Optimal neoantigen prediction and selection: Where's the sweet spot?

To improve upon the current efficacy rate of cancer therapeutics, a careful approach should be taken to identify epitopes (neoantigens) capable of guiding a broad immune response against the tumour. Exploration of the alterations uniquely detected in tumour tissue leads to dense information with sparse actionability, where identification of the most relevant target epitopes is the main challenge. This whitepaper describes how the optimal neoantigen prediction pipeline looks like by thorough discussion of the crucial steps in the neoantigen prediction set-up. Also a comparison is made of the most commonly used prediction pipelines to define the most advanced pipeline currently available.

Building a robust neoantigen identification pipeline requires a fine balance between having a **comprehensive detection approach** and an **optimised prioritisation strategy**. The former is especially important in cold tumours, where the mutational load is limited and expansion to novel kinds of mutational events (such as gene fusion and alternative splicing events) is needed to establish a sufficiently broad neoantigen pool. However, such an integrative approach carries the risk of finding too many potential candidates and thus requires a fine-tuned prioritisation filtering approach to compile the best final selection. Too strict filtering can reduce sensitivity and might drastically limit the available neoantigens, with a negative impact on the detection of subclonal mutations. In contrast, using relaxed filtering settings has the potential of including false positive neoantigens in the final pool, negatively affecting therapy efficiency and/or increasing the need for additional validation steps.

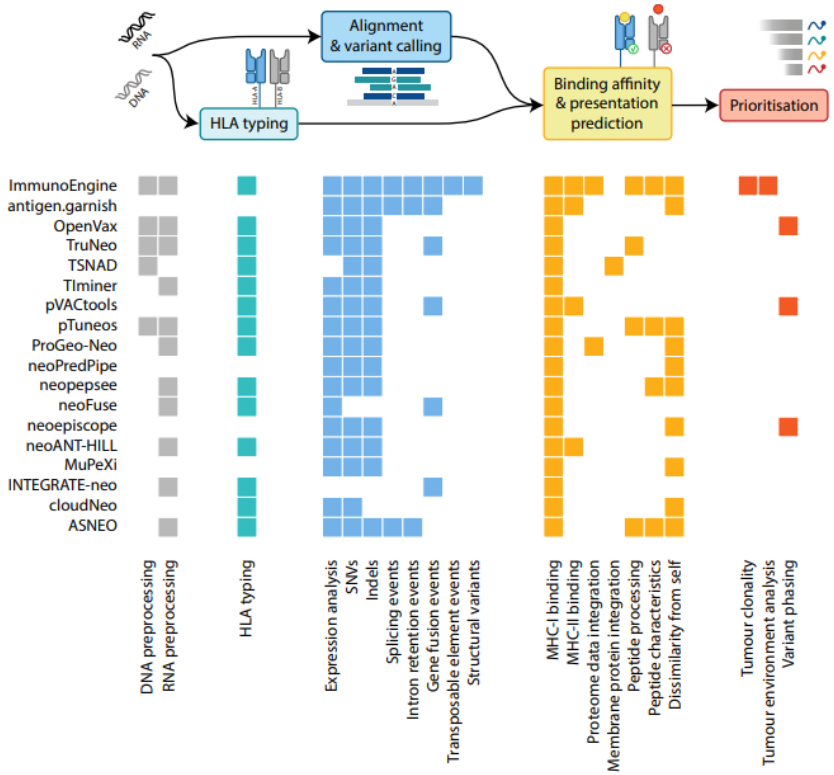

Figure 1. Schematic illustration of all necessary steps for optimal neoantigen prediction and selection, with an overview of the adopted features for each of the studied neoantigen prediction pipelines

A standard neoantigen prediction setup typically consists of 4 steps: raw sequencing data pre-processing, followed by (1) **HLA typing**; (2) **read mapping and neoantigen identification**; (3) **prediction of MHC binding and presentation**; and (4) **neoantigen prioritisation and selection** (**Figure 1**). The primary input for any prediction pipeline is usually matching normal and tumour DNA based whole exome sequencing (WES) data. In some cases, WES can be replaced by more expensive whole genome

sequencing (WGS) to increase the range of detectable mutational events (e.g. gene fusion events and chromosomal rearrangements). This can further be extended by RNA sequencing data, for confirmation and expression level analysis of elucidated mutations and to identify a novel set of transcript-level alterations. Additionally, ribosome-sequencing, mass spectrometry (MS), and MS-ligandome analysis can be used to further increase the number of tumour-specific alterations taken into consideration and to confirm variants predicted based on DNA and RNA sequencing datasets. In the following paragraphs, each of the aforementioned 4 steps will be described in more detail, followed by an in-depth comparison of the currently available prediction pipelines. Since most pipelines are restricted to MHC-I-based predictions, this whitepaper mainly focuses on the latter MHC-type. For information regarding MHC-II neoantigen prediction, we refer to publicly available literature[1–3].

HLA TYPING

The first prerequisite for developing an effective cancer vaccine is the selection of peptides/neoantigens able to be efficiently bound to and presented by major histocompatibility complex (MHC) molecules, since only such peptide/MHC (pMHC) complexes can elicit an immunogenic response through T-cell recognition. The HLA genes, coding for these complexes, are extremely patient-specific. Actionable neoantigen prediction is therefore largely dependent on a correct **inference of the individual HLA haplotype**. In most pipelines, MHC allele typing is performed through mapping of the DNA reads, which has been shown to result in **highly accurate predictions** (~99%)[4,5]. In contrast, RNA-based HLA typing is less performant, but an integrative RNA/DNA approach using several algorithms helps to further increase precision. Finally, it is recommended to consider expression levels and mutational events of the various HLA genes, since such events have shown to strongly impact neoantigen presentation and often cause immune evasion[6,7].

READ-MAPPING AND NEOANTIGEN IDENTIFICATION

The next step starts with mapping the sequencing reads on the human reference genome, with an appropriate aligner. This is followed by the most delicate process of the neoantigen discovery pipeline, namely characterising the tumour-specific variants. It involves the screening of exome or genome-wide tumoural mutations, combined with the detection of all other events known to affect the antigen repertoire.

Single nucleotide variants (SNVs) and **indels** are the two most commonly used variant types in prediction pipelines. Although the majority of pipelines focuses on a single variant caller – selected from the wide range of publicly available tools – to detect these events, it is recommended to only retain **variants identified by multiple calling algorithms** to further increase variant calling accuracy. In comparison to SNVs, indels frequently result in a shifted reading frame causing the ensuing peptides to be heavily impacted and thus deliver more immunogenic epitopes than the former. Despite the fact that indels are more desirable, they are often more difficult to characterise, thus furthering the need for ensemble approaches of various variant callers.

Finding immunogenic neoantigens for cancer vaccination in **cold tumours** with **low tumour mutational burden** (TMB) is challenging due to the low mutational frequency of such tumours, which means mutations beyond these standard variant types must be considered[8]. **Gene fusions** for example are a potent new neoantigen source and have been reported to drive development of approximately 16% of all cancers[9]. A large set of tools is readily available for the elucidation of gene fusion events based on either WGS or RNA-seq data. Unfortunately, such tools still result in quite a high number of false positives, which could be addressed by only retaining predictions supported by multiple data types or methods. A similar observation can be made for **alternative splicing events**. These have been shown to be more common in tumours than in healthy samples (up to 30%) and thus constitute another large pool of potential new neoantigen candidates[10]. Splicing aberrations can be detected using RNA-seq through straight-forward differential exon/intron expression analysis or with the help of specialised event-based tools. As for indels, gene fusion and alternative splicing events as well as neoisoforms, and transposable element activity have the added value of potentially resulting in more immunogenic neoantigens due to their huge sequence-altering effect.

In addition to the mutation-driven events discussed above, alterations in the tumour caused by **aberrant translation** of normal, unmutated parts of the genome should be examined as well. This requires further extending the neoantigen search domain with non-canonical events such as transcripts with **alternative start codons**, **uncanonical open reading frames** (ORFs), and **small ORFs** in for example long non-coding RNAs and 5' untranslated regions (UTRs)[11–15]. These events are often deemed more interesting than intracellular changes caused by mutational processes, as they are not dependent on the random occurrence of a specific tumoural mutation. Neoantigens derived from such events are **more likely to be shared**, thus requiring a less individualised therapeutic approach. Moreover, it has been shown that nonmutated aberrantly expressed tumour-specific antigen (aeTSA) events greatly outnumber mutated TSA events[8]. As such, taking both types in consideration in any prediction pipeline is likely to have a positive impact on the neoantigen detection success rate, even in tumours with low TMB.

PREDICTION OF MHC BINDING AND PRESENTATION

Presentation of neoantigens on the patient's MHC molecules is **crucial for efficient T-cell recognition**. As such, a lot of attention has been given to developing predictors able to reliably estimate the likelihood for a neoantigen to be presented at the cell surface. Earlier algorithms were mostly trained on data obtained through ***in vitro* peptide-MHC binding affinity assays**, and missed the impact of, for example, peptide/MHC (pMHC) stability, peptide degradation, and transportation. This results in **inaccurate predictions** of overall peptide presentation. Strong MHC binders, predicted by this type of tools, can still result in low presentation due to, for example, inadequate proteasomal degradation and TAP transport. That is why, when using models trained using *in vitro* affinity data in a prediction workflow, it is important to also include tools trained on *in vitro* proteasome digestion data, among others, to assess these upstream processes[16]. Unfortunately, this brings the risk of decreasing accuracy by aggregating potential prediction errors from different sources/tools. More recently developed methods include ***in vivo* ligandome data**, generated from pMHC immunoprecipitation followed by mass spectrometry (MS). As such, the overall peptide presentation process is considered at once, giving rise to more accurate presentation estimates.

As over 90% of MHC-presented peptides are unable to trigger a strong, directed immune response against the neoantigens, it is imperative to assess the **immunogenicity potential** of every predicted neoantigen to increase the effectiveness of the resulting cancer vaccine[17]. Unfortunately, research on *in silico* prediction of immunogenicity, i.e., the likelihood that a T-cell will recognise and react to a peptide presented on MHC-I, is still lacking. Most prediction workflows try to do this by determining the **dissimilarity to self** of the predicted neoantigens. This is based on the observation that neoantigens, similar in sequence to native peptides, are likely to be subjected to immunogenic tolerance, while neoantigens without native counterpart are more easily identified as foreign by the immune system. However, **other peptide and receptor-specific features** have shown to contribute to immunogenicity as well. For example, the amino acid sequence of the peptide by itself is believed to significantly affect pMHC-TCR recognition and the degree of immunogenicity. Similarly, other studies have reported immunogenic peptides to be enriched in hydrophobic amino acids at TCR interaction sites and have stressed the importance of peptide structure and amino acid weight, size, and charge in establishing pMHC-TCR complexes[17–20]. These findings indicate that modelling pMHC-TCR interactions should be feasible and should be included in prediction pipelines to increase the success rate of selecting truly immunogenic neoantigens. Unfortunately, validated pMHC-TCR interaction data is still limited at this time resulting in many such tools generating inaccurate predictions.

NEOANTIGEN SELECTION

Neoantigen prioritisation is the final step of the neoantigen discovery process. The vast majority of alterations in the tumour appears to have no immunogenic effect and prediction pipelines are still returning **lots of false positive, non-immunogenic neoantigens**. It is therefore crucial to filter out these unsuitable neoantigens and select a subset of highly actionable candidates. This is further substantiated by the fact that, in ICI responders the effector T-cell response is commonly driven by only a limited set of neoantigen-specific T-cell expansions, favouring the hypothesis that only a restricted number of neoantigens mediate

anti-tumour immune responses[21]. In addition, it is expected that inclusion of subdominant neoantigens, i.e., antigens that require immunisation as they do not naturally induce an immune response, into any therapeutic strategy will lead to a decrease of the selection of antigen loss variants, which is particularly of interest in the context of clonal tumour evolution.

Also in the context of **tumour heterogeneity** and **subclonality**, it is essential to implement *in silico* filtering and selection that prefers functionally important tumour alterations (driver mutations) with high allelic fraction to increase the probability of retaining **highly clonal epitopes** and **prevent immune escape**[22]. To aid in the selection of such clonal mutations, while discarding subclonal variants present in only a subset of the cancer cells, various bioinformatic methods have been developed over the last years. These clonal evolution assessments base predictions on different factors such as copy number alterations (CNA), variant allele fractions and tumour purity.

COMPARISON OF AVAILABLE PREDICTION PIPELINES

In general, the performance of any neoantigen prediction pipeline largely depends on the type of mutational events it can detect and the downstream neoantigen quality assessments it performs. Both are equally important. The optimal pipeline takes into account multiple variant types, confirmed by multiple variant callers, and uses various multi-angle models to filter out potential non-immunogenic hits. When looking at the most commonly used neoantigen prediction pipelines, some clear trends and core components can be observed (**Figure 1**). The majority of the fifteen pipelines studied here do not natively perform WES/WGS pre-processing and variant calling. Instead, they **rely on the user to generate a list of variants** using a set of widely available methods. Only pTuneos, TSNAD, OpenVax, TruNeo and myNEOs ImmunoEngine handle raw sequencing input[23–26]. A large number of workflows allow some sort of RNA-seq pre-processing, however, most of the time this relates to expression analysis and neoantigen prioritisation, with only a few pipelines performing RNA-seq based read alignment and variant calling.

Integrated **HLA typing is also widely adopted** by most of the pipelines, only Neoepiscope, neoPredPipe, and MuPeXI expect a user-supplied list of HLA alleles[27–29]. With some exceptions, all pipelines **focus on SNVs and indels**. NeoFuse, pVACtools, INTEGRATE-NEO, TruNeo and the myNEO ImmunoEngine can handle fusion events, while ASNEO and the myNEO ImmunoEngine are also compatible with splicing and intron retention events[25,30–33]. Variant phasing, i.e., the ability to cope with multiple variants in close vicinity into a single mutant peptide, has only been integrated by neoepiscope and pVACtools[27,31]. It should be noted that variant phasing at the epitope level is an event with a very low probability of occurrence, making this a less crucial step. Finally, the **importance of MHC binding** is further stressed by the fact that all the examined pipelines include this feature. In this context, the myNEO ImmunoEngine goes one step further by adding the ability to accurately predict peptide presentation on top of its binding affinity assessment. MHC-II compatibility, however, is limited to neoANT-HILL, pVACtools and myNEO the ImmunoEngine[31,34]. Inclusion of proteomic data and additional peptide-related features are only performed by a selection of the pipelines.

Overall, most prediction pipelines abide to the four most important steps outlined at the beginning of this whitepaper, namely HLA typing, variant identification, immunogenicity prediction and neoantigen presentation. In addition, immunogenicity assessment is commonly performed through MHC binding and dissimilarity to self analyses, with a widespread focus on SNVs and indels. In contrast, only a few pipelines expand to gene fusion and splicing events. Upon comparison of the most commonly used pipelines, the myNEO ImmunoEngine appears to be the most comprehensive pipeline. It stands out by its ability to consider variants from multiple sources (SNV/indel and structural variants, gene fusion, splicing transposable element events, …) and its ability to assess neoantigen presentation and immunogenicity by means of proteomics supported, state-of-the-art prediction models. To conclude, although the identification of tumour specific neoantigens have come a long way in the context of specificity and sensitivity, a wide adoption of high-throughput proteomics data and advances in the field of pMHC/TCR modelling are expected to result in a significant increase in prediction accuracy, further improving cancer vaccine efficacy.

About myNEO

myNEO (Ghent, Belgium) developed a platform enabling genomic-informed drug discovery in the key therapeutic areas of oncology and immunology. The data-driven ImmunoEngine identifies the most efficacious targets (epitopes) for each cancer patient, uniquely presented on the tumour cells and capable of redirecting a patient's immune system, leading to elimination of the cancer cells. The discovery platform enables targets to be identified even in hard-to-treat tumours with a cold/lowly mutated profile. Similarly, the company has applied its technology to identify immunogenic sequences in infectious diseases, capable of protecting populations with strong broad immune responses. myNEO is one of the companies that emerged from the Novalis biotech incubator fund at the end of 2018, founded by two leading entrepreneurs already known for several successes in the biotech industry: Wim Van Criekinge, professor of computational biology at Ghent University, and childhood friend Jan Van den Berghe.

Contact us

Interested in more information about myNEO? Contact us!

Lien Lybaert, PhD
Scientific Alliance Manager
lien.lybaert@myneotx.com

---

1. Reynisson, B., Alvarez, B., Paul, S., Peters, B. & Nielsen, M. NetMHCpan-4.1 and NetMHCIIpan-4.0: Improved predictions of MHC antigen presentation by concurrent motif deconvolution and integration of MS MHC eluted ligand data. *Nucleic Acids Res.* (2021) doi:10.1093/NAR/GKAA379.

2. Racle, J. *et al.* Robust prediction of HLA class II epitopes by deep motif deconvolution of immunopeptidomes. *Nat. Biotechnol.* (2019) doi:10.1038/s41587-019-0289-6.

3. Chen, B. *et al.* Predicting HLA class II antigen presentation through integrated deep learning. *Nat. Biotechnol.* (2019) doi:10.1038/s41587-019-0280-2.

4. Kiyotani, K., Mai, T. H. & Nakamura, Y. Comparison of exome-based HLA class i genotyping tools: Identification of platform-specific genotyping errors. *J. Hum. Genet.* (2017) doi:10.1038/jhg.2016.141.

5. Bauer, D. C., Zadoorian, A., Wilson, L. O. W. & Thorne, N. P. Evaluation of computational programs to predict HLA genotypes from genomic sequencing data. *Brief. Bioinform.* (2018) doi:10.1093/bib/bbw097.

6. Shukla, S. A. *et al.* Comprehensive analysis of cancer-associated somatic mutations in class i HLA genes. *Nat. Biotechnol.* (2015) doi:10.1038/nbt.3344.

7. McGranahan, N. *et al.* Allele-Specific HLA Loss and Immune Escape in Lung Cancer Evolution. *Cell* (2017) doi:10.1016/j.cell.2017.10.001.

8. Laumont, C. M. *et al.* Noncoding regions are the main source of targetable tumor-specific antigens. *Sci. Transl. Med.* (2018) doi:10.1126/scitranslmed.aau5516.

9. Gao, Q. *et al.* Driver Fusions and Their Implications in the Development and Treatment of Human Cancers. *Cell Rep.* (2018) doi:10.1016/j.celrep.2018.03.050.

10. Kahles, A. *et al.* Comprehensive Analysis of Alternative Splicing Across Tumors from 8,705 Patients. *Cancer Cell* 34, 211-224.e6 (2018).

11. Koster, J. & Plasterk, R. H. A. A library of Neo Open Reading Frame peptides (NOPs) as a sustainable resource of common neoantigens in up to 50% of cancer patients. *Sci. Rep.* (2019) doi:10.1038/s41598-019-42729-2.

12. Ouspenskaia, T. *et al.* Thousands of novel unannotated proteins expand the MHC I immunopeptidome in cancer. *bioRxiv* (2020) doi:10.1101/2020.02.12.945840.

13. Orr, M. W., Mao, Y., Storz, G. & Qian, S. B. Alternative ORFs and small ORFs: Shedding light on the dark proteome. *Nucleic Acids Res.* (2021) doi:10.1093/NAR/GKZ734.

14. Kochetov, A. V. Alternative translation start sites and their significance for eukaryotic proteomes. *Molecular Biology* (2006) doi:10.1134/S0026893306050049.

15. Liu, J. *et al.* Initiation of translation from a downstream in-frame AUG codon on BRCA1 can generate the novel isoform protein ΔBRCA1 (17aa). *Oncogene* (2000) doi:10.1038/sj.onc.1203599.

16. Calis, J. J. A., Reinink, P., Keller, C., Kloetzel, P. M. & Keşmir, C. Role of peptide processing predictions in T cell epitope identification: contribution of different prediction programs. *Immunogenetics* (2015) doi:10.1007/s00251-014-0815-0.

17. Wells, D. K. *et al.* Key Parameters of Tumor Epitope Immunogenicity Revealed Through a Consortium Approach Improve Neoantigen Prediction. *Cell* 183, 818-834.e13 (2020).

18. Wang, S. *et al.* Analyzing the effect of peptide-HLA-binding ability on the immunogenicity of potential CD8+ and CD4+ T cell epitopes in a large dataset. *Immunol. Res.* (2016) doi:10.1007/s12026-016-8795-9.

19. Calis, J. J. A. *et al.* Properties of MHC Class I Presented Peptides That Enhance Immunogenicity. *PLoS Comput. Biol.* (2013) doi:10.1371/journal.pcbi.1003266.

20. Chowell, D. *et al.* TCR contact residue hydrophobicity is a hallmark of immunogenic CD8+ T cell epitopes. *Proc. Natl. Acad. Sci. U. S. A.* (2015) doi:10.1073/pnas.1500973112.

21. Roh, W. *et al.* Integrated molecular analysis of tumor biopsies on sequential CTLA-4 and PD-1 blockade reveals markers of response and resistance. *Sci. Transl. Med.* (2017) doi:10.1126/scitranslmed.aah3560.

22. Williams, M. J. *et al.* Quantification of subclonal selection in cancer from bulk sequencing data. *Nat. Genet.* (2018) doi:10.1038/s41588-018-0128-6.

23. Zhou, Z. *et al.* TSNAD: An integrated software for cancer somatic mutation and tumour-specific neoantigen detection. *R. Soc. Open Sci.* 4, (2017).

24. Zhou, C. *et al.* pTuneos : p rioritizing tu mor neo antigens from next-generation s equencing data. 1–17 (2019).

25. Tang, Y. *et al.* TruNeo: an integrated pipeline improves personalized true tumor neoantigen identification. *BMC Bioinformatics* 21, 1–16 (2020).

26. Kodysh, J. & Rubinsteyn, A. OpenVax: An open-source computational pipeline for cancer neoantigen prediction. in *Methods in Molecular Biology* (2020). doi:10.1007/978-1-0716-0327-7_10.

27. Wood, M. Neoepiscope improves neoepitope prediction with multi-variant phasing. 23, 2019 (2019).

28. Schenck, R. O., Lakatos, E., Gatenbee, C., Graham, T. A. & Anderson, A. R. A. NeoPredPipe: High-throughput neoantigen prediction and recognition potential pipeline. *BMC Bioinformatics* 20, 1–6 (2019).

29. Bjerregaard, A. M., Nielsen, M., Hadrup, S. R., Szallasi, Z. & Eklund, A. C. MuPeXI: prediction of neo-epitopes from tumor sequencing data. *Cancer Immunol. Immunother.* 66, 1123–1130 (2017).

30. Fotakis, G., Rieder, Di., Haider, M., Trajanoski, Z. & Finotello, F. NeoFuse: Predicting fusion neoantigens from RNA sequencing data. *Bioinformatics* 36, 2260–2261 (2020).

31. Hundal, J. *et al.* PVACtools: A computational toolkit to identify and visualize cancer neoantigens. *Cancer Immunol. Res.* 8, 409–420 (2020).

32. Zhang, J., Mardis, E. R. & Maher, C. A. INTEGRATE-neo: A pipeline for personalized gene fusion neoantigen discovery. *Bioinformatics* (2017) doi:10.1093/bioinformatics/btw674.

33. Zhang, Z. *et al.* ASNEO: Identification of personalized alternative splicing based neoantigens with RNA-seq. *Aging (Albany. NY).* 12, 14633–14648 (2020).

34. Coelho, A. C. M. F. *et al.* NeoANT-HILL: An integrated tool for identification of potential neoantigens. *BMC Med. Genomics* 13, 1–8 (2020).