (TESLA_12 and TESLA_16), where three out of eight validated peptides were identified, reporting anyhow at least one immunogenic neoantigen even in the worst performing scenario. Interestingly, ENEO did not select 246 of the 501 (~49.3%) originally predicted neoepitopes, which were assayed as not immunogenic. This resulted in a set of neoepitopes that exhibited notable sensitivity, as indicated by the recall over increasing fractions of the proposed list (Fig. 3B). We compared our approach with various standard methods used by teams in the TESLA consortium, adopting the metrics defined in the original publication. These were defined as top 20 immunogenic fraction (TTIF) and fraction ranked (FR) and are used for comparisons within individual patients. The first is a proxy of the specificity, as it approximates the number of immunogenic peptides in the first 20th positions of the ranked list, considering 20 as an adequate testing size for therapeutic application [22]. The second could be intended as a sensitivity metric, reporting the number of immunogenic peptides over the total positive tested in the first 100th positions of the ranked list. All the participant teams used T/N WES for variant calling and, consequently, the list of experimentally tested peptides is expected to come from variants detectable from WES. ENEO outperformed on average other teams for all melanoma patients in both FR and TTIF (one-sided t-test, $P < 1e-4$; Supplementary Fig. S5). For NSCLC patients, ENEO demonstrated superior FR for patient TESLA_16 than average (one-sided t-test, $P < 1e-4$; Supplementary Fig. S5), while for patient TESLA_12, the result was not significantly different from other teams' submissions on average (two-sided t-test, $P = .5$; Supplementary Fig. S5). Of note, in two patients we identified a previously characterized cancer neoantigen (RVWDVSGLRK) with strong predicted binding affinity, known to come

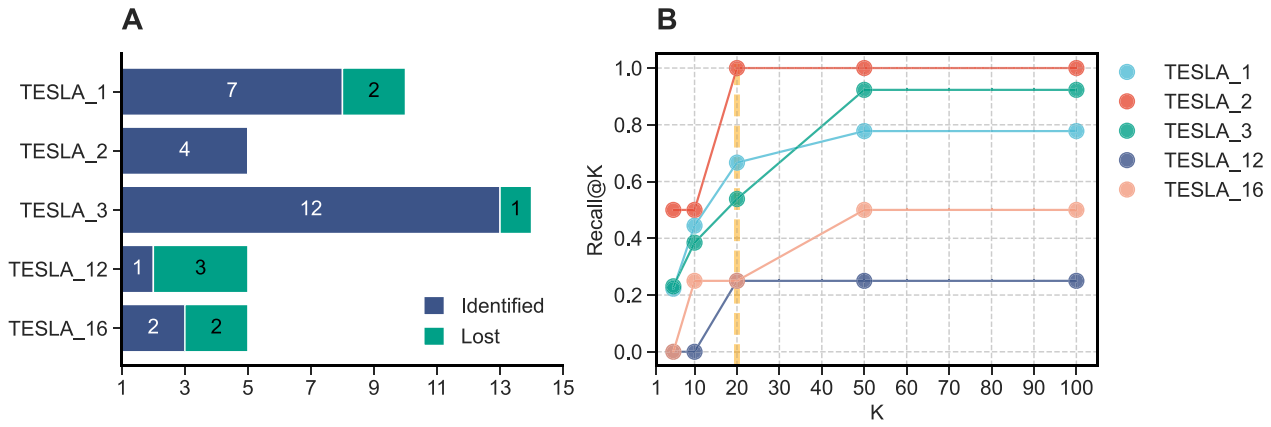

Figure 3. Performances on immunogenic pMHC identification from the TESLA benchmark. (**A**) Number of positive experimentally validated pMHC detected or non-detected by ENEO predictions. (**B**) Patient-wise recall-at-K, defined as the fraction of immunogenic pMHCs identified for each patient (recall) over increasing fractions (k) of the proposed ranked list. The scattered line is fixed at the 20th position, preserving the choice adopted in the original TESLA publication.

from an A-to-I RNA editing event occurring in the protein COPA in the 164th aminoacid, responsible for a substitution of an isoleucine with a valine. This event was recently proposed as a driver of metastasis in colorectal cancer through ER stress [56], and was eluted in melanoma samples following MHC-purified mass spectrometry and demonstrated to induce IFN_{γ} secretion [57, 58].

Neoantigen prediction is consistent across different tumor types and mutational burden levels

To verify the robustness of our approach, we collected tumor RNA-seq from two publicly available cohorts that applied a similar screening protocol in two different tumor types. The first one, here referred to as "Gros," examined the antigen specificity of T-cell responses detected from peripheral blood lymphocytes in three metastatic melanoma patients [23]. The second one, here referred to as "NCI" (National Cancer Institute), is composed of patient with metastatic gastrointestinal cancers characterized by microsatellite stability and mismatch repair proficiency [5, 6]. In this group, neoantigen specificity was determined from the T-cell response of cultured tumor infiltrating lymphocytes. These two studies shared the screening protocol, consisting in somatic variant calling from T/N WES and, after transcript expression confirmation and patient-specific variant selection, screening via synthesis of 25-mer constructs and transfection in autologous dendritic cells using tandem mini genes [59].

For the Gros cohort, we processed tumor RNA-seq reads with ENEO and merged the resulting candidate pMHCs into 25-mer frames, resulting in the production of 612 constructs (Supplementary Table S3). Our analysis identified 30, 31, and 22 out of all tested 25-mers, successfully detecting six out of the seven positively tested 25-mers. Furthermore, we examined the minimal epitope immunogenicity of positively tested 25-mers to determine how many positive pMHCs were reported by ENEO and their relative rankings. We identified eight out of nine positive pMHCs, with at least one immunogenic peptide detected per patient (Fig. 4A). Importantly, among the tested constructs, immunogenic neoantigens were placed by ENEO within the top 30 ranked pMHCs (Fig. 4B). Owing to the patient-specific ranked list, all the positive tested pMHCs were at least included in the upper half (i.e. 50th percentile) of the candidates (Fig. 4B). Notably, the peptide FVVPYMIYLL, immunogenic for patient Mel_3998, derives from two phased somatic variants (p.Y1000F and p.H1007Y) in the PDS5A gene. Its detection was possible due to the retention of haplotype information by the variant calling process and its use in subsequent variant peptide generation (methods).

Given the availability of somatic variants profiled by authors via T/N WES, we could investigate origin of the missed neoantigen. The variant underlying the peptide KVDPIGHVY, responsible for the p.E168K transition in MAGEA6 (a known melanoma-associated antigen [60]), was discarded due to insufficient coverage (<5 RNA-seq reads). This is in accordance with Gartner *et al.*'s findings [24], which reported no distinct evidence for the alternative allele in the transcriptomics data. As the experimental procedure followed by authors was developed on WES variants, RNA-exclusive variants detected by our approach were not subject to the original immunological screening. We reported an average of 147 25-mer constructs generated from just as many RNA-exclusive variants, which leads to the proposal of an average of ~301 candidate pMHCs for each patient (Supplementary Table S3).

For the "NCI" cohort, we resembled the experimental procedure adopted for the "Gros" one. This group of patients represents an even more challenging experimental setup for testing the robustness of ENEO, as these tumors have low burden of somatic mutations and are less responsive to checkpoint inhibitor therapy [61]. Tumor RNA-seq analysis led to an average of ~271 constructs, with an extended difference between patients (Supplementary Table S3). We successfully identified 11 out of the total 13 positively tested 25-mer, detecting at least one immunogenic construct for all but one patient (Fig. 4A). In the ranked list of pMHCs generated out of the tested 25-mers, at least one immunogenic neoantigen was found within the top 10 ranked entries, corresponding to the first 20th percentile, for seven out of the nine patients under investigation (Fig. 4B). ENEO remained robust to the heterogeneity of this dataset, in term of sequencing platform, read length and, most importantly, to the overall number of reads (Supplementary Table S4). As somatic variants profiled via T/N WES are available [6], we investigate events

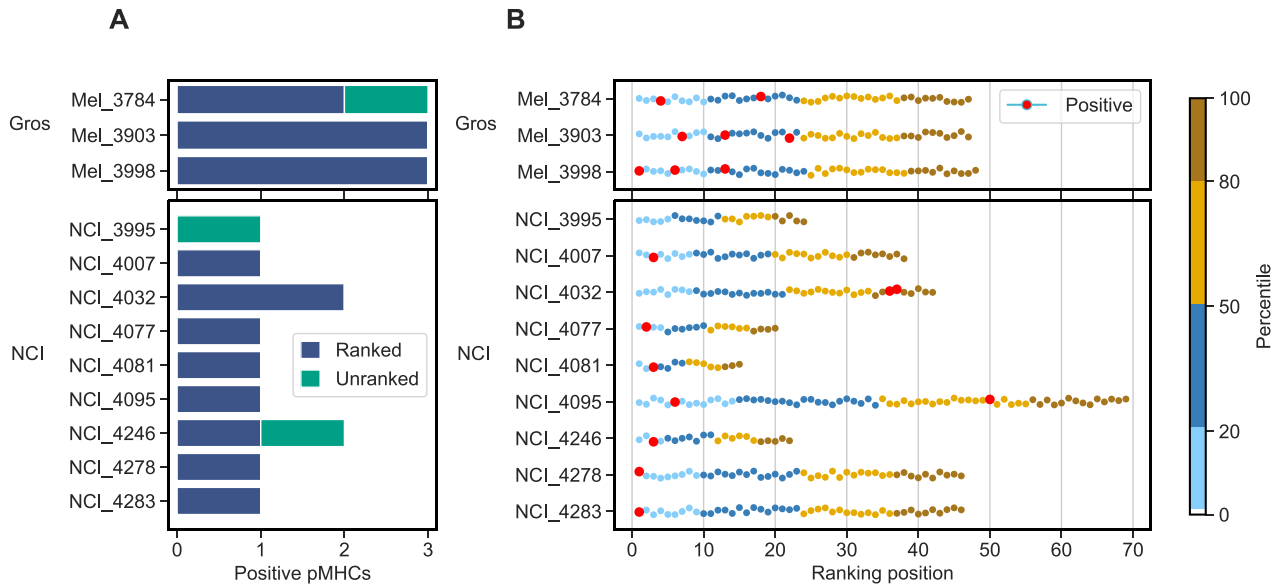

Figure 4. Identification of pMHCs in external cohorts. (**A**) Number of positive pMHCs identified by ENEO for each patient. (**B**) Patient-wise ranking of candidate pMHCs for tested 25-mer identified by ENEO, colored by their percentile among the ranked list.

behind missing neoantigens. For the patient NCI_3995, the p.G12D transition in KRAS detected from T/N WES is reported as the source for the single immunogenic peptide reported. In the RNA-seq data we failed to observe consistent evidence for the mutated transcript, as the mutated allele is supported by only two reads out of the total eight mapped in this region. A similar scenario is reported for the patient NCI_4246, where the unidentified peptide arises from a SNV in the gene ARMC9, encoding for a melanocyte-specific antigen (KU-MEL1). For this variant, we found that in the RNA-seq data the mutated allele is supported only by two out of the total six reads mapped. In both cases, the alteration is filtered out during the variant calling process due to the low coverage, resulting in a poor genotype quality. For this group of patients, as for the Gros cohort, the analysis of RNA-seq data leads to the proposal of 264 untested candidate 25-mers, originating from the same number of RNA-specific variants. The binding affinity estimation and subsequent filtering of ENEO results in the generation of ~1,150 candidate pMHCs (Supplementary Table S3).

Taken together, these results indicate that the tumor-only approach implemented in ENEO permits the identification of cancer neoantigens in different histological and technical conditions, without requiring neither a matched control nor concurrent DNA-sequencing.

Discussion

In this study, we investigated the feasibility and limitations of predicting immunogenic neoantigens using only the tumor RNA-seq data. We compared our approach with state-of-art methods that additionally require the tumor/normal (T/N) DNA sequencing on a benchmark dataset, and validated it on external cohorts encompassing different tumor histologies and experimental setups. We accomplished this by developing ENEO, an easy-to-use and automatized computational workflow designed to predict tumor neoantigens using RNA-seq data without a matched-control sample. Being built with the Snakemake workflow, it is designed to be modular, extensible, and scalable to different platforms and infrastructures. Importantly, ENEO addresses the challenge to distinguish germline variants in the absence of a matched control sample through a Bayesian probabilistic model. The model leverages variant population allele frequencies and genotype likelihood and proves to be able to remove most germline variants out of the detected RNA-variants in the TESLA cohort. We tested its ability to detect immunogenic neoantigens in public studies, encompassing different tumor types and experimental validation, proving its clinical utility even in very low mutated tumors like MMR-proficient gastrointestinal cancers. Harnessing the growing evidence pointing to the contribution of transcriptional specific alterations to the plethora of MHC-presented peptides, we demonstrated that a tumor RNA-based approach like the one implemented in ENEO preserves resolution on canonical mutated peptides originated from expressed genomic alterations, proposing at the meanwhile candidate pMHCs specifically arising from RNA-specific mutations. Despite the encouraging performances, the reported approach is not exempt of limitations. Due to the nature of an RNA-seq experiment, the resulting reads are limited to the actively transcribed regions and proportional to the amount of sampled messenger RNA molecules and to their length. In a typical short read RNA-seq experiment, the number of produced reads is often set to a desirable threshold, which is defined *ad hoc* to satisfy study requirements. This introduces a significant limitation in the resolution of the generated experiment which, although being still adequate for a general gene expression analysis, will not confer enough power to the detection of a large repertoire [62] of somatic mutations. Indeed, in our tests/experiments, we consistently observed that missed neoepitopes are generated from variants falling in under-represented regions, where the minimum coverage requirement was not met, causing uncertainty in the genotyping process. As most of the tested datasets were generated with a relatively low depth protocol (<50M reads), recalling previous reports [62], we speculated that increas-

ing the number of sequencing reads could dramatically boost the sensitivity of an RNA-seq based approach, enabling the detection of tumor specific alterations even in low expressed transcripts.

Another important consideration must be accounted due to the lack of a matched control sample. While we reported that the use of population genetic databases and genotype likelihoods via the Bayesian germline error model of ENEO is fruitful, rare and ultrarare events may be still prone to misclassification. Additionally, population databases are known to be still under-representative of many ethnicities, suggesting that this may negatively affect the performances on them. These limitations can potentially be mitigated using larger population databases, which, at time of writing, are becoming increasingly representative of previously poor characterized ethnicities [63].

Overall, ENEO opens to the applicability of neoantigen detection using solely the tumor RNA-seq data, providing a faster and cost-effective way to interrogate the transcribed regions of the human genome. Its robustness across different tumor histologies and mutational burdens highlights its potential in various translational and experimental setups, reinforcing the importance of high-resolution transcriptional profiling in the precision oncology.

Acknowledgements

The authors thank the anonymous reviewers for their valuable suggestions. We acknowledge the CINECA award under the ISCRA initiative, for the availability of high-performance computing resources and support.

Author contributions: **DT** Conceptualization, Methodology, Software, Writing - Original Draft. **MD**: Conceptualization, Software. **GB**: Methodology. **GG, MC**: Conceptualization. **RDA**: Conceptualization, Supervision, Funding acquisition, Writing - Review & Editing. All authors reviewed and approved the final manuscript.

Supplementary data

Supplementary data is available at NAR Genomics & Bioinformatics online.

Conflict of interest

None declared.

Funding

This work was supported by European Union-Next Generation EU, in the context of PNRR, Investment 1.5 Ecosystems of Innovation, Project Tuscany Health Ecosystem (THE), ECS00000017, Spoke 3 CUP: B83C22003930001 (to RDA), and PNRR, Missione 4 "Istruzione e Ricerca"- Componente C2, Investimento 1.1 PRIN 20227NHWH2 Computational methods for third generation cancer genomics, CUP: B53D23007820006 (to RDA), Advanced ERC grant Vaccibiome 834634 (to GG and MD), and specific funding from Regione Toscana/Istituto per lo Studio, la Prevenzione e la Rete Oncologica (ISPRO) (to MC).

Data availability

ENEO is freely available at the https://github.com/ctglab/ENEO and https://doi.org/10.5281/zenodo.14931773.

References

1. Schumacher TN, Schreiber RD. Neoantigens in cancer immunotherapy. *Science* 2015;**348**:69–74. https://doi.org/10.1126/science.aaa4971
2. Lu YC, Robbins PF. Cancer immunotherapy targeting neoantigens. *Semin Immunol* 2016;**28**:22–7. https://doi.org/10.1016/j.smim.2015.11.002
3. Ott PA, Hu Z, Keskin DB *et al.* An immunogenic personal neoantigen vaccine for patients with melanoma. *Nature* 2017;**547**:217–21. https://doi.org/10.1038/nature22991
4. Hu Z, Leet DE, Allesøe RL *et al.* Personal neoantigen vaccines induce persistent memory T cell responses and epitope spreading in patients with melanoma. *Nat Med* 2021;**27**:515–25. https://doi.org/10.1038/s41591-020-01206-4
5. Tran E, Ahmadzadeh M, Lu YC *et al.* Immunogenicity of somatic mutations in human gastrointestinal cancers. *Science* 2015;**350**:1387–90. https://doi.org/10.1126/science.aad1253
6. Parkhurst MR, Robbins PF, Tran E *et al.* Unique neoantigens arise from somatic mutations in patients with gastrointestinal cancers. *Cancer Discov* 2019;**9**:1022–35. https://doi.org/10.1158/2159-8290.CD-18-1494
7. Keskin DB, Anandappa AJ, Sun J *et al.* Neoantigen vaccine generates intratumoral T cell responses in phase Ib glioblastoma trial. *Nature* 2019;**565**:234–9. https://doi.org/10.1038/s41586-018-0792-9
8. Lang F, Schrörs B, Löwer M *et al.* Identification of neoantigens for individualized therapeutic cancer vaccines. *Nat Rev Drug Disc* 2022;**21**:261–82. https://doi.org/10.1038/s41573-021-00387-y
9. Roudko V, Greenbaum B, Bhardwaj N. Computational prediction and validation of tumor-associated neoantigens. *Front Immunol* 2020;**11**:27. https://doi.org/10.3389/fimmu.2020.00027
10. Tran E, Robbins PF, Rosenberg SA. 'Final common pathway' of human cancer immunotherapy: targeting random somatic mutations. *Nat Immunol* 2017;**18**:255–62. https://doi.org/10.1038/ni.3682
11. Chong C, Coukos G, Bassani-Sternberg M. Identification of tumor antigens with immunopeptidomics. *Nat Biotechnol* 2022;**40**:175–88. https://doi.org/10.1038/s41587-021-01038-8
12. Laumont CM, Perreault C. Exploiting non-canonical translation to identify new targets for T cell-based cancer immunotherapy. *Cell Mol Life Sci* 2018;**75**:607–21. https://doi.org/10.1007/s00018-017-2628-4
13. Smart AC, Margolis CA, Pimentel H *et al.* Intron retention is a source of neoepitopes in cancer. *Nat Biotechnol* 2018;**36**:1056–8. https://doi.org/10.1038/nbt.4239
14. Ouspenskaia T, Law T, Clauser KR *et al.* Unannotated proteins expand the MHC-I-restricted immunopeptidome in cancer. *Nat Biotechnol* 2022;**40**:209–17. https://doi.org/10.1038/s41587-021-01021-3
15. Chong C, Müller M, Pak H *et al.* Integrated proteogenomic deep sequencing and analytics accurately identify non-canonical peptides in tumor immunopeptidomes. *Nat Commun* 2020;**11**:1293. https://doi.org/10.1038/s41467-020-14968-9
16. Liao H, Barra C, Zhou Z *et al.* MARS an improved de novo peptide candiyear selection method for non-canonical antigen target discovery in cancer. *Nat Commun* 2024;**15**:661. https://doi.org/10.1038/s41467-023-44460-z
17. Tretter C, de Andrade Krätzig N, Pecoraro M *et al.* Proteogenomic analysis reveals RNA as a source for tumor-agnostic neoantigen identification. *Nat Commun* 2023;**14**:4632. https://doi.org/10.1038/s41467-023-39570-7
18. Yizhak K, Aguet F, Kim J *et al.* RNA sequence analysis reveals macroscopic somatic clonal expansion across normal tissues.

Science 2019;**364**:eaaw0726. https://doi.org/10.1126/science.aaw0726
19. Coudray A, Battenhouse AM, Bucher P *et al.* Detection and benchmarking of somatic mutations in cancer genomes using RNA-seq data. *PeerJ* 2018;**6**:e5362. https://doi.org/10.7717/peerj.5362
20. Rubinstein J, Nicolson N, Rottmann D *et al.* Choice of control tissue impacts designation of germline variants in a cohort of papillary thyroid carcinoma patients. *Ann Oncol* 2020;**31**:815–21.
21. Piskol R, Ramaswami G, Li J. Reliable identification of genomic variants from RNA-seq data. *Am J Hum Genet* 2013;**93**:641–51. https://doi.org/10.1016/j.ajhg.2013.08.008
22. Wells DK, van Buuren MM, Dang KK *et al.* Key parameters of tumor epitope immunogenicity revealed through a consortium approach improve neoantigen prediction. *Cell* 2020;**183**:818–34. https://doi.org/10.1016/j.cell.2020.09.015
23. Gros A, Parkhurst MR, Tran E *et al.* Prospective identification of neoantigen-specific lymphocytes in the peripheral blood of melanoma patients. *Nat Med* 2016;**22**:433–8. https://doi.org/10.1038/nm.4051
24. Gartner JJ, Parkhurst MR, Gros A *et al.* A machine learning model for ranking candiyear HLA class I neoantigens based on known neoepitopes from multiple human tumor types. *Nat Cancer* 2021;**2**:563–74. https://doi.org/10.1038/s43018-021-00197-6
25. Chen S, Zhou Y, Chen Y *et al.* fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* 2018;**34**:i884–90. https://doi.org/10.1093/bioinformatics/bty560
26. Li H, Durbin R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* 2009;**25**:1754–60. https://doi.org/10.1093/bioinformatics/btp324
27. Auwera GAVd, O'Connor BD. *Genomics in the cloud: using Docker, GATK, and WDL in Terra. First edition*. O'Reilly, 2020.
28. Benjamin D, Sato T, Cibulskis K *et al.* Calling somatic SNVs and indels with Mutect2. bioRxiv, https://doi.org/10.1101/861054, 2 December 2019, preprint: not peer reviewed.
29. Poplin R, Chang PC, Alexander D *et al.* A universal SNP and small-indel variant caller using deep neural networks. *Nat Biotechnol* 2018;**36**:983–7. https://doi.org/10.1038/nbt.4235
30. Krusche P, Trigg L, Boutros PC *et al.* Best practices for benchmarking germline small-variant calls in human genomes. *Nat Biotechnol* 2019;**37**:555–60.
31. Olson ND, Wagner J, McDaniel J *et al.* PrecisionFDA Truth Challenge V2: calling variants from short and long reads in difficult-to-map regions. *Cell Genom* 2022;**2**:100129.
32. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 2010;**26**:841–2.
33. McLaren W, Gil L, Hunt SE *et al.* The Ensembl variant effect predictor. *Genome Biol* 2016;**17**:122. https://doi.org/10.1186/s13059-016-0974-4
34. Mölder F, Jablonski KP, Letcher B *et al.* Sustainable data analysis with Snakemake. *F1000Research* 2021;**10**:33. https://doi.org/10.12688/f1000research.29032.2
35. Dobin A, Davis CA, Schlesinger F *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* 2013;**29**:15–21. https://doi.org/10.1093/bioinformatics/bts635
36. Kim S, Scheffler K, Halpern AL *et al.* Strelka2: fast and accurate calling of germline and somatic variants. *Nat Methods* 2018;**15**:591–4. https://doi.org/10.1038/s41592-018-0051-x
37. Pedersen BS, Layer RM, Quinlan AR. Vcfanno: fast, flexible annotation of genetic variants. *Genome Biol* 2016;**17**:118. https://doi.org/10.1186/s13059-016-0973-5
38. Lek M, Karczewski KJ, Minikel EV *et al.* Analysis of protein-coding genetic variation in 60,706 humans. *Nature* 2016;**536**:285–91. https://doi.org/10.1038/nature19057
39. Karczewski KJ, Francioli LC, Tiao G *et al.* The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* 2020;**581**:434–43. https://doi.org/10.1038/s41586-020-2308-7
40. Phan L, Jin Y, Zhang H *et al.* ALFA: allele frequency aggregator. In: *National Center for Biotechnology Information*. US National Library of Medicine, 2020, **10**. https://www.ncbi.nlm.nih.gov/snp/docs/gsr/alfa/#citing-this-project
41. Robert CP, Casella G. The Metropolis—Hastings Algorithm. In: *Monte Carlo Statistical Methods*. Springer, 2004, 267–320. https://doi.org/10.1007/978-1-4757-4145-2_7
42. Virtanen P, Gommers R, Oliphant TE *et al.* SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat Methods* 2020;**17**:261–72. https://doi.org/10.1038/s41592-019-0686-2
43. Stadick S. *Perbase*; 2023
44. Patro R, Duggal G, Love MI *et al.* Salmon: fast and bias-aware quantification of transcript expression using dual-phase inference. *Nat Methods* 2017;**14**:417–9. https://doi.org/10.1038/nmeth.4197
45. Soneson C, Love MI, Robinson MD. Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences. *F1000Research* 2015;**4**:1521. https://doi.org/10.12688/f1000research.7563.2
46. Song L, Bai G, Liu XS *et al.* Efficient and accurate KIR and HLA genotyping with massively parallel sequencing data. *Genome Res* 2023;**33**:923–31. https://doi.org/10.1101/gr.277585.122
47. Wood MA, Nguyen A, Struck AJ *et al.* neoepiscope improves neoepitope prediction with multivariant phasing. *Bioinformatics* 2020;**36**:713–20. https://doi.org/10.1093/bioinformatics/btz653
48. Reynisson B, Alvarez B, Paul S *et al.* 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;**48**:W449–54. https://doi.org/10.1093/nar/gkaa379
49. Reardon B, Koşaloğlu-Yalçın Z, Paul S *et al.* Allele-specific thresholds of eluted ligands for T-Cell epitope prediction. *Mol Cell Proteomics* 2021;**20**:100122. https://doi.org/10.1016/j.mcpro.2021.100122
50. Vita R, Mahajan S, Overton JA *et al.* The Immune Epitope Database (IEDB): 2018 update. *Nucleic Acids Res* 2019;**47**:D339–43. https://doi.org/10.1093/nar/gky1006
51. Marcu A, Bichmann L, Kuchenbecker L *et al.* HLA Ligand Atlas: a benign reference of HLA-presented peptides to improve T-cell-based cancer immunotherapy. *J Immun Cancer* 2021;**9**:e002071.
52. Alexandrov LB, Kim J, Haradhvala NJ *et al.* The repertoire of mutational signatures in human cancer. *Nature* 2020;**578**:94–101. https://doi.org/10.1038/s41586-020-1943-3
53. Castle JC, Loewer M, Boegel S *et al.* Mutated tumor alleles are expressed according to their DNA frequency. *Sci Rep* 2014;**4**:4743. https://doi.org/10.1038/srep04743
54. Wang Q, Shashikant CS, Jensen M *et al.* Novel metrics to measure coverage in whole exome sequencing datasets reveal local and global non-uniformity. *Sci Rep* 2017;**7**:885.
55. Wang L, Wang S, Li W. RSeQC: quality control of RNA-seq experiments. *Bioinformatics* 2012;**28**:2184–5. https://doi.org/10.1093/bioinformatics/bts356
56. Wang SY, Zhang LJ, Chen GJ *et al.* COPA A-to-I RNA editing hijacks endoplasmic reticulum stress to promote metastasis in colorectal cancer. *Cancer Lett* 2023;**553**:215995. https://doi.org/10.1016/j.canlet.2022.215995
57. Pritchard AL, Hastie ML, Neller M *et al.* Exploration of peptides bound to MHC class I molecules in melanoma. *Pigm Cell Melanoma Res* 2015;**28**:281–94. https://doi.org/10.1111/pcmr.12357
58. Zhang M, Fritsche J, Roszik J *et al.* RNA editing derived epitopes function as cancer antigens to elicit immune responses. *Nat Commun* 2018;**9**:3919. https://doi.org/10.1038/s41467-018-06405-9
59. Tran E, Turcotte S, Gros A *et al.* Cancer immunotherapy based on mutation-specific CD4+ T cells in a patient with epithelial cancer. *Science* 2014;**344**:641–5. https://doi.org/10.1126/science.1251102
60. Rogner UC, Wilke K, Steck E *et al.* The melanoma antigen gene (MAGE) family is clustered in the chromosomal band Xq28.

Genomics 1995;**29**:725–31. https://doi.org/10.1006/geno.1995.9945
61. Le Dung T, Uram Jennifer N, Wang Hao *et al.* PD-1 blockade in tumors with mismatch-repair deficiency. *New Engl J Med* 2015;**372**:2509–20. https://doi.org/10.1056/NEJMoa1500596
62. Quaglieri A, Flensburg C, Speed TP *et al.* Finding a suitable library size to call variants in RNA-seq. *BMC Bioinformatics* 2020;**21**:553. https://doi.org/10.1186/s12859-020-03860-4
63. Sun KY, Bai X, Chen S *et al.* A deep catalog of protein-coding variation in 985,830 individuals. bioRxiv, https://doi.org/10.1101/2023.05.09.539329, 10 May 2023, preprint: not peer reviewed.

Received: September 4, 2024. **Revised:** January 21, 2025. **Editorial Decision:** February 21, 2025. **Accepted:** March 11, 2025