frontiers

Supplementary Material

1 Supplementary Data

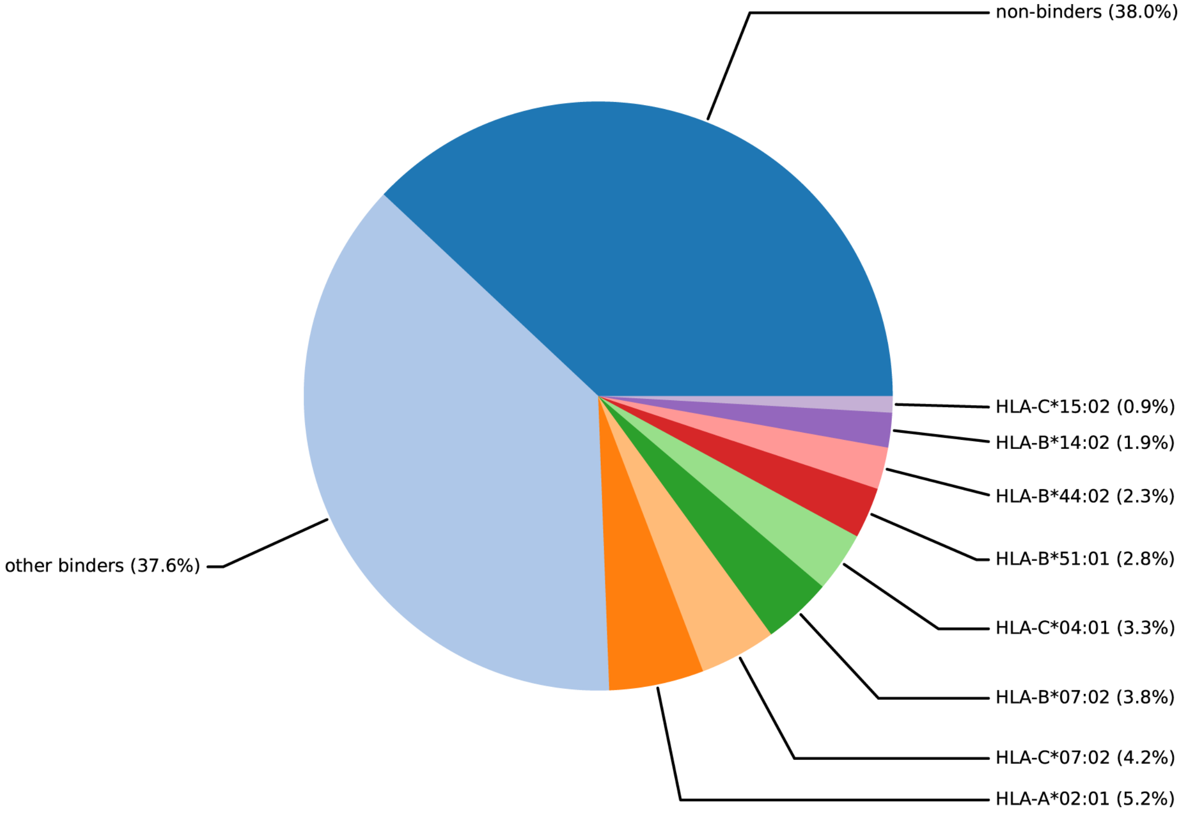
1.1 Pseudocode for Peptide rank boosting

```
For i in 1, 2, 3
    ∀ candidate in candidates
        boost =  W_npa * boost_npa + \
                 W_nph * boost_nph + \
                 W_nMHC * boost_nMHC + \
                 W_TPM * boost_TPM + \
                 W_ovlp * boost_ovlp
        new_rank = old_rank * (1-boost)
```

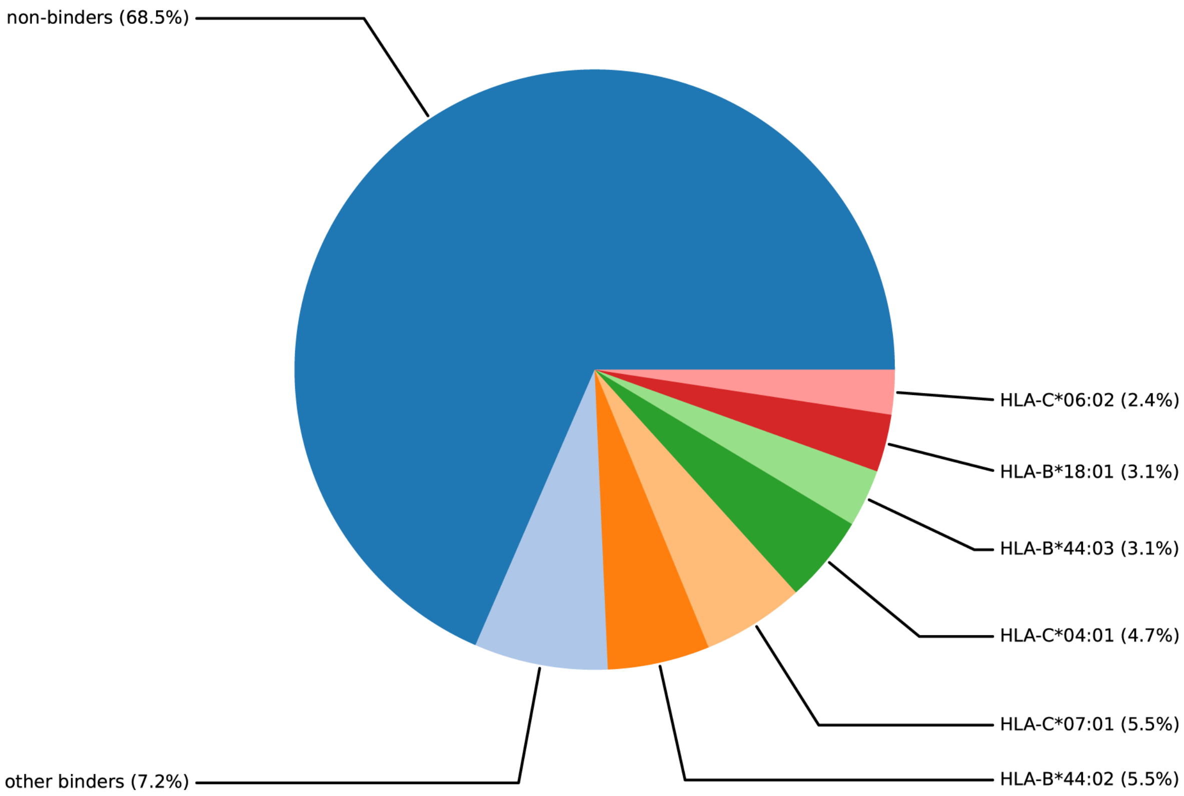
Where W_x describes the weight for covariate x, boost_x describes the score for the candidate x from 0-1, npa=number of peptides constituting an IAR, nph=number of strongly binding peptides constituting the IAR, nMHC=number of MHCs predicted to recognize a neoepitope from this IAR, TPM=expression of the gene harboring the IAR, and ovlp=number of events where a 9-mer and 10-mer overlap and are predicted to bind to the same MHC (only valid for MHCI). The individual scores for each of the covariates was calculated empirically based on internal test data. The Ranking algorithm is available at https://github.com/arkal/rankboost.

2 Supplementary Figures, Tables, and Files

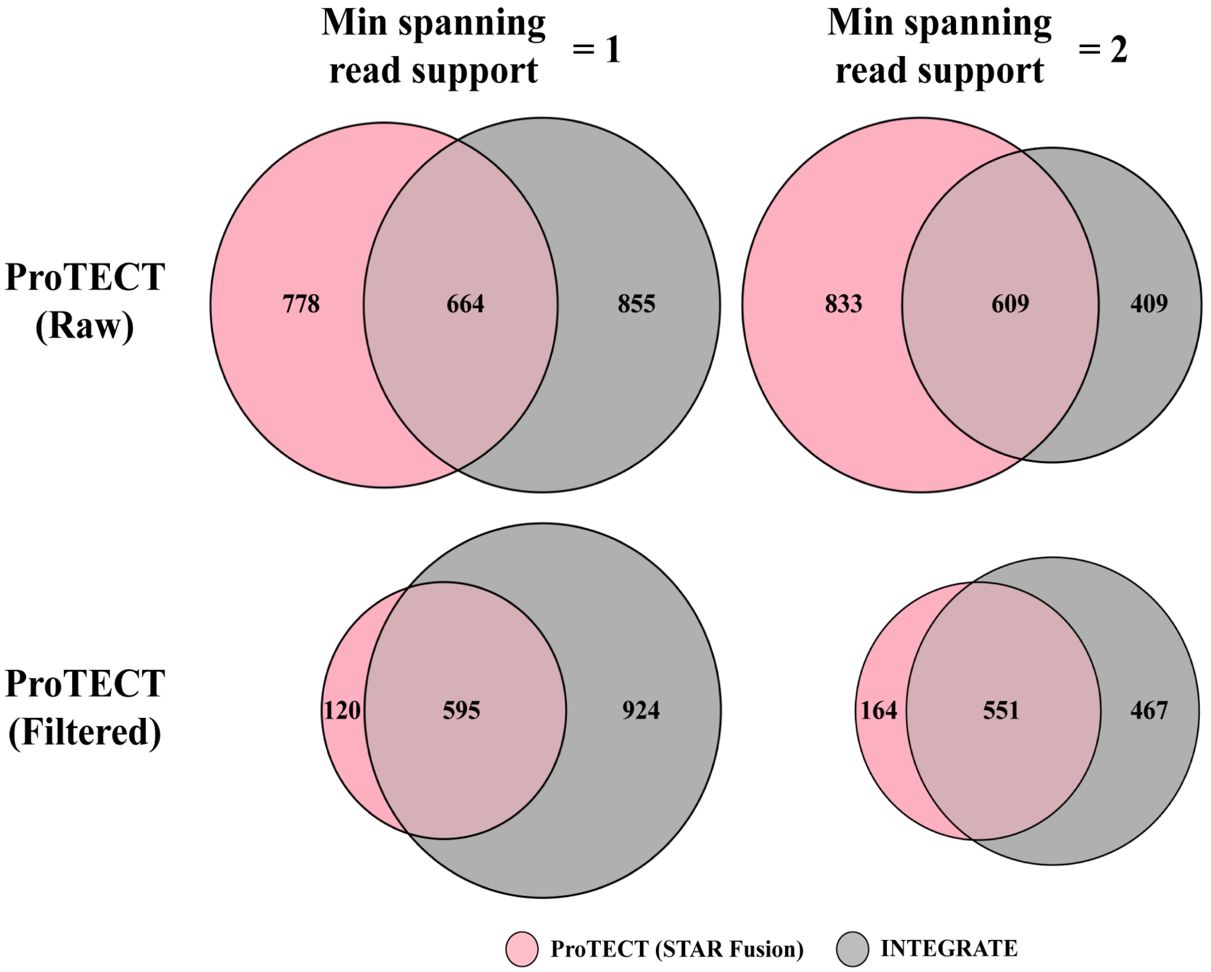
2.1 Supplementary Figures



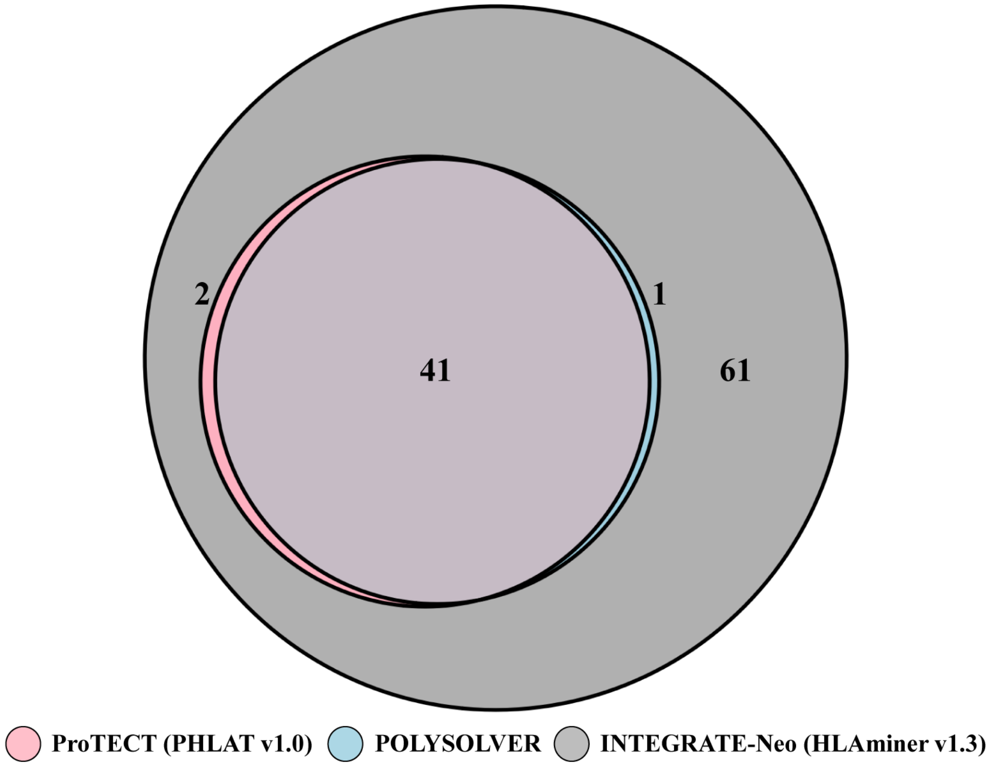
Supplementary Figure 1. Distribution of interesting MHC alleles called in samples harboring the TMPRSS2-ERG fusion IAR DNSKMALNSEALSVVSED at chr21:41498119-chr21:38445621



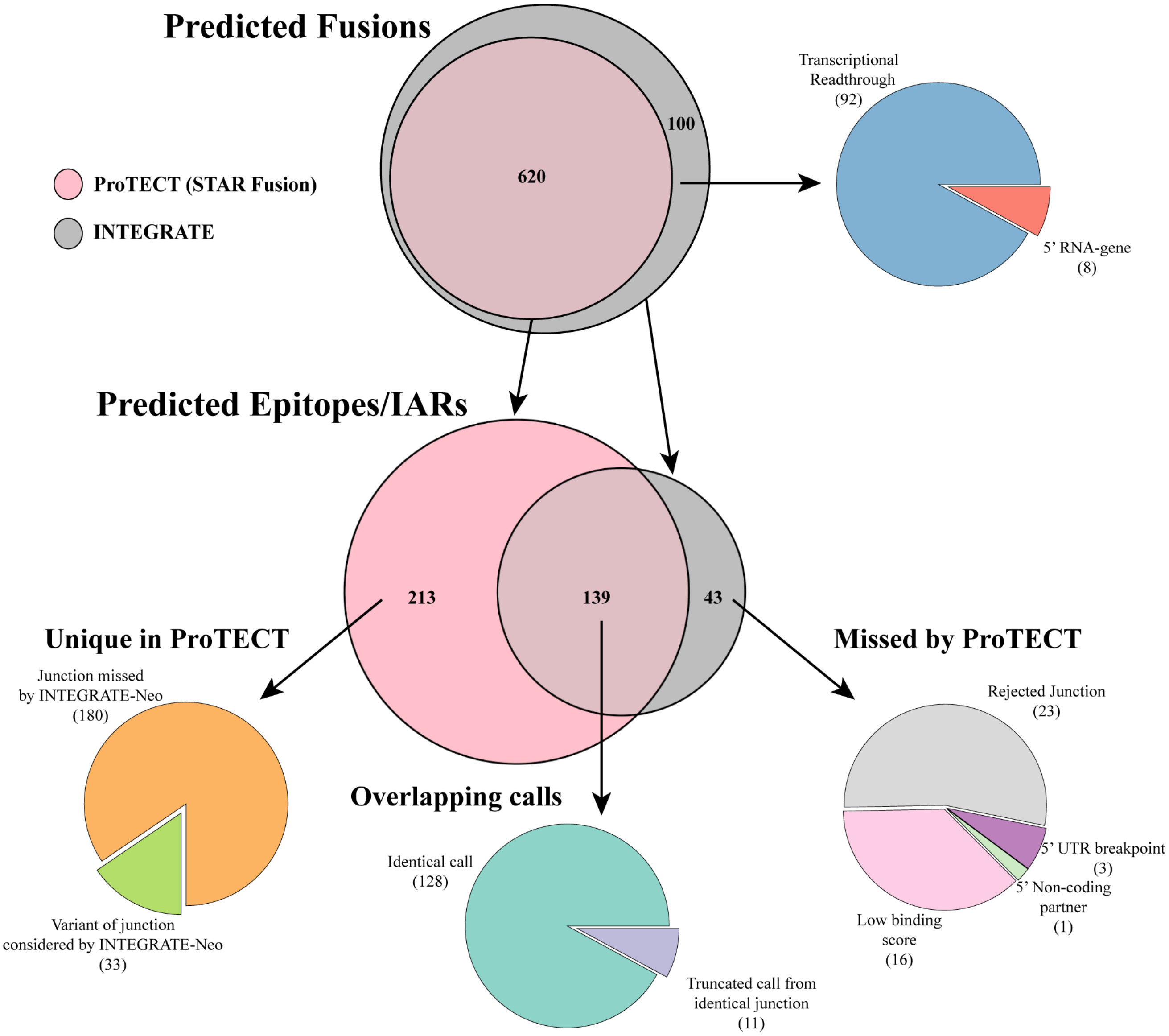
Supplementary Figure 2. Distribution of interesting MHC alleles called in samples harboring the TMPRSS2-ERG fusion IAR SGCEERGAAGSLISCE at chr21:41507950-chr21:38445621.



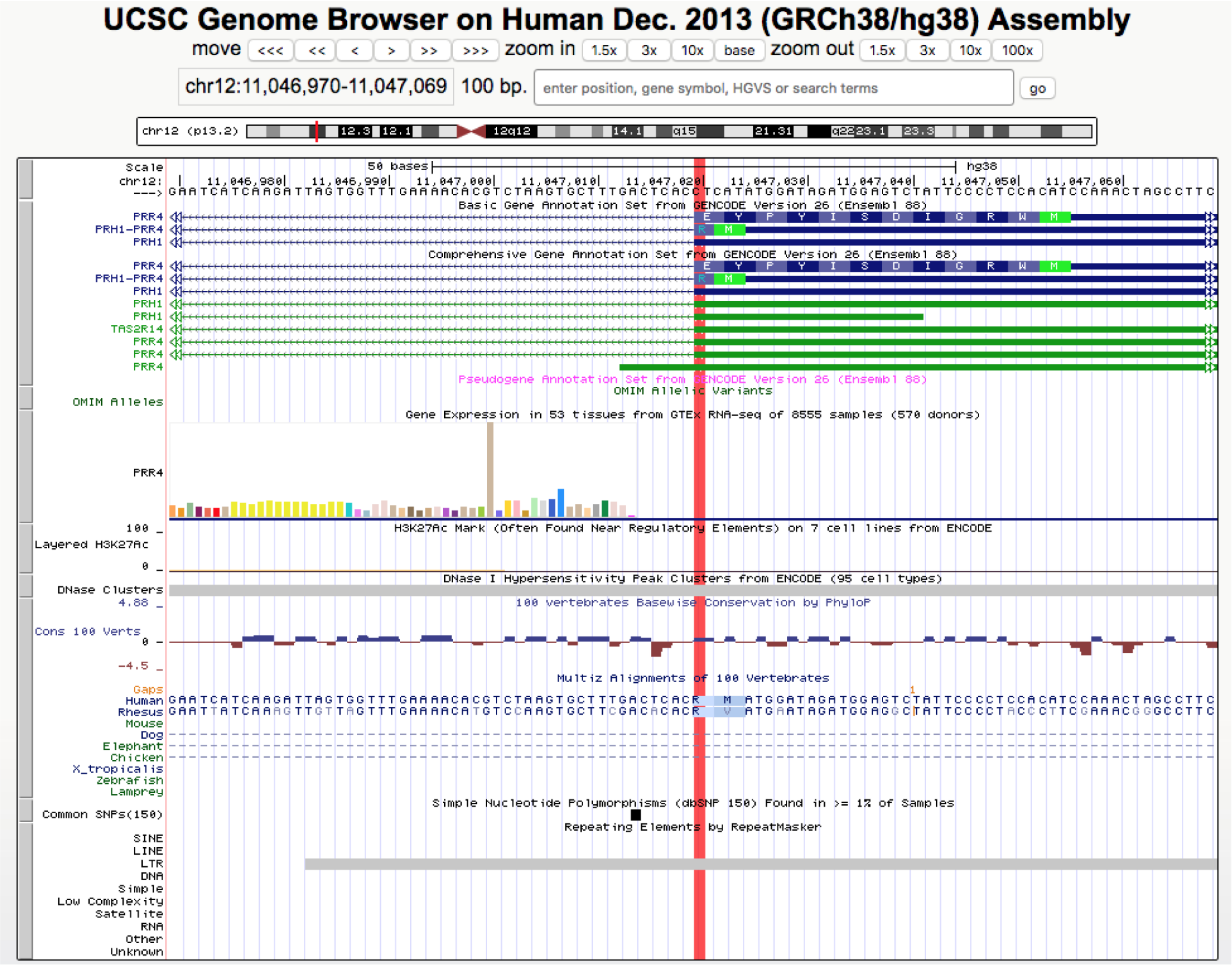
Supplementary Figure 3. Overlap between the fusion events called by ProTECT and INTEGRATE on a cohort of 240 prostate cancer samples. **Top**: There is low overlap between the raw ProTECT calls and INTEGRATE. Filtering INTEGRATE calls to retain only events with >= 2 supporting spanning reads (SSR) discards 33% of all calls but some of these events are called by ProTECT with >=2 SSR. **Bottom**: The filtered ProTECT results show better overlap with INTEGRATE calls using >=1 SSR and >=2 SSR. The number of ProTECT calls remains constant across rows since ProTECT uses a minimum of 2 SSR by default. The number of INTEGRATE calls remains constant across columns since there is no filtering applied to the INTEGRATE calls.



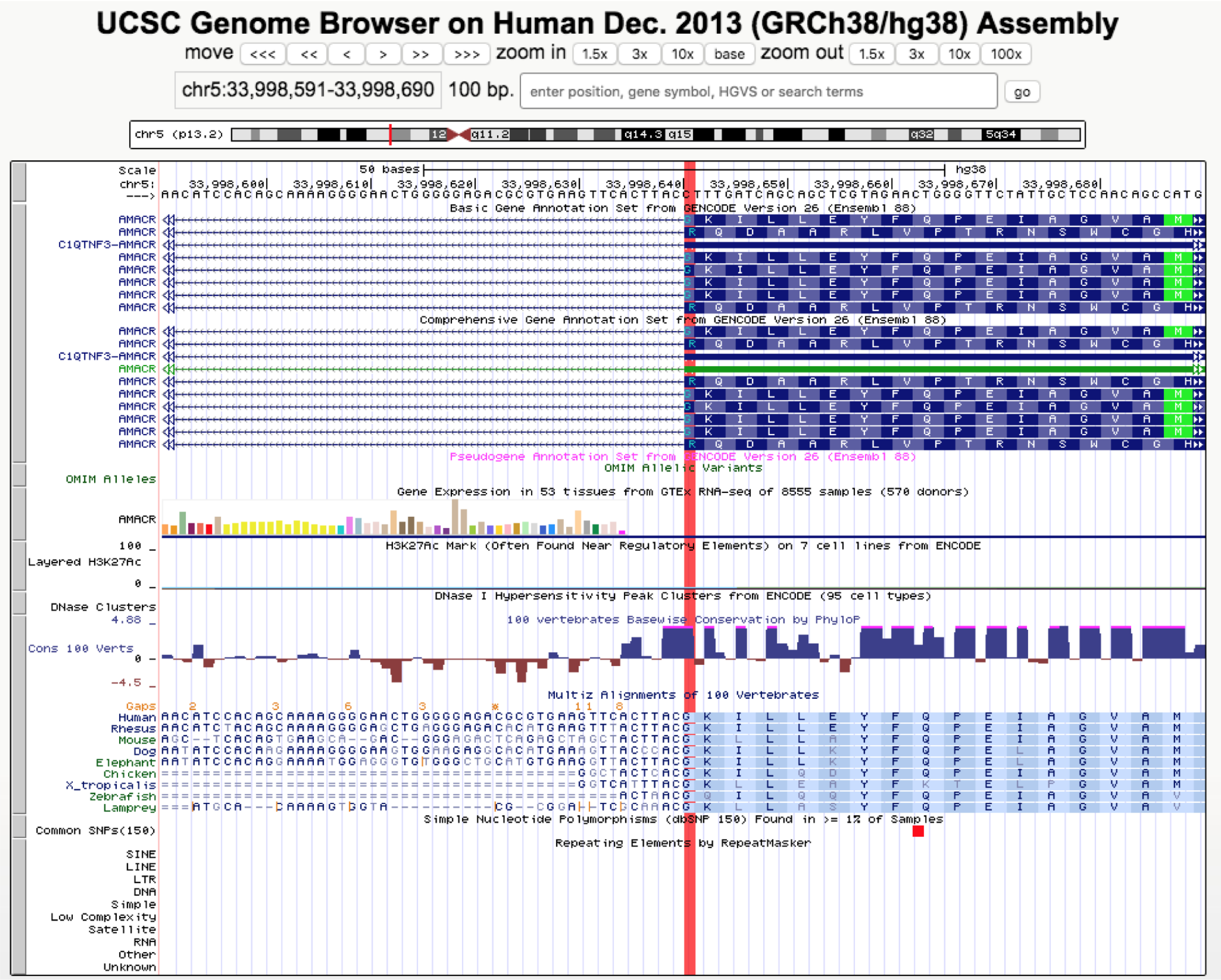
Supplementary Figure 4. HLA haplotypes called by HLAMiner in the INTEGRATE-Neo paper have very low overlap with ProTECT and POLYSOLVER.



Supplementary Figure 5: A schematic comparing ProTECT and INTEGRATE-Neo, using the fusion and MHC calls from the INTEGRATE-Neo paper (for maximal comparability). Protect rejects 100/720 calls from 83 patients and calls 352 IARs from the rest. A majority (23/43) of the epitopes missed by ProTECT arose from the rejected junctions. Of the 139 overlapping IARs, a small fraction (11) were not identical to the INTEGRATE call, but had significant overlap with the INTEGRATE-Neo expected epitope.


UCSC Genome Browser on Human Dec. 2013 (GRCh38/hg38) Assembly
move <<< << < > >> >>> zoom in 1.5x 3x 10x base zoom out 1.5x 3x 10x 100x
chr12:11,046,970-11,047,069 100 bp.
enter position, gene symbol, HGVS or search terms
go
chr12 (p13.2)
12.3
12.1
12q12
14.1
q15
21.31
q22
23.1
23.3
Scale
50 bases
hg38
chr12:
11,046,980
11,046,990
11,047,000
11,047,010
11,047,020
11,047,030
11,047,040
11,047,050
11,047,060
--->
GAATCATCAAGATTAGTGGTTTGAAAACACGTCTAAGTGCTTTGACTCACCTCATATGGATAGATGGAGTCTATTCCCCTCCACATCCAAACTAGCCTTC
Basic Gene Annotation Set from GENCODE Version 26 (Ensembl 88)
PRR4
PRH1-PRR4
PRH1
E Y P Y I S D I G R W M
R M
Comprehensive Gene Annotation Set from GENCODE Version 26 (Ensembl 88)
PRR4
PRH1-PRR4
PRH1
PRH1
PRH1
TAS2R14
PRR4
PRR4
PRR4
E Y P Y I S D I G R W M
R M
Pseudogene Annotation Set from GENCODE Version 26 (Ensembl 88)
OMIM Allelic Variants
OMIM Alleles
Gene Expression in 53 tissues from GTEx RNA-seq of 8555 samples (570 donors)
PRR4
100 _
Layered H3K27Ac
0 _
H3K27Ac Mark (Often Found Near Regulatory Elements) on 7 cell lines from ENCODE
DNase I Hypersensitivity Peak Clusters from ENCODE (95 cell types)
DNase Clusters
4.88 _
100 vertebrates Basewise Conservation by PhyloP
Cons 100 Verts
0 -
-4.5 _
Multiz Alignments of 100 Vertebrates
Gaps
1
Human GAATCATCAAGATTAGTGGTTTGAAAACACGTCTAAGTGCTTTGACTCACR M ATGGATAGATGGAGTCTATTCCCCTCCACATCCAAACTAGCCTTC
Rhesus GAATTATCAAAGTTGTTAGTTTGAAAACATGTCCAAGTGCTTCGACACACR V ATGAATAGATGGAGGCTATTCCCCTACCCTTCGAAACGGGCCTTC
Mouse
Dog
Elephant
Chicken
X_tropicalis
Zebrafish
Lamprey
Simple Nucleotide Polymorphisms (dbSNP 150) Found in >= 1% of Samples
Common SNPs(150)
Repeating Elements by RepeatMasker
SINE
LINE
LTR
DNA
Simple
Low Complexity
Satellite
RNA
Other
Unknown

Supplementary Figure 6. A UCSC genome browser showing the 5' breakpoint (highlighted) for the two missed epitopes from the ENSG00000231887-ENSG00000003056 fusion. The 5' partner was reported as PRH1 but breakpoint is in the 5' UTR for PRH1. The overlapping PRR4 gene contains the breakpoint sequence predicted by INTEGRATE.



chr5:33,998,591-33,998,690 100 bp.

UCSC Genome Browser on Human Dec. 2013 (GRCh38/hg38) Assembly
move <<< << < > >> >>> zoom in 1.5x 3x 10x base zoom out 1.5x 3x 10x 100x
enter position, gene symbol, HGVS or search terms
go
chr5 (p13.2)
Scale
50 bases
hg38
chr5:
33,998,600
33,998,610
33,998,620
33,998,630
33,998,640
33,998,650
33,998,660
33,998,670
33,998,680
Basic Gene Annotation Set from GENCODE Version 26 (Ensembl 88)
AMACR
C1QTNF3-AMACR
Comprehensive Gene Annotation Set from GENCODE Version 26 (Ensembl 88)
Pseudogene Annotation Set from GENCODE Version 26 (Ensembl 88)
OMIM Allelic Variants
OMIM Alleles
Gene Expression in 53 tissues from GTEx RNA-seq of 8555 samples (570 donors)
H3K27Ac Mark (Often Found Near Regulatory Elements) on 7 cell lines from ENCODE
Layered H3K27Ac
DNase I Hypersensitivity Peak Clusters from ENCODE (95 cell types)
DNase Clusters
100 vertebrates Basewise Conservation by PhyloP
Cons 100 Verts
Multiz Alignments of 100 Vertebrates
Gaps
Human
Rhesus
Mouse
Dog
Elephant
Chicken
X_tropicalis
Zebrafish
Lamprey
Simple Nucleotide Polymorphisms (dbSNP 150) Found in >= 1% of Samples
Common SNPs(150)
Repeating Elements by RepeatMasker
SINE
LINE
LTR
DNA
Simple
Low Complexity
Satellite
RNA
Other
Unknown

Supplementary Figure 7. A UCSC genome browser showing the 5' breakpoint (highlighted) for the two missed epitopes from the ENSG00000273294-ENSG00000164182 fusion. The 5' partner was reported as the readthrough transcript C1QTNF3-AMACR but the breakpoint is in the 5' UTR for C1QTNF3-AMACR. The overlapping AMACR contains the epitope predicted by INTEGRATE-Neo.

2.2 Supplementary Tables

Supplementary Table 1. The overall metrics for the 326 samples PRAD ProTECT run.This table includes all metrics captured for each file including input file sizes, and numbers of mutations called and accepted, peptides generated, and IARs predicted.

Supplementary Table 2. The ProTECT ranks for all 7 vaccine candidates tested by Carreno et.al. (1).

Supplementary Table 3. Ranked IARs predicted from all 8 samples described in Carreno et.al.(1).

Supplementary Table 4. Detailed reasons for rejecting INTEGRATE-Neo(2) predicted neoepitopes. Table A describes 4 breakpoints that were missed due to issues with the 5' partner, Table B describes a breakpoint that should have failed similar to the case in Table 1, but were rescued by the prediction of a second, overlapping breakpoint, and Table C describes the INTEGRATE-Neo- and ProTECT-predicted binding affinity for the epitopes rejected as poor binders.

2.3 Supplementary Files

Supplementary File 1: A tarball containing all the results from running ProTECT on 326 samples in the TCGA PRAD cohort.

Supplementary File 2: A tarball containing all the results from running ProTECT on 8 Melanoma samples described by Carreno et.al. (1).

Supplementary File 3: A tarball containing all the results from running ProTECT on 83 samples from the TCGA PRAD cohort, described by Zhang et.al (2). Specifically, this contains the results from running ProTECT using the fusions and HLA calls described in Zhang et.al.

3 references

1. Carreno BM, Magrini V, Becker-Hapak M, Kaabinejadian S, Hundal J, Petti AA, Ly A, Lie W-R, Hildebrand WH, Mardis ER, et al. A dendritic cell vaccine increases the breadth and diversity of melanoma neoantigen-specific T cells. Science (2015). doi:10.1126/science.aaa3828

2. Zhang J, Mardis ER, Maher CA. INTEGRATE-neo: a pipeline for personalized gene fusion neoantigen discovery. Bioinformatics (2017) 33:555–557. doi:10.1093/bioinformatics/btw674