Establishing HLA Typing at CUBI and Benchmarking Tools

Clemens Messerschmidt
Manuel Holtgrewe

13.10.2015

Abstract

The human leucocyte antigen (HLA) gene cluster plays an important role in adaptive immunity. HLA class I genes encode for proteins on every nucleated cell that present peptides from inside the cell.

Introduction

In clinical use, HLA typing often requires the use of special techniques for the enrichment of the HLA gene cluster. As this requires a commitment of resources and time, it would be desirable to gather this information from already existing information, e.g. whole exome sequences (WES).

Tools

- Szolek et al. (2014) presented OptiType.
- Bai et al. (2014) PHLAT
- Li (unpublished) BWAkit
- Wittig et al. (2015) HLAssign

Table 1: Tool applicability based on type of NGS data

input	hlassign	optitype	bwakit
HLA panel	✓	✓	
WES	?	✓	?
WGS		✓(yara)	✓

Datasets

To benchmark the presented tools, multiple dataset were gathered, covering all 3 types once. For the benchmark, we rely on NGS datasets with known HLA types, i.e. the individuals were additionally HLA typed using conventional methods like Sanger sequencing.

HLA Panel

Wittig et al. (2015) presented HLAssign, a tool to predict HLA types together with a custom panel to enrich for the HLA gene cluster. The supplementary material for the publication includes sequencing datasets from

different sequencing system, i.e. Hiseq, Miseq.For our benchmark, we analyzed 311 Hiseq panel datasets, which were sequenced paired-end, 100 bp.

CAUTION: The datasets only contain reads that have a perfect match seed of 30 against the database. Even though the reads are paired, the fastq files are *not*. This can introduce problems with aligners like Bowtie2, which is used by PHLAT. In this case, reads were treated as single-end. HLAssign, OptiType and bwakit identify reads pairs by their name.

SAMPLE SWAPS In a pilot study we were able to discover sample swaps in the data, i.e. the FASTQ files were mixed up for ~ 10 % of all samples. This leads to a decrease in performace, as correctly predicted HLA types will be classified as false as they do not match the reference.

WES

Sample set 1: 11 datasets. The exome-seq datasets are from the 1000 Genomes Project (ftp-trace.ncbi.nih.gov/1000genomes/ftp/data/) and were used by Liu et al. (2013) (Supplementary table 2 http://nar.oxfordjournals.org/content/suppl/2013/05/11/gkt481.DC1/nar-00605-met-k-2013-File004_update.pdf). Samples were prepared using Agilent SureSelect V2 and provide 96-98 % exome coverage (>= 10X), run on Illumina Genome Analyzer IIx.

Sample set 2: 182 datasets. First used by Major et al. (2013) and later by Szolek et al. (2014), this dataset contains 182 exomes, also from the 1000 Genomes Project.

From Major et al. (2013):

> There are 270 Coriell IDs in the HapMap database, but at the end of QC check there were only 31 Coriell IDs (41 samples) for whole-genome experiments and 131 Coriell IDs (182 samples) for whole-exome experiments left – many Coriell cell lines were sequenced more than once. Figure 1 explains the details of this filtering process.

WGS

20 samples of low-coverage WGS data (~ 30X) of the HapMap Project, also used by Warren et al. (2012), were used for benchmarking on WGS. This data are 2 x 102 bp reads, sequenced on Illumina HiSeq 2000.

Benchmarks

Performance was evaluated using measures from classification theory with the problem treated as multi-label, multi-class. Each item, i.e. a sample or patient can be assigned with labels signalling one HLA allele each by each method. Duplicates are allowed for homozygous predictions. Lables are than compared to the reference HLA type.

TP_i is number of alleles classified as $type_i$ in reference and prediction

FP_i is number of alleles classified as $type_i$ in the prediction but not in the reference

FN_i is number of alleles classified as $type_i$ in the reference but not in the prediction

The definitions for recall (ρ), precision (π) and F-measure are taken from Özgür, Özgür, and Güngör (2005):

$\rho_i = \frac{TP_i}{TP_i + FN_i}$

$\pi_i = \frac{TP_i}{TP_i + FP_i}$

The F-measure is a measure for a test's accuracy. It considers both precision and recalls, traditionally their harmonic mean (F1-measure) and will take values between 0 and 1. The overall F-measure over all categories (HLA types) can be computed in two fashions. Here, we opt for the micro-averaged F-measure, which gives equal weight to each sample. It tends to be dominated by the classifiers (i.e. HLA typing method) performance on common categories (e.g. A*02:01) rather than giving equal weight to categories. The latter would mean that the performance on rare HLA types would be equally important as on frequent ones.

The micro-averaged F-Measure for N categories is computed as follows:

$\rho = \frac{TP}{TP+FN} = \frac{\sum_i^N TP_i}{\sum_i^N TP_i+FN_i}$

$\pi = \frac{TP}{TP+FP} = \frac{\sum_i^N TP_i}{\sum_i^N TP_i+FP_i}$

$F_1 = \frac{2\rho\pi}{\rho+\pi}$

Performance on Public Datasets

The following section will list the performance results for the presented measures on each of the datasets. If a method is missing from in the table no results were generated as of the writing of this report. If necessary, specifics for the results of a given method on a dataset will be reported, if necessary for the correct interpretation of performance.

HLA Panel

	Recall	Precision	Misclassification rate	F_micro measure
optitype	0.89	0.89	0.11	0.89
bwakit	0.84	0.84	0.16	0.84
hlassign	0.88	0.89	0.11	0.88
phlat	0.70	0.73	0.27	0.72

REMARK: For OptiType, only 306 out of 311 panels were typed because of missing results (unsolved memory issues).

WES: Sample set 1 (11 datasets)

	Recall	Precision	Misclassification rate	F_micro measure
optitype	1.00	1.00	0.00	1.00
bwakit	0.98	0.98	0.02	0.98
hlassign	0.83	0.83	0.17	0.83

WES: Sample set 2 (182 datasets)

	Recall	Precision	Misclassification rate	F_micro measure
optitype	0.98	0.98	0.02	0.98
bwakit	0.47	0.55	0.45	0.51

WGS (20 datasets)

	Recall	Precision	Misclassification rate	F_micro measure
bwakit	0.08	0.28	0.72	0.13
phlat	0.70	0.70	0.30	0.70

Internal Results

References

Bai, Yu, Min Ni, Blerta Cooper, Yi Wei, and Wen Fury. 2014. "Inference of High Resolution HLA Types Using Genome-Wide RNA or DNA Sequencing Reads." *BMC Genomics* 15 (1): 325. doi:10.1186/1471-2164-15-325.

Liu, Chang, Xiao Yang, Brian Duffy, Thalachallour Mohanakumar, Robi D. Mitra, Michael C. Zody, and John D. Pfeifer. 2013. "ATHLATES: Accurate Typing of Human Leukocyte Antigen Through Exome Sequencing." *Nucleic Acids Research* 41 (14): e142–42. doi:10.1093/nar/gkt481.

Major, Endre, Krisztina Rigó, Tim Hague, Attila Bérces, and Szilveszter Juhos. 2013. "HLA Typing from 1000 Genomes Whole Genome and Whole Exome Illumina Data." *PLoS ONE* 8 (11): e78410. doi:10.1371/journal.pone.0078410.

Özgür, Arzucan, Levent Özgür, and Tunga Güngör. 2005. "Text Categorization with Class-Based and Corpus-Based Keyword Selection." In *Computer and Information Sciences - ISCIS 2005*, edited by pInar Yolum, Tunga Güngör, Fikret Gürgen, and Can Özturan, 606–15. Lecture Notes in Computer Science 3733. Springer Berlin Heidelberg. http://link.springer.com/chapter/10.1007/11569596_63.

Szolek, András, Benjamin Schubert, Christopher Mohr, Marc Sturm, Magdalena Feldhahn, and Oliver Kohlbacher. 2014. "OptiType: Precision HLA Typing from Next-Generation Sequencing Data." *Bioinformatics* 30 (23): 3310–16. doi:10.1093/bioinformatics/btu548.

Warren, René L., Gina Choe, Douglas J. Freeman, Mauro Castellarin, Sarah Munro, Richard Moore, and Robert A. Holt. 2012. "Derivation of HLA Types from Shotgun Sequence Datasets." *Genome Medicine* 4 (12): 95. doi:10.1186/gm396.

Wittig, Michael, Jarl A. Anmarkrud, Jan C. Kässens, Simon Koch, Michael Forster, Eva Ellinghaus, Johannes R. Hov, et al. 2015. "Development of a High-Resolution NGS-Based HLA-Typing and Analysis Pipeline." *Nucleic Acids Research* 43 (11): e70–70. doi:10.1093/nar/gkv184.