Bioinformatics, 33(19), 2017, 3110–3112
doi: 10.1093/bioinformatics/btx375
Advance Access Publication Date: 12 June 2017
Applications Note

OXFORD

Sequence analysis

CloudNeo: a cloud pipeline for identifying patient-specific tumor neoantigens

Preeti Bais,[1,†] Sandeep Namburi,[1,†] Daniel M. Gatti,[2] Xinyu Zhang[1] and Jeffrey H. Chuang[1,3,*]

[1]The Jackson Laboratory for Genomic Medicine, Farmington, CT 06030, USA, [2]The Jackson Laboratory, Bar Harbor, ME 04609, USA and [3]Department of Genetics and Genome Sciences, University of Connecticut Health, Farmington, CT 06032, USA

*To whom correspondence should be addressed.
[†]The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.
Associate Editor: John Hancock

Received on November 14, 2016; revised on March 17, 2017; editorial decision on June 5, 2017; accepted on June 7, 2017

Abstract

Summary: We present CloudNeo, a cloud-based computational workflow for identifying patient-specific tumor neoantigens from next generation sequencing data. Tumor-specific mutant peptides can be detected by the immune system through their interactions with the human leukocyte antigen complex, and neoantigen presence has recently been shown to correlate with anti T-cell immunity and efficacy of checkpoint inhibitor therapy. However computing capabilities to identify neoantigens from genomic sequencing data are a limiting factor for understanding their role. This challenge has grown as cancer datasets become increasingly abundant, making them cumbersome to store and analyze on local servers. Our cloud-based pipeline provides scalable computation capabilities for neoantigen identification while eliminating the need to invest in local infrastructure for data transfer, storage or compute. The pipeline is a Common Workflow Language (CWL) implementation of human leukocyte antigen (HLA) typing using Polysolver or HLAminer combined with custom scripts for mutant peptide identification and NetMHCpan for neoantigen prediction. We have demonstrated the efficacy of these pipelines on Amazon cloud instances through the Seven Bridges Genomics implementation of the NCI Cancer Genomics Cloud, which provides graphical interfaces for running and editing, infrastructure for workflow sharing and version tracking, and access to TCGA data.

Availability and implementation: The CWL implementation is at: https://github.com/TheJacksonLaboratory/CloudNeo. For users who have obtained licenses for all internal software, integrated versions in CWL and on the Seven Bridges Cancer Genomics Cloud platform (https://cgc.sbgenomics.com/, recommended version) can be obtained by contacting the authors.

Contact: jeff.chuang@jax.org

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Mutations in tumor genomes create specific peptide changes that can be recognized by the immune system and influence sensitivity to immunotherapy (van der Most *et al.*, 1996; van Rooij *et al.*, 2013). The mechanism of action involves binding of native major histocompatibility complex (MHC) class I and II molecules, a.k.a. human leukocyte antigen (HLA) complex I and II molecules, to the novel peptide sequences that result from protein-changing somatic mutations in cancer cells. Cells presenting these neoantigens are recognized as foreign by T-cells, which then selectively destroy them.

With the arrival of new next generation sequencing platforms, it has become possible to interrogate the genomes of patient tumors and computationally predict T-cell reactivity against putative mutation-derived neoantigens (Schumacher *et al.*, 2015) by estimating the binding of MHC class I molecules to each new peptide sequence.

Several bioinformatics tools are routinely used to predict tumor neoantigen—MHC class I binding from sequencing data. For example, HLAMiner (Warren *et al.*, 2012) and Polysolver (Shukla *et al.*, 2015) are software tools that can predict patient-specific HLA classes I and II typing from RNA sequencing data, and netMHCpan (Nielsen *et al.*, 2016) predicts HLA-peptide binding. Prior studies in cancer immunotherapy have successfully used these tools to predict the efficacy of immuno-oncological therapies in a patient-specific manner (Rizvi *et al.*, 2015; Van Allen *et al.*, 2015), demonstrating the importance of making such methods easily available to the general research community. However, the cost of developing and maintaining the bioinformatics infrastructure to perform this type of analysis is substantial. In particular, research groups are generating increasing amounts of custom sequencing data or investigating massive consortium datasets such as The Cancer Genome Atlas (Weinstein *et al.*, 2013), for which data transfer and scalability of computing can be significant obstacles to analysis on local compute clusters. To resolve these problems, we have developed a cloud-based analysis pipeline for tumor neoantigen detection.

2 Description

We developed the CloudNeo pipeline on the Seven Bridges cloud platform as part of the National Cancer Institute's Cancer Genomics Cloud [http://www.cancergenomicscloud.org/] (CGC), which uses Docker containers to execute the tasks in the workflow. Briefly, CloudNeo takes a vcf file (for mutations) and bam file (for HLA typing) as inputs and then outputs HLA binding affinity predictions for all mutated peptides (see Supplementary Fig. S1). A first input to CloudNeo is a list of non-synonymous mutations in vcf file format. There are multiple somatic mutation calling pipelines that can be used to generate and filter this vcf file (Alioto *et al.*, 2015), including several which are available through the CGC. The genomic variants are translated into amino acid changes using the VEP tool (McLaren *et al.*, 2010) and a custom R script that we have created called Protein_Translator. The output of the custom tool is a list of N-amino-acid-long peptide sequences in a fasta format, such that the single peptide change is in the middle of the N-mer. In parallel, Protein_Translator generates another fasta file for the homologous N-mers with no peptide mutation. Users have options to calculate the HLA types using either HLAminer or Polysolver. Six HLA types are predicted, namely the top two predictions for each of HLA-A, HLA-B and HLA-C. The final step in the pipeline is the NetMHCpan tool, which uses the HLA types and the N-mer mutant peptide sequences to calculate the binding affinities for potential neoantigens. Affinities between the two HLA-A, two HLA-B, and two HLA-C molecules and each of the ([N/2]+1)mer peptide subsequences within the N-mers are computed. The output of the pipeline is a list of peptide subsequences along with the MHC binding affinity scores for each of the six HLA types. Similar results are generated for the homologous unmutated peptide sequences as a comparison.

To test this pipeline, we analyzed 23 melanoma tumor samples (Hugo *et al.*, 2016) as described earlier using both the HLAminer and Polysolver versions of the pipeline. We then predicted neoantigens based on criteria of strong mutant-MHC binding affinity (NetMHCpan score < 500), non-zeroexpression of the transcript containing the mutation, and lack of strong affinity between the non-mutated sequence and the MHC (NetMHCpan score for the non-mutant sequence ≥ 500). For each sample we merged the set of neoepitopes predicted across the six HLA types. The neoepitope load ranged from 0 to 1244 with an average of 107.89 using the HLAminer version of the pipeline. For the Polysolver version of the pipeline, the same filtering criteria were used and the neoepitope load was from 0 to 1417 with an average load of 133.53. The differences in the two pipeline results were due to differing HLA type predictions by Polysolver and HLAminer. 16 HLAtype predictions by the tools overlapped with each other, and there were 102 unique HLA predictions from Polysolver and 122 unique predictions from HLAminer. While our HLA type predictions were based on RNA-seq data, CloudNeo can also use DNA data as inputs for HLA calling. The average wall time required to run the pipeline for a given tumor on CGC was 8 h and 2 min for the HLAminer version and 7 h and 25 min for the Polysolver version (see Supplementary Material 'Pipeline Performance').

3 Discussion

Other recent methods, such as (Hundal *et al.*, 2016), are similar to CloudNeo in providing a computational pipeline for neoantigen prediction. However, to our knowledge CloudNeo is the only such pipeline that has been developed for cloud computing. This allows users to realize advantages of cloud analysis, including massive computing scalability and access to large datasets on the CGC such as TCGA, as these can be reached without downloading to a local server. This cloud approach also makes CloudNeo easy to match to time and budget restrictions on demand, providing a flexible computational approach for the research community. A version of the CloudNeo pipeline is openly available at the Github site as a Common Workflow Language (CWL) implementation that can be run using Rabix (Kaushik *et al.*, 2016), allowing for running on systems including AWS, Google Compute Engine and Azure. Licenses for academically licensed software (HLAminer and NetMHCpan) must be obtained by users, but simple instructions to do so are provided at the Github site. Users with licenses can also contact the authors to request a version with all software integrated. Full versions are available either in CWL or as a workflow on the Seven Bridges implementation of the CGC. The CGC version is recommended, as this provides additional functionality including graphical interfaces for running and editing, simple workflow sharing and version tracking, improved calling of multiple cloud instances, and access to TCGA data. Full details and docs are at https://github.com/TheJacksonLaboratory/CloudNeo.

Acknowledgement

We thank G. Kaushik for assistance with the CGC platform.

Funding

J.H.C. was supported by the National Cancer Institute of the National Institutes of Health under awards [R21CA191848] and supplement [R21CA191848-01A1S1]. Research was also partially supported by the National Cancer Institute under award [P30CA034196]. D.M.G. was supported by National Institute of General Medical Sciences under award [R01 GM07068308].

Conflict of Interest: none declared.

References

Alioto,T.S. *et al.* (2015) A comprehensive assessment of somatic mutation detection in cancer using whole-genome sequencing. *Nat. Commun.*, **6**, 10001.

Hugo,W. *et al.* (2016) Genomic and transcriptomic features of response to anti-PD-1 therapy in metastatic melanoma. *Cell*, **165**, 35–44.

Hundal,J. *et al.* (2016) pVAC-Seq: a genome-guided in silico approach to identifying tumor neoantigens. *Genome Med.*, **8**, 11.

Kaushik,G. *et al.* (2016) Rabix: an open-source workflow executor supporting recomputability and interoperability of workflow descriptions. *Pac. Symp. Biocomput.*, **22**, 154–165.

McLaren,W. *et al.* (2010) Deriving the consequences of genomic variants with the Ensembl API and SNP effect predictor. *Bioinformatics*, **26**, 2069–2070.

Nielsen,M. *et al.* (2016) NetMHCpan-3.0; improved prediction of binding to MHC class I molecules integrating information from multiple receptor and peptide length datasets. *Genome Med.*, **8**, 33.

Rizvi,N.A. *et al.* (2015) Cancer immunology. Mutational landscape determines sensitivity to PD-1 blockade in non-small cell lung cancer. *Science*, **348**, 124–128.

Schumacher,T.N., and Schreiber,R.D. (2015) Neoantigens in cancer immunotherapy. *Science*, **348**, 69–74.

Shukla,S.A. *et al.* (2015) Comprehensive analysis of cancer-associated somatic mutations in class I HLA genes. *Nat Biotechnol.*, **33**, 1152–1158.

Van Allen,E.M. *et al.* (2015) Genomic correlates of response to CTLA-4 blockade in metastatic melanoma. *Science*, **350**, 207–211.

van der Most,R.G. *et al.* (1996) Analysis of cytotoxic T cell responses to dominant and subdominant epitopes during acute and chronic lymphocytic choriomeningitis virus infection. *J. Immunol.*, **157**, 5543–5554.

van Rooij,N. *et al.* (2013) Tumor exome analysis reveals neoantigen-specific T-cell reactivity in an ipilimumab-responsive melanoma. *J. Clin. Oncol.*, **31**, e439–e442.

Warren,R.L. *et al.* (2012) Derivation of HLA types from shotgun sequence datasets. *Genome Med*, **4**, 95.

Weinstein,J.N. *et al.* (2013) The Cancer Genome Atlas Pan-Cancer analysis project. *Nat. Genet.*, **45**, 1113–1120.