Strategy for the identification and prioritization of candidate neoantigens from large-scale NGS data

Javier Lanillos Manchón

Master's Program in Molecular Techniques in Life Sciences

2018

Valtteri Wirta, PhD
Hassan Foroughi, PhD
Clinical Genomics Unit | Science for Life Laboratory

Ola Nilsson, PhD
Center for Molecular Medicine (CMM) | Karolinska Institutet

Index

TABLE OF FIGURES

Abstract

***The clinical potential for personalized medicine in cancer immunotherapy** is nowadays becoming a plausible option in many advanced cases. These approaches exploit the natural ability of the immune system to patrol, identify, target and destroy tumor cells. Specifically, some strategies base on tumor-specific antigens or neoantigens, which direct patient-specific T-cells to react against the tumor.*

This project is involved in one of the very first step towards this personalized medicine approach: the robust identification of suitable neoantigens only present in the tumor sample by bioinformatics analysis of Next Generation Sequencing (NGS) or Massive Parallel Sequencing (MPS) technologies data. This report first describes the Whole Exome and RNA Sequencing data generation strategy using Twist Biosciences Solution-Hybrid Selection library preparation and Illumina Sequencing technologies. Second, a bioinformatics workflow, which combines Vardict, Strelka and Mutect2 variant callers, is proposed to identify nonsynonymous somatic mutations. Next, patient-specific HLA alleles and candidate neoantigens are predicted using HLAminer and pVACseq tools. Finally, neoantigens are prioritized based on variant allele frequency, gene expression analysis, MHC-binding affinity prediction and data visualization. Results from the analysis of matched-pair tumor and normal samples from two urinary bladder cancer patients have shown features to take into account such as the sample quality, the performance of variant callers, the importance to ensure a correct gene expression analysis and the search for validation of predicted neoantigens.

Introduction

***Cancer is a worldwide and urgent problem**.* By 2030, 13 million people will die worldwide due to cancer (1). Cancer can affect to any individual, although its prevalence in the population varies among different groups, based on factors such as gender, lifestyle and economic income. For example, cancer is more prevalent among elderly as well as among smokers (1.5 million deaths due to cancer are linked to tobacco) and individuals with an unhealthy diet or with high alcohol consumption (2). The incidence rate for cancer will double in low-middle-income countries in comparison to high-income countries in 2030 (3). According to global studies, the most common cancer type among women in high-income countries is breast cancer, while in low-income countries cervical cancer is the most prevalent (3). Just in 2010, the estimated cost in cancer medicine was US$ 1.16 trillion (4). Much research has been focused on the development of new and more efficient drugs and treatments for various types of cancer (5). Nevertheless, cancer remains one of the major causes of death. Successful treatment of cancer is not only dependent on the availability of effective treatments, but also on the ability to detect the cancer during early stages and to accurately define each patient's status and cancer profile to offer the best treatment. In fact, a key to success when fighting cancer is the so-called precision medicine (6), although it is still mostly in research level. These achievement and milestones have been possible through the establishment of several technologies and specialties into the clinic. One of those technologies is Massive Parallel Sequencing (MPS), also known as Next Generation Sequencing (NGS) whose costs have been significantly dropping (7,8), thus enabling a patient-focused sequencing profiling. And one of those specialties is bioinformatics, which is the field that covers the necessity to analyze MPS-derived data (9).

***Screenshot of cancer**.* Cancer is a group of diseases involving abnormal cell proliferation of cancer cells with the potential to invade or spread to other parts of the body beyond normal tissue boundaries (10). This atypical cell growth appears in cells whose genomes have accumulated permanent deleterious changes or mutations, which break homeostasis and contributes the onset of cancer. It is known that cancer begins at a single cell level and in any part of the body (10,11); a normal cell must undergo several subsequent transformations in the genome to become a malignant cancer cell, particularly on genes involved in cell growth regulation and differentiation, which must be altered in order to acquire a succession of hallmark capabilities (10). Malignancy of cells is also raised by the formation and over-expression of novel oncogenes (genes with the potential to cause cancer disease) and by the under expression or disabling of tumor suppressor genes (genes that preserve normal cells to undergo steps towards their transformation into cancer cells) (10). The accumulation of multiple changes is often required to turn a normal cell into a cancerous cell (11). The classification of cancer is currently being rewritten (12) to improve characterization, for example, by trying to quantify the clinical impact of each tumor type by looking into mutational signatures and mutation burden (13). However, statistics nowadays classify cancer according to the affected tissue, where lung, breast, colorectal, stomach and liver cancer are the most frequent in the population (4,14).

***Cancer Immunotherapy*.** Surgical, chemo, radio, targeted and hormone therapy are several strategies that have emerged historically (15) and have been well-implemented into the clinics to treat cancer. However, their effectiveness is not guaranteed, as each patient will respond differently to one given treatment (16). The immune system has been described as a competent agent in finding and destroying cancer cells. Immunotherapy-based treatments aim to vitalize the self-patient’s cells of the immune system to bring out a response against tumor cells and eradicate them (17). The stimulation of the immune system to successfully treat cancer is a considered a holy grail amongst researchers and clinicians since it is well known that cancers develop mechanisms to evade the attack of T-cells (18). Diverse immunotherapy drugs and their combinations have demonstrated the clinical benefit of targeting tumor cells via the immune system (19,20); i.e. checkpoint blockade inhibitors, antibody-drug conjugates, cancer vaccines, engineered CAR T-cells and T-cell transplants (19). Nowadays, much attention is put on neoantigens, which arise due to nonsynonymous somatic mutations as suitable candidates to develop personalized and targeted immunotherapies (21,22). In addition, MPS technologies enable screening of the specific tumor profile of each patient to detect the best neoantigen targets (23).

Context

This master thesis project is motivated by the extreme necessity of enabling a bioinformatics pipeline to identify neoantigens from MPS data and apply them to develop a new cancer immunotherapy approach based on adoptive T-cell transfer therapy (21). TCER AB is a Swedish private company with a close collaboration with the Therapeutic Immune Design research group at the Centre for Molecular Medicine (CMM) held at Karolinska Institutet. The Clinical Genomics (CG) Unit (24) is allocated at the Science for Life Laboratory and provides with MPS technology infrastructure and bioinformatics services for clinical and research projects within a strong national and international collaboration. This unit assists in the innovative translation of sequencing-based tools into the clinics (rare diseases, cancer, microbiome) hosting interdisciplinary and highly trained staff. This master thesis gathers the collaboration between TCER and CG in order to solve the Neoantigen Prediction step highlighted in *Figure 1*. The project workflow (*Figure 1*) starts with the recruitment of urinary bladder cancer patients at the hospital. Under the pertinent ethical approval and patient consent, surgeons perform surgical extraction of the tumor sample, and both tumor DNA and RNA material are extracted in the laboratory. A blood sample is also taken to isolate Peripheral Monoclonal Blood Cells for extraction of normal DNA and RNA sample material (normal-matched). During a second surgery where the patient has the urinary bladder removed, tumor-draining lymph nodes are also excised for the isolation of the immune cells that are later used during the T-cell therapy in development.

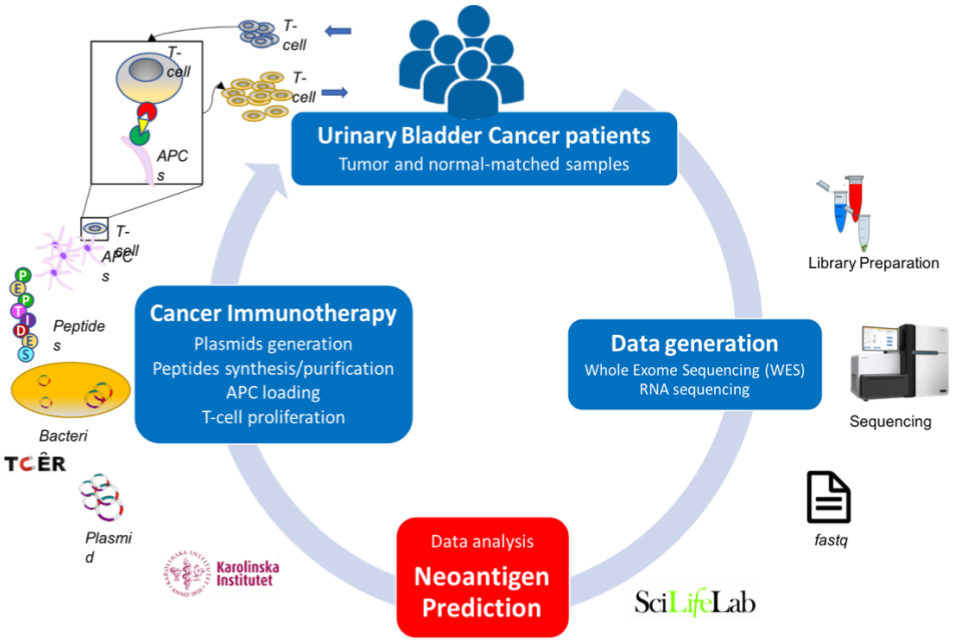


Figure 1. Project workflow. The goal of the master thesis project is the MPS data analysis (highlighted in red)

During this master thesis, tumor and normal samples are prepared for sequencing and subsequently analyzed using bioinformatics. This step aims to reveal candidate neoantigens that will be serve as starting point for the technology development in TCER AB. The most suitable neoantigens are cloned by plasmid designs, transformed into bacteria, expressed and purified afterwards. The newly produced neoantigens are used for *ex-vivo* stimulation of Antigen Presenting Cells (APCs), which in turn present shorter peptides in MHC-class I and II molecules to T-cells in the same culture well. Any T-cell with a positive immune activation to any of the neoantigens will undergo expansion and proliferation into a specific T-cell population. Thus, the last step of the workflow would include preparation and inoculation of a proliferated T-cell population in the patient, which hopefully will target and react against specific tumor cells expressing the neoantigens predicted. Therefore, **the detection of suitable neoantigens is crucial first step for the success** of this project.

Goals

1. Decision of a data generation strategy compatible with the scientific questions
2. Quality Control (QC) assessment of data generated
3. Establish variant calling, prioritization and annotation strategy
4. Assessment of the value of incorporating RNA sequencing data
5. Neoantigen prioritization strategy for generating peptides
6. Bioinformatics workflow construction of the steps 2-4

Background

This section reviews key concepts in sequencing, the biology behind the generation of a neoantigen and surveys the state-of-the-art in neoantigen prediction and prioritization.

Sequencing

Whole Exome Sequencing (WES) has been successfully used to identify variants in the protein-coding genes (25,26). Most of the sequencing approaches to identify tumor neoantigens are based on WES (27–29), which has proven to be the most followed sequencing strategy in neoantigen prediction (30). This approach reduces sequencing costs with respect to Whole Genome Sequencing (WGS) and it is suitable for this project in which only actual transcribed and translated genes (exome ~1% of the total genome) are the target of interest. However, there are some trade-offs to bear in mind using WES approach like off-target read (31). The final sequencing strategy needs to take into account important aspects related to the library preparation and sequencing depth to achieve enough read depth coverage in the data. Achieving the required coverage to overcome any specific research question in sequencing-based cancer projects is another handicap. Some studies have addressed this issue and report that 30x WGS might not be enough to get enough data as to cover all coding-proteins (32). Regarding WES coverage, it has been reported that 120x could detect 100% of variants present in the exome (33).

Targeted capture library preparation using gene panels would only interrogated a limited amount of several hundred genes. As different expressed mutant peptides expressed may vary widely among patients, there is no a golden standard list of reference candidates (34), targeted capture of a gene panel would significantly decrease the chance to find a list of candidates. Therefore, WES represents the most suitable DNA sequencing strategy for mutation detection in protein-coding genes.

Cancer Immunology: Neoantigens to elicit T-cell response

Neoantigens. The term neoantigen refers to a mutated peptide derived from passenger and driver nonsynonymous somatic mutations, with potential to elicit a non-self-response by the immune system (18,35).

There is a distinction between tumor-associated antigens (TAAs) and tumor-specific antigens (TSAs) or neoantigens (36). TAAs can be expressed in both normal and tumor cells (although tumor cells over-express these class of antigens (37)) and recognized by specific T-cells; their immunogenicity is highly dependent on overexpressed levels to provoke certain immunogenic mechanisms (38). On the other hand, TSAs (neoantigens) are exclusively formed in tumor cells (39). The immune system plays an essential role in our daily recognition of newly synthesized cancer cells in order to keep the homeostasis in the body and stop carcinogenesis (40). Neoantigens are proposed to provide a new window for potentially using the immune system to target them and destroy tumors (36,41,42).

Cancer is patient-specific. The tumor landscape and mutational signatures are being studied deeper and deeper, concluding that, aside from mutational signatures/patterns in the population, each individual carries his/her own cancer profile (28). In fact, the mutational landscape is highly fluctuating among patients, tumors from the patient him/herself, and between distinct spots of a single tumor (43).

The estimated average of nonsynonymous tumor mutations found in humans is within the range of 100 and 120 (28,34). However, this number differs depending on the type of tumor and mutation (44). For example, melanoma and lung cancer show the highest number of mutations whereas the lowest ones go to several types of leukemia (13). The mutation burden and prognosis for the response to treatment has been correlated in previous studies (45). Also, deficiency in DNA repair mechanisms is being correlated to neoantigen load (46). An estimation of the likelihood of neoantigen development shows that tumors with higher mutation rate (number of somatic mutations per Mb) are prone to derive neoantigens. Melanoma is the only tumor type with “frequent” formation and others such as lung, colorectal, stomach, uterus, bladder and liver cancer have a more “regular” likelihood (34). The type of mutation should also be taken into account. Most of the attention for neoantigen prediction has been focused on Single Nucleotide Variants (SNVs) leading to nonsynonymous mutations (44). However, indel-type mutations could elicit stronger immunogenic response due to their ability to introduce frameshifts and new Open Reading Frames (ORFs) (44).

This heterogeneous outlook points towards patient-specific analysis to potentially reveal the neoantigen landscape. In fact, it was not until the combination between MPS technologies and bioinformatics when patient-specific neoantigen predictions became doable (28,47). Sequencing data analysis has proved its potential to find suitable neoantigens for cancer immunotherapy (48).

Antigen processing, binding, presentation and T-cell recognition. WES analysis in cancer has successfully revealed some somatic mutations that derived into CD8+ T cells mutant antigens (28,49). In this context, the Major Histocompatibility Complex (MHC, also known as Human Leukocyte Antigen or HLA in humans) in the patient plays an essential role, and concomitant studies of antigen-MHC-epitope binding predictions from sequencing data have proven their potential to find suitable neoantigens in cancer immunotherapy (28). Other relevant biological events related to antigen processing, binding and presentation involves protein cleavage, peptide-MHC binding and peptide-MHC-TCR (T-Cell Receptor) immunogenic reaction (35).

The path that peptides needs to undergo for antigen presentation includes firstly their cleavage or lysis of a protein in the cell-cytosol by proteases. The peptide chunks are carried and processed in the endoplasmic reticulum of the cells, where they will bind a maturing MHC molecule to form the peptide-MHC complex. This complex will be finally exposed at the surface of the cell membrane (50). Antigen presentation to the immune system that elicits subsequent T-cell activation takes place between the MHC at the cell surface of APCs and T-cell Receptors (TCRs) with the presence of a peptide fragment (or antigen) held in the peptide-binding groove of the MHC (51). These final peptides fragments bound to the MHC and located in the binding groove are usually 8-11 amino acids long (52). Each person shares a personal set of genes encoding for the MHC proteins. Most of these genes are located on chromosome 6 (6p21.31) and they represent one of the most variable polygenic and promiscuous regions (highly polymorphic) in the human genome (53). MHC is classified into class I (HLA-A, HLA-B, HLA-C) and II (HLA-DRB1, HLA-DQB1, HLA-DQA1, HLA-DPB1, etc.) genes, which are present in almost every nucleated cell; MHC Class II genes are mainly present on some specialized cells. In the case of MHC-I, it will mainly elicit a CD8+ T-cell response whereas MHC-II mainly interacts with CD4+ T-cell response (54).

Neoantigen prediction

The detection of neoantigens in the past has been arduous and has focused on a limited number of genes (55,56), until the emergence and well-establishment of MPS technologies (36). Several clinical trials have been conducted in order to use predicted and validated neoantigens (20). Nowadays, although the evidence for neoantigen immunogenic response highly depends on biological assays, the power of bioinformatics and its combination with MPS enables to interrogate the whole set of genes identify neoantigens from each tumor (57).

A definite strategy to predict neoantigens with high confidence is in development (58). However, neoantigen predictions enable to narrow the large amount of possibilities (34). The common framework proposed for neoantigen prediction using high-throughput NGS data is prominently illustrated by three main features: robust somatic mutations detection followed by estimation of some biologically relevant parameters with computational methods and confirmation of the mutated predicted peptides by gene expression with RNA sequencing data analysis (39) or other methods (57). Biological relevant parameters are being a key element in the neoantigen prioritization. Results validation has been done with data visualization, biological assays (57) and other experiments like “immuno-proteomics-based” with Mass Spectrometry (59).

Somatic mutation identification from MPS data. Cancer genomes can display heterogeneous and complex changes in the genome (60) and each patient will present diverse mutational landscape (13). In addition, predicted neoantigens must be tumor specific in order to avoid possible side-effects like autoimmunity (30). Therefore, accurate and optimized mutation detection pipeline in NGS data is essential and ground step in the neoantigen prediction workflow (61,62). The detection and discrimination of germline and somatic mutations in tumor aligned sequencing data can be performed using variant calling tools (63–65). Despite some attempts to identify germline and somatic mutations only from tumor sequencing data (66), normal-matched sample is mostly required in a clinical setting; for example, patient-matched blood DNA sample (60).

The coverage of each position interrogated needs to be high enough for calling variants with high confidence, certain sequencing read depth is required (67). Somatic variants detected by callers used in the clinics are passed through stringent filters to reduce the number of false positives (60). In order to increase automation and avoid manual labor albeit keeping high accuracy and precision, the combination of variant callers to find the somatic variants has been recurrently applied (68). However, it must be considered the low consensus among different variant callers when applying these approaches into the clinics (69).

Relevant biological parameters. Several studies on neoantigen prediction implement selection criteria based on the binding affinity between the peptide and HLA proteins (57). Others have used some parameters related to antigen processing and presentation, physical features of the peptide sequences and immunogenicity (35): proteolysis, peptide transport, amino acid composition (hydrophobicity, charge, polarity, size), peptide stability and TCR activity. Estimation of these parameters are based on both physical models, machine learning algorithms and larger and larger databases collecting experimental data (70,71). As mentioned earlier, most antigens derived from tumors are usually bound to MHC-class I, but MHC-class II is not discarded. Therefore, the knowledge of MHC-class I and II genes of the patient is a mandatory condition to estimate likely binding predictions. There are several methods to figure patient's HLA, being qPCR-based test one of the most reliable (60). Interestingly, MHC-class I and II alleles can be predicted from WGS, WES or RNA-sequencing data using diverse bioinformatics approaches (72–74).

Despite *in-silico* calculation of these parameters, they do not inform of the actual context of each predicted neoantigen for each specific case and patient, but rather help during the prioritization criteria by cutting down the number of candidates and finally choosing the best candidates (57). In other words, finding relevant biological parameters in this scenario (like the highest probability of binding to the HLA complex and/or inducing tumor-specific T-cell response) is an attempt to improve neoantigen prediction and reduce the number of laborious, time consuming and costly biological assays with non-potent neoantigens that unable to initiate an immune response (75).

Gene expression of tumor-specific antigens. A set of mutations translated into mutated amino acid sequence and predicted to bind the MHC of APCs and elicit certain T-cell immune response is not useful if those genes are not truly expressed in the tumor. Therefore, selection of neoantigens predicted to be expressed in the tumor is a logical step and RNA sequencing (RNAseq) data enables such feat (29). Indeed, RNAseq analysis faces several challenges related to the dynamic nature in the cells and intrinsic complexity of RNA (76). Once the RNAseq data is mapped to a reference transcriptome and post processed, some methods to normalize and count gene expression in paired-end sequencing are Fragments Per Kilobase Million reads (FPKM for paired end reads and RPKM for single end reads) or Transcripts per kilobase Million (TMP) (77,78). It has been found that using any of these values accordingly is a good method to pick only the expressed transcripts in the tumor (45,79).

Bioinformatics pipelines for neoantigen prediction. Several publicly available tools integrate some features described above and perform in-silico neoantigen prediction (35,72,79). Commonly, these pipelines translate the mutant peptide sequences from an input list of somatic mutations (VCF file) and require the patient's HLA alleles. Then, they execute certain MHC-binding affinity predictors (80). During this step onwards, their strategies and criteria to output a list of neoantigen candidates are diverse. Currently, **there is no benchmarking available** since some of them like **Neopepsee** (35) have been recently published (2017/18) and their features differ substantially. For example, Neopepsee includes the power of machine learning classification built on Support Vector Machine (SVM) model to input nine different biological parameters. It also includes gene expression analysis. However, there are two practical **disadvantages with**

Neopepsee: due to the tool recently being published, the **documentation is not fully described** and it also only runs with an **outdated and deprecated Immune Epitope Database (IEDB)**(70)[1]. Another tool is **pVACseq** (79), developed as part of a whole package of bioinformatics tools (pVACtools) for cancer immunotherapy. It requires input VCF annotated by Variant Effect Predictor (81). This annotation **allows automated detection of nonsynonymous somatic mutations**, in comparison to Neopepsee, which requires more curated variants. pVACseq is often **updated** and its **flexible** (optional input files: variant Allele Frequency (AF), transcript count, read coverage) usage allows to trim several parameters (MHC-binding prediction methods, peptide length, peptide stability and chopping predictions, among others).

Materials and Methods

Sample collections and DNA/RNA extraction. Tumor bladder tissue sample from *two patients (Sample IDs: **EEW50, HRIU59**)* was resected and fresh frozen for DNA and RNA extraction afterwards. As normal matching sample, Peripheral Monoclonal Blood Cells (PBMC) were isolated from blood. Tumor/normal DNA and tumor/normal RNA were extracted with QIAGEN Allprep kit, following manufacturer's indications. Sample concentration was assessed using Qubit BroadRange (BR) double-stranded DNA (dsDNA) and RNA kits. Fragment size distribution (only DNA) was assessed with gDNA TapeStation 2200. RIN value for RNA samples was provided by an external report and two RNA control samples were included in the analysis.

Whole-exome and RNA library preparation and sequencing. DNA was stored at -20° until library preparation. 50 ng tumor and 50 ng normal input DNA was prepared using Twist Human Core Exome Kit (Twist Bioscience), following manufacturer's protocol. Adapter ligation was done with IDT Dual Index UMI adapters (i5 is 70nt and the i7 is 74nt long). Pre- and post-hybridization PCR was done with KAPA HiFi HotStart ReadyMix (Roche) and xGen® Library Amp Primer (IDT). Hybridization included the Twist Human Core Exome kit set of baits (120 bp dsDNA probes covering ~33 Mb of conserved protein-coding regions), xGen® Universal Blockers for an eight-sample pool during 16 hours incubation at 70°C; tumor:normal DNA were pooled together with a proportion of 3:1 (60). QC was performed before hybridization and before sequencing using Qubit High-Sense dsDNA kit and Tapestation 2200. RNA raw samples were stored at -80° until library preparation. RNA sequencing-ready libraries were prepared using TruSeq Stranded mRNA Library Prep Kit (Illumina). Both DNA and RNA library preparations incorporated indexing to allow samples pooling during sequencing. Tumor DNA, normal DNA and tumor RNA libraries were paired-end sequenced on a Hiseq 2500 Rapid Run with 100 base pairs (bp) paired-end mode to 30 million read pairs per sample.

WES data QC, alignment and post processing for analysis ready reads. Exome sequencing data was processed by an internal cancer analysis pipeline. Sequencing data quality was analyzed using FastQC[2] (version 0.11.5). Low quality bases and Illumina adapter sequence were trimmed from both read-ends using Cutadapt (version 1.16) (82) and mapped to human genome reference (GRCh37) using the Burrows-Wheeler Aligner software (bwa-mem, version 0.7.15-r1140) (83). Output Sequence SAM files were converted into

[1] Very recently (May 2018), the IEDB database used by *Neopepsee* was updated.
[2] FastQC: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

BAM format and sorted with Samtools (version 1.7) (84). BED file provided by the manufacturer was modified (padding=100) and used accordingly ("chr" letters deleted from the first column and entry "Un_gl00028" removed, which belongs to an unlocalized contig not included in the reference genome). Duplicates were marked and removed using Picard (version 2.18.3)(85). Later, indel local realignment and base recalibration was performed using the Genome Analysis ToolKit (GATK version 3.7) (86). Output BAM files were ready for variant calling, filtering and annotation. Coverage, duplication rate and alignment metrics were calculated using Picard *Collecthsmetrics* and *Collectmultiplemetrics* tools (85).

RNA-seq data QC, mapping, post processing and transcript analysis. QC and trimming of RNA sequencing reads was performed just as in WES data. Read mapping was done with Hisat2 (version 2.05) (87) (important flags: *-q --rg-id <RG> --rg <Seq Platform> --dta-cufflinks --rna-strandness RF -p 8*). The human reference indexed transcriptome for the alignment was downloaded from the Hisat2 website (H. Sapiens GRChr37, *genome_snp_tran*, 4.5 GB) and built using the bash script included (*make_grch37.sh)*. Output SAM files were converted into BAM and sorted and duplicated reads were removed as described. Experiment strand-specificity and gene body coverage was assessed using *infer_experiment.py* and *geneBody_coverage.py* (hg19_housekeeping.bed.gz file as reference) scripts from RSeQC (version 2.6.4). Gene expression counts (FPKM) was calculated using Cufflinks (version 2.2.1) (88) with a GTF annotated transcriptome reference (Homo_sapiens.GRCh37.75.gtf, Ensembl). Protein-coding genes[3] with non-zero gene expression were extracted from *gene.fpkm.tracking* output file (experiment type: *--library-type fr-firststrand*). Coverage, duplication rate and alignment metrics were calculated as described above.

Variant calling, filtering and annotation. The strategy for the detection of somatic mutations included the parallel matched-pair (tumor and normal DNA) analysis of the aligned and post-processed reads combining results from Vardict(65), Mutect2 (GATK version 3.7) (89) and Strelka (90) with default parameters, excepting (Vardict): *minimum Allele Frequency (AF) = 0.01, p-value threshold = 0.9, max mismatches allowed per read = 4.5*. AF in Strelka was calculated as described in Strelka Github User Guide[4] (Variants with no reads supporting the reference allele in the tumor data were skipped). A first filtering step of the variants was performed on the output VCF from each variant caller: from Vardict, only "*PASS*" + "*StrongSomatic*" variants; from Mutect2 and Strelka, only "*PASS*" variants in the info field. "*PASS*" indicates that all the filters applied by each variant caller on each mutation has been passed. The term "*StrongSomatic*" applied by Vardict are for variants which are exclusively reported in the tumor sample. A second filtering step removed variants which did not comply three more conditions (60): AF > 0.01; number of reads supporting the variant > 5; and (AF of the variant in the tumor / AF of the same variant in the normal sample) > 5x. Finally, new VCF files only containing information about the tumor sample were created (pVACseq accepts VCF files with only one sample) were annotated using the Variant Effect Predictor (VEP) Tool v91 (81).

HLA typing and neoantigen prediction. Patient-specific HLA allele typing was predicted from sorted and de-duped normal sample WES data with HLAminer tool (v1.3.1, *Read Alignment (HPRA) WGS* method). The alleles with highest confidence and score were chosen from the HLAminer prediction output list (see Results). **pVACseq** was installed and used as part of the pVACtools pipeline (version 1.0.3) for the neoantigen prediction, with default parameters except the following: *-e 8,9,10,11 --net-chop-method cterm -*

[3] HUGO Gene Nomenclature Committee: https://www.genenames.org/cgi-bin/statistics
[4] Strelka Github User Guide: https://github.com/Illumina/strelka/tree/master/docs/userGuide

-netmhc-stab --top-score-metric=lowest -d 40. All the available affinity binding predictors in pVACseq (*NNalign,NetMHC,NetMHCIIpan, NetMHCcons, NetMHCpan, PickPocket, SMM, SMMPMBEC,SMMalign)* were used to analyze the peptides against the patient-specific HLA alleles. Default binding-affinity threshold in pVACseq was kept (500 nM). Bam-readcount coverage files and gene expression (Cufflinks) were not used as input in pVACseq. Gene expression and allele frequency was included later. pVACseq output was filtered according to gene expression values, retaining only those genes with an FPKM > 0.
Data visualization. Genomic regions of interest for mutation visualization were extracted and indexed from de-duped and sorted BAM files using Samtools (version 0.1.19) and visualized with IGV onto GRCh37 genome reference (version 2.4.5) (91).

Results

Pipeline overview. *Figure 2* summarizes the bioinformatics workflow. After variant calling, the normal sample must be removed from the output VCF file, since pVACseq only requires tumor sample. Variant calling consensus was analyzed creating one VCF file for each overlapping between two or more variant callers. The *Ensembl* variant consequence (defined by Sequence Ontology[5]) annotation by the Variant Effect Predictor (VEP) (81) is required by pVACseq in order to process nonsynonymous mutations (79) (missense, inframe indel and frameshift mutation)[6]. By requirement of pVACseq, the wild type peptide sequence and downstream peptide changes were annotated in the VCF files by VEP. **Importantly, due to reasons yet to be investigated, pVACseq could not process the variants from the Strelka-only input VEP annotated VCF file**.

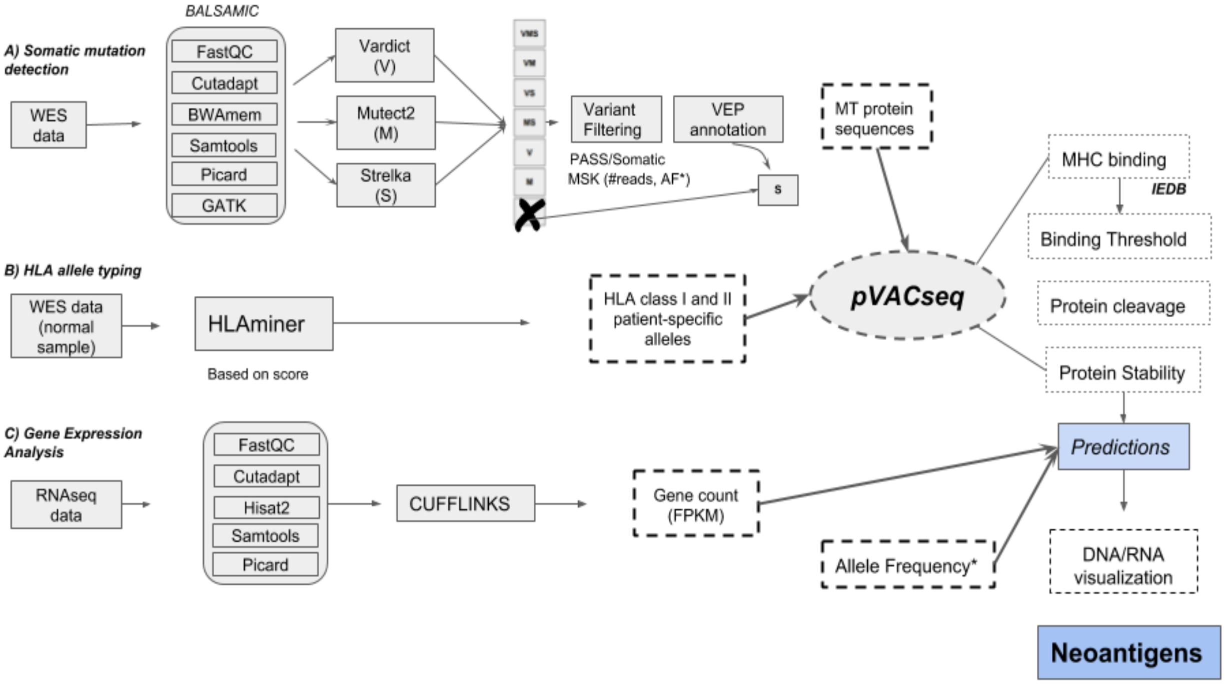


Figure 2. Bioinformatics workflow. Somatic mutation detection and HLA allele prediction are the two main inputs for pVACseq. RNAseq analysis for gene quantification and later visualization is added to the predicted neoantigens for prioritization.

[5] The Sequence Ontology: http://www.sequenceontology.org/
[6] Ensembl "consequence" annotation: https://www.ensembl.org/info/genome/variation/predicted_data.html

HLA alleles typing prediction was performed by HLAminer using normal EEW50 and HRIU59 WES data; the input fastq file from HRIU59 was down sampled with seqtk (92) to 35 M read pairs and the original fast files from EEW50 were used without downsampling. HLA alleles predicted for each sample are summarized in

Table 4 (ANNEX). The selection criteria were based on ranking score and confidence value, reported by HLAminer (73). Later, pVACseq would search what alleles are available in each of the MHC-binding algorithms invoked and analyze only the possible allele-algorithm combinations.

pVACseq splits each input mutated protein sequence into shorter chunks and launches MHC-binding affinity predictions against them. The output from pVACseq reports the HLA-allele found to bind each peptide chunk and their corresponding predicted binding affinity values passing the input threshold. Protein cleavage site and protein stability predictions were also included for further work. Gene expression and AF calculations were performed and added externally to pVACseq.

Library preparation and sequencing data quality assessment, alignment metrics and coverage report

Figure 3 shows the fragment length distribution for input tumor and normal DNA EEW50 and HRIU59 samples. The quality of the tumor EEW50 was significantly lower than the rest of samples, indicated by the absence of a peak in High Molecular Weight (HMW) DNA molecules (>60,000 bp) (*Figure 3*a). However, new DNA isolation was not possible since no more extracted tumor sample was available. The concentrations and initial volumes ensured the required input amount of input DNA for the library preparation (*Figure 4*c).

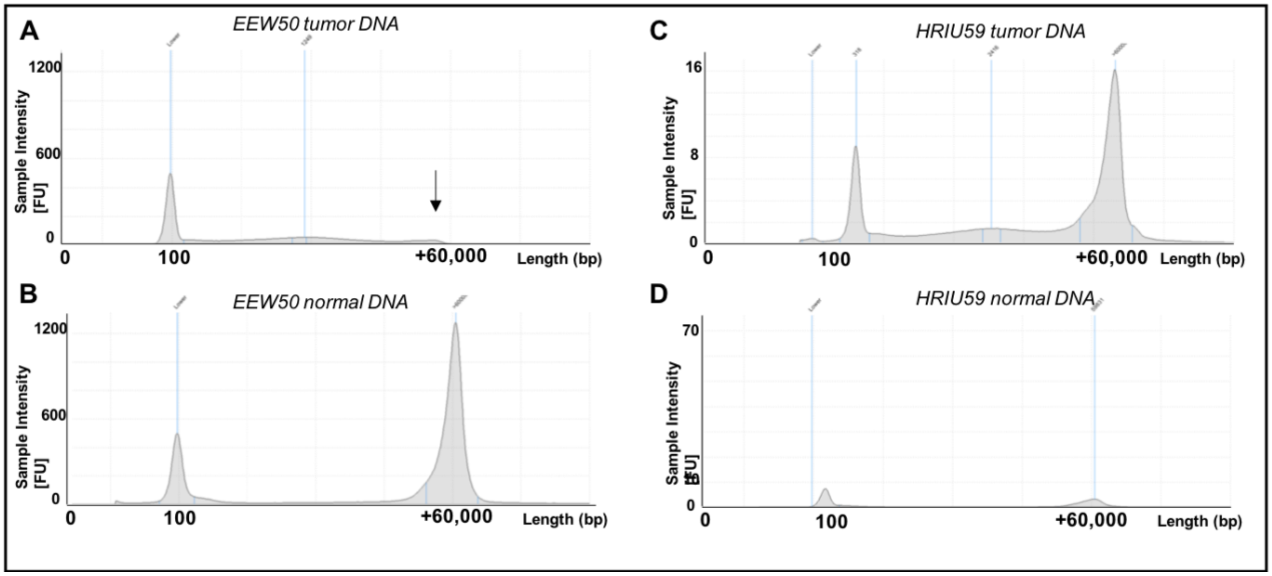
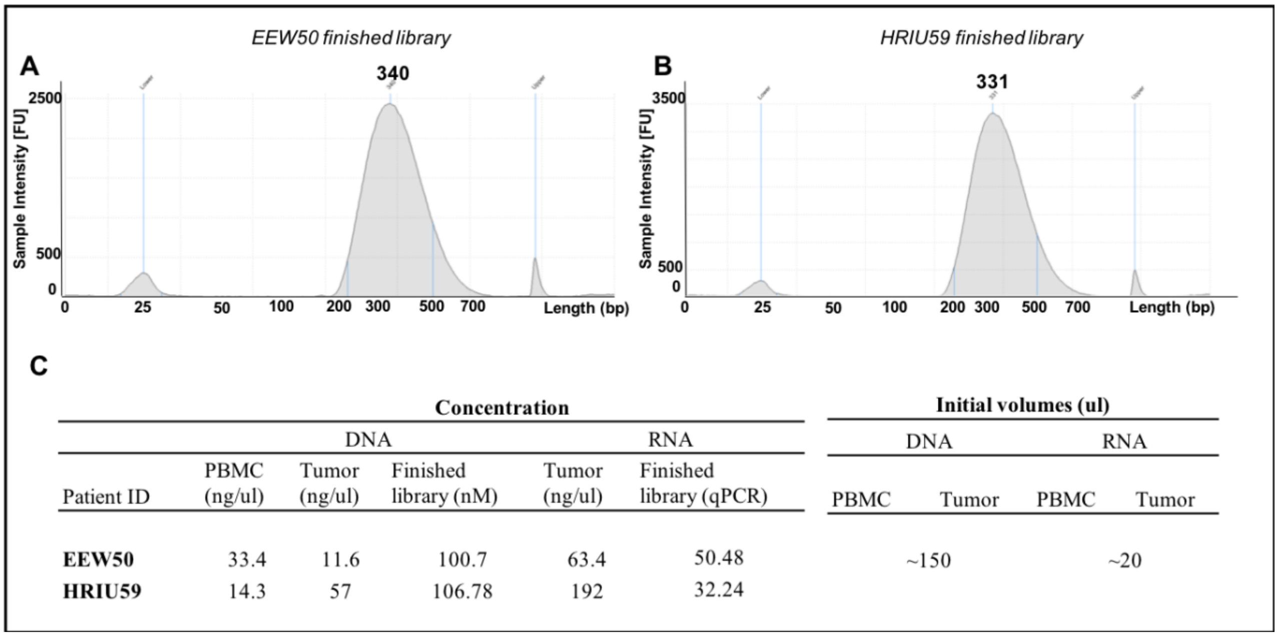


Figure 3. QC of tumor and normal EEW50 and HRIU59 input DNA samples. The absence of HMW DNA molecules in EEW50 tumor DNA (A, black arrow) shows its degradation and thus low quality.

Sample EEW50 was prepared, pooled and hybridized with other standard samples to have a total amount eight samples in the same tube (three parts EEW50 tumor DNA and one EEW50 normal DNA plus four other samples). The final library ready for sequencing was 340 bp long in average and its concentration was 100,7 nM. Sample HRIU59 followed the same routine with some standard samples to ensure 3:1 tumor/normal ratio and got an average fragment size of 331 bp and 106.78 nM concentration (*Figure 4*).



	Concentration					Initial volumes (ul)			
	DNA			RNA		DNA		RNA	
Patient ID	PBMC (ng/ul)	Tumor (ng/ul)	Finished library (nM)	Tumor (ng/ul)	Finished library (qPCR)	PBMC	Tumor	PBMC	Tumor
EEW50	33.4	11.6	100.7	63.4	50.48	~150		~20	
HRIU59	14.3	57	106.78	192	32.24				

Figure 4. DNA and RNA finished libraries QC. The average fragment size and concentration values passed the requirements to be sequenced.

After sequencing, raw WES data quality assessment was performed (*Figure 5*). The GC content analysis suggested slight systematic bias towards 60% in sample EEW50 (*Figure 5*a). *Figure 5*b in shows that the phred score per base was higher than 30 for both paired-end sequencing reads (Read 1 and 2) of the four WES samples with a read length of 101 bp, indicating that base sequencing quality was high and concordantly to the number of sequencing cycles. Adapter content analysis showed presence of the Illumina universal sequence adapter (not shown), trimmed afterwards by Cutadapt. RNA samples also passed GC content, phred score and adapter content (not shown).

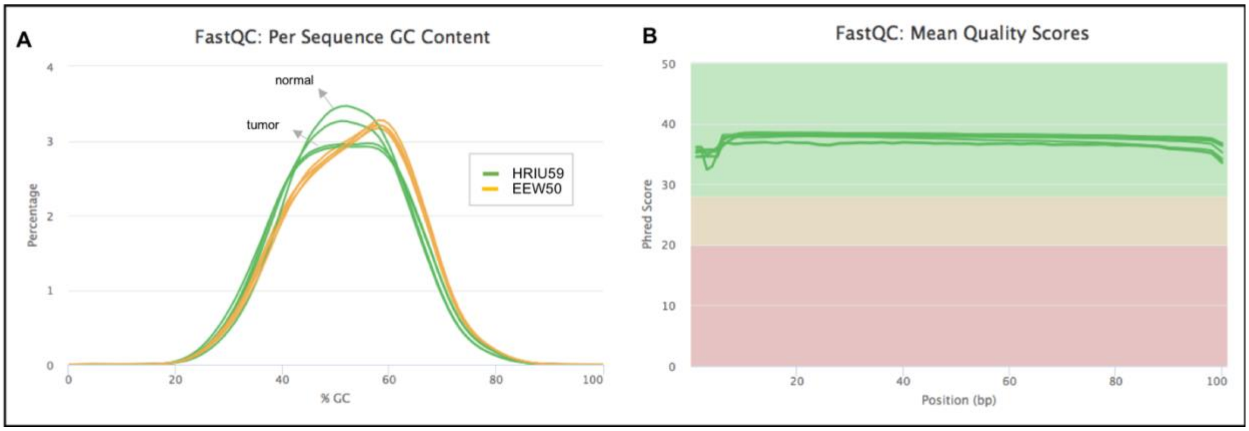
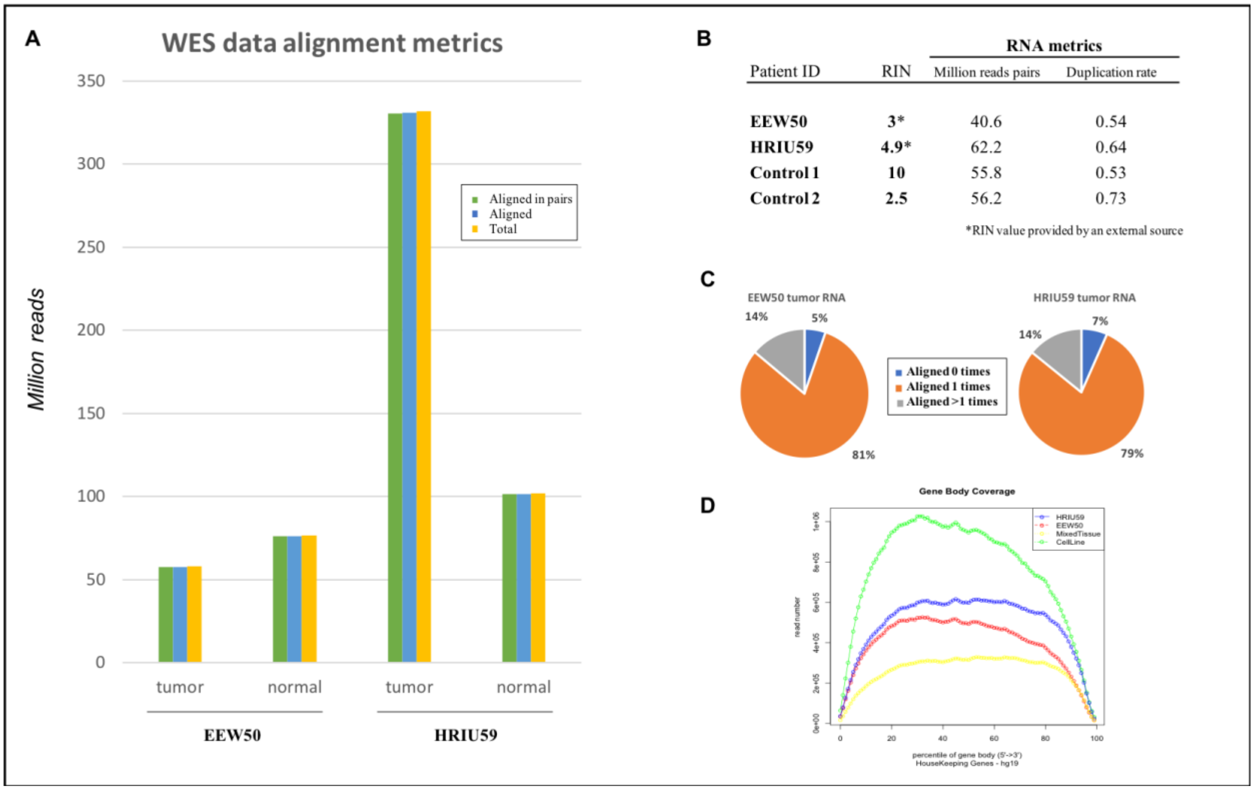


Figure 5. FastQC metrics of WES data

Read alignment in both patients (both WES and RNA data) is summarized in *Figure 6*. EEW50 WES sample resulted in 28.87M (million) and 38.08M aligned read-pairs in tumor and normal, respectively. On the other hand, tumor and normal HRIU59 WES sample got 165.15M and 50.72M aligned read-pairs, respectively. Both EEW50 and HRIU59 WES samples got 100% of passed-filters reads aligned. Despite the fact that the *tumor:normal* pooling ratio was 3:1 in hybridization, WES EEW50 sample did not yield a similar sequencing reads ratio. In addition, WES HRIU59 sample yielded much higher number of reads because its library preparation was sequenced alone in a single flow cell in Rapid mode. Rapid mode only has 2 lanes in the flow cell, and each lane yields between 150M and 200M read pairs. Thus, approximately 400M read pairs were available during that sequencing run. WES sample EEW50 was also sequenced in Rapid mode but pooled together with other library preparations in the same flow cell.

The attempted RNA sequencing depth was 30M read pairs. The output number of million read pairs from sequencing was higher than 30M for both EEW50 and HRIU59 samples (40M and 62M read-pairs, respectively) as shown in *Figure 6*b. The duplication rate (*Figure 6*b) was very high in comparison to other works (93), suggesting the low quality of the input RNA samples. Experiment strand-specificity was assessed for RNA sequencing data to confirm it as first-stranded (data not shown), concordantly with the cDNA library preparation protocol. First-stranded means that the first read sequenced direction of the read pair is reversed complementary to the mRNA molecule. *Figure 6*d shows the gene body coverage estimated for EEW50, HRIU59 and control RNA samples. This analysis did not report clear 3’ bias as expected from the RIN values, and a final conclusion on RNA degradation cannot be drawn without further investigation of the housekeeping genes list used in the calculation and application of normalization among samples.



		RNA metrics	
Patient ID	RIN	Million reads pairs	Duplication rate
EEW50	**3***	40.6	0.54
HRIU59	**4.9***	62.2	0.64
Control 1	**10**	55.8	0.53
Control 2	**2.5**	56.2	0.73

*RIN value provided by an external source

Figure 6. WES and RNA alignment metrics and gene body coverage. The tumor:normal ratio 3:1 in number of reads is only observed in sample HRU59 (A). Duplication rate in RNA samples is higher than 0.5 (B) and gene body coverage (D) does not show definite degradation evidence, suggesting to improve this analysis.

The duplication rate and coverage calculation of WES samples assess the performance of the exome capture protocol and library preparation. *Table 1* shows duplication within 2-4 % for WES samples, although it was higher (~7%) in the tumor HRIU59 WES sample, suggesting that higher sequencing depth increases the number of duplicates sequenced. The mean target coverage in sample HRIU59 was the highest, supporting the sequencing depth. The mean target coverage of EEW50 tumor WES sample was lower than its normal pair, suggesting low quality of the input tumor sample, since a higher coverage was expected.

Table 1. Duplication and coverage metrics of WES samples

	Duplication Rate	**Coverage**		
EEW50		Bait	Target	Max[7]
Tumor	0.029	61	67	1300
Normal	0.020	79	84	1891
HRIU59				
Tumor	0.073	537	381	6535
Normal	0.035	165	118	2358

[7] Maximum Base Target Coverage

Somatic variants detection and filtering combining three variant callers

The three variant callers interrogated matched-normal and tumor WES samples EEW50 and HRIU59. Somatic variants with "PASS" and "StrongSomatic" tags and AF > 0.1 were selected (*Table 2*). The main observation is the low consensus between different variant callers (*Figure 7*), as previously reported (64,69). Another observation is the comparison between the number of variants found by each tool (*Figure 7*). Interestingly, the number of somatic variants identified by Mutect2 was two orders of magnitude lower than Strelka (64). Vardict found more variants than Strelka in patient EEW50, whereas the opposite situation was given in patient HRIU59 (3027 and 588 in Strelka and Vardict, respectively). The comparison between the number of variants found in both samples by Vardict suggests lower sensitivity to the coverage. AF distribution is shown in *Figure 7* (right and left borders). Vardict, Mutect2 and Strelka found most of the variants in the AF range from 0 to 0.2 (99, 100, 69%, respectively); notably, the distribution in Strelka showed wider ranges. Strelka does not report the AF directly, but it can be calculated. The calculation of the AF in Strelka was slightly different than Vardict, since not perfect correlation between values of variants found in both Strelka and Vardict was found (*Figure 7*a, top-left histogram).

In order to rise the confidence in the mutations found, more stringent filtering criteria was applied based on two more conditions (60):

1) AF reported in the tumor for a given variant needs to be 5x the AF reported for the same variant reported in the normal-matched sample (if reported)
2) At least 6 reads support the somatic variant in the tumor sample

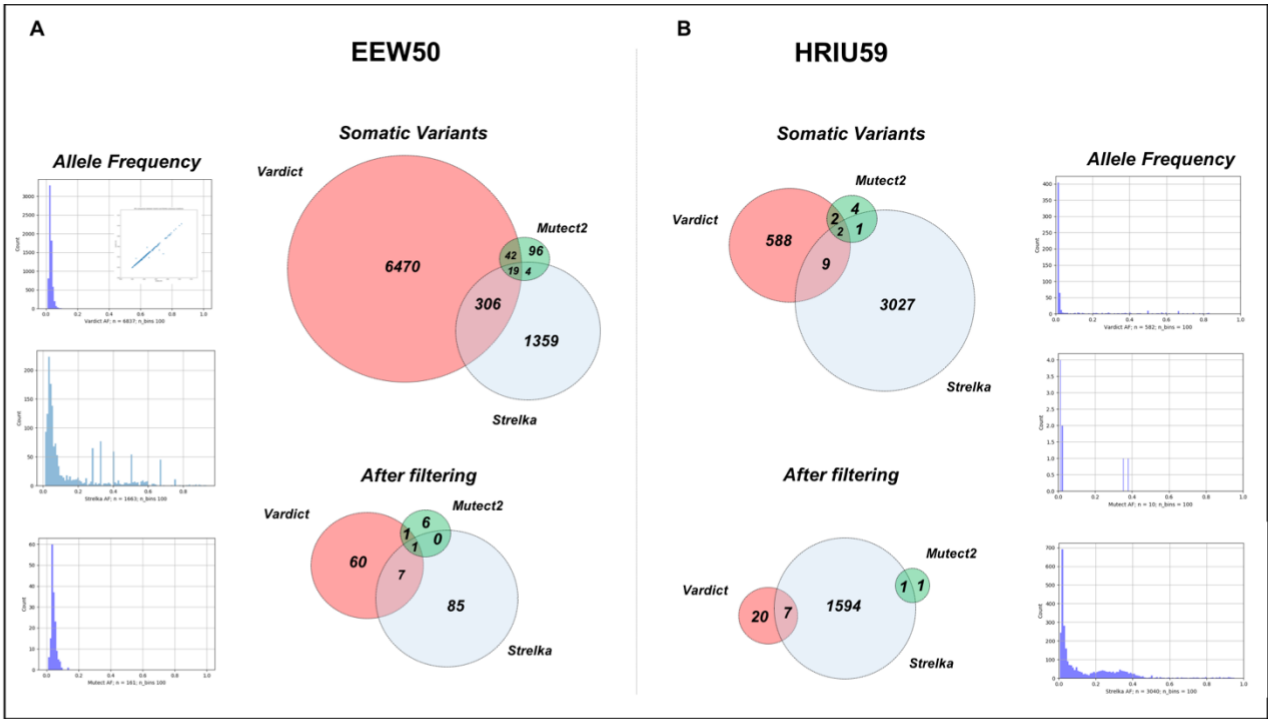


Figure 7. Somatic mutations detected by Vardict, Strelka and Mutect2 and AF distribution[8]. Venn diagrams show the number of mutations found by each variant caller and the overlapping results among them, both before (up) and after (down) the filtering criteria. All mutations passing the filters were used in neoantigen prediction.

The impact of a more stringent filtering criteria was visible (*Table 2*). The number of variants was significantly decreased in sample EEW50 (98, 93 and 88% of total variants in Vardict, Mutect2 and Strelka, respectively) due to the second condition. This is not the case of sample HRIU59, in which more variants found by Strelka were filtered by the first condition (32%).

Table 2. Number of variants before and after applying the filtering conditions. The first condition was more restrictive.

	Filtered variants				After filters
EEW50	Total	#reads<5	tAF/nAF<5	both	Total
Vardict	6837	**6767**	2371	2370	69
Mutect2	161	**150**	47	44	8
Strelka	1661	**1466**	324	222	93
HRIU59					
Vardict	601	**546**	465	437	27

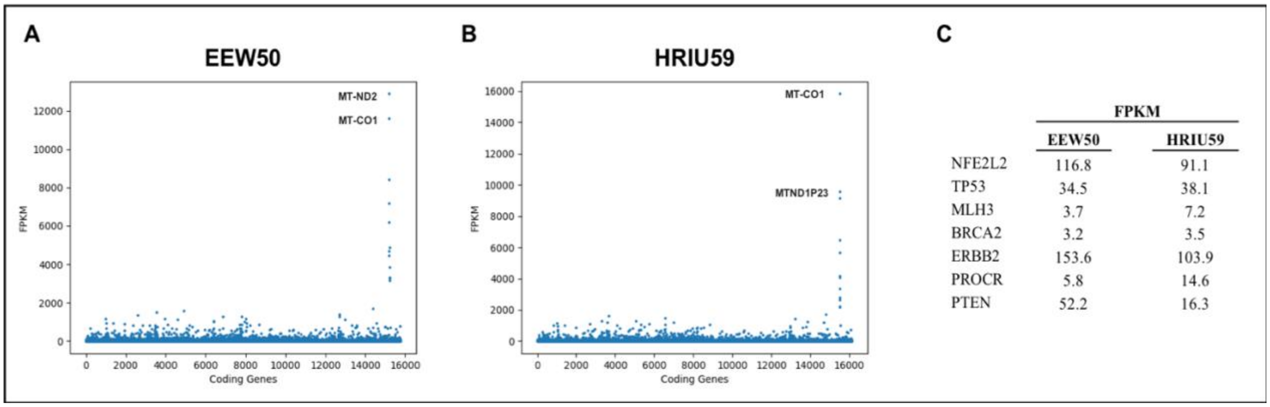
[8] For a better visualization of the AF distribution, see ANNEX

Mutect2	9	7	7	7	2
Strelka	3039	686	**1000**	249	1602

Predicted neoantigens load and prioritization based on variant type, MHC-binding affinity and gene expression

All selected variants were annotated using the *Variant Effect Predictor* and input into pVACseq. However, not all the input variants were finally processed by pVACseq because their annotated consequence was other different than "missense", "inframe indel", or "frameshift mutations" (81). Very important notice, as mentioned before, none of the mutations from Strelka were processed by pVACseq. This phenomenon occurred in both EEW50 and HRIU59. After tracking the pipeline back, it was found that pVACseq interpreted the output VEP annotated VCF file from Strelka as "empty". This fact points to an issue with the formatting of the Strelka output VCF files.

In parallel, gene expression analysis in samples EEW50 and HRIU59 using *Cufflinks* reported the FPKM values for each protein-coding gene. As shown in *Figure 8*, FPKM values lie in the same range for both samples. Values higher than 4000 FPKM mostly belong to mitochondrial genes or gene-coding microRNAs (for example, MT-ND1/MT-ND2, MTND1P23, MT-CO1/MT-CO3, MTND2P28, Hsa-mir-6723, MT-ATP6). Only 23378 and 24539 out of 63785 (coding and non-coding) *Cufflinks* gene entries had FPKM higher than 0 for EEW50 and HRIU59, respectively. Out of those, 15728 and 16102 were coding genes. A high percentage of these genes showed FPKM values within 0 and 1 FPKM (42.3% and 48.8% in EEW50 and HRIU59, respectively).



	FPKM	
	EEW50	HRIU59
NFE2L2	116.8	91.1
TP53	34.5	38.1
MLH3	3.7	7.2
BRCA2	3.2	3.5
ERBB2	153.6	103.9
PROCR	5.8	14.6
PTEN	52.2	16.3

Figure 8. Non-zero gene expression of protein coding genes in tumor RNA

The output list of candidates from pVACseq was annotated with FPKM. One caveat is that pVACseq reports several results for the same gene and peptide chunk if different valid transcripts are available in the MHC-binding predictors' database. In this case, all transcripts from one gene were grouped together and the best scored results was annotated (although all transcripts normally produced the same exact result).

Neoantigens predicted in sample EEW50. 47 and 39 neoantigens were predicted to bind to MHC-class I and II, respectively (*Figure 9*). All predictions were associated to four MHC-class I and four MHC-class II different patient-specific alleles. Interestingly, some neoantigens were predicted to bind both MHC-class I and II predicted alleles. In addition, most of the mutated peptides were predicted from variants found by Vardict and overlapping between Vardict and the other two callers, as expected from the significant difference between the number of variant reported by each caller and the lack of results for only-Strelka variants.

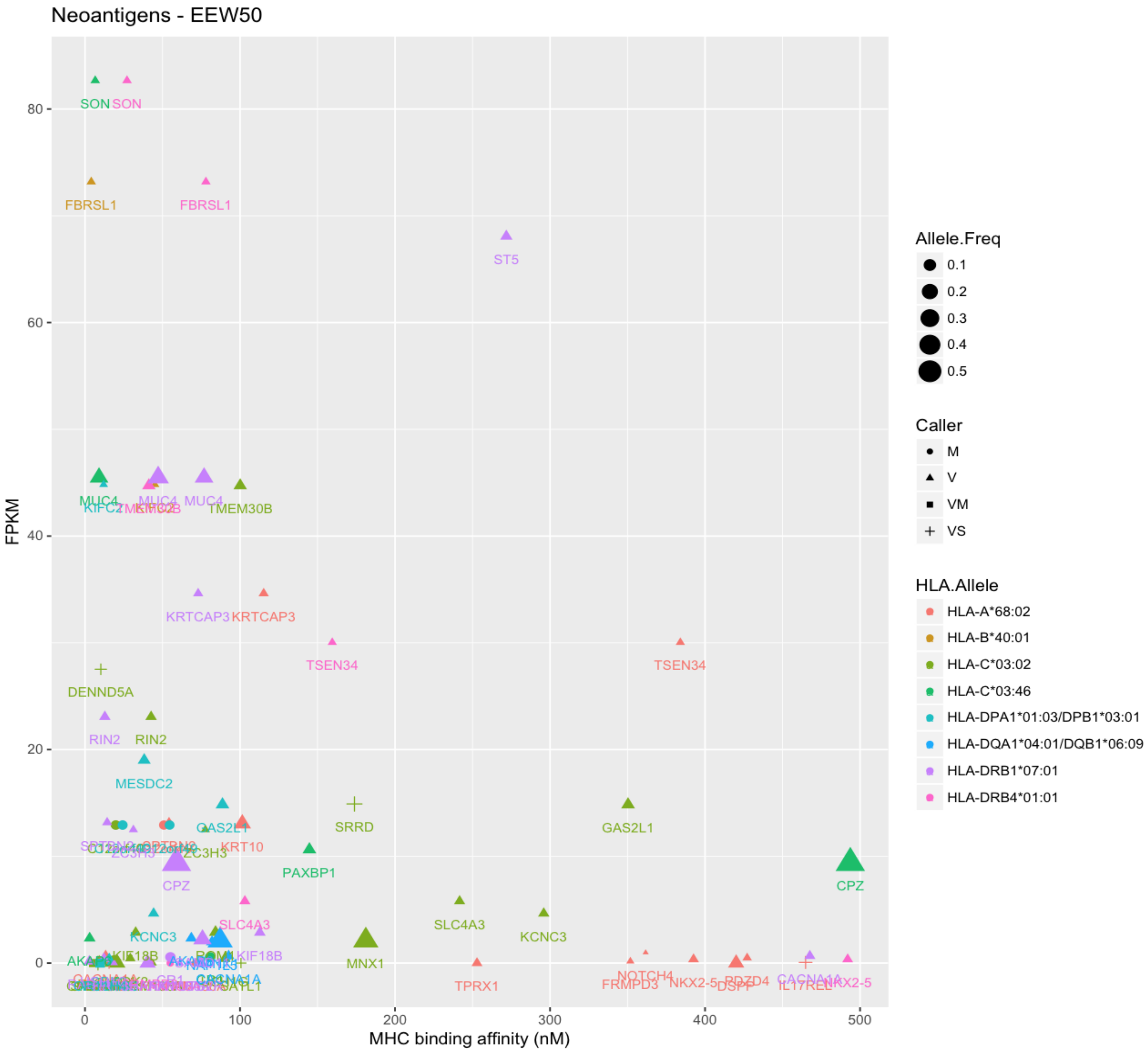


Figure 9. Neoantigens predicted in sample EEW50. In order to help with the prioritization, gene quantification values are plotted versus predicted MHC binding affinity of the neoantigens predicted, as well as the allele frequency (dot size), patient's specific HLA binding allele (dot color), and variant caller combination that found the corresponding mutation. Note that one neoantigen can be predicted to bind two HLA alleles (even from MHC-class I and II)

Vardict is able to identify "complex type" mutations; indeed, a big fraction of predicted neoantigens derived from this type of mutations. *Figure 10*a-c shows several examples of mutations that derived into predicted neoantigens and their visualization (*Figure 10*d). Complex-type mutations predicted by Vardict are more difficult to figure out from their visualization, since they can be a combination of different mutational processes (65). On the other hand, substitutions are easy to visualize and indel-type mutations such as deletions can be clearly observed (*Figure 10*b-c).

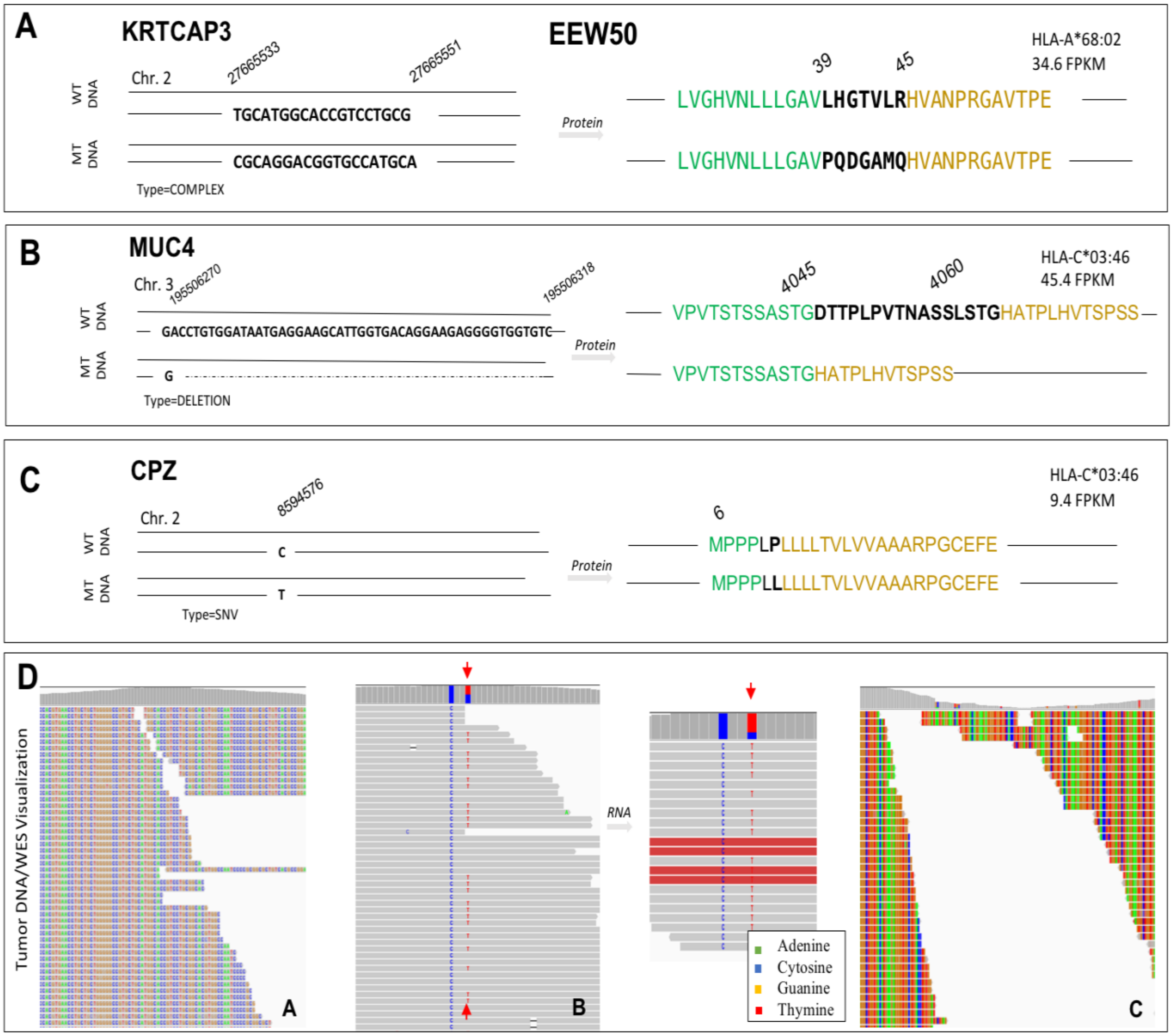


Figure 10. Visualization examples of variants found in EEW50 sample.

BAM files visual inspection of selected neoantigens was also performed over the RNAseq data of the tumor samples to confirm variants predicted with WES data (20). In this sample, only one predicted neoantigen was confirmed this way (*Figure 10*d(b)).

Neoantigens predicted in sample HRIU59 (*Figure 11*). This result did not include predictions from variants detected by Strelka, which in fact was the variant caller that found most mutations in sample HRIU59 (*Figure 7*). Similar to some findings in sample EEW50, 6 neoantigens were predicted to bind different alleles of both

MHC-class families (I and II) (*Figure 11*). Interestingly, all these predicted neoantigens were related to key genes related to cancer: TP53, BRCA2, PTEN, ERBB2, PROCR, FGFR3, MLH3, NFE2L2 (94).

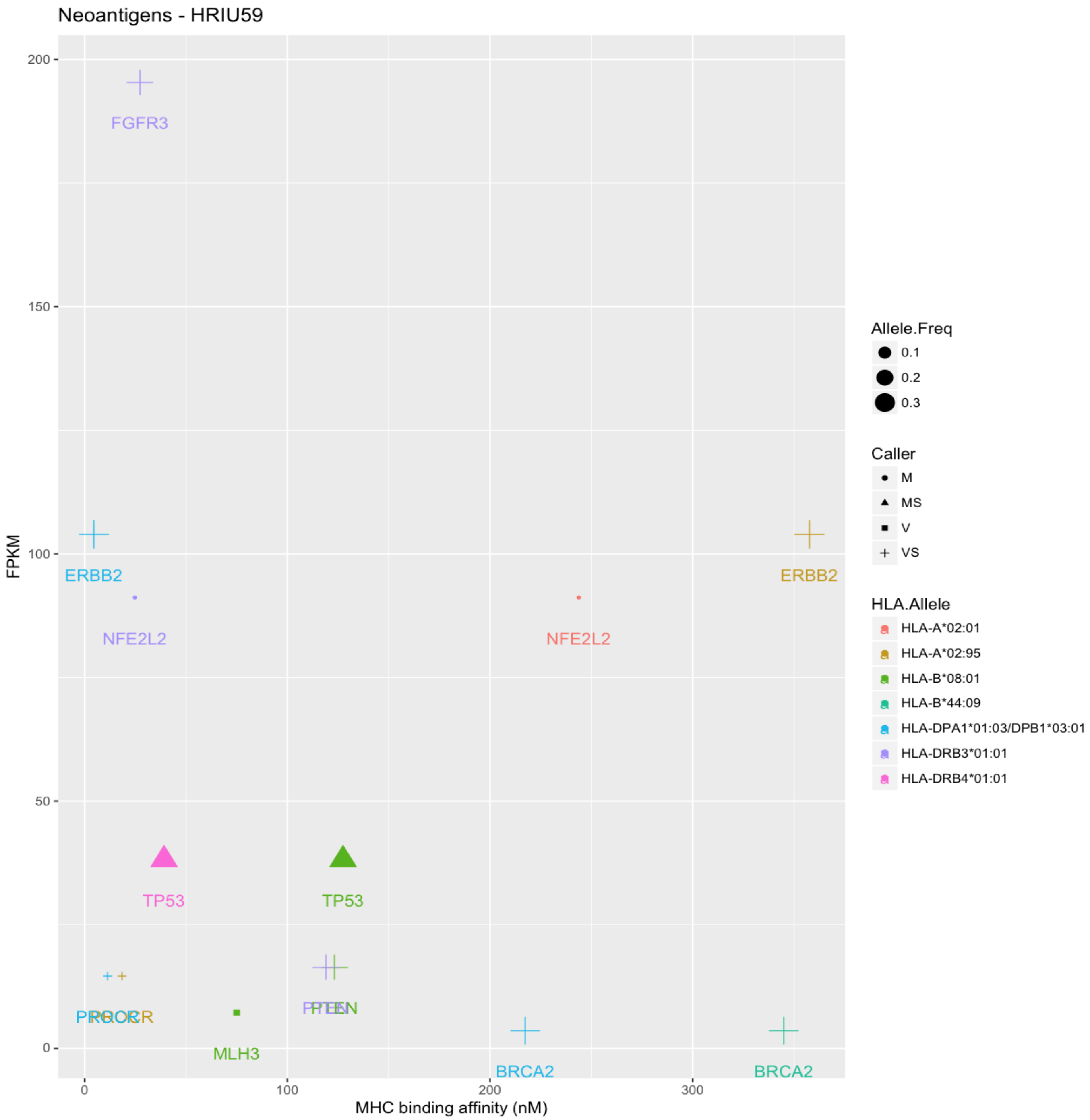


Figure 11. Neoantigens predicted in sample HRIU59. 14 neoantigens was initial the outcome by pVACseq for sample HRIU59(see Correction Note). These findings showed well-associated genes to cancer and the mutations found are supported by more than one variant caller in most cases. Same as in EEW50, one same neoantigen can be predicted to bind both HLA-class I and II.

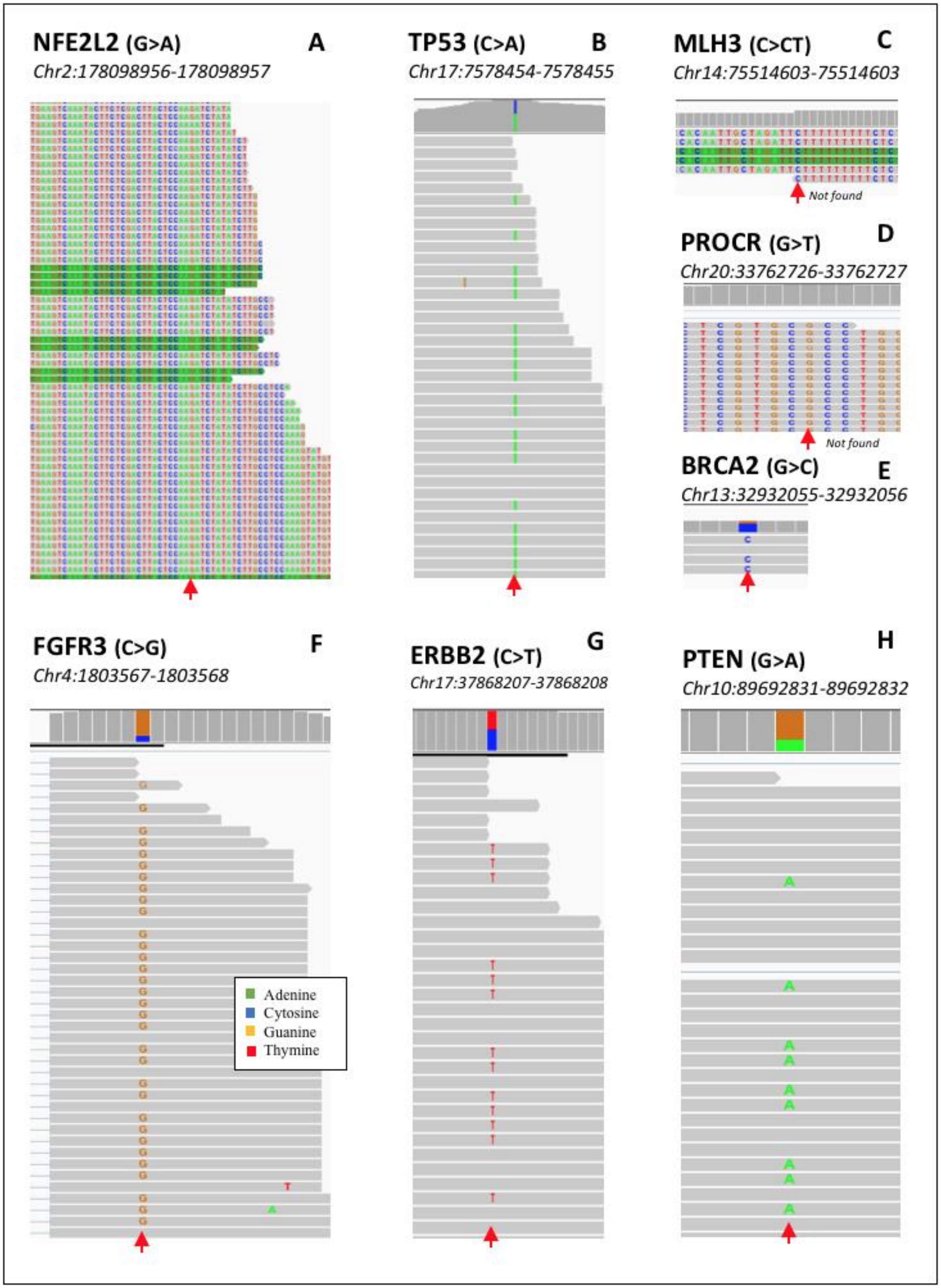


Figure 12. RNA visualization of the variants in sample HRIU59

Visual inspection of tumor RNAseq revealed the presence of the reported mutations in all the genes predicted except PROCR and MLH3 gene (*Figure 12*). However, the coverage of the SNV in PROCR (G→ T) and the insertion in MLH3 (C→ CT) genes was the lowest with 16x and 6x, respectively (*Table 3*).

Table 3. Neoantigens predicted in sample HRIU59

							RNA			
Gene	*Chr*	*Start-Stop*	*Ref*	*Alt*	*Change*	*Pos*	*Observed?*	*#ref*	*#alt*	*AF*
NFE2L2	2	178098956-7	G	A	L/F	14	Yes	321	21	0.06
FGFR3	4	1803567-8	C	G	S/C	249	Yes	81	357	0.82
PTEN	10	89692831-2	G	A	E/K	106	Yes	28	11	0.28
BRCA2	13	32932055-6	G	C	E/Q	2599	Yes	1	3	0.75
MLH3	14	75514603	C	CT	-/X	585-6	No	6	0	0
TP53	17	7578454-5	C	A	A/S	159	Yes	49	63	0.59
ERBB2	17	37868207-8	C	T	P/S	85	Yes	305	288	0.49
PROCR	20	33762726-7	G	T	R/L	98	No	16	0	0

Discussion

In this project, WES data analysis from MPS technologies has revealed patient-specific nonsynonymous mutations, that together with gene expression quantification analysis of tumor RNA sample, HLA allele typing from WES data and MHC-peptide binding estimation have enabled the identification of candidate neoantigens for cancer immunotherapy. Matched-pair urinary bladder cancer samples from two patients were analyzed and the sample quality was found as a sensitive point for the good performance of the project. Regarding to the bioinformatics analysis, somatic mutation detection, RNA sequencing analysis, HLA prediction and MHC-binding prediction have been found to be crucial parameters along the pipeline, which will require a deeper understanding in order to counteract the limitations and provide biologically meaningful information.

First, the **evaluation of the sample quality** was made by means of QC tests at the laboratory side to anticipate the success of the subsequent steps. This checking point showed that the EEW50 sample had severe signs of degradation, since the fragment size distribution was very low for HMW molecules (>60 kb). This fact compromised the rest of the work, since one of the first steps in the library preparation is fragmentation, and the protocols are optimized to work with HMW molecules to achieve an optimal fragment length distribution. Concomitantly, the quality of the RNA samples was poor by means of concentration and fragment size especially in EEW50 and it would normally not pass the requirements to be prepared and guarantee high quality data. The reason could be due to the sample management after surgery, low yielded amount of material, or, most probably, the choice in the extraction and isolation protocol of the genetic material (95,96). Therefore, **input sample quality should be always assessed to avoid reagents and sequencing waste and ensure success in the results.**

We took the decision on data generation and how to perform the proper data analysis. The WES approach has been used widely before in similar cancer immunotherapy projects (30), in order to interrogate the protein-coding genes of each patient. In this matter, the kit used in this project developed by Twist Bioscience yielded notable results regarding on target coverage and percent of duplicates. We attempted to achieve 30M read pairs per sample, as described in the materials and methods section. The EEW50 tumor sample got approximately that estimated number of million read pairs although tumor sample HRIU59 got 150M read pairs due to the reasons mentioned previously. As described before, hybridization consisted of an 8-sample pooled preparation with a tumor:normal-matched sample ratio of 3:1 from hybridization onwards. Hence, a higher number of read pairs was expected in the tumor sample, which was the desired outcome. In accordance, when comparing among the paired-matched sample (tumor and normal), the number of million reads obtained by the tumor sample HRIU59 was three times higher than normal HRIU59 sample. However, this event did not occur in EEW50, which points to a difference in the quality of the samples again. These comparison between the two samples was also reflected in the final coverage which might influence the variant calling. The exome panel used in this project covered a total region of ~33 Mb. Theoretically, a sequencing depth of 30 M read-pairs should yield around 180x average coverage, following the Lander/Waterman equation (97). However, that is the ideal case, and some factors reduce this number significantly, like the capture efficiency, alignment performance, non-constant coverage distribution along the whole exome design and the number of duplicates. More data will mean stronger evidence on the mutations found, but also higher sequencing and computational cost; thus, the differences in sequencing depth among the two samples was taken into account in the overall performance of the neoantigen prediction. In other words, the more data, the better expectations to find candidate neoantigens. The sensitivity/specificity trade-off must be adjusted according to the costs in sequencing and the minimum required amount of data for the proposed challenge. The aim of 30M read-pairs sequencing depth with the aforementioned library preparation protocol seems suitable middle point between cost and yield as long as the sample quality is sufficient.

Despite contrary indications supported by a low sample quality, **RNA samples were sequenced** together with two more RNA control samples **for gene expression quantification analysis**. The number of aligned reads was sufficient to run gene expression analysis, although the duplication rate was very high, probably related to the aforementioned low sample quality (98). It has been also shown that high duplication rate might be due to some highly expressed genes which transcribe many natural duplicates (99). However, the probability that two “twin” cDNA molecules are fragmented the same way during library preparation is too low (98), and thus, duplicates were removed from the gene expression analysis in this project. The gene body coverage analysis of these four samples did not suggest 3’ or 5’ bias (93,100) despite the probably low RIN values of both RNA samples EEW50 and HRIU59 as is indicated by the external QC; we suspect that further work is necessary to validate this analysis. Gene expression analysis was included after pVACseq analysis and helps to drive neoantigen prioritization. In conclusion, the quality of the RNA samples was far from ideal and importantly, data quality and duplicate removal analysis leaves an opened window for future review (98,100).

Variant calling enabled somatic mutation detection, **which is a core step in neoantigen prediction**. The combination of several variant callers seems a helpful strategy to enhance the accuracy in mutation detection. **Three variants callers** (Vardict, Strelka, Mutect2) were combined and **low consensus** among them was observed (63,64,69) due to their different algorithmic approaches, alignment method, post-processing and optimization (69). The overlap among variant callers provided higher confidence that those mutations were present in the samples. In order to **increase their confidence**, **two more conditions were applied**, similarly to the mutation detection pipeline of the MSK-IMPACT project (60). In fact, one such condition has already been implemented in neoantigen prediction (30). This approach **dramatically reduced the number variants**, being as much as two orders of magnitude in Strelka and Vardict. **The read coverage in both normal and tumor samples plays an important role in variant calling**. More specifically, at least 80x is required in WES data (63,67,101). All samples reached these specifications except the WES EEW50 tumor sample, which had the lowest quality since quality control before library preparation. Some discussion is being focused on the type of mutations and what neoantigens can be predicted from them. As mentioned before, missense SNVs have been more widely used in neoantigen detection, perhaps because of the ease in their detection and understanding in comparison to indels. However, identification of indel mutations for neoantigen prediction is very interesting as these neoantigens create frameshifts and new ORFs that could elicit stronger immune response than SNVs (44). In this sense, Strelka, Vardict and Mutect2 could detect deletion and insertion indels type (44). A deeper study for indel type mutations could be done in order to assess their reliability in the detection and later the potential to become a neoantigen. In fact, the performance of Vardict in detection of “complex” mutation types should be better addressed by accuracy assessment (using well-characterize real or simulated data) and how “complex” mutations are affected with sample quality, sequencing depth and coverage (different sets of samples and sequencing experiments) in order to include this type of mutations since they would represent potential neoantigen candidates.

Visual inspection of detected mutations showed a good matching between WES and RNAseq data only in HRIU59. However, this consensus poorly performed in sample EEW50. This observation suggests again the higher sample quality of DNA and RNA sample in HRIU59 with respect to EEW50. However, if the list of mutations became larger, this approach would become tedious without a more automatic method to visualize both WES and RNAseq data. Instead, variant detection and comparison between both RNAseq and WES data could be proposed in order to assess the agreement between both types of data. What if we could use only RNAseq data for both mutation detection (102) and expression analysis? However, most of the publications reviewed in this project related to neoantigen prediction from MPS data made use of WES; WES analysis provides more stable scenario of the AF distribution, which is useful for neoantigen prioritization, whereas RNAseq is dynamic and its analysis and understanding represents a harder challenge in this matter (102).

The final **list of candidates was extracted from the predicted neoantigens by pVACseq and MHC-binding affinity, HLA predicted binding allele, FPMK values and AF were visualized** together. Each of these parameters (AF, MHC-binding affinity and FPKM) added one more level of understanding for the final personalized prioritization in each patient. AF values are widely used to reveal the mutational load of each variant in the tumor and thus, if a mutation is spread in a high percentage along the whole tumor, it could represent a suitable target for immunotherapy. The AF threshold (AF=0.01) followed recommended specifications from MSK-IMPACT project (60), which is not focused on neoantigen prediction. Indeed, studies in neoantigen prediction have focused their attention in other AF thresholds, like 0.04 (57) and even 0.1 (45), based on two totally different criteria: 0.04 was set after running a Youden index statistical analysis and 0.1 was based on the variant caller recommended AF thresholds. In HRIU59, we have found neoantigens whose mutations had an AF higher than 0.2 (like TP53, ERBB2 and BRCA2 in HRIU59) whereas other interesting finding showed less than 0.05 AF (NFE2L2 and PROCR in HRIU59). Tumor heterogeneity, truncating mutations and cancer evolution (103,104) must shape the AF distribution but from the bioinformatics side, deeper understanding of the number of reads supporting a specific genomic position could enhance the confidence in AF calculation. RNAseq analysis and visualization provided essential clues about what mutations were transcribed and highly expressed in sample HRIU59. Therefore, **prioritization based on expression is one of the most robust criteria**. However, gene expression analysis should be revisited and transcript-specific analysis added to the pipeline. On the other hand, **MHC-binding parameter must not be considered determinant but useful to narrow down the list of output candidates** (58). In fact, MHC-binding prediction has also relied in the HLA allele typing predictions from sequencing data in this project. Validation of such predicted alleles by means of other bioinformatics approaches or biological assays must be performed to increase the rate of success.

As **future work**, the most urgent short-term goal is fixing the format incompatibility between pVACseq and Strelka, since more than 1500 variants are pending to be analyze. The second short-term goal is writing the detailed documentation of the code and steps which runs the pipeline, allowing reproducibility by any user. Later, the presented results will be validated from the performance of the subsequent steps in the whole project. In addition, a bioinformatics validation could consist of reproducing this workflow with another independent set of data that has already been used for similar purposes. pVACseq represents a main element in this project and further studies can be developed for better characterization of its functionalities, in addition, the use of its potential to analyze different transcript isoforms should be taken included. HLA alleles typing prediction from WES data and inexpert selection was certainly a limitation in this study. In fact, it has been reported that the performance of WES data is less sensitive than RNAseq and WGS (73) for this purpose. On the other hand, an emphasis on bladder cancer biology would increase the rate of success in this project (30), including functional pathway, Gene Ontology, mutational signatures and tumor profiling analysis. The available list of experimentally and clinically relevant proven neoantigens is limited and uncertain so far (34), combined to the lack of robust MHC binding affinity and T-cell response characterization. Therefore, joint efforts between different fields (MPS, bioinformatics, Mass Spectrometry, clinical interpretation) are vital for the success of new approaches in cancer immunotherapy. In the bioinformatics side, larger neoantigen databases would improve attempts to classify new neoantigens with the help of machine learning methods (35).

In conclusion, this work aims to ease the path of a neoantigen-based cancer immunotherapy by providing with the bioinformatics support to analyze NGS data. Deep literature review, implementation of tools concordant to each challenge described and evaluation of a final list of neoantigens under the selected approach has been performed to achieve the established goals. Although far from the real scenario (58), the reduction in costs and time of MPS-technology based approaches opens an optimistic gap for personalized cancer immunotherapy. The establishment of a robust sequencing data-based neoantigen bioinformatics prediction pipeline would evade the need for some costly and laborious wet lab assays to find or confirm neoantigens. This way, cancer immunotherapy strategies in development aim to become consolidated and positively impact where is now urgently needed. *After all, nothing new has to be invented, but our immune system rediscovered*.

Ethical reflection

Ethical approval for sample sequencing was necessary to start this project. Patient data was processed pseudonymised. It must be taken into very responsible consideration the value of the sample material and the critical situation for the patients' health status. Also, any partial or advanced results during the development of the technology must be revisited carefully, so that any eventual issues may be avoided when the platform is implemented in the clinic. These patients suffer an extremely severe life-threatening disease with poor prognosis and they do not have any more standard treatment possibilities, thus, a cancer immunotherapy approach could provide chance where other therapies have failed. In order to increase the probability of success for this new technology, extra pressure to be fast but accurate through the pipeline should be taken, since any new solution could mean an absolute revolution. Noteworthy, this technology might provide with a new possibility to achieve a cure for cancer. However, we could point two issues related to global economic and personal information integrity. First, the costs for cancer treatments are not affordable by everyone, which creates an unfair distribution and gives more privilege to only certain part of the population. This fact is reflected by the unequal number of cancer cases between developed and developing countries. A sustainable and affordable medicine in cancer is far from becoming real, should not we be concern about that? Second, clinical data protection is a must and more and more governments, hospitals and other entities are getting aware with the expansion of NGS; data privacy and handling are an essential issue to address whenever people's information is used by not only medical workers, such as bioinformaticians, or even companies which can profit from big data. Avoiding wrong medical conclusions from the data and decide what is the right amount of information that the patient needs to know will always be two points to consider by physicians and the understanding between other expertise fields and is essential.

Acknowledgements

I could not have been more thankful for having experienced this great training under your supervision, Valtteri, which has given me priceless opportunities and inspiration to enjoy what I do and drive my future steps. The second mention is to Hassan, for your key supervision and advice. I would not have made it without your magic in bioinformatics (a chess game is pending though).

The next mention goes to my lab mates and friends at Clinical Genomics. A thousand thanks to the unparalleled wet lab team that prepared the samples: Anna G., Anna E., Keyvan E., Anna Z. and Anna L. Thank you so much to Emma S. and Emilia O., who performed the last steps in the data generation. I feel that there are more than twenty people left to mention that should be included in this paragraph. So, a little message for a gigantic group: Clinical Genomics, you rock.

It's time to say a sincere THANKS to Ola Nilsson and Hans Grönlund et al. from CMM. I really enjoyed being enrolled and having your fantastic feedback. It has definitely introduced me into the field of cancer immunotherapy as an exciting challenge for the future. I so wish all the best for your success in this project because it would mean better chances for many people which suffers this terrible disease. Speaking of nice feedback, the one from Adnane Achour and Tanya Sandalova, who have made me reflect about the huge impact of bringing different "scientific cultures" together.

And of course, to my friends and classmates, who are my family in Stockholm. Thank you Cris, I think we have built a beautiful bridge from Madrid to Stockholm. Finally, the ones who are always there, no matter where I am, no matter what I do: my family, my parents and my brother.

Stockholm, 21st of May, 2018.

Bibliography

1. Ferlay J, Soerjomataram I, Ervik M, Dikshit R, Eser S, Mathers C, Rebelo M, Parkin DM, Forman D, Bray F. Cancer Incidence and Mortality Worldwide: IARC CancerBase No. 11. Vol. GLOBOCAN 2, International Agency for Research on Cancer. 2013.
2. Global, regional, and national comparative risk assessment of 79 behavioural, environmental and occupational, and metabolic risks or clusters of risks, 1990-2015: a systematic analysis for the Global Burden of Disease Study. 2016.
3. Organization WH. Projections of mortality and causes of death, 2015 and 2030 [Internet]. 2015. Available from: http://www.who.int/healthinfo/global_burden_disease/projections/en/
4. Stewart BW, Wild CP. World cancer report 2014. World Heal Organ [Internet]. 2014;1–2. Available from: http://www.videnza.org/wp-content/uploads/World-Cancer-Report-2014.pdf
5. Begley CG, Ellis LM. Drug development: Raise standards for preclinical cancer research. Nature. 2012;483(7391):531–3.
6. Shin SH, Bode AM, Dong Z. Precision medicine: the foundation of future cancer therapeutics. npj Precis Oncol [Internet]. 2017;1(1):12. Available from: http://www.nature.com/articles/s41698-017-0016-z
7. Muir P, Li S, Lou S, Wang D, Spakowicz DJ, Salichos L, et al. The real cost of sequencing: Scaling computation to keep pace with data generation. Genome Biol [Internet]. 2016;17(1):1–9. Available from: http://dx.doi.org/10.1186/s13059-016-0917-0
8. Wetterstrand KA. DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP).
9. Shameer K, Badgeley MA, Miotto R, Glicksberg BS, Morgan JW, Dudley JT. Translational bioinformatics in the era of real-time biomedical, health care and wellness data streams. Brief Bioinform. 2017;18(1):105–24.
10. Hanahan D, Weinberg RA. The hallmarks of cancer. Cell [Internet]. 2000;100(1):57–70. Available from: http://www.ncbi.nlm.nih.gov/pubmed/10647931
11. Hanahan D, Weinberg RA. Hallmarks of cancer: The next generation. Cell [Internet]. 2011;144(5):646–74. Available from: http://dx.doi.org/10.1016/j.cell.2011.02.013
12. Ding L, Bailey MH, Porta-Pardo E, Thorsson V, Colaprico A, Bertrand D, et al. Perspective on Oncogenic Processes at the End of the Beginning of Cancer Genomics. Cell. 2018;173(2):305–320.e10.
13. Alexandrov LB, Nik-Zainal S, Wedge DC, Aparicio SAJR, Behjati S, Biankin A V., et al. Signatures of mutational processes in human cancer. Nature. 2013;500(7463):415–21.
14. Galon J, Pagès F, Marincola FM, Thurin M, Trinchieri G. The immune score as a new possible approach for the classification of cancer / J. Galon [et al.]. // J. Transl. Med. – 2012. – Vol. 10, № 1. – P. 1-4. 2012;2–5.
15. Sudhakar a. History of Cancer, Ancient and Modern Treatment Methods Akulapalli. J Cancer Sci Ther [Internet]. 2010;1(2):1–4. Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2927383/pdf/nihms226784.pdf
16. Apellaniz-Ruiz M, Tejero H, Inglada-Perez L, Sanchez-Barroso L, Gutierrez-Gutierrez G, Calvo I, et al. Targeted sequencing reveals low-frequency variants in EPHA genes as markers of paclitaxel-induced peripheral neuropathy. Clin Cancer Res. 2017;23(5):1227–35.
17. Mellman I, Coukos G, Dranoff G. Cancer immunotherapy comes of age. Nature [Internet]. 2011;480(7378):480–9. Available from: http://dx.doi.org/10.1038/nature10673
18. Aldous AR, Dong JZ. Personalized neoantigen vaccines: A new approach to cancer immunotherapy. Bioorganic Med Chem [Internet]. 2017; Available from: https://doi.org/10.1016/j.bmc.2017.10.021
19. Sathyanarayanan V, Neelapu SS. Cancer immunotherapy: Strategies for personalization and combinatorial approaches. Mol Oncol. 2015;9(10):2043–53.
20. Ott PA, Hu Z, Keskin DB, Shukla SA, Sun J, Bozym DJ, et al. An immunogenic personal neoantigen vaccine for patients with melanoma. Nature [Internet]. 2017;547(7662):217–21. Available from: http://dx.doi.org/10.1038/nature22991
21. Wayteck L, Breckpot K, Demeester J, De Smedt SC, Raemdonck K. A personalized view on cancer immunotherapy. Cancer Lett [Internet]. 2014;352(1):113–25. Available from: http://dx.doi.org/10.1016/j.canlet.2013.09.016
22. Vormehr M, Diken M, Boegel S, Kreiter S, Türeci Ö, Sahin U. Mutanome directed cancer immunotherapy. Curr Opin Immunol. 2016;39:14–22.
23. Ku CS. Clinical relevance of cancer genome sequencing. World J Gastroenterol [Internet]. 2013;19(13):2011. Available from: http://www.wjgnet.com/1007-9327/full/v19/i13/2011.htm
24. Laboratory) CGU (Science for L. No Title [Internet]. Available from: https://www.scilifelab.se/facilities/clinical-genomics-stockholm/
25. Choi M, Scholl UI, Ji W, Liu T, Tikhonova IR, Zumbo P, et al. Genetic diagnosis by whole exome capture and massively parallel DNA sequencing. Proc Natl Acad Sci [Internet]. 2009;106(45):19096–101. Available from: http://www.pnas.org/cgi/doi/10.1073/pnas.0910672106
26. Timmermann B, Kerick M, Roehr C, Fischer A, Isau M, Boerno ST, et al. Somatic mutation profiles of MSI and MSS colorectal cancer identified by whole exome next generation sequencing and bioinformatics analysis.

PLoS One. 2010;5(12):1–10.

27. Yadav M, Jhunjhunwala S, Phung QT, Lupardus P, Tanguay J, Bumbaca S, et al. Predicting immunogenic tumour mutations by combining mass spectrometry and exome sequencing. Nature. 2014;515(7528):572–6.
28. Castle JC, Kreiter S, Diekmann J, Löwer M, Van De Roemer N, De Graaf J, et al. Exploiting the mutanome for tumor vaccination. Cancer Res. 2012;72(5):1081–91.
29. Reactivity NT. Tumor exome analysis reveals neoantigen-specigic T-cell reactivity in an ipilimumab-responsive melanoma. 2013;31(32):439–42.
30. Van Buuren MM, Calis JJA, Schumacher TNM. High sensitivity of cancer exome-based CD8 T cell neo-antigen identification. Oncoimmunology. 2014;3(5).
31. Clark MJ, Chen RR, Lam HYK, Karczewski KJ, Chen RR, Euskirchen G, et al. Performance comparison of exome DNA sequencing technologies. Nat Biotechnol [Internet]. 2011;29(10):908–14. Available from: http://www.ncbi.nlm.nih.gov/pubmed/21947028%5Cnhttp://www.nature.com.ezp-prod1.hul.harvard.edu/nbt/journal/v29/n10/full/nbt.1975.html
32. Ajay SS, Parker SCJ, Abaan HO, Fuentes Fajardo K V., Margulies EH. Accurate and comprehensive sequencing of personal genomes. Genome Res. 2011;21(9):1498–505.
33. Kim K, Seong M-W, Chung W-H, Park SS, Leem S, Park W, et al. Effect of Next-Generation Exome Sequencing Depth for Discovery of Diagnostic Variants. Genomics Inform [Internet]. 2015;13(2):31. Available from: http://genominfo.org/journal/view.php?doi=10.5808/GI.2015.13.2.31
34. Schumacher TN, Schreiber RD. Neoantigens in cancer immunotherapy. Sci Mag. 2015;348(6230).
35. Kim S, Kim HS, Kim E, Lee MG, Shin E, Paik S, et al. Neopepsee: accurate genome-level prediction of neoantigens by harnessing sequence and amino acid immunogenicity information. Ann Oncol [Internet]. 2018;(May):1030–6. Available from: http://academic.oup.com/annonc/advance-article/doi/10.1093/annonc/mdy022/4817339
36. Gubin MM, Artyomov MN, Mardis ER, Schreiber RD. Tumor_neoantigens__building_a_.PDF. J Clin Invest. 2015;125(9):3413–21.
37. Cheever MA, Allison JP, Ferris AS, Finn OJ, Hastings BM, Hecht TT, et al. The prioritization of cancer antigens: A National Cancer Institute pilot project for the acceleration of translational research. Clin Cancer Res. 2009;15(17):5323–37.
38. Pardoll D. Does the immune system see tumors as foreign or self? Annu Rev Immunol [Internet]. 2003;21(1):807–39. Available from: http://www.annualreviews.org/doi/10.1146/annurev.immunol.21.120601.141135
39. Lu YC, Robbins PF. Cancer immunotherapy targeting neoantigens. Semin Immunol. 2016;28(1):22–7.
40. Schreiber RD. Cancer Immunoediting : Integrating Suppression and Promotion. 2014;1565(2011):1565–71.
41. Lennerz V, Fatho M, Gentilini C, Frye RA, Lifke A, Ferel D, et al. The response of autologous T cells to a human melanoma is dominated by mutated neoantigens. Proc Natl Acad Sci [Internet]. 2005;102(44):16013–8. Available from: http://www.pnas.org/cgi/doi/10.1073/pnas.0500090102
42. Zhou J, Dudley ME, Rosenberg SA, Robbins PF. Persistence of Multiple Tumor-Specific T-Cell Clones Is Associated with Complete Tumor Regression in a Melanoma Patient Receiving Adoptive Cell Transfer Therapy. J Immunother [Internet]. 2005;28(1):53–62. Available from: http://content.wkhealth.com/linkback/openurl?sid=WKPTLP:landingpage&an=00002371-200501000-00007
43. Marco Gerlinger, M.D., Andrew J. Rowan, B.Sc., Stuart Horswell, M.Math., James Larkin, M.D. PD, David Endesfelder, Dip.Math., Eva Gronroos, Ph.D., Pierre Martinez, Ph.D., Nicholas Matthews BS, Aengus Stewart, M.Sc., Patrick Tarpey, Ph.D., Ignacio Varela, Ph.D., Benjamin Phillimore, B.Sc., Sharmin Begum, M.Sc., Neil Q. McDonald, Ph.D., Adam Butler, B.Sc., David Jones, M.Sc., Keiran Raine, M.Sc., Calli Latimer, B.Sc., Claudio R. Santos, Ph.D., Ma PD. Intratumor Heterogeneity and Branched Evolution Revealed by Multiregion Sequencing. NEnglJMed2. 2012;366(20):1859–69.
44. Turajlic S, Litchfield K, Xu H, Rosenthal R, McGranahan N, Reading JL, et al. Insertion-and-deletion-derived tumour-specific neoantigens and the immunogenic phenotype: a pan-cancer analysis. Lancet Oncol [Internet]. 2017;18(8):1009–21. Available from: http://dx.doi.org/10.1016/S1470-2045(17)30516-8
45. Lauss M, Donia M, Harbst K, Andersen R, Mitra S, Rosengren F, et al. Mutational and putative neoantigen load predict clinical benefit of adoptive T cell therapy in melanoma. Nat Commun [Internet]. 2017;8(1):1–10. Available from: http://dx.doi.org/10.1038/s41467-017-01460-0
46. Germano G, Lamba S, Rospo G, Barault L, Magri A, Maione F, et al. Inactivation of DNA repair triggers neoantigen generation and impairs tumour growth. Nature [Internet]. 2017;552(7683):1–5. Available from: http://dx.doi.org/10.1038/nature24673
47. Matsushita H, Vesely MD, Koboldt DC, Rickert CG, Uppaluri R, Magrini VJ, et al. Cancer exome analysis reveals a T-cell-dependent mechanism of cancer immunoediting. Nature [Internet]. 2012;482(7385):400–4. Available from: http://dx.doi.org/10.1038/nature10755
48. Reactivity NT. J OURNAL OF C LINICAL O NCOLOGY Tumor Exome Analysis Reveals. 2013;31(32):439–42.
49. Klein G, Sjögren HO, Klein E, Hellstrom KE. Epitope landscape in breast and colorectal cancer. Cancer Res. 1960;20(11):1561–72.
50. Lázaro S, Gamarra D, Del Val M. Proteolytic enzymes involved in MHC class I antigen processing: A guerrilla

army that partners with the proteasome. Mol Immunol [Internet]. 2015;68(2):72–6. Available from: http://dx.doi.org/10.1016/j.molimm.2015.04.014

51. Behl JD, Verma NK, Tyagi N, Mishra P, Behl R, Joshi BK. The Major Histocompatibility Complex in Bovines: A Review. ISRN Vet Sci [Internet]. 2012;2012(2):1–12. Available from: http://www.hindawi.com/journals/isrn/2012/872710/

52. Hoof I, Peters B, Sidney J, Pedersen LE, Sette A, Lund O, et al. NetMHCpan, a method for MHC class i binding prediction beyond humans. Immunogenetics. 2009;61(1):1–13.

53. Janeway CA Jr, Travers P, Walport M et al. The major histocompatibility complex and its functions [Internet]. 5th ed. New york: Garland Science; 2001. Available from: https://www.ncbi.nlm.nih.gov/books/NBK27156/

54. Calis JJA, Maybeno M, Greenbaum JA, Weiskopf D, De Silva AD, Sette A, et al. Properties of MHC Class I Presented Peptides That Enhance Immunogenicity. PLoS Comput Biol. 2013;9(10).

55. Sharkey MS, Lizée G, Gonzales MI, Patel S, Topalian SL. CD4+ T-Cell Recognition of Mutated B-RAF in Melanoma Patients Harboring the V599E Mutation. Cancer Res. 2004;64(5):1595–9.

56. Kubuschok B, Schmits R, Hartmann F, Cochlovius C, Breit R, König J, et al. Use of Spontaneous Epstein-Barr Virus-Lymphoblastoid Cell Lines Genetically Modified to Express Tumor Antigen as Cancer Vaccines: Mutated p21 *ras* Oncogene in Pancreatic Carcinoma as a Model. Hum Gene Ther [Internet]. 2002;13(7):815–27. Available from: http://www.liebertonline.com/doi/abs/10.1089/10430340252898993

57. Karasaki T, Nagayama K, Kuwano H, Nitadori JI, Sato M, Anraku M, et al. Prediction and prioritization of neoantigens: integration of RNA sequencing data with whole-exome sequencing. Cancer Sci. 2017;108(2):170–7.

58. The problem with neoantigen prediction. Nat Biotechnol. 2017;35(2):97.

59. Bräunlein E, Krackhardt AM. Identification and Characterization of Neoantigens As Well As Respective Immune Responses in Cancer Patients. Front Immunol. 2017;8(NOV):1–8.

60. Zehir A, Benayed R, Shah RH, Syed A, Middha S, Kim HR, et al. Mutational landscape of metastatic cancer revealed from prospective clinical sequencing of 10,000 patients. Nat Med. 2017;23(6):703–13.

61. Bohnert R, Vivas S, Jansen G. Comprehensive benchmarking of SNV callers for highly admixed tumor data. PLoS One. 2017;12(10):1–18.

62. Nielsen R, Paul JS, Albrechtsen A, Song YS. Genotype and SNP calling from next-generation sequencing data. Nat Rev Genet [Internet]. 2011;12(6):443–51. Available from: http://dx.doi.org/10.1038/nrg2986

63. Xu C. A review of somatic single nucleotide variant calling algorithms for next-generation sequencing data. Comput Struct Biotechnol J [Internet]. 2018;16:15–24. Available from: https://doi.org/10.1016/j.csbj.2018.01.003

64. Cai L, Yuan W, Zhang Z, He L, Chou KC. In-depth comparison of somatic point mutation callers based on different tumor next-generation sequencing depth data. Sci Rep [Internet]. 2016;6(April):1–9. Available from: http://dx.doi.org/10.1038/srep36540

65. Lai Z, Markovets A, Ahdesmaki M, Chapman B, Hofmann O, Mcewen R, et al. VarDict: A novel and versatile variant caller for next-generation sequencing in cancer research. Nucleic Acids Res. 2016;44(11).

66. Kalatskaya I, Trinh QM, Spears M, McPherson JD, Bartlett JMS, Stein L. ISOWN: Accurate somatic mutation identification in the absence of normal tissue controls. Genome Med. 2017;9(1):1–18.

67. Meynert AM, Bicknell LS, Hurles ME, Jackson AP, Taylor MS. Quantifying single nucleotide variant detection sensitivity in exome sequencing Quantifying single nucleotide variant detection sensitivity in exome sequencing. 2013;

68. Sieradzka K., Leszczorz K., Garbulowski M. PA. Consensus Approach for Detection of Cancer Somatic MutationsNo Title. In: Man-Machine Interactions 5 ICMMI 2017 Advances in Intelligent Systems and Computing, vol 659 Springer, Cham. 2018.

69. O'Rawe J, Jiang T, Sun G, Wu Y, Wang W, Hu J, et al. Low concordance of multiple variant-calling pipelines: Practical implications for exome and genome sequencing. Genome Med. 2013;5(3).

70. Vita R, Overton JA, Greenbaum JA, Ponomarenko J, Clark JD, Cantrell JR, et al. The immune epitope database (IEDB) 3.0. Nucleic Acids Res. 2015;43(D1):D405–12.

71. Flower DR. Immunoinformatics: Predicting Immunogenicity In Silico. Methods in Molecular Biology; 2007.

72. Snyder A, Makarov V, Merghoub T, Yuan J, Zaretsky JM, Desrichard A, et al. Genetic Basis for Clinical Response to CTLA-4 Blockade in Melanoma. N Engl J Med [Internet]. 2014;371(23):2189–99. Available from: http://www.nejm.org/doi/10.1056/NEJMoa1406498

73. Warren RL, Choe G, Freeman DJ, Castellarin M, Munro S, Moore R, et al. Derivation of HLA types from shotgun sequence datasets. Genome Med. 2012;4(12).

74. Hosomichi K, Jinam TA, Mitsunaga S, Nakaoka H, Inoue I. Phase-defined complete sequencing of the HLA genes by next-generation sequencing. BMC Genomics. 2013;14(1).

75. Sjöblom T, Jones S, Wood LD, Parsons DW, Lin J, Barber TD, et al. The consensus coding sequences of human breast and colorectal cancers. Science (80-) [Internet]. 2006;314(5797):268–74. Available from: http://www.ncbi.nlm.nih.gov/pubmed/16959974

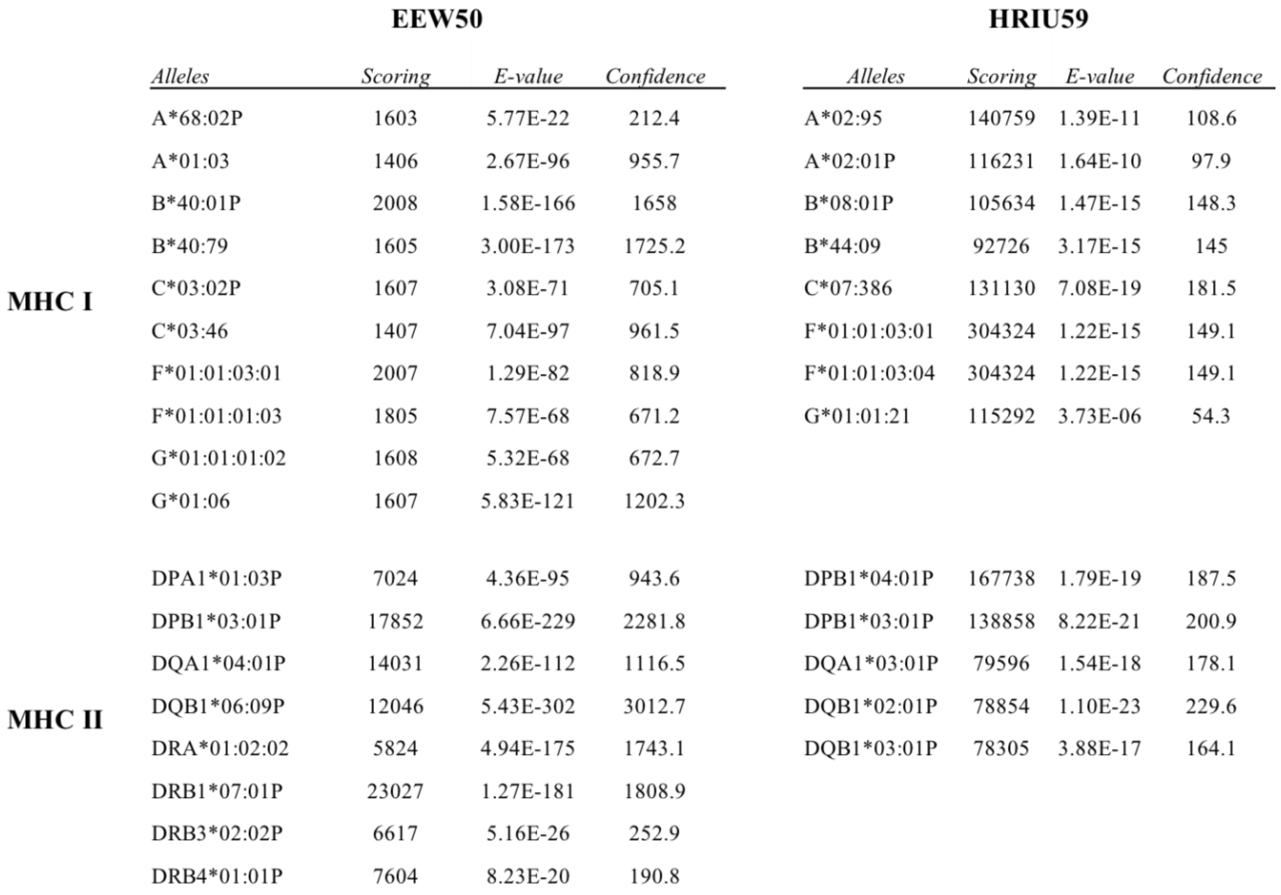
76. Fatih Ozsolak and Patrice M. Milos. RNA sequencing: advances, challenges and opportunities. Nat Rev Genet. 2011;

77. Conesa A, Madrigal P, Tarazona S, Gomez-Cabrero D, Cervera A, McPherson A, et al. A survey of best

practices for RNA-seq data analysis. Genome Biol. 2016;17(1):1–19.
78. Saal LH, Vallon-Christersson J, Häkkinen J, Hegardt C, Grabau D, Winter C, et al. The Sweden Cancerome Analysis Network - Breast (SCAN-B) Initiative: A large-scale multicenter infrastructure towards implementation of breast cancer genomic analyses in the clinical routine. Genome Med. 2015;7(1):1–12.
79. Hundal J, Carreno BM, Petti AA, Linette GP, Griffith OL, Mardis ER, et al. pVAC-Seq: A genome-guided in silico approach to identifying tumor neoantigens. Genome Med [Internet]. 2016;8(1):1–11. Available from: http://dx.doi.org/10.1186/s13073-016-0264-5
80. DTU-Bioinformatics Prediction Tools.
81. McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GRS, Thormann A, et al. The Ensembl Variant Effect Predictor. Genome Biol [Internet]. 2016;17(1):1–14. Available from: http://dx.doi.org/10.1186/s13059-016-0974-4
82. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal [Internet]. 2011;17(1):10. Available from: http://journal.embnet.org/index.php/embnetjournal/article/view/200
83. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009;25(14):1754–60.
84. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009;25(16):2078–9.
85. Picard Tools.
86. McKenna et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010;20:1297–303.
87. Kim D, Langmead B, Salzberg SL. HISAT: A fast spliced aligner with low memory requirements. Nat Methods. 2015;12(4):357–60.
88. Trapnell C, Williams B a, Pertea G, Mortazavi A, Kwan G, van Baren MJ, et al. Transcript assembly and abundance estimation from RNA-Seq reveals thousands of new transcripts and switching among isoforms. Nat Biotechnol [Internet]. 2011;28(5):511–5. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3146043&tool=pmcentrez&rendertype=abstract
89. Cibulskis K, Lawrence MS, Carter SL, Sivachenko A, Jaffe D, Sougnez C, et al. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. Nat Biotechnol [Internet]. 2013;31(3):213–9. Available from: http://dx.doi.org/10.1038/nbt.2514
90. Saunders CT, Wong WSW, Swamy S, Becq J, Murray LJ, Cheetham RK. Strelka: Accurate somatic small-variant calling from sequenced tumor-normal sample pairs. Bioinformatics. 2012;28(14):1811–7.
91. IGV (Integrative Genomic Viewer). Integrative Genomics Viewer. Broad Inst [Internet]. 2013;29(1):24–6. Available from: https://www.broadinstitute.org/igv/node/250
92. Institute B. Seqtk [Internet]. Available from: https://github.com/lh3/seqtk
93. Sigurgeirsson B, Emanuelsson O, Lundeberg J. Sequencing degraded RNA addressed by 3′ tag counting. PLoS One. 2014;9(3).
94. Lawrence MS, Stojanov P, Polak P, Kryukov G V., Cibulskis K, Sivachenko A, et al. Mutational heterogeneity in cancer and the search for new cancer-associated genes. Nature. 2013;499(7457):214–8.
95. Sellin Jeffries MK, Kiss AJ, Smith AW, Oris JT. A comparison of commercially-available automated and manual extraction kits for the isolation of total RNA from small tissue samples. BMC Biotechnol. 2014;14(1):1–13.
96. Hussing C, Kampmann ML, Mogensen HS, Børsting C, Morling N. Quantification of massively parallel sequencing libraries - A comparative study of eight methods. Sci Rep. 2018;8(1):1–9.
97. Port E, Sun F, Martin D, Waterman MS. Genomic mapping by end-characterized random clones: a mathematical analysis. Genomics. 1995;26(1):84–100.
98. Li X, Nair A, Wang S, Wang L. Quality Control of RNA-Seq Experiments. Vol. 1269, RNA Bioinformatics. 2015. 137-146 p.
99. Bansal V. A computational method for estimating the PCR duplication rate in DNA and RNA-seq experiments. BMC Bioinformatics. 2017;18(Suppl 3).
100. Gallego Romero I, Pai AA, Tung J, Gilad Y. RNA-seq: Impact of RNA degradation on transcript quantification. BMC Biol. 2014;12(1):1–13.
101. Sims D, Sudbery I, Ilott NE, Heger A, Ponting CP. Sequencing depth and coverage: Key considerations in genomic analyses. Nat Rev Genet [Internet]. 2014;15(2):121–32. Available from: http://dx.doi.org/10.1038/nrg3642
102. Piskol R, Ramaswami G, Li JB. Reliable identification of genomic variants from RNA-seq data. Am J Hum Genet [Internet]. 2013;93(4):641–51. Available from: http://dx.doi.org/10.1016/j.ajhg.2013.08.008
103. Burrell RA, McGranahan N, Bartek J, Swanton C. The causes and consequences of genetic heterogeneity in cancer evolution. Nature. 2013;501(7467):338–45.
104. Merlo LMF, Pepper JW, Reid BJ, Maley CC. Cancer as an evolutionary and ecological process. Nat Rev Cancer. 2006;6(12):924–35.

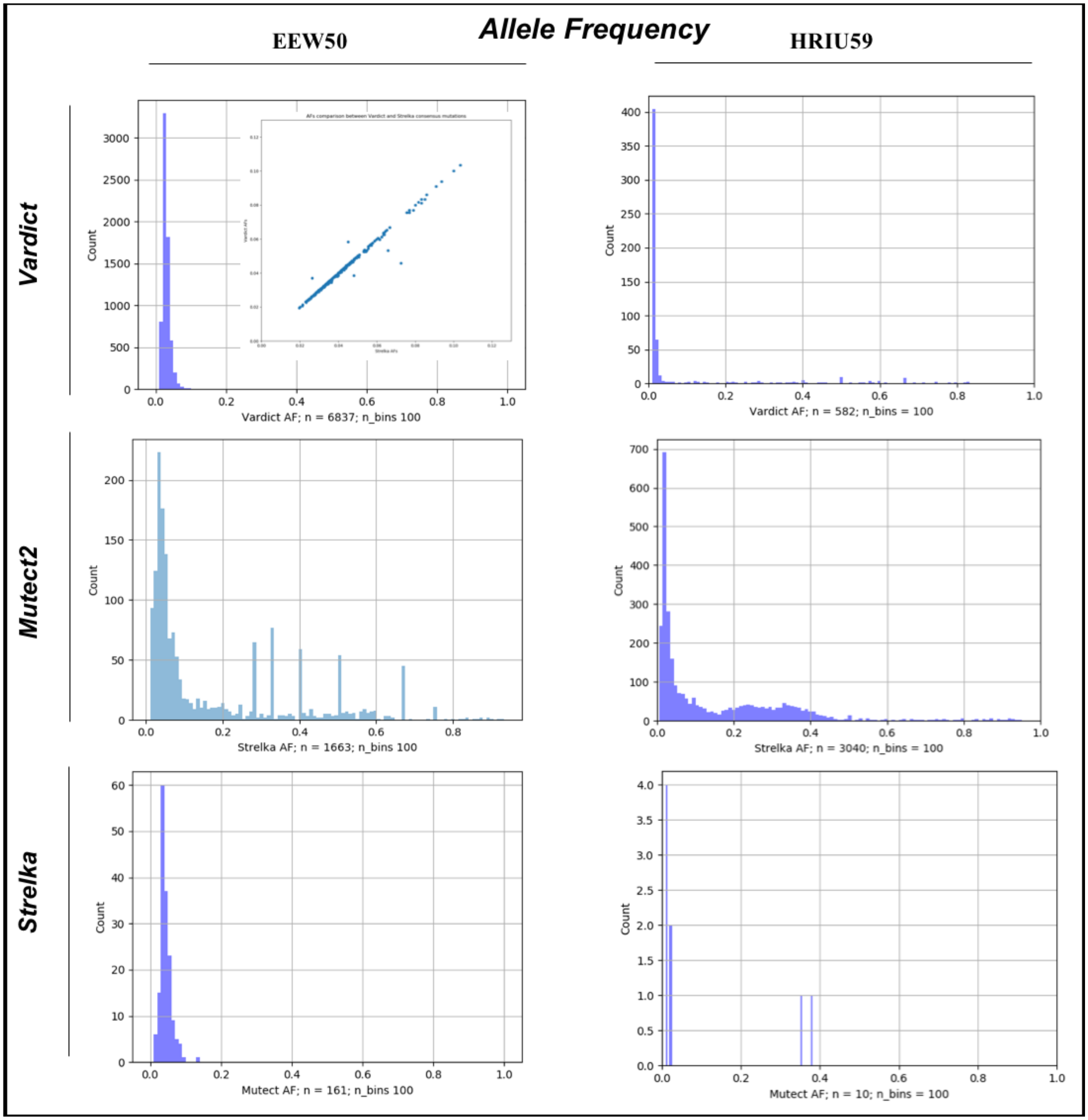
ANNEX

Table 4. HLA allele predictions chosen for each patient



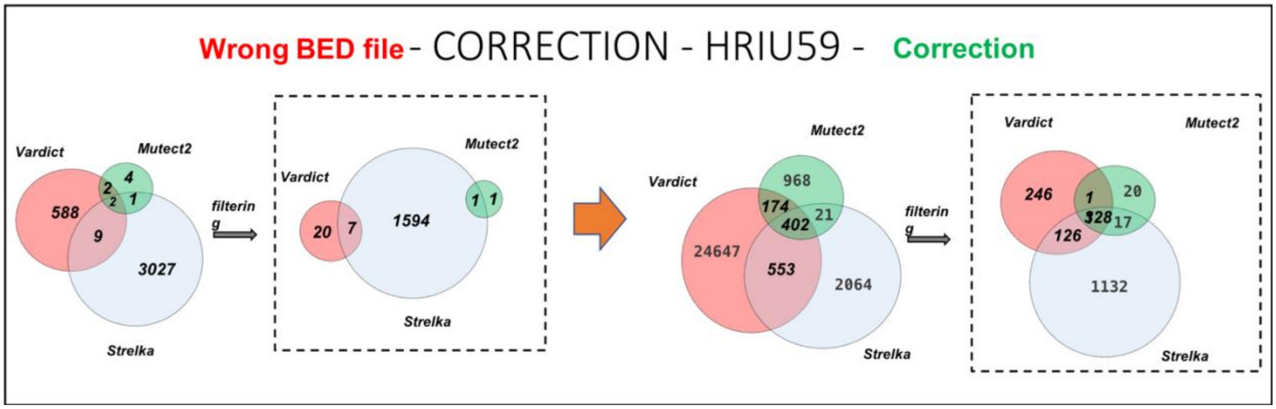
	EEW50				HRIU59			
	Alleles	*Scoring*	*E-value*	*Confidence*	*Alleles*	*Scoring*	*E-value*	*Confidence*
MHC I	A*68:02P	1603	5.77E-22	212.4	A*02:95	140759	1.39E-11	108.6
	A*01:03	1406	2.67E-96	955.7	A*02:01P	116231	1.64E-10	97.9
	B*40:01P	2008	1.58E-166	1658	B*08:01P	105634	1.47E-15	148.3
	B*40:79	1605	3.00E-173	1725.2	B*44:09	92726	3.17E-15	145
	C*03:02P	1607	3.08E-71	705.1	C*07:386	131130	7.08E-19	181.5
	C*03:46	1407	7.04E-97	961.5	F*01:01:03:01	304324	1.22E-15	149.1
	F*01:01:03:01	2007	1.29E-82	818.9	F*01:01:03:04	304324	1.22E-15	149.1
	F*01:01:01:03	1805	7.57E-68	671.2	G*01:01:21	115292	3.73E-06	54.3
	G*01:01:01:02	1608	5.32E-68	672.7				
	G*01:06	1607	5.83E-121	1202.3				
MHC II	DPA1*01:03P	7024	4.36E-95	943.6	DPB1*04:01P	167738	1.79E-19	187.5
	DPB1*03:01P	17852	6.66E-229	2281.8	DPB1*03:01P	138858	8.22E-21	200.9
	DQA1*04:01P	14031	2.26E-112	1116.5	DQA1*03:01P	79596	1.54E-18	178.1
	DQB1*06:09P	12046	5.43E-302	3012.7	DQB1*02:01P	78854	1.10E-23	229.6
	DRA*01:02:02	5824	4.94E-175	1743.1	DQB1*03:01P	78305	3.88E-17	164.1
	DRB1*07:01P	23027	1.27E-181	1808.9				
	DRB3*02:02P	6617	5.16E-26	252.9				
	DRB4*01:01P	7604	8.23E-20	190.8				

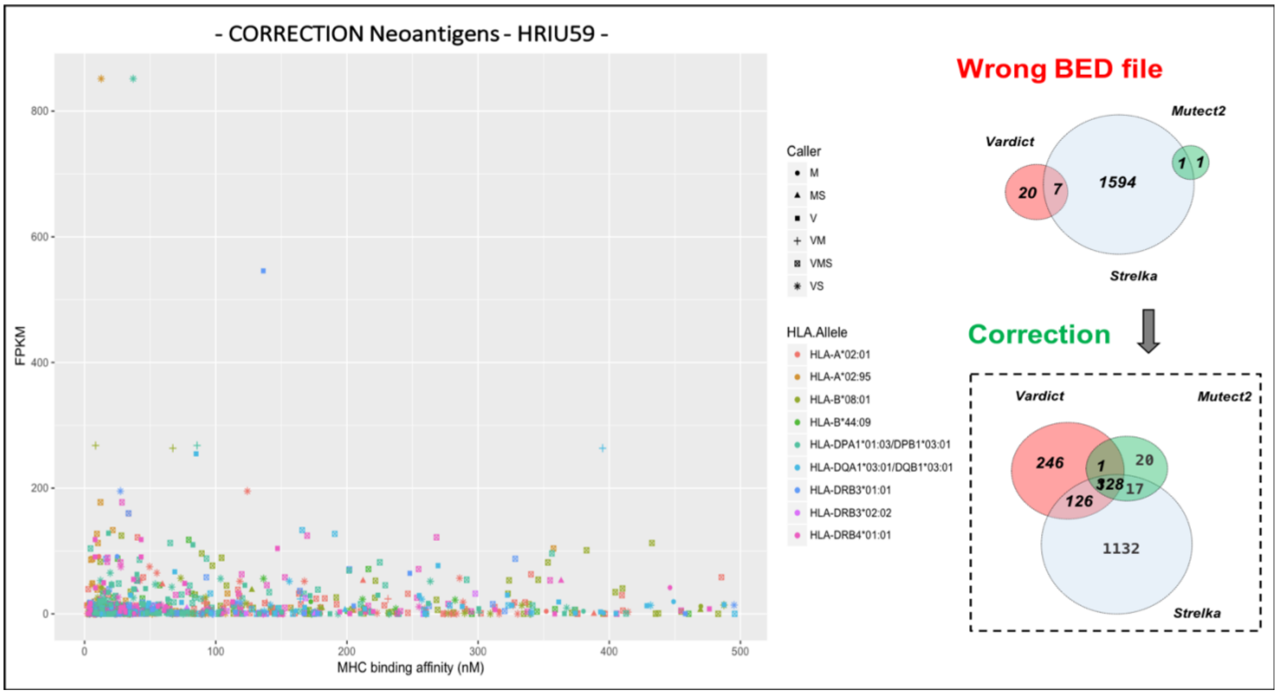
Supplementary Figure 7. Allele frequency distribution of all mutations found by each variant caller shown in Figure 7.



Correction Note – Mutations and neoantigens in sample HRIU59

This section aims to report a correction on the number of mutations and thus predicted neoantigens in HRIU59 sample. The initial number of neoantigens found was 14 as reported by Figure 11 (within the Results subsection called "***Neoantigens predicted in sample HRIU59***"). However, the lower number of mutations found in sample HRIU59 with respect to sample EEW50, albeit the great difference in amount of data, led to its investigation during last week of the project. It was found that the BED file reporting the regions of interest during variant calling analysis was not the right one, but this file spanned a smaller fraction of the human exome. Therefore, variant calling was missing some regions and thus not reporting existing mutations (Figure 7, mutations detected in HRIU59 is not complete). It is important to clarify that the neoantigens already reported in sample HRIU59 (Figure 11) are corrected identified. After a second analysis round using the right BED file, we must then report two figures as correction to the ones mentioned above. As observed, the number of mutations increases several orders of magnitude and the overlapping between the three variant callers improves. In addition, 792 neoantigen results came out from pVACseq, which will need further observation and prioritization in order to choose a list of suitable candidates.


Wrong BED file - CORRECTION - HRIU59 - Correction
Vardict
Mutect2
588
2
4
2
1
9
3027
Strelka
filterin
g
Vardict
Mutect2
20
7
1594
1
1
Strelka
Vardict
Mutect2
968
174
21
402
24647
553
2064
Strelka
filterin
g
Vardict
Mutect2
246
1
20
328
17
126
1132
Strelka


- CORRECTION Neoantigens - HRIU59 -
FPKM
MHC binding affinity (nM)
Caller
● M
▲ MS
■ V
+ VM
⊠ VMS
✱ VS
HLA.Allele
HLA-A*02:01
HLA-A*02:95
HLA-B*08:01
HLA-B*44:09
HLA-DPA1*01:03/DPB1*03:01
HLA-DQA1*03:01/DQB1*03:01
HLA-DRB3*01:01
HLA-DRB3*02:02
HLA-DRB4*01:01
Wrong BED file
Vardict
Mutect2
Strelka
Correction
Vardict
Mutect2
Strelka