

## Gene expression

# Detecting known neopeptides, gene fusions, transposable elements, and circular RNAs in cell-free RNA

Mayank Mahajan<sup>1</sup>  and Martin Hemberg<sup>1,\*</sup> 

<sup>1</sup>Gene Lay Institute of Immunology and Inflammation, Brigham and Women's Hospital, Massachusetts General Hospital and Harvard Medical School, Boston, MA 02115, United States

\*Corresponding author. Gene Lay Institute of Immunology and Inflammation, Brigham and Women's Hospital, Massachusetts General Hospital and Harvard Medical School, Hale Building of Transformative Medicine, 60 Fenwood Road, Boston, MA 02115, United States. E-mail: mhemberg@bwh.harvard.edu.

Associate Editor: Christina Kendzierski

### Abstract

**Motivation:** Cancer is the second leading cause of death worldwide, and although there have been advances in treatments, including immunotherapies, these often require biopsies which can be costly and invasive to obtain. Due to lack of pre-emptive cancer detection methods, many cases of cancer are detected at a late stage when the definitive symptoms appear. Plasma samples are relatively easy to obtain, and they can be used to monitor the molecular signatures of ongoing processes in the body. Profiling cell-free DNA is a popular method for monitoring cancer, but only a few studies have explored the use of cell-free RNA (cfRNA), which shows the recent footprint of systemic transcription.

**Results:** Here, we developed FastNeo, a computational method for detecting known neopeptides in human cfRNA. We show that neopeptides and other biomarkers detected in cfRNA can discern Hepatocellular carcinoma patients from the healthy patients with a sensitivity of 0.84 and a specificity of 0.79. For colorectal cancer we achieve a sensitivity of 0.87 and a specificity of 0.8. An important advantage of our cfRNA based approach is that it also reports putative neopeptides which are important for therapeutic purposes.

**Availability and implementation:** The FastNeo package is available at <https://github.com/yashumayank/FastNeo> and <https://zenodo.org/records/11521368>. The benchmark pipelines to detect Immune Epitope database and Tumor-Specific Neoantigen database neopeptides using HaploTypeCaller, bcftools, and Lofreq, and to run FastNeo with STAR instead of Bowtie2 are also available in the above github repository.

## 1 Introduction

Cell-free nucleic acids are either released by dying cells or actively secreted by live cells into body fluids, such as plasma. These nucleic acids can be acquired via liquid biopsy and sequenced to study the active processes in the body, and by detecting mutations or epigenetic modifications one can infer the presence of a tumor (Corcoran Ryan and Chabner Bruce 2018, Cristiano *et al.* 2019). Consequently, cell-free DNA (cfDNA) is emerging as a tool for disease monitoring, and detection of early-stage tumors (Bronkhorst *et al.* 2019, Georgakopoulos-Soares *et al.* 2021b). By contrast, cell-free RNA (cfRNA) is less well studied even though it can reveal actively transcribed genes, mutations, repetitive elements, and exosomal RNA from the whole body (Tan *et al.* 2019, Zhu *et al.* 2021, Chen *et al.* 2022, Chang *et al.* 2024, Reggiardo *et al.* 2023). cfRNA shows a current footprint of the systemic RNA expression, however the tissue of origin of the cfRNA cannot be ascertained easily. It has been estimated that up to 95% of the cfRNA in plasma is derived from other tissues or the microbiome, with only a small minority originating from the tumor (Larson *et al.* 2021). Furthermore, most cfRNA is fragmented by nucleases and is expected to produce short reads. Thus, deriving clinically relevant information from cell-free RNA is challenging.

Epitopes are short peptide fragments derived from proteins that are expressed in a cell. Some of these epitopes bind to the

human leukocyte antigen (HLA) class I or class II molecules and are then presented on the cell surface for immune surveillance. Mutations in the coding sequence can give rise to neopeptides that are missing in the proteome of healthy germline cells. Cancer-specific neopeptides are ideal targets for immunotherapies since they are unique to malignant cells, and they are hence highly sought after (Ott *et al.* 2017, Lang *et al.* 2022).

Although most methods for detecting neopeptides focus on ones created through point mutations, they can emerge from any type of mutation. Gene fusions have been shown to be recurrent in some cancer types and they can be used as diagnostic biomarkers (Stransky *et al.* 2014, Taniue and Akimitsu 2021). Gene fusions can produce neopeptides which have been explored in various studies (Fotakis *et al.* 2020, Zhang *et al.* 2020, Kumar *et al.* 2024). Importantly, T cell response against these gene-fusion neopeptides can be used for personalized immunotherapies to target cancers that are driven by gene fusions (Biernacki *et al.* 2020, Weber *et al.* 2022).

Genome-nullomers are nucleotide kmers that are not present in the human genome, likewise consensus coding sequence (CCDS)-nullomers are nucleotide kmers that are not present in the human CCDS (Georgakopoulos-Soares *et al.* 2021a). Genome-nullomers identified in both cfDNA (Georgakopoulos-Soares *et al.* 2021b) and cfRNA have been used to distinguish cancer patients from healthy controls

(Montgomery *et al.* 2024). Building on this work we present FastNeo, a tool for quick and reliable detection of known human neopeptides and gene fusions in the cfRNA. An important difference is that FastNeo uses CCDS-nullomers, which provide higher sensitivity for detection of neopeptides in the coding transcripts, and it is customized for transcriptomic data with high fragmentation and low allele coverage.

We also quantify the expression of transposable elements (TEs) and circular RNAs (circRNAs) in the cfRNA. Retrotransposons are typically silenced in mammalian cells, nonetheless 6%–30% of mammalian transcripts have been shown to initiate within repetitive elements (Faulkner *et al.* 2009). Disruption of the epigenetic landscape in cancer can result in upregulation, and expression of long interspersed nuclear elements (LINE) and human endogenous retrovirus (HERV) have been observed in various cancer types (Büscher *et al.* 2005, Rodić *et al.* 2014). CircRNAs are regarded as reliable biomarkers for cancer (Memczak *et al.* 2013, Bachmayr-Heyda *et al.* 2015, Vo *et al.* 2019). They are resistant to exoribonuclease (RNase R) and their closed structure provides higher stability compared to their linear counterparts (Memczak *et al.* 2013). circRNAs have been detected in circulating exosomes to show their potential as biomarkers for cancer diagnosis (Li *et al.* 2015, Roy *et al.* 2022). We demonstrate that by combining neopeptide detection through CCDS-nullomers, TEs, and circRNAs, it is possible to accurately distinguish cancer patients from healthy controls.

## 2 Materials and methods

### 2.1 Nullomers associated with neopeptide databases

Neopeptides described in Immune Epitope database (IEDB) and Tumor-Specific Neoantigen database (TSNAdb) were acquired (Wu *et al.* 2018, Dhanda *et al.* 2019). The “epitope full v3.csv” file was downloaded from IEDB webpage [www.iedb.org](http://www.iedb.org), on 19 September 2022. Neopeptides and wild-type epitopes with the following characteristics were retrieved from the above file: “RelatedObject EpitopeRelationship” is “neopeptide,” “RelatedObject OrganismName” contains the word “sapien,” “Epitope Description” only has letters A to Z, “RelatedObject ParentProteinIRI” has the uniprot ID. The uniprot IDs in IEDB were converted to transcript IDs using “HUMAN 9606 idmapping selected.tab,” which was downloaded from [www.uniprot.org](http://www.uniprot.org) on 30 June 2022. 114 uniprot IDs in IEDB did not have a corresponding transcript ID. Next, “frequent\_neoantigen\_ICGC\_4.0\_adj.txt” and “frequent\_neoantigen\_TCGA\_4.0\_adj.txt” were downloaded from TSNAdb webpage <http://biopharm.zju.edu.cn> on 19 September 2022, and the neopeptides were retrieved along with the corresponding wildtype epitopes and transcript IDs.

The neopeptides were then aligned to the wild-type epitopes using the Smith-Waterman local alignment algorithm and the aligned sequences were used to detect the sites of amino acid insertion, deletion, and substitution. 521 unique neopeptides from IEDB that had an insertion or deletion of amino acid, or >2 amino acid substitutions were excluded (Supplementary Table S1). The TSNAdb neopeptides consist of single nucleotide substitutions only.

To find the position of neopeptides on the human CCDS, each wild-type epitope was searched on the protein translation of the corresponding CCDS. The coding sequence of

each wild-type epitope was retrieved and all possible DNA conformations corresponding to each neopeptide were generated using all the possible combinations of codons at the sites of mutation. The sequence around the mutated sites in each DNA conformation was scanned for CCDS-nullomers of length 16, which are nucleotide kmers of length 16 that are not present in the human CCDS. The length of the nullomers was chosen to provide a balance between sensitivity and specificity of the detection algorithm and to provide faster search speed (Georgakopoulos-Soares *et al.* 2021b). The CCDS-nullomers can be present in non-coding parts of the genome, such as the repeat elements, and their transcription can result in false positive matches. Fortunately, 99.9% of the neopeptide-derived nullomers occur <139 times in the human genome sequence (Supplementary Fig. S1a and b), and to improve specificity, nullomers that appear >300 times in the human genome sequence were removed. Finally, we created a mapping file for FastNeo, that maps each known neopeptide to the corresponding gene and the list of CCDS-nullomers that span across the neopeptide causing mutation/mutations.

Unless otherwise mentioned, Ensembl GRCh38.p13 genome release 107 was used for human genome sequence, CCDS, and proteins.

### 2.2 Neopeptide binding affinity to HLA complexes

netMHCpan and netMHCIIpan were used to predict the binding affinity of both neopeptides and wildtype epitopes to various MHC complexes (Jurtz *et al.* 2017, Reynisson *et al.* 2020). Since TSNAdb epitopes are 8–11 amino acids (aa) long, netMHCpan was used to predict the binding affinity to MHC I complexes. We used the default list of MHC I alleles in the TSNAdb and added 3 more alleles, HLA-C\*08:01, HLA-A\*33:03, HLA-B\*08:01, which have high population frequencies as per the Allele Frequency Net Database (Wu *et al.* 2018, Gonzalez-Galarza *et al.* 2020).

IEDB has both long and short neopeptides. The binding affinity of neopeptides of length  $\leq 11$  aa was predicted against the MHC I complex using the same approach that was used for the TSNAdb neopeptides. In addition, netMHCIIpan was used to predict the binding affinity of neopeptides of length > 11 aa to the MHC II complex. We predicted their binding affinity against the following MHC II alleles that were chosen for high frequency: DRB1\*0701, DRB1\*1501, DRB1\*0301, DRB1\*1101, DRB1\*0101, DRB1\*1302, DRB1\*1301, DRB1\*1502, DRB1\*0401, DRB1\*1201, DRB1\*0403, DRB4\*0101, DRB3\*0202, DRB3\*0101, DRB5\*0101, DRB3\*0301, HLA-DQA1\*0102-DQB1\*0501, HLA-DQA1\*0103-DQB1\*0501, HLA-DQA1\*0501-DQB1\*0201, HLA-DPA1\*0103-DPB1\*0201, HLA-DPA1\*0103-DPB1\*0402 (Hamed *et al.* 2018, Gonzalez-Galarza *et al.* 2020, Satapornpong *et al.* 2020, Khan *et al.* 2022, Pedersen *et al.* 2023).

As advised in the netMHCpan documentation, we select candidate binders based on %Rank, which is the rank of the predicted affinity compared to a set of random natural peptides, rather than nM Affinity. Only neopeptides with “%Rank” below 2% were retained, and the HLA alleles and binding affinities were added to the respective neopeptides or wildtype epitopes in the mapping file used by FastNeo, which is described below. In case of multiple predictions per epitope, all the predictions are shown.

## 2.3 Neopeptide frequency in germline of healthy population

The germline variants from various genetic ancestry groups were extracted from the exome sequencing data in gnomAD v4.0.0 (Chen *et al.* 2024). Allele frequency in females (AF\_XX), males (AF\_XY), Africans/African Americans (AF\_afr), admixed Americans (AF\_amr), Ashkenazi Jews (AF\_asj), East Asians (AF\_eas), Finnish (AF\_fin), Middle Eastern (AF\_mid), non-Finnish Europeans (AF\_nfe), South Asians (AF\_sas), combined allele frequency (AF), and the genetic ancestry group with highest proportion of the allele (grpmax), were retrieved from the gnomAD database for the allelic variants using the following criteria: AF >1e-07, VEP: Consequence = “missense\_variant,” VEP: BIOTYPE = “protein\_coding,” VEP: Feature consists of a transcript ID starting with “ENST.” The population frequency of each germline allelic variant that produces one of neopeptides described in the IEDB or TSNEdb was retrieved and added to the respective neopeptides in the mapping file used by FastNeo, which is described below.

## 2.4 Nullomers associated with gene fusions

Gene fusion file “ChimerKB4.xlsx” was downloaded from ChimerDB on May 19, 2023 and a subset of fusions with “ChimerPub” column = “Pub” were selected (Jang *et al.* 2020). The nucleotide sequences were extracted from the same human genome sequence that has been used to generate the ChimerDB (downloaded from hgdownload.cse.ucsc.edu/goldenpath/hg19/bigZips/). 500 nucleotides were extracted from upstream of the 3′ end, and downstream of the 5′ end, of each gene fusion junction and were fused at the junction. These fused sequences are used to map the reads by the FastNeo, which is described below. We searched for CCDS-nullomers that spanned the fusion junctions and removed the ones that were repeated > 300 times in the genome. In addition, we scanned neopeptides at the fusion junctions where (i) both 5′ and 3′ junctions are inside the coding sequence, (ii) the 5′ junction does not overlap the start codon, and (iii) the 3′ gene is still in its original frame after fusion. Nine aa-long peptides that were absent from the reference human protein sequences were classified as neopeptides. We ignored neopeptides produced via fusions in the intron and UTRs regions as it is non-trivial to predict the neopeptides produced by such fusions. We created a mapping file for FastNeo that maps each gene fusion to the corresponding fused gene sequence, the putative neopeptides on the fusion junction, and a list of CCDS-nullomers that span across the fusion junction.

## 2.5 Neopeptide and fusion detection pipeline

The FastNeo pipeline uses stranded RNA-seq data from cfrRNA and CCDS-nullomers as input and outputs the putative neopeptides and gene fusions detected in the cfrRNA (Fig. 1).

**Step 1:** The CCDS-nullomers corresponding to the neopeptides and gene fusions are searched on the first read, and reverse complement sequences of those CCDS-nullomers were searched on the second read of the paired-end RNA-seq sequencing data.

**Step 2 (neopeptides):** Read pairs containing nullomer/nullomers associated to neopeptides are mapped to the coding sequences of the human genome GRCh38 using Bowtie2 with parameters (–very-sensitive-local) (Langmead and Salzberg 2012). Duplicate reads are then removed using

samtools rmdup (Danecek *et al.* 2021). Additionally, reads with more than 33.3% soft-clipped bases are removed. All reads with MAPQ score >10, or the alignment score (AS) greater than a minimum expected score (MES) are retained.

MES is a linear function of aligned length (AL) defined as,  $MES = \text{slope} * (AL - 35) + 64$ , where slope is calculated using the desired AS when the AL is between 35 and 150 nucleotides. We used AS =64 for AL =35, which allows 1 medium quality mismatch (penalty=4) and AS =277 for AL =150, which allows ~4 medium quality mismatches (penalty=15); hence, slope = (277-64)/(150-35) = 1.85.

**Step 2 (gene fusions):** Reads pairs containing nullomer/nullomers associated with gene fusions are mapped to the fused sequences created above using Bowtie2 with parameters (–very-sensitive-local) (Langmead and Salzberg 2012). Duplicate reads are then removed using samtools rmdup (Danecek *et al.* 2021). Additionally, reads with more than 33.3% soft-clipped bases are removed and only reads with 5 or more bases mapped on both sides of the fusion junction are retained. MAPQ scores are less relevant when reads are mapped to a very small set of sequences. Hence, only reads with MAPQ score >30 or AS greater than the MES, as described above, are retained.

**Step 3:** The read coverage per nullomer is calculated as the total reads that contain the nullomer sequence and are mapped to the corresponding neopeptide coding sequence or gene fusion junction. The read coverage of each neopeptide and gene fusion is calculated as the read coverage of the most covered corresponding nullomer. The total nullomers found per neopeptide or gene fusion are also reported. The neopeptides produced by the same mutation are reported together in the same row. The mapping files described in the above sections are used to map nullomers to corresponding neopeptides and gene fusions. The mapping files also contain meta-data, which is reported in the output.

**Output (neopeptides):** The neopeptide output file shows (i) the fastq file name, (ii) gene id, (iii) HGNC symbol, (iv) most covered nullomer per mutation/mutations, (v) neopeptides associated to the nullomer, (vi) number of mapped reads, (vii) number of nullomers on the read with most nullomers, (viii) database name, (ix) gene function, (x) HLA binding affinities of wildtype epitope, (xi) neopeptide, and (xii) the frequency of the neopeptide producing mutations in healthy individuals of various genetic ancestry groups.

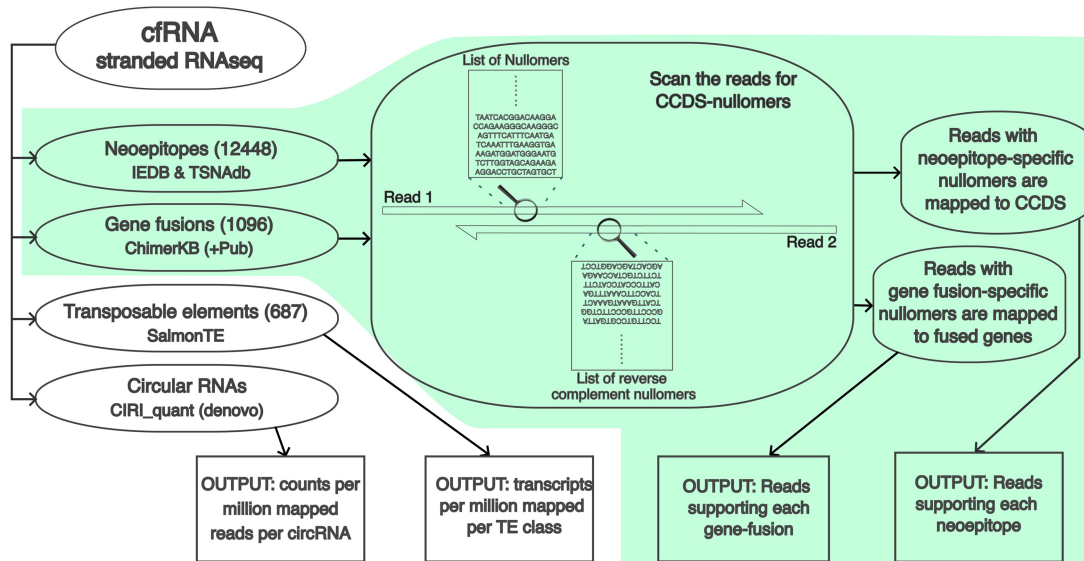
**Output (gene fusions):** The gene fusion output file shows (i) fastq file name, (ii) ChimerKB ID, (iii) most covered nullomer per gene fusion, (iv) neopeptides associated with the gene fusion, (v) number of mapped reads, (vi) number of nullomers on the read with most nullomers, (vii) gene function, (viii) genomic loci of 5′ junction, and (ix) genomic loci of 3′ junction.

The read coverage of variants is low in the cfrRNA sequencing data, hence low threshold of filters was used. Neopeptides with ≥3 mapped reads were used for all the downstream analyses. Fusions with ≥2 mapped reads and ≥2 detected nullomers were used for all the downstream analyses.

## 2.6 Benchmarking neopeptide detection against other variant calling methods

Known neopeptides were detected in the cfrRNA samples in the GSE142987 and GSE136651 datasets using GATK HaplotypeCaller (McKenna *et al.* 2010), samtools/bcftools





**Figure 1.** Overview of the complete workflow. The FastNeo workflow is shown with a green background.

mpileup (Li 2011), and Lofreq (Wilm *et al.* 2012), which specifically suited to low coverage datasets (Alosaimi *et al.* 2021). We mapped all the reads to the human genome using STAR 2.7.10a (Dobin *et al.* 2013) with the following options: -outSAMmultNmax 3 -outFilterMultimapNmax 15 -outFilterMismatchNmax 15 -alignSJDBoverhangMin 2 -alignIntronMax 1000000 -peOverlapNbasesMin 10 Values used for outFilterMultimapNmax, alignIntronMax and alignSJDBoverhangMin were inspired from the section “ENCODE options” in the STAR manual. Option peOverlapNbasesMin is said to improve mapping accuracy for paired-end libraries with short insert sizes. GATK “MarkDuplicates” tool was used to remove the duplicate reads (<http://broadinstitute.github.io/picard/>).

For bcftools, the mapped reads were passed to “bcftools mpileup” with options “-q 10 -Q 30,” and “bcftools call” was used to call variants. The variants supported by <3 reads were filtered out. For Lofreq and HaplotypeCaller, base qualities of mapped reads were re-estimated by running “SplitNCigarReads,” “BaseRecalibrator” with option “-known-sites 1000GENOMES-phase\_3.vcf,” and “ApplyBQSR” in that order. These tools are part of the GATK and “1000GENOMES-phase\_3.vcf” file containing allele frequencies from 1000 genomes phase 3 populations was downloaded from the ensembl ftp site. The mapped reads with recalculated base quality were then passed to “lofreq call” with default options, and “HaplotypeCaller” with option “-dbsnp” “1000GENOMES-phase\_3.vcf.” For HaplotypeCaller, variants with FS > 30.0, QD < 3.0 and AD < 3 were filtered out. Ensembl Variant Effect Predictor (VEP) (McLaren *et al.* 2016) was used to find missense and frameshift variations among the variants discovered by HaplotypeCaller, bcftools, and Lofreq. The known neoepitopes in IEDB/TSNAdB were detected around the missense and frameshift mutations using custom scripts (available from the FastNeo github page).

Each tool was run on each sample using two cores, making use of both cores in the steps that permitted it. The run-time of all the tools were compared on GSE142987 dataset, with average 8 million bases per sample, and GSE136651 dataset, with average 0.7 million bases per sample.

To benchmark the results of FastNeo while using STAR mapping instead Bowtie2 mapping, the reads that contained CCDS-nullomers were mapped to the human genome using STAR 2.7.10a (Dobin *et al.* 2013) with same options as above. Duplicate reads were removed using GATK “MarkDuplicates” tool (<http://broadinstitute.github.io/picard/>). The sam files with the remaining mapped reads were transformed from genome coordinates to the CCDS coordinates using “mudskipper shuffle” and “mudskipper bulk -p” (<https://github.com/OceanGenomics/mudskipper>).

## 2.7 Data sets

We used four plasma-derived cfRNA datasets (Supplementary Table S1). The first consists of 30 healthy donors and 35 Hepatocellular carcinoma (HCC) patients (GEO id: GSE142987) (Zhu *et al.* 2021). The second dataset consists of 20 healthy donors, 8 HCC patients, and 10 multiple myeloma patients (GEO id: GSE182824) (Roskams-Hieter *et al.* 2022). The second dataset also includes samples from pre-cancerous conditions that were ignored in this study. The third dataset consists of 46 healthy donors and 54 colorectal, 37 stomach, 27 HCC, 35 lung, and 31 esophageal cancer patients (GEO id: GSE174302) (Chen *et al.* 2022). The fourth dataset consists of 11 of pancreatic cancer patients and 22 healthy donors (GEO id: GSE136651) (Reggiardo *et al.* 2023). All data was de-identified, and there is no metadata about the treatment these patients were undergoing.

All datasets were sequenced using stranded RNA-seq libraries. Data from GSE142987, GSE182824, and GSE174302 has been sequenced using Illumina HiSeq with similar throughputs of ~9 million bases, while data from GSE136651 has been sequenced using Illumina NextSeq with a much lower throughput of ~0.7 million bases (Supplementary Table S1). Hence the data from GSE136651 was analysed separately, while the other three were analysed together.

## 2.8 Discriminative neoepitopes

FastNeo was used to quantify neoepitope expression in cfRNA datasets. The read coverage of each neoepitope was

normalized by the total reads in the corresponding sample and then scaled by  $\log_{10}$  to estimate the mean normalized expression per neoepitope. Neoepitopes that were present  $>20\%$  of all the cancer patients,  $>20\%$  healthy donors in the merged dataset, and had a mean normalized expression  $>0.1$  in both groups of donors were categorized as generic neoepitopes. Generic neoepitopes were removed from all the analyses and figures (Supplementary Table S2). Neoepitopes with twice the mean expression in cancer patients as compared to the healthy donors, or vice-versa were categorized as discriminative neoepitopes. In addition, discriminative neoepitope should be detected in cfRNA of  $\geq 3$  individuals from the group they are enriched in. For ease of visualization, the discriminative neoepitopes that are shown in the Fig. 2 and Supplementary Fig. S3 were detected in cfRNA of  $\geq 5$  individuals from the group they are enriched in. The heatmaps of discriminative neoepitopes were plotted using pheatmap library in R and the neoepitopes were clustered using “ward.D2” method and the euclidean distance.

## 2.9 Discriminative TEs

SalmonTE (parameters: quant—reference=hs) was used to quantify TEs in cfRNA datasets (Jeong *et al.* 2017). SalmonTE orders human TEs into 687 manually curated classes, and the expression is measured for each class of TE. SalmonTE output consists of effective length, reads mapped, and transcript per million (TPM) for each of the 687 TE classes (Fig. 1). Reads per kilobase per million mapped reads (RPKM) values were calculated to measure the expression of each TE class in each sample using effective TE length, reads mapped to the TE, and the total reads in the sample. Hence,  $RKPM = (\text{mapped\_reads\_per\_TE} / (\text{total\_reads\_in\_sample} / 1000000)) / \text{effective\_TE\_length}$ .

The pairwise Wilcoxon test and Bonferroni correction was performed using  $\log_{10}$  of RKPM of each TE class to detect differentially expressed TEs in the specific cancer type versus healthy donors. TEs were highly expressed in cfRNA samples, hence the TEs expressed in  $<80\%$  of healthy donors or  $<80\%$  of cancer patients were removed before performing the test. The  $P$ -value based cut-offs used to select most discriminative TEs.

## 2.10 Discriminative circRNAs

CIRIquant 1.1.3 was installed with default settings and genome hg19 as specified in the CIRIquant documentation (Zhang *et al.* 2020), and was run with parameter “—library-type 1” to quantify the circRNA in cfRNA datasets. CIRIquant reports circRNA expression in count per million (CPM), which is already normalized against overall RNA expression. circRNAs with  $<2$  reads mapped to the back splicing junction were removed from all downstream analysis.

The pairwise Wilcoxon test and Bonferroni correction was performed using  $\log_{10}$  of CPM scores of each gene to detect differentially expressed circRNA genes in the specific cancer type versus healthy donors. circRNA genes expressed in  $<40\%$  of the samples were removed before performing the test. In case of multiple circRNAs per gene, the circRNA with highest CPM score was used, or in a separate analysis the circRNA from each combination of exons was considered separately. The  $P$ -value based cut-offs were used to select the most discriminative circRNAs.

## 2.11 Classification of cancer patients

The classifiers were trained in 10 independent randomized iterations using 5-fold cross validation on general linear models with Ridge and Lasso regularizations, support vector machine (SVM) and Random Forests. The feature selection was done separately for each training set in each fold. The selected neoepitopes were required to have twice the mean expression in cancer patients as compared to the healthy donors, or vice-versa, and must be detected in cfRNA of  $\geq 3$  individuals of the group they are enriched in. The generic neoepitopes, as described above, were removed. Five most discriminative TEs with lowest  $P$ -values, and  $P$ -value  $< 1e-7$  were selected using pairwise Wilcoxon test and Bonferroni correction. TEs expressed in  $<80\%$  of healthy donors or  $<80\%$  of patients of the specific cancer type were removed before performing the pairwise Wilcoxon test. Five circRNAs with lowest  $P$ -values, and  $P$ -value  $< 1e-4$  were selected using pairwise Wilcoxon test and Bonferroni correction. circRNAs expressed in  $<40\%$  of healthy donors or patients of the specific cancer type were removed before performing the pairwise Wilcoxon test.

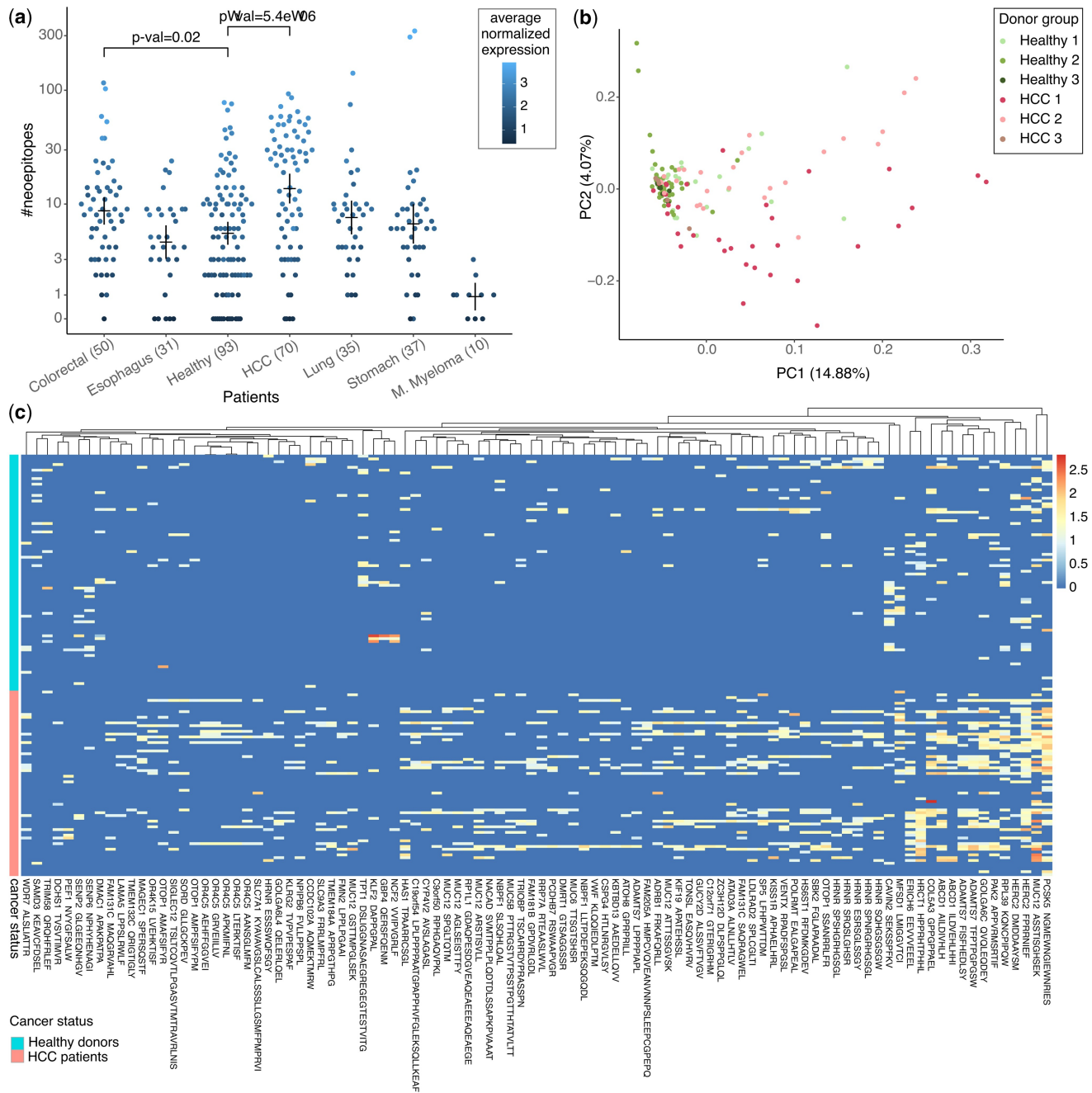
The classifiers were trained in 10 independent randomized iterations using 5-fold cross validation on general linear models with Ridge and Lasso regularizations, support vector machine (SVM) and Random Forests. The performance metrics achieved by each method was averaged across all folds first and then across all iterations. The R library “glmnet” was used to train linear regression models using Ridge and Lasso regularizations with options family=“binomial.” Lambda values were optimized separately for Ridge and Lasso using “cv.glmnet.” All values of lambda between  $10^2$  and  $10^{-2}$  with step change of  $10^{-0.1}$  were tested. Eventually, lambda=0.6 for Ridge and lambda=0.04 for Lasso was used for classification using only neoepitopes, and lambda=0.1 for Ridge and lambda=0.03 for Lasso was used for classification using other feature sets. R library “randomForest” was used to train random forest models with option nodesize=2. R library “E1071” was used to train the SVM models with options kernel = linear, cost = 2 and scale = FALSE. The cost = 2 was used for SVM to prevent overfitting.

## 3 Results

### 3.1 Neoepitope and gene-fusion associated nullomers

To increase the likelihood of identifying bona fide neoepitopes we restrict our search to the ones described in the TSNadb and IEDB (Wu *et al.* 2018, Vita *et al.* 2019). TSNadb consists of neoepitopes that are present in four or more cancer patients in the cancer genome atlas (TCGA) and international cancer genome consortium (ICGC) databases. IEDB consists of neoepitopes that have been experimentally characterized in published studies and many of them have also been tested for immune response using methods such as qualitative binding to HLA complexes and assays to monitor the release of IFN $\gamma$ , TNF $\alpha$ , and various interleukins. The experimental data associated with IEDB neoepitopes can be explored using the interactive platform CEDAR (Koşaloğlu-Yalçın *et al.* 2023).

To identify putative neoepitopes and gene fusions from cfRNA, we utilize nullomers, kmers that are not present in the human reference sequence. Here, we use the CCDS-



**Figure 2.** Putative neoepitopes detected in cfRNA. (a) Number of known neoepitopes detected in plasma cfRNA of patients grouped by cancer type and healthy. Each dot is coloured by average normalized expression for all the putative neoepitopes in the respective donor.  $P$ -values show the significance of the difference in distribution of healthy donors versus patients from two cancer types with most donors, calculated using the Wilcoxon test. Means and error-bars are shown for each group. (b) First two components of a PCA using expression of all putative neoepitopes in HCC patients and healthy donors. Suffixes 1, 2, and 3 in the donor groups correspond to the different datasets in [Supplementary Table S1](#). (c) Heatmap showing the expression of discriminative neoepitopes detected in the cfRNA of cancer patients and healthy donors (see Section 2). Each column represents a neoepitope, and each row represents a patient or a healthy donor. Column label shows one of the neoepitopes together with the gene symbol. See [Supplementary Table S5](#) for a full list of discriminative neoepitopes with descriptions.

nullomers, kmers that are absent in the CCDS ([Georgakopoulos-Soares et al. 2021a](#)). As CCDS-nullomers are not expected in the CCDS, they can be deployed for a fast, yet reliable search of neoepitope producing mutations and cancer detection ([Georgakopoulos-Soares et al. 2021b](#)). Most neoepitopes and gene-fusions result in one or more CCDS-nullomers being created, and thus, they are well suited for nullomer-based detection.

As we scanned the wildtype epitopes corresponding to the neoepitopes in TSNAdB and IEDB, some of them were missing in the corresponding CCDS of human genome GRCh38. p13 ([Supplementary Table S2](#)). We recovered 12 526 out of 13 582 (~92%), and 79 871 out of 85 185 (~94%) wildtype-neoepitopes pairs from IEDB and TSNAdB, respectively ([Supplementary Table S3a](#)). Most of these neoepitopes were lost while mapping them to the recent version of



genome, while <0.3% and <0.9% of the neoepitopes in the IEDB and TSNEdB, respectively lacked CCDS-nullomers of length 16 in their coding sequence (Supplementary Table S2).

To reduce false positives, we characterized the neoepitopes associated with germline variants from the genetic ancestry groups described in the gnomAD database (Brown *et al.* 2014, Chen *et al.* 2024). 645 and 9811 neoepitopes in IEDB and TSNAdB, respectively, were associated with germline variants, of which the african/african american (AFR), non-Finnish European (NFE) and south asian (SAS) were three most frequent ancestry groups (Supplementary Table S3a).

IEDB had 3661 unique neoepitopes of length  $\leq 11$  aa and 8787 unique neoepitopes of length 12–37 aa (Supplementary Table S2). The binding affinity of the shorter neoepitopes was predicted against the MHC I complex (see Section 2), while binding affinity of the longer neoepitopes was predicted against the MHC II complex. 2833 out of 3661 neoepitopes in IEDB and 69 095 out of 79 871 TSNAdB were predicted to have binding affinity to one or more MHC I alleles (Supplementary Table S3b). HLA-A\*02:01, HLA-C\*01:02 and HLA-C\*04:01 were predicted as the three most frequent HLA alleles that bind to IEDB neoepitopes, while HLA-C\*07:01, HLA-C\*07:02, and HLA-C\*04:01 were predicted as the three most frequent HLA alleles that bind to TSNAdB neoepitopes (Supplementary Table S3b). Alas, only 13 out of 8787 neoepitopes in IEDB were predicted to have binding affinity to one or more MHC II alleles. This was not surprising since MHC II binding prediction is still an emerging field and has not been studied as much as MHC I binding.

Similar to the neoepitopes, we restrict our search to the gene fusions described in the ChimerKB dataset and are supported by published literature. We were able to retrieve 1096 unique gene fusion junctions. Nine or more CCDS-nullomers of length 16 were retrieved from each of the 1096 gene fusion junctions, and 182 of these junctions were found to generate putative neoepitopes.

### 3.2 Neoepitopes in cfRNA

Nullomers were used to detect putative neoepitopes in the plasma cfRNA of cancer patients and healthy donors in five public datasets that were sequenced using stranded RNA-seq libraries and contained  $\geq 10$  samples per condition. cfRNA for most cancer types in this study is derived from GSE174302, except multiple myeloma, which is derived from GSE182824, and hepatocellular carcinoma (HCC), which is merged from three different datasets (Supplementary Table S1). The healthy donors were also merged from three different datasets (Supplementary Table S1).

A median of only one putative neoepitope was discovered in cfRNA of multiple myeloma patients. By contrast, a median of 21 putative neoepitopes were discovered in cfRNA of HCC patients (Fig. 2a). This difference was not surprising as tumor mutational burden is generally higher in HCC patients (Klebanov *et al.* 2019, Gabbia and De Martin 2023). Surprisingly, a median of 6 putative neoepitopes were discovered in the cfRNA of healthy donors. Nonetheless, the neoepitopes from healthy cells are not expected to be immunogenic due to various factors, such as antigen processing, antigen presentation, and co-inhibitory signals. The number of putative neoepitopes discovered in HCC patients was noticeably higher than in healthy donors, the distributions of the two groups largely overlapped, suggesting that merely counting is not sufficient to distinguish the cancer

patients from healthy donors (Fig. 2a). As expected of the plasma derived cfRNA, the number of detected neoepitopes dropped quickly when higher read coverage cutoff was used (Supplementary Fig. S2). A median of two and five putative neoepitopes were detected in the plasma cfRNA of pancreatic cancer patients and healthy donors, respectively, in the GSE136651 dataset, which is sequenced at only 1/10th the coverage as compared to the GSE174302 dataset (Supplementary Fig. S3a).

Next, we used principal component analysis (PCA) to evaluate what drives the variance in the neoepitope repertoire of the HCC patients relative to the healthy donors from the three merged datasets. The first principal component (PC) primarily separates the HCC samples from the healthy ones, suggesting that the two groups differ in terms of their putative neoepitope repertoire (Fig. 2b).

We investigated which of the putative neoepitopes were differentially expressed in the cfRNA of cancer patients compared to the healthy donors. We first excluded generic neoepitopes that were highly expressed in cfRNA of cancer patients as well as healthy donors (Supplementary Table S4). Reassuringly, the generic neoepitopes mostly consisted of peptides derived from mutations in the mitochondrial genes (Supplementary Table S4). Then putative neoepitopes with  $\geq 2$ -fold differential expression in the cancer patients versus healthy donors were retrieved (see Section 2). Among the neoepitopes that were differentially expressed in HCC, nine (six shown) came from MUC12, “mucin 12 cell surface associated” protein, seven (six shown) came from HRNR, “hornerin” protein and seven (five shown) came from OR4C5, “olfactory receptor family 4 subfamily C member 5’ protein (Fig. 2c, Supplementary Table S5a). Further characteristics of these neoepitopes can be obtained from their binding data. For example, the MUC12 neoepitope, STISGHSEK, was found in cfRNA from 30 HCC patients and has a predicted binding affinity of 8.35 nM to HLA-A\*11:01. In comparison, the corresponding wild-type, STTSGHSEK, had a predicted binding affinity of 33.12 nM to HLA-A\*11:01. Fewer putative neoepitopes were enriched in the healthy donors, but surprisingly one of them was from TPT1, “tumor protein translationally controlled 1” protein. Similarly, discriminative neoepitopes were detected in cfRNA of other cancer types (Supplementary Figs S3 and S4 and Supplementary Table S5). Among the 250 putative neoepitopes that were differentially expressed in patients from all cancer types versus healthy, only 11 were differentially expressed in all 6 cancer types (Supplementary Table S5g and Supplementary Fig. S5). neoepitopes.

### 3.3 Benchmarking the tools for detecting known neoepitopes

As there is no tool available to discover the known neoepitopes from the epitope databases, we constructed a pipeline that uses variant calling tools and the Variant Effect Predictor to predict the known neoepitopes in the cfRNA data (Section 2). We benchmarked the nullomer approach for neoepitope detection against the two most widely used variant discovery tools, gatk haplotypeCaller (McKenna *et al.* 2010), and bcftools mpileup (Li 2011). In addition, we benchmarked against Lofreq (Wilm *et al.* 2012), which is especially well suited for low frequency variants. Interestingly, there was little overlap between the putative neoepitopes detected by the four methods with haplotypeCaller and

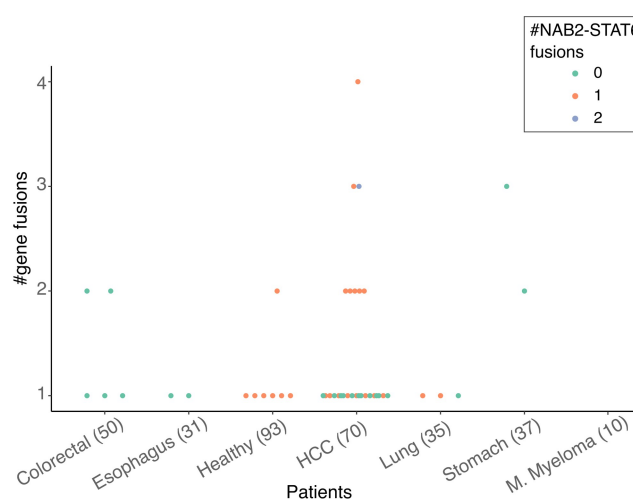
bcftools having the most overlap. (Supplementary Fig. S6a). There were only 71 unique putative neopeptides detected by all four methods in the GSE142987 dataset, and 12 in the GSE136651 dataset (Supplementary Table S6). This is not a surprise as the nullomer approach is much different than the other three, which share most of the components of their pipelines.

Notable differences include the fact that our approach uses the MES filter, which is designed to be more inclusive of short reads but removes reads with more mismatches than a dynamic threshold, which is estimated based on the mapped length of the read. For example, in this study we have allowed 1 medium quality mismatch (penalty = 4) for aligned length of 35 bases and ~4 medium quality mismatches (penalty = 15) for aligned length of 150 bases. Our approach is sensitive to read orientation, it ignores the antisense reads, and maps the reads to the CCDS using bowtie2 (Langmead and Salzberg 2012). All other tools shown here for the comparison use STAR (Dobin *et al.* 2013), which does a transcriptome-aware genome alignment. To test the difference in the mapping strategies, we used STAR to map the nullomer containing reads and then converted the genome coordinates to CCDS coordinates to estimate the read coverage of neopeptides using nullomers. Using STAR mapping increased the common neopeptides that were detected by all tools by 24% in GSE142987 and 120% in GSE136651 (Supplementary Fig. S6c and d). 51%–53% of the neopeptide-donor pairs that were detected using bowtie were also detected using STAR.

Another notable difference is that most tools filter out reads mapped with a low MAPQ score, which is highly biased against short reads. As cfRNA is mostly fragmented, the short reads are especially important. Hence, we devised a customized alignment-score based filter, which adjusts to the variations in read length. It is required that the reads with the nullomer must map to the neopeptide coding sequence or gene fusion junction corresponding to that nullomer. FastNeo can search for known neopeptides arising from up to two missense variants in the same read. There were about ten times fewer putative neopeptides detected by bcftools and haplotypeCaller in the low throughput data (Supplementary Fig. S6b) than in the high throughput data (Supplementary Fig. S6a). By contrast, FastNeo was able to detect the most putative neopeptides in the low throughput data (Supplementary Fig. S6b). As expected, the nullomer based approach is much faster than any of the other approaches (Supplementary Fig. S7).

### 3.4 Gene fusions in cfRNA

Next, we used the nullomer approach to detect putative gene fusions in the plasma cfRNA of cancer patients and healthy donors. Similar to before, we restrict our search to the gene fusions published and curated in the ChimerDB database (Jang *et al.* 2020). The total number of putative gene fusions detected was small, with no more than four events in any of the patients (Fig. 3). The known clinically relevant gene fusions discovered in the cfRNA data consisted of EML4-ALK fusion in one HCC patient and ETV6 fusions in one stomach cancer, nine HCC and four colorectal cancer patients (Ford *et al.* 1998, Iijima *et al.* 2000, Bohlander 2005, Taniue and Akimitsu 2021). The number of fusions detected was low and dropped quickly when higher read coverage cut-off was used (Supplementary Fig. S8). The low numbers were



**Figure 3.** Number of putative gene fusions per patient grouped by healthy donors and cancer type. Patients with NAB2-STAT6 gene fusions are coloured as per legend.

not surprising as it has been shown that gene fusions drive cancer progression in only 16.5% of cases (Gao *et al.* 2018). Overall, gene fusions were more common in the cfRNA of cancer patients as compared to healthy individuals, consistent with previous studies (Gao *et al.* 2018). NAB2 and STAT6 gene fusion in chr12 (between 57486364: + strand and 57490916, – strand) was detected in healthy donors and HCC patients from two of the three merged datasets. Interestingly, seven of the eight gene fusion events detected in the healthy individuals involved the fusion of NAB2 and STAT6 (Fig. 3). The third common fusion was YPEL5-PPP1CB, which was detected in nine different cancer patients (Supplementary Table S7) and has also been described as a “recurrent reciprocal RNA chimera” (Velusamy *et al.* 2013).

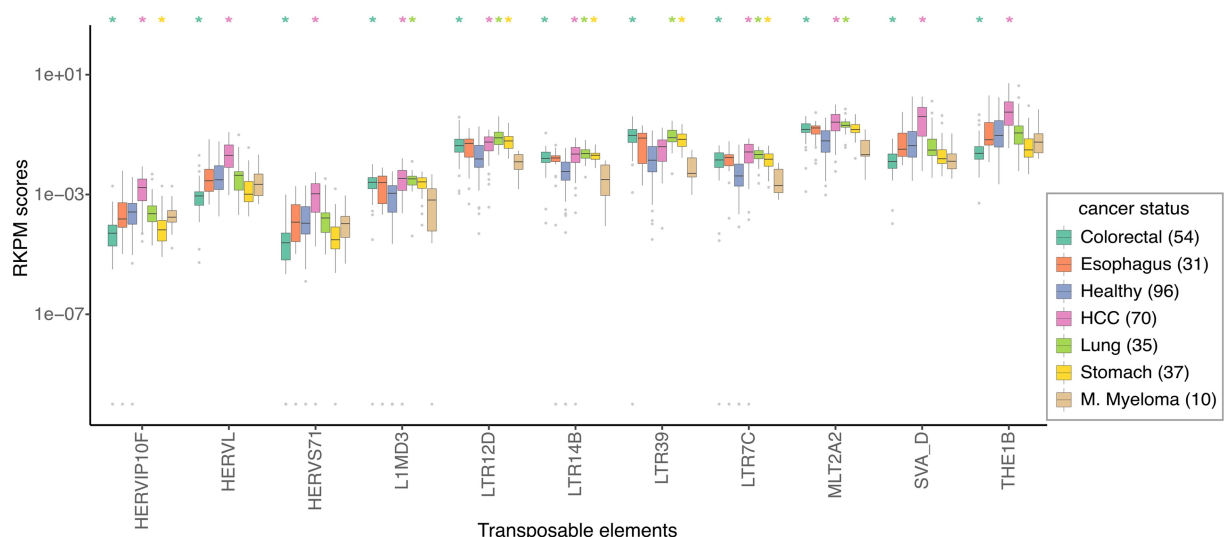
### 3.5 TEs in cfRNA

The plasma cfRNA from both healthy donors and cancer patients was found to be rich in TEs. Among the 687 TE classes quantified by SalmonTE (Jeong *et al.* 2017), 279 were found to be differentially expressed between healthy donors and cancer patients (Supplementary Table S8). The selected TEs, shown in Fig. 4, are among the most differentially expressed in multiple cancer types. Interestingly, 35, 45, and 5 Alu TEs were found to be differentially expressed in the cfRNA of pancreatic cancer, HCC, and colorectal cancer patients, respectively (Supplementary Fig. S9 and Supplementary Table S8). Similarly, 11, 13, 60, and 19 long terminal repeat (LTR) TEs were differentially expressed in cfRNA of stomach cancer, lung cancer, HCC, and colorectal cancer patients, respectively (Supplementary Table S8a). Finally, 26, 11, and 1 HERV TEs were differentially expressed in HCC and colorectal cancer, and stomach cancer patients, respectively (Supplementary Table S8a). Expression of some TE classes, such as HERVIP10F, were significantly lower in colorectal cancer and stomach cancer patients as compared to the healthy donors (Fig. 4, Supplementary Data: Supplementary Table S8a).

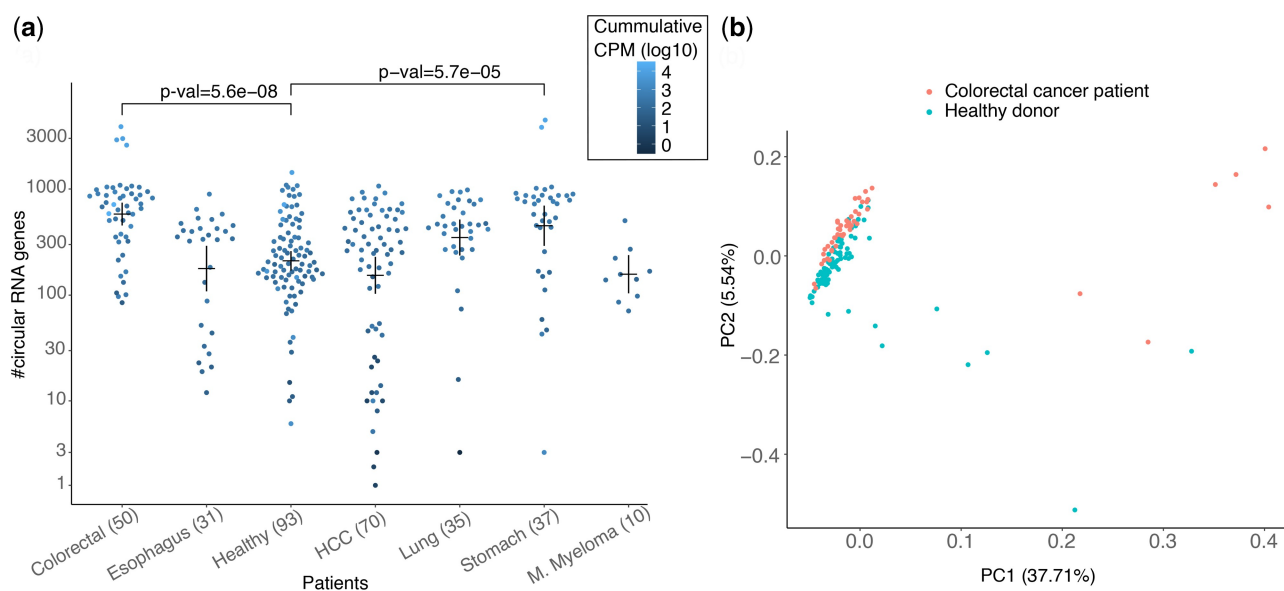
### 3.6 circRNAs in cfRNA

CIRI-quant was used for de-novo detection of circRNAs in the cfRNA (Zhang *et al.* 2020). The number of circRNAs detected in the cfRNA of HCC patients varied from 1 to





**Figure 4.** A box and whisker plot showing the top TE that are differentially expressed (Wilcoxon test corrected  $P$ -value  $< 1e-08$ ) in the cfRNA of healthy donors versus patients of any two cancer types. Only 4 of the 10 such LTR TE have been shown (see [Supplementary Table S8](#)). The cancer-types with significantly different distribution of the TE expression are marked on top with the colour matched asterisk. Outliers with RKPM score  $\leq 1e-10$  were assumed to have RKPM score  $= 1e-10$ . The horizontal line corresponds to the median, two hinges correspond to the 25th and 75th percentiles, and two whiskers correspond to the largest and smallest value no further than 1.5 times the distance between 25th and 75th percentiles.



**Figure 5.** (a) Number of predicted circRNAs in the cfRNA of patients grouped by cancer type and healthy. Each dot is coloured by cumulative CPM of all the circRNAs in the respective cfRNA sample.  $P$ -values of two cancer types with most significant differences in distribution of circRNAs in healthy donors versus patients are shown.  $P$ -values were calculated using the Wilcoxon test with Bonferroni correction. Means and error-bar are shown for each group. (b) First two components of a PCA using log of CPM scores of all circRNAs detected in the colorectal cancer patients and healthy donors.

~1000 and the circRNAs detected in healthy donors followed a similar distribution ([Fig. 5a](#)). The number of circRNAs detected in the cfRNA of Colorectal cancer patients were significantly higher than healthy donors ([Fig. 5a](#)), and the first two PCs were able to partially segregate the cluster of colorectal cancer patients from the healthy donors ([Fig. 5b](#)). Even while using moderate cutoffs for  $P$ -value and recurrence in groups of donors, only a few circRNAs were found to be differentially expressed in cancer patients against the healthy controls ([Supplementary Table S9](#)). 77, 25, 38 differentially expressed circRNAs were only found in colorectal, stomach, and lung cancer, and ZCCHC7 was the only one such circRNA that was found in HCC

([Supplementary Table S9a](#)). In the above analysis, only the most expressed circRNAs were used for each gene. Assuming that each combination of the back-spliced exons functions differently and can be uniquely linked to a disease, we repeated the above analysis using all circRNAs per gene. In this analysis, 22 additional circRNAs were differentially expressed in colorectal cancer, and 9 of the genes had 2–3 circRNAs per gene that were differentially expressed in colorectal cancer ([Supplementary Table S9a and b](#)). In both cases, circRNA from the ZCCHC7 locus was the only differentially expressed circRNA in HCC. The variation in circRNA counts in the cfRNA of different cancer patients and the healthy was very similar in both analysis ([Fig. 5](#), [Supplementary Fig. S10](#)).

Thus, for the classifiers in the next section, only the circRNA with highest CPM score from each gene was used.

### 3.7 Clustering and supervised learning

To demonstrate that it is possible to discriminate between healthy and cancer samples based on the biomarkers discovered in cfRNA, we developed classifiers using expression of putative neopeptides TEs and circRNAs as features. Gene fusions were not used for classification due to low recurrency among the samples used in this study, but they can easily be incorporated into our framework. The classifiers trained using all three modalities performed the best (Supplementary Fig. S11). Using 5-fold cross validation, random forest classifiers reached a balanced accuracy of 0.81 and 0.84 in HCC and colorectal cancer, respectively (Table 1; Supplementary Tables S10). The classifiers were able to identify cancer patients with high sensitivity and specificity. Among 96 healthy donors and 66 HCC patients in the merged dataset, a selected TE was expressed in average 112.52 (median 126) cfRNA donors, while a selected neopeptide was only expressed in average 3.24 (median 1) cfRNA donors. Thus, the classifier must be able to handle this difference in recurrence of TE and neopeptide expression. The performance of Ridge and Lasso classifiers suffers while training on a merged feature-set, as the optimal value of regularization parameter lambda differs for the TEs and neopeptides. The random forest classifier performed overall better than SVM, Lasso and Ridge regression (Supplementary Tables S10–S13). We also observed that samples with little or no expression of the selected features were the ones that were prone to wrong classification.

The expression level of human and microbe derived transcripts in plasma cfRNA has been shown to achieve an AUROC > 90% for distinguishing HCC and colorectal cancer patients from the healthy controls (Chen *et al.* 2022). In another study, expression of 3 ncRNAs (SNORD3B-1, circ-0080695, and miR-122) obtained the highest average AUROC of 89.4% (Zhu *et al.* 2021). In comparison, our classifier trained on putative neopeptides, TEs, and circRNAs in plasma cfRNA achieved an average AUROC of 87% for HCC and 92% for colorectal cancer (Fig. 6, Supplementary Table S10e and j).

Cross-validation can result in an overestimate of the classifier's ability to assess new samples. To test the classifier performance across batches, a random forest classifier was trained for HCC detection using donors from two of the

datasets at a time and tested on the remaining one dataset. The merged set of features was used with the same selection strategy as above. The best performing classifier was the one trained on the features detected in cfRNA of 35 HCC patients, and 66 healthy donors in the GSE182824 and GSE174302 datasets. It achieved a balanced accuracy of 0.68 when tested on the 35 HCC patients, and 30 healthy donors in the GSE142987 dataset (Supplementary Table S15). 20 out of 103 putative neopeptides that were selected to train the above classifier were not expressed in any of the samples in the test-set. Overall, the results suggest that batch specific noise is present in the cfRNA, however, features that are recurrent in a disease-specific manner are also present.

## 4 Discussion

Cancer causes systemic changes, some of which can be detected in the cell free nucleic acids from body fluids. In this study we have characterized some of the biomarkers that can be derived from cfRNA and used to detect cancer. We used nullomers to identify putative neopeptides, and we demonstrated that the composition of putative neopeptides can be used to distinguish cancer patients from the healthy controls. Nullomers were also utilized to identify putative gene fusions, which are hard to detect in RNAseq data (Haas *et al.* 2019), and cannot be easily distinguished from other chimeric RNA (Velusamy *et al.* 2013). Although the number of detected gene fusions was small, they were mostly found in cancer patients.

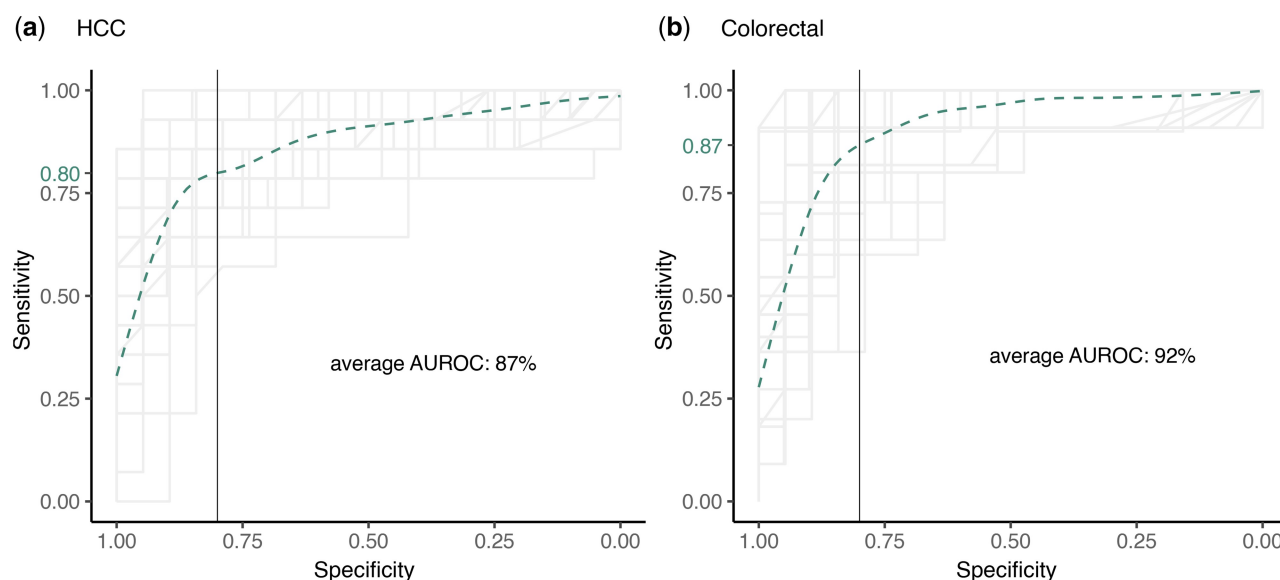
We showed that the TEs and circRNAs in the cfRNA are also rich in cancer-specific biomarkers. We combined these two independent modalities with the neopeptides, to obtain a classifier with a balanced accuracy of 0.84 for colorectal cancer, and 0.81 for HCC. Among these three modalities, the cancer classifiers trained using only TE expression performed best, and the classifiers trained using only circRNAs performed the worst. Currently available tools for circRNA prediction are known for low sensitivity and require RNase P treated samples to improve the accuracy (Vromman *et al.* 2025).

We introduce a nullomer-based approach for rapid detection of known neopeptides and gene fusions directly in the cfRNA data. The advantage of using the nullomer-based approach is that it does not require any genotype information, is inclusive of fragmented short reads, uses stringent criteria to filter variants arising due to technical artifacts and filters

**Table 1.** Random Forest classifier performance averaged from 10 iterations of 5-fold cross-validation.<sup>a</sup>

Cancer type	Features	F1	Balanced accuracy	Specificity	Sensitivity
HCC	Neopeptides	0.80	0.75	0.63	0.86
HCC	TEs	0.85	0.81	0.74	0.87
HCC	circRNA	0.72	0.65	0.56	0.74
HCC	Neopeptides +TEs	0.84	0.81	0.80	0.83
HCC	Neopeptides +TEs +circRNA	0.84	0.81	0.79	0.84
CRC	neopeptides	0.82	0.76	0.70	0.83
CRC	TEs	0.85	0.78	0.71	0.86
CRC	circRNA	0.83	0.75	0.63	0.86
CRC	Neopeptides +TEs	0.87	0.82	0.76	0.88
CRC	Neopeptides +TEs +circRNA	0.88	0.84	0.80	0.87

<sup>a</sup> As the performance metrics were estimated for three different sets of features, samples with no expressed features were included.



**Figure 6.** Overlapped ROCs of all 5 folds of the 10 iterations of 5-fold cross validation of the random forest classifier while classifying (a) HCC patients, and (b) colorectal cancer patients among the healthy donors. The classifier was trained using expression of neoepitopes, circRNAs and TEs in the cfRNA samples. Each ROC line was plotted with  $\alpha = 0.1$  and the specificity of 0.8 is marked with a solid vertical line. Sensitivity modeled by fitting a generalized additive model to all the ROCs is shown as a dotted green curve, and the sensitivity predicted at a specificity of 0.8 is shown as a green label on the y-axis.

germline variants. In summary, the nullomer approach is fast, has minimal dependencies, and finds neoepitopes that existing tools tend to miss. The work presented here demonstrates how multiple signals from cfRNA can be leveraged to provide robust cancer detection across a variety of cancer types. This includes the use of our newly developed FastNeo method which can provide therapeutically relevant information through the identification of putative neoepitopes. It also shows the binding affinity of the epitopes against the known HLA variants, and the frequency of the neoepitope in the known human germline variants.

Some promising patterns were observed but substantially more cfRNA data is needed from both healthy donors and patients from multiple cancer types and stages to be able to more confidently assess the accuracy of our method and distinguish the recurrent biomarkers from noise. One major challenge is the relatively small number of cfRNA samples available to us. Hence, we were unable to apply a more stringent test-train-validation scheme which would have allowed us to also highlight which features are the most informative for discriminating patients from healthy controls.

Compared to the classifiers developed by the authors of the dataset that we analysed, we found that our method has similar performance. A key difference is that our framework is generally applicable across cancer types. Moreover, our method provides additional, clinically relevant information in the form of putative neoepitopes. However, the cells that produce the discriminative neoepitopes, and their ability to activate immune cells against tumors must be verified experimentally. The growing repertoire of experimental validated neoepitopes and gene fusions may also improve the predictions of this method. An important validation experiment to follow up on this study is to profile the tumor as well, ideally with mass-spectrometry, to validate that the predicted neoepitopes are indeed displayed on the malignant cells. To the best of our knowledge no such dataset exists in the public domain today, so both blood and tumor samples would have to be first collected and profiled.

Another advantage of our method is that it is independent of other modalities, and as such it can be combined with approaches that consider DNA or protein biomarkers in a liquid biopsy to detect cancer. Together with RNA from peripheral immune cells and RNA from circulating intact tumor cells cfRNA can provide additional biomarkers that could be tracked during cancer diagnosis and progression. It has become clear that raw tumor mutation burden is a poor predictor of cancer diagnosis, tumor foreignness and response to immunotherapy however, biologically relevant subsets of the tumor mutational burden, such as neoepitope producing mutations might provide a more reliable predictive metrics (Wood *et al.* 2020). Overall, detecting relevant biomarkers in cfRNA and tracking them over a period could be beneficial in disease detection and prognosis.

## Author contributions

Mayank Mahajan (Data curation [equal], Formal analysis [equal], Methodology [equal], Software [equal], Validation [equal], Visualization [equal], Writing—original draft [equal], Writing—review & editing [equal]) and Martin Hemberg (Conceptualization [equal], Funding acquisition [equal], Methodology [equal], Software [equal], Supervision [equal], Writing—original draft [equal], Writing—review & editing [equal])

## Supplementary data

[Supplementary data](#) are available at *Bioinformatics* online.

Conflict of interest: None declared.

## Funding

M.M. and M.H. were funded by startup funds from the Evergrande Center.



## References

- Alsosaimi S, van Biljon N, Awany D *et al.* Simulation of African and non-African low and high coverage whole genome sequence data to assess variant calling approaches. *Brief Bioinform* 2021;22:bbaa366.
- Bachmayr-Heyda A, Reiner AT, Auer K *et al.* Correlation of circular RNA abundance with proliferation—exemplified with colorectal and ovarian cancer, idiopathic lung fibrosis and normal human tissues. *Sci Rep* 2015;5:8057.
- Biernacki MA, Foster KA, Woodward KB *et al.* CBFB-MYH11 fusion neoantigen enables T cell recognition and killing of acute myeloid leukemia. *J Clin Invest* 2020;130:5127–41.
- Bohlander SK. ETV6: a versatile player in leukemogenesis. *Seminars Cancer Biol* 2005;15:162–74.
- Bronkhorst AJ, Ungerer V, Holdenrieder S *et al.* The emerging role of cell-free DNA as a molecular marker for cancer management. *Biomol Detect Quantif* 2019;17:100087.
- Brown SD, Warren RL, Gibb EA *et al.* Neo-antigens predicted by tumor genome meta-analysis correlate with increased patient survival. *Genome Res* 2014;24:743–50.
- Büscher K, Trefzer U, Hofmann M *et al.* Expression of human endogenous retrovirus K in melanomas and melanoma cell lines. *Cancer Res* 2005;65:4172–80.
- Chang A, Loy C, Eweis-LaBolle D *et al.* Circulating cell-free RNA in blood as a host response biomarker for the detection of tuberculosis. *Nat Comms* 2024;15:4949.
- Chen S, Francioli LC, Goodrich JK *et al.* A genomic mutational constraint map using variation in 76,156 human genomes. *Nature* 2024;625:92–100. <https://doi.org/10.1038/s41586-023-06045-0>
- Chen S, Jin Y, Wang S *et al.* Cancer type classification using plasma cell-free RNAs derived from human and microbes. *Elife* 2022;11:e75181.
- Corcoran Ryan B, Chabner Bruce A. Application of cell-free DNA analysis to cancer treatment. *N Engl J Med* 2018;379:1754–65.
- Cristiano S, Leal A, Phallen J *et al.* Genome-wide cell-free DNA fragmentation in patients with cancer. *Nature* 2019;570:385–9.
- Danecek P, Bonfield JK, Liddle J *et al.* Twelve years of SAMtools and BCFtools. *Gigascience* 2021;10:giab008.
- Dhanda SK, Mahajan S, Paul S *et al.* IEDB-AR: immune epitope database—analysis resource in 2019. *Nucleic Acids Res* 2019;47:W502–6.
- Dobin A, Davis CA, Schlesinger F *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* 2013;29:15–21.
- Faulkner GJ, Kimura Y, Daub CO *et al.* The regulated retrotransposon transcriptome of mammalian cells. *Nat Genet* 2009;41:563–71.
- Ford AM, Bennett CA, Price CM *et al.* Fetal origins of the TEL-AML1 fusion gene in identical twins with leukemia. *Proc Natl Acad Sci USA* 1998;95:4584–8.
- Fortakis G, Rieder D, Haider M *et al.* NeoFuse: predicting fusion neoantigens from RNA sequencing data. *Bioinformatics* 2020;36:2260–1.
- Gabbia D, De Martin S. Tumor mutational burden for predicting prognosis and therapy outcome of hepatocellular carcinoma. *Int J Mol Sci* 2023;24:3441.
- Gao Q, Liang W-W, Foltz SM *et al.* Cancer Genome Atlas Research Network. Driver fusions and their implications in the development and treatment of human cancers. *Cell Rep* 2018;23:227–38.e3.
- Georgakopoulos-Soares I, Yizhar-Barnea O, Mouratidis I *et al.* Absent from DNA and protein: genomic characterization of nullomers and nullpeptides across functional categories and evolution. *Genome Biol* 2021a;22:245.
- Georgakopoulos-Soares I, Barnea O, Mouratidis I *et al.* Leveraging sequences missing from the human genome to diagnose cancer. medRxiv, 2021b.
- Gonzalez-Galarza FF, McCabe A, Santos EJMD *et al.* Allele frequency net database (AFND) 2020 update: gold-standard data classification, open access genotype data and new query tools. *Nucleic Acids Res* 2020;48:D783–8.
- Haas BJ, Dobin A, Li B *et al.* Accuracy assessment of fusion transcript detection via read-mapping and de novo fusion transcript assembly-based methods. *Genome Biol* 2019;20:213.
- Hamed CT, Meiloud G, Veten F *et al.* HLA class I (-A, -B, -C) and class II (-DR, -DQ) polymorphism in the Mauritanian population. *BMC Med Genet* 2018;19:2.
- Iijima Y, Ito T, Oikawa T *et al.* A new ETV6/TEL partner gene, ARG (ABL-related gene or ABL2), identified in an AML-M3 cell line with a t(1; 12)(q25; p13) translocation. *Blood* 2000;95:2126–31.
- Jang YE, Jang I, Kim S *et al.* ChimerDB 4.0: an updated and expanded database of fusion genes. *Nucleic Acids Res* 2020;48:D817–24.
- Jeong H-H, Yalamanchili H, Guo C *et al.* An ultra-fast and scalable quantification pipeline for transposable elements from next generation sequencing data. In: Altman R, Dunker K, Ritchie M *et al.* (eds), *Biocomputing* 2018. Singapore: World Scientific, 2017, 168–79.
- Jurtz V, Paul S, Andreatta M *et al.* NetMHCpan-4.0: improved peptide-MHC class I interaction predictions integrating eluted ligand and peptide binding affinity data. *J Immunol* 2017;199:3360–8.
- Khan T, Rahman M, Ahmed I *et al.* Human leukocyte antigen class II gene diversity tunes antibody repertoires to common pathogens. *Front Immunol* 2022;13:856497.
- Klebanov N, Artomov M, Goggins WB *et al.* Burden of unique and low prevalence somatic mutations correlates with cancer survival. *Sci Rep* 2019;9:4848.
- Koşaloğlu-Yalçın Z, Blazeska N, Vita R *et al.* The cancer epitope database and analysis resource (CEDAR). *Nucleic Acids Res* 2023;51:D845–52.
- Kumar H, Luo R, Wen J *et al.* FusionNeoAntigen: a resource of fusion gene-specific neoantigens. *Nucleic Acids Res* 2024;52:D1276–88.
- Lang F, Schrörs B, Löwer M *et al.* Identification of neoantigens for individualized therapeutic cancer vaccines. *Nat Rev Drug Discov* 2022;21:261–82.
- Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. *Nat Methods* 2012;9:357–9.
- Larson MH, Pan W, Kim HJ *et al.* A comprehensive characterization of the cell-free transcriptome reveals tissue- and subtype-specific biomarkers for cancer detection. *Nat Commun* 2021;12:2357.
- Li H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* 2011;27:2987–93.
- Li Y, Zheng Q, Bao C *et al.* Circular RNA is enriched and stable in exosomes: a promising biomarker for cancer diagnosis. *Cell Res* 2015;25:981–4.
- McKenna A, Hanna M, Banks E *et al.* The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* 2010;20:1297–303.
- McLaren W, Gil L, Hunt SE *et al.* The ensembl variant effect predictor. *Genome Biol* 2016;17:122.
- Memczak S, Jens M, Elefsinioti A *et al.* Circular RNAs are a large class of animal RNAs with regulatory potency. *Nature* 2013;495:333–8.
- Montgomery A, Tsiatsianis GC, Mouratidis I *et al.* Utilizing nullomers in cell-free RNA for early cancer detection. *Cancer Gene Ther* 2024;31:861–70.
- Ott PA, Hu Z, Keskin DB *et al.* An immunogenic personal neoantigen vaccine for patients with melanoma. *Nature* 2017;547:217–21.
- Pedersen MB, Assmusen SR, Sarfelt FM *et al.* Integration of HLA-DR linkage disequilibrium to MHC class II predictions. bioRxiv:2023.
- Reggiardo RE, Maroli SV, Peddu V *et al.* Profiling of repetitive RNA sequences in the blood plasma of patients with cancer. *Nat Biomed Eng* 2023;7:1627–35.
- Reynisson B, Barra C, Kaabinejadian S *et al.* Improved prediction of MHC II antigen presentation through integration and motif deconvolution of mass spectrometry MHC eluted ligand data. *J Proteome Res* 2020;19:2304–15.
- Rodić N, Sharma R, Sharma R *et al.* Long interspersed element-1 protein expression is a hallmark of many human cancers. *Am J Pathol* 2014;184:1280–6.
- Roskams-Hieter B, Kim HJ, Anur P *et al.* Plasma cell-free RNA profiling distinguishes cancers from pre-malignant conditions in solid and hematologic malignancies. *NPJ Precis Onc* 2022;6:1–11.
- Roy S, Kanda M, Nomura S *et al.* Diagnostic efficacy of circular RNAs as noninvasive, liquid biopsy biomarkers for early detection of gastric cancer. *Mol Cancer* 2022;21:42.

- Satapornpong P, Jinda P, Jantararoungtong T *et al.* Genetic diversity of HLA class I and class II alleles in Thai populations: contribution to genotype-guided therapeutics. *Front Pharmacol* 2020;**11**:78.
- Stransky N, Cerami E, Schalm S *et al.* The landscape of kinase fusions in cancer. *Nat Commun* 2014;**5**:4846.
- Tan C, Cao J, Chen L *et al.* Noncoding RNAs serve as diagnosis and prognosis biomarkers for hepatocellular carcinoma. *Clin Chem* 2019;**65**:905–15.
- Taniue K, Akimitsu N. Fusion genes and RNAs in cancer development. *Non-Coding RNA* 2021;**7**:10.
- Velusamy T, Palanisamy N, Kalyana-Sundaram S *et al.* Recurrent reciprocal RNA chimera involving YPEL5 and PPP1CB in chronic lymphocytic leukemia. *Proc Natl Acad Sci USA* 2013;**110**:3035–40.
- Vita R, Mahajan S, Overton JA *et al.* The immune epitope database (IEDB): 2018 update. *Nucleic Acids Res* 2019;**47**:D339–43.
- Vo JN, Cieslik M, Zhang Y *et al.* The landscape of circular RNA in cancer. *Cell* 2019;**176**:869–81.e13.
- Vromman M, Anckaert J, Bortoluzzi S *et al.* Large-scale benchmarking of circRNA detection tools reveals large differences in sensitivity but not in precision. *Nat Methods* 2025;**22**:448–1169.
- Weber D, Ibn-Salem J, Sorn P *et al.* Accurate detection of tumor-specific gene fusions reveals strongly immunogenic personal neo-antigens. *Nat Biotechnol* 2022;**40**:1276–84.
- Wilm A, Aw PPK, Bertrand D *et al.* LoFreq: a sequence-quality aware, ultra-sensitive variant caller for uncovering cell-population heterogeneity from high-throughput sequencing datasets. *Nucleic Acids Res* 2012;**40**:11189–201.
- Wood MA, Weeder BR, David JK *et al.* Burden of tumor mutations, neoepitopes, and other variants are weak predictors of cancer immunotherapy response and overall survival. *Genome Med* 2020;**12**:33.
- Wu J, Zhao W, Zhou B *et al.* TSNAdb: a database for tumor-specific neoantigens from immunogenomics data analysis. *Genomics Proteomics Bioinf* 2018;**16**:276–82.
- Zhang J, Chen S, Yang J *et al.* Accurate quantification of circular RNAs identifies extensive circular isoform switching events. *Nat Commun* 2020;**11**:90.
- Zhu Y, Wang S, Xi X *et al.* Integrative analysis of long extracellular RNAs reveals a detection panel of noncoding RNAs for liver cancer. *Theranostics* 2021;**11**:181–93.