Background Currently, most neoantigen pipelines often focus on the detection of neoantigens derived from mutations in the coding regions of the genome. However, in some cancer indications, the number of mutations detectable in tumours can be very low (low tumour mutational burden). This limits the number of actionable neoantigens and results in so-called ‘cold’ tumours. In these cases, non-canonical neoantigens resulting from alterations in non-coding regions of the human genome could represent a high potential alternative for treatment.

Indeed, recent research has revealed that previously presumed non-coding regions of the human genome, such as long non-coding RNAs (lncRNAs), can contain translatable small open reading frames (smORFs) generating micropeptides. Some of these micropeptides have already been shown to be involved in cancer development, but these small peptides could also represent a high potential source of non-canonical neoantigens for personalised therapy.

Materials and Methods Here, we present smORFin, a machine learning algorithm specifically trained to identify smORFs in transcripts and to assess their coding potential. While most tools are focused on longer sequences, smORFin is specifically developed to target small ORFs (<303 nucleotides). Furthermore, smORFin also accounts for smORFs with alternative initiation codons, thereby improving its sensitivity for the detection of novel unannotated smORFs.

In addition, the impact of mutations in allegedly non-coding regions of tumour genomes and its influence on the neoantigen repertoire, was evaluated through integration of smORFin in a neoantigen identification pipeline targeting lncRNA-derived mutated epitopes; lncRNAtranscripts.

Results The smORFin model reaches a precision of 0.98 and an accuracy of 0.95 on its testing dataset. Using this new prediction tool, a library of human smORFs was assembled, the so-called smORFome. This library of smORFs, and their associated proteins, was evaluated as a reference for spectrum to peptide matching in mass spectrometry data (MS) analysis. Indeed, the evaluation of seven MS samples revealed and validated the presence of smORFome-related micropeptides and HLA-I associated epitopes originating from smORFs.

Furthermore, it was observed that lncRNA-derived epitopes only represent a minor fraction of the total neoantigen load. Strikingly, when only focusing on tumours with a low neoantigen load, lncRNAtranscripts represented up to 27% of the total neoantigen load. This indicates that for tumours with a low TMB, and therefore with a low neoantigen load, lncRNAtranscripts allows to significantly expand the neoantigen repertoire. Biological in vivo/in vitro validation remains necessary for the detection of novel unannotated lncRNAtranscripts.

Conclusions A novel random forest-based algorithm was developed to address the need for reliable identification of lncRNA-born smORFs. Furthermore, the integration of this prediction algorithm in a neoantigen pipeline allowed the identification of lncRNA-derived neoantigens and marks them as a potential novel source for personalised immunotherapy.

Disclosure Information C. Bogaert: None. L. Van Oudenhove: None. B. Fant: None.