BIOINFORMATICS ORIGINAL PAPER

Vol. 30 no. 23 2014, pages 3310–3316
doi:10.1093/bioinformatics/btu548

Sequence analysis

Advance Access publication August 20, 2014

OptiType: precision HLA typing from next-generation sequencing data

András Szolek[1,*,†], Benjamin Schubert[1,†], Christopher Mohr[1,†], Marc Sturm[2], Magdalena Feldhahn[3] and Oliver Kohlbacher[1]

[1]Applied Bioinformatics, Center for Bioinformatics, Quantitative Biology Center, and Department of Computer Science, University of Tübingen, [2]Institute of Medical Genetics and Applied Genomics, University of Tübingen, and [3]CeGaT GmbH, 72076 Tübingen, Germany

Associate Editor: Michael Brudno

ABSTRACT

Motivation: The human leukocyte antigen (HLA) gene cluster plays a crucial role in adaptive immunity and is thus relevant in many biomedical applications. While next-generation sequencing data are often available for a patient, deducing the HLA genotype is difficult because of substantial sequence similarity within the cluster and exceptionally high variability of the loci. Established approaches, therefore, rely on specific HLA enrichment and sequencing techniques, coming at an additional cost and extra turnaround time.

Result: We present OptiType, a novel HLA genotyping algorithm based on integer linear programming, capable of producing accurate predictions from NGS data not specifically enriched for the HLA cluster. We also present a comprehensive benchmark dataset consisting of RNA, exome and whole-genome sequencing data. OptiType significantly outperformed previously published *in silico* approaches with an overall accuracy of 97% enabling its use in a broad range of applications.

Contact: szolek@informatik.uni-tuebingen.de

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

Received on January 30, 2014; revised on August 1, 2014; accepted on August 10, 2014

1 INTRODUCTION

The human leukocyte antigen (HLA) cluster located on chromosome 6 is one of the most polymorphic regions of the human genome and encodes for several genes involved in functions of the immune system, including HLA classes I and II. Both HLA classes comprise three major loci (HLA-I: A, B, C; HLA-II: DP, DQ, DR), which are co-dominantly expressed. HLA-I/II molecules present intracellular and extracellular peptides, respectively, and interact with other immune cells to induce an adaptive immune response. Thus, HLA-I/II molecules play an important role in many medical areas, such as vaccinology (Haralambieva *et al*., 2013; Ovsyannikova and Poland, 2011), regenerative and transplantation medicine (Bradley, 1991; Opelz *et al*., 1999) and autoimmune diseases (Thorsby and Lie, 2005; Undlien *et al*., 2001).

Over 7300 different HLA-I and 2200 HLA-II alleles are known to date [IMGT/HLA Release 3.14.0, July 2013 (Robinson *et al*., 2013)]. In addition to this vast allelic variation, HLA alleles display a high degree of sequence similarity even across different loci, which drastically increases the complexity of uniquely identifying a genotype using short-read sequencing techniques. Established HLA typing approaches make use of labor-intensive and time-consuming probing techniques, such as sequence-specific oligonucleotide probe hybridization, PCR amplification with sequence specific primers or serotyping techniques, which often lead to ambiguous genotyping results (Liu *et al*., 2013). HLA typing can be done with different degrees of resolution, with two-digit and four-digit types distinguishing HLA allele families and distinct HLA protein sequences, respectively. In 2009, Gabriel *et al*. (2009) and Bentley *et al*. (2009) demonstrated the use of targeted next-generation sequencing (NGS) for HLA typing to overcome the problems mentioned above. Several new protocols have recently been established based on NGS technologies (Lank *et al*., 2010, 2012; Moonsamy *et al*., 2013; Shiina *et al*., 2012). These methods are still accompanied by labor-intensive preparations and remain time consuming. More recently, Danzer *et al*. (2013) published an automated protocol based on GS 454 Junior sequencing allowing a high-resolution typing with a turnaround time of 2 days. To reduce time and cost expense even further, *in silico* approaches have been developed. In 2011, Erlich *et al*. published an approach based on posterior probability of allele pairs and integrated it into a 454 GS FLX Titanium sequencing pipeline (Erlich *et al*., 2011).

Common to the above approaches is the dedicated generation of NGS data for the sole purpose of HLA typing. Routine sequencing of patient exomes or whole genomes has been established in many larger clinical centers, and it should be possible to determine the HLA type from these data through purely computational means. Using existing data can save both money and time; however, because of the high variability of the HLA loci, the typical read mapping and variant calling-based analysis of NGS data is not suitable to determine the HLA genotype. Warren *et al*. (2012) proposed an algorithm (HLAminer) based on allele-specific scoring for whole genome, exome and

*To whom correspondence should be addressed.
†The authors wish it to be known that, in their opinion, the first three authors should be regarded as Joint First Authors.

transcriptome sequencing to solve this problem. It assembles reads *de novo* and aligns the resulting contigs against an HLA reference database. A score for each HLA allele is then calculated based on properties of the aligned contigs, and the highest scoring alleles for each locus are selected. In 2013, Boegel *et al.* suggested a greedy algorithm (seq2HLA) based on read count maximization for RNA-Seq data (Boegel *et al.*, 2013). After read mapping, in an initial round, the algorithm determines the allele with the highest number of mapped reads for each locus individually. After discarding the selected alleles and already assigned reads, second alleles are selected accordingly. ATHLATES, published by Liu *et al.* (2013), uses an HLA reference database to filter for relevant reads, which are used for contig construction. The reference HLA sequences are decomposed into exons, and the best mapping contig for each exon is determined. For one HLA locus at a time, each allele is scored based on its overall Hamming distance to all aligned exons. A candidate allele list is generated by applying different filtering criteria. Using this candidate list, the most probable HLA allele pairs per locus are determined based on the minimal Hamming distance to the variable positions of each exon. One of the most recent approaches, published by Kim *et al.*, uses a tree-based top-down greedy algorithm (HLAforest) to predict the HLA genotypes based on RNA-Seq data (Kim and Pourmand, 2013). The algorithm generates an HLA alignment tree for each read based on the mapping results against an HLA reference database whose leaf nodes indicate four-digit alleles and inner nodes represent allele families to which the read could be mapped. Then, alignment probabilities and node weights are distributed in the trees. Based on the sum of weights of all trees, the highest scoring allele family is selected. After reweighting the nodes based on the selected allele family, the four-digit HLA allele is selected similarly. A different approach by Major *et al.* (2013) applies various filtering criteria and optimizes coverage depth and base coverage. The first filter criterion enforces certain sequence coverage of exons 2 and 3 of the HLA alleles, as these exons are the most polymorphic regions that also encode for the binding core of the HLA molecule. Additionally, reads are filtered based on mismatches and alignment orientation of the paired reads. Subsequently, alleles are sorted and filtered based on their coverage depth and sequence coverage of their alignment. Finally, allele pairs are selected such that coverage depth and sequence coverage are optimized.

Yet, these methods do not yield sufficiently accurate predictions, especially in terms of clinical usability. Boegel *et al.*'s approach is capable of only two-digit genotyping; Kim *et al.* and Warren *et al.* could achieve only 85–90% correctly predicted four-digit HLA genotypes on RNA-Seq data, and for short-read RNA-Seq and whole-genome sequencing (WGS) data, the accuracy was even lower. Major *et al.* could accomplish an accuracy of 94% on exome sequencing samples that fulfilled all their filtering criteria, but out of the 217 samples they have considered, only 161 could be fully typed.

A possible cause for low typing accuracy in the aforementioned approaches might be the independent consideration of each locus. Sequence homology between loci can lead to ambiguous read alignments where reads map to alleles of multiple loci equally well. Another reason for suboptimal performance could be explained by disregarding intronic information in exome or WGS data. However, including intronic regions is not trivial, as the intron sequences of the majority of HLA alleles are unknown. In fact, 94.6% of HLA sequences contained in the IMGT database lack parts of their exonic or intronic sequences.

To tackle these issues, we developed a new method named OptiType, which considers all major and minor HLA-I loci simultaneously. OptiType works on the premise that the correct genotype explains the source of more reads than any other genotype, where an allele is said to explain a read if the read is aligned to it with no more mismatches than to any other allele. Hence, the method finds an allele combination, which maximizes the number of reads they explain. The method consists of three key steps (Fig. 1). First, reads are mapped against a carefully constructed HLA allele reference (Fig. 1A). Because only exon 2 and 3 subsequences are available for all alleles, these regions are considered during read mapping so that no allele is disadvantaged because of incomplete sequence information. Additionally, for exome and genome sequencing data, we included flanking intronic regions and developed a method to impute missing sequence data based on phylogenetic information. Second, from the initial read mapping results, a binary matrix is generated indicating which alleles a specific read could be aligned to with the least number of mismatches (Fig. 1B). Finally, based on this matrix, a special case of the set cover problem (Karp, 1972) is formulated as an integer linear program (ILP) that selects up to two alleles for each locus simultaneously, maximizing the number of mapped reads that can be explained by the predicted genotype (Fig. 1C). Besides the major HLA-I alleles A, B and C, minor alleles G, H and J are considered during optimization, as long subsequences of these minor loci show high similarity with major loci, occasionally causing ambiguous read alignments.

Furthermore, we present a comparison of OptiType against previously published methods on RNA sequencing, exome sequencing and WGS datasets, and evaluate its performance in a clinical setting on in-house lymphoblastic leukemia patient data. Additionally, we investigate the influence of coverage depth on prediction performance using a sample specifically enriched for the HLA region and simulated sequencing data. Finally, we summarize and discuss the results, and give an outlook on the possible applications of OptiType.

2 METHODS

2.1 Reference construction from phylogenetic information

HLA nucleotide coding DNA sequences (CDS), genomic nucleotide sequences and feature annotation for all HLA-I alleles have been obtained from the IMGT/HLA database [Release 3.14.0, July 2013 (Robinson *et al.*, 2013)] for read mapping reference sequence construction.

Reference sequences for RNA-Seq data were built by concatenating exon 2 and 3 coding sequences, which were available for all alleles in the database. Mapping DNA sequencing data, however, required taking the intron sequences flanking exons 2 and 3 into consideration as well, despite the fact that they were not available for the majority of HLA alleles. To this end, OptiType uses reconstructed intron sequences for partially sequenced alleles. We impute the missing sequence data by replacing it

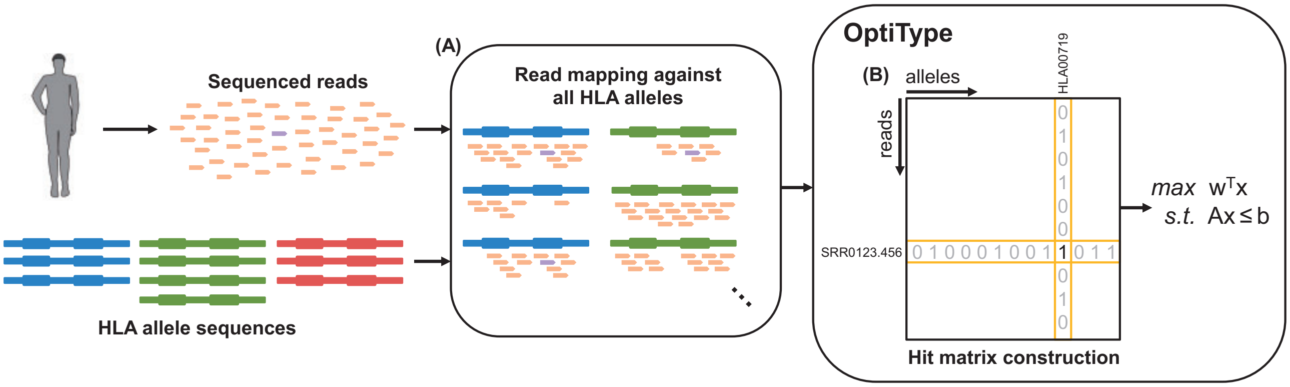

Fig. 1. OptiType's four-digit HLA typing pipeline. Reference libraries for genomic and CDS are generated by extracting exons 2 and 3 from each known HLA-I allele. For genomic sequences, flanking intronic regions are also extracted. If some of these regions are missing, phylogenetic information is used to reconstruct the missing segments from the closest relative HLA-I allele. NGS reads are mapped against the so-constructed HLA allele reference (**A**). From the mapping result a binary hit matrix $\boldsymbol{C}^{\boldsymbol{R}\times\boldsymbol{A}}$ is constructed for all reads $\boldsymbol{r} \in \boldsymbol{R}$ mapping to at least one allele $\boldsymbol{a} \in \boldsymbol{A}$ of the reference with $\boldsymbol{C}_{\boldsymbol{r},\boldsymbol{a}} = \mathbf{1}$ if read $\boldsymbol{r}$ could be mapped to allele $\boldsymbol{a}$; otherwise, $\boldsymbol{C}_{\boldsymbol{r},\boldsymbol{a}} = \mathbf{0}$ (**B**). Based on this hit matrix, an ILP is formulated that optimizes the number of explainable reads by selecting up to two alleles (columns of the hit matrix) for each HLA-I locus (**C**). The selected alleles represent the most probable genotype

with its closest neighbor with respect to sequence similarity from among the complete allele sequences. The procedure, therefore, attempts to reconstruct partial allele sequences based on their closest phylogenetic relatives with known intron sequences, using the fact that intronic variability in HLA is characterized by highly systematic mutations reflecting the ancestral lineage of the alleles (Blasczyk *et al.*, 1997).

Sequence similarity values were obtained from full distance matrices computed with Clustal Omega 1.2.0 (Sievers *et al.*, 2011). Partial alleles were partitioned into sets according to their exons with known sequences to ensure sequence similarity calculation using maximal sequence information available. All complete alleles were added to every set, followed by computing distance matrices between set members' concatenated exon sequences. Partial alleles have shown to have 1.66 (± 1.04) nearest neighbors with unique intron sequences on average. Sequences of partial alleles with multiple nearest neighbors were reconstructed with each of the nearest neighbors, resulting in 10 779 reconstructed sequences for 6489 partial alleles.

The quality of sequence reconstruction was validated in a leave-one-out fashion. Introns 1, 2 and 3 for each of the fully sequenced alleles were discarded and reconstructed using the remaining alleles, considering only exon 2 and 3 sequences for nearest neighbor identification. The reconstructed intron sequences were compared with their original counterparts and showed a sequence similarity of 99.89% (± 0.43%), corresponding to an average 1.2 edit distance error on the three introns combined. For comparison, sequence similarity between introns of the same loci was found to be 97.36% (± 2.15%), corresponding to 29 nt differences on average. The used reference sequences can be found in the supplementary material (S12).

2.2 Read mapping

Read mapping was performed by RazerS3 3.1, which is part of the open source C + + library project SeqAn (Döring *et al.*, 2008; Weese *et al.*, 2012). RNA-Seq data were mapped against the nucleotide CDS reference library; exome sequencing and WGS data were mapped against the genomic nucleotide reference library. All best alignments for every read with a sequence identity of at least 97% were taken into account (*--percent-identity 97 --distance-range 0*). The maximum number of reported best matches (*--max-hits*) was set to infinity. All read matches fulfilling those criteria were reported in SAM file format.

2.3 Hit matrix construction

A binary hit matrix $C^{R\times L}$ was constructed for all reads $r \in R$ mapping to at least one allele $a \in L$ of the reference with $C_{r,a} = 1$ if read r mapped to allele a; otherwise, $C_{r,a} = 0$. Columns of rare alleles whose four-digit subtypes were not reported in allelefrequencies.net (Gonzalez-Galarza *et al.*, 2011) or dbMHC (NCBI Resource Coordinators, 2013) at all were removed from the matrix. To reduce the size of the matrix, reads with the same mapping profile (i.e. identical rows) were combined and reflected in a row weight vector o_r. Columns corresponding to alleles that were *covered* by other alleles were also dropped, where allele b covering a is defined as $(C_{:,a}^T C_{:b} = |C_{:,a}|) \wedge (|C_{:,a}| < |C_{:,b}|)$ with $a, b \in L$ and reflects that all reads mapping to a also map to b, with b having additional mapping reads. The remaining rows and columns were used for model construction.

For paired-end read data, the full hit matrices were constructed for both read pairs individually. Rows corresponding to matching pairs of reads were combined with a point-wise *AND* operation, and all reads without mapping mate reads were discarded.

2.4 Optimization problem

We base our approach on the assumption that the correct HLA genotype explains the highest number of mapped reads. Therefore, we are searching for the best HLA allele combination of up to six major and six minor HLA-I alleles, which maximizes the number of reads potentially originating from this selection, under the biological constraints that at least one and at most two alleles are selected per locus [constraints (1) and (2)]. This type of problem can be conveniently formulated as an ILP. In contrast to sufficiently complex probabilistic models capturing uncertanties in the data, the conditional joint distribution of alleles and further considerations, an ILP formulation can guarantee an optimal solution at the expense of modeling uncertainty. Solving an ILP finds an optimal solution to a linear objective function subject to linear constraints and integrality requirements on the variables (Schrijver, 1998). In the following, we state the problem of finding the best HLA allele combination as an ILP.

For each allele $a \in L$, a binary variable x_a was introduced with $x_a = 1$, indicating that a is an element of the solution set $S \subseteq L$. Additionally, another binary variable y_r for each read $r \in R$ was assigned to represent if read r is explained by one of the selected alleles $a \in S$. For this effect, the binary hit matrix $C^{R\times L}$ was used to construct constraints forcing y_r to take on $y_r = 1$ if read r could be explained by the current solution set [constraint (3)]. The resulting ILP, maximizing the number of explained reads, could then be defined as follows:

Objective		
$\boldsymbol{max_{S\subseteq L}}$	$\sum_{r\in R} \boldsymbol{o_r \cdot y_r}$	Maximize the number of explained reads
Subject to		
(1) $\forall \boldsymbol{X} \in \{\boldsymbol{A,B,C,G,H,J}\}$	$\sum_{a\in X} x_a \le \tau^{max}$	Ensures that each locus is represented by at most τ^{max} alleles
(2) $\forall \boldsymbol{X} \in \{\boldsymbol{A,B,C,G,H,J}\}$	$\sum_{a\in X} x_a \ge \tau^{min}$	Ensures that each locus is represented by at least τ^{min} alleles
(3) $\forall \boldsymbol{r} \in \boldsymbol{R}$:	$\sum_{a\in L} x_a \cdot C_{r,a} \ge y_r$	Ensures that $y_r = 1$ only if read r originates from one of the selected alleles

with o_r being the number of previously collapsed rows with the same mapping profile, and A, B, C, G, H and J the sets of alleles for the major loci HLA-A, B, C and the minor loci HLA-G, H, J. τ^{max} and τ^{min} represent the maximum ($\tau^{max} = 2$) and minimum ($\tau^{min} = 1$) number of selected alleles per locus reflecting the diploid nature of the human genome and allowing homozygosity in the genotype.

As this formulation favors heterozygous allele combinations because of spurious hits (e.g. from sequencing errors), the objective function was extended with a regularization term accounting for homozygosity

$$g(r) = \begin{cases} \sum_{a\in L} x_a - n^{loci}, & y_r = 1 \\ 0, & otherwise \end{cases},$$

where n^{loci} describes the number of loci (here $n^{loci} = 6$). The regularization term is weighted by a constant β representing the proportion of reads that have to be additionally explained by an allele combination to choose a heterozygous solution over a homozygous one. The regularization term can be directly translated into an ILP formulation, for which an additional integer variable g_r for each read $r \in R$ and constraints (4) to (6) had to be introduced. Additionally, a small penalization term γ was added to prioritize alleles with full sequence information over reconstructed alleles contributing to equally good solutions. The ILP is thus defined as:

Objective		
$\boldsymbol{max_{S\subseteq L}}$	$\sum_{r\in R} \boldsymbol{o_r \cdot (y_r - \beta \cdot g_r)} - \sum_{a\in L^R} \boldsymbol{\gamma \cdot x_a}$	Maximize the number of explained reads
Subject to		
(1) $\forall \boldsymbol{X} \in \{\boldsymbol{A,B,C,G,H,J}\}$	$\sum_{a\in X} x_a \le \tau^{max}$	Ensures that each locus is represented by at most τ^{max} alleles
(2) $\forall \boldsymbol{X} \in \{\boldsymbol{A,B,C,G,H,J}\}$	$\sum_{a\in X} x_a \ge \tau^{min}$	Ensures that each locus is represented by at least τ^{min} alleles
(3) $\forall \boldsymbol{r} \in \boldsymbol{R}$:	$\sum_{a\in L} x_a \cdot C_{r,a} \ge y_r$	Ensures that $y_r = 1$ only if read r originates from one of the selected alleles
(4) $\forall \boldsymbol{r} \in \boldsymbol{R}$:	$g_r \le \tau^{loci} \cdot y_r$	Limits g_r to 0 if read r can not be explained by the current solution
(5) $\forall \boldsymbol{r} \in \boldsymbol{R}$:	$g_r \le \sum_{a\in L} x_a - n^{loci}$	Limits g_r to the number of heterozygous loci
(6) $\forall \boldsymbol{r} \in \boldsymbol{R}$:	$g_r \ge (\sum_{a\in L} x_a - n^{loci}) - n^{loci} \cdot (1 - y_r)$	Enforces g_r to take on one of the limit values stipulated by (4) and (5) depending on y_r

where $L^R \subseteq L$ is the set of reconstructed alleles, and γ a small constant factor penalizing the use of reconstructed alleles ($\gamma = 0.01$).

Evaluation of different values for β was carried out by performing a nested 5-fold cross-validation stratified for evenly distributed heterozygous and homozygous cases on 253 runs of the 1000 Genomes Project. Accuracy has been analyzed in terms of percentage of correctly typed alleles. Different values in the range from 0.000 to 0.050 with a step size of 0.001 have been tested for β, showing best performance with $\beta = 0.009$.

2.5 NGS datasets

To permit comparison with previously published approaches, the same publicly available NGS datasets have been used, for which PCR-verified HLA genotypes were available.

Sixteen samples of a colorectal cancer RNA-Seq study [SRP010181 (Warren *et al.*, 2012)] and 20 samples of low-coverage WGS data of the HapMap Project (The International HapMap Consortium, 2005) used by Warren *et al.* and Kim *et al.* have been obtained from the NCBI Sequence Read Archive (NCBI Resource Coordinators, 2013). Both datasets contained 2 × 100 to 2 × 102 bp long reads produced by Illumina HiSeq 2000.

For comparison with Boegel *et al.* and Kim *et al.*, 37 nt long paired-end RNA-Seq reads generated by Illumina Genome Analyzer II originating from 50 lymphoblastic cell line samples of CEU HapMap individuals [ERA002336 (Montgomery *et al.*, 2010)] have been obtained from the European Nucleotide Archive (Leinonen *et al.*, 2011).

Furthermore, OptiType was validated on two datasets, which have been used by Major *et al.* They benchmarked their method on a HapMap WGS dataset consisting of 41 runs, partly overlapping with those used by Warren *et al.*, and an exome sequencing dataset consisting of 182 runs of 1000 Genomes Project samples. Only samples for which Major *et al.* predicted full genotypes were considered, resulting in 12 HapMap WGS and 161 1000 Genomes Project datasets. We expanded this benchmark set by including additional data from the 1000 Genomes Project (The 1000 Genomes Project Consortium, 2012) consisting of all 253 Illumina HiSeq 2000 and Genome Analyzer II exome sequencing runs.

To compare OptiType with ATHLATES, we used the publicly available subset of their benchmark dataset consisting of 11 samples from the 1000 Genomes Project (Liu *et al.*, 2013). To assess the method on clinical samples, we included an in-house generated dataset, which cannot be made publicly available because of privacy concerns. The dataset consisted of 10 exome sequenced acute lymphoblastic leukemia (ALL) patients with experimentally determined HLA types. Exome enrichment of the samples was performed using the SureSelect Human All Exon V2 kit (Agilent Technologies; Böblingen, Germany) or the SeqCap EZ Human Exome Library V2 kit. The resulting libraries were sequenced on an Illumina Genome Analyzer IIx using paired-end mode with 76 bp per read. On average, 94 million reads were produced per sample, resulting in an average coverage of 90× on the whole exome. Furthermore, two samples of a single patient were used, one of them enriched with a SureSelectXT Human All Exon V5 kit (Agilent Technologies; Böblingen, Germany), and the other with a custom SureSelect HLA kit provided by Michael Wittig (Institute of Clinical Molecular Biology, Christian-Albrechts-University of Kiel, Germany) and Agilent Technologies to enrich the HLA loci. Both samples were sequenced with an Illumina HiSeq 2500 sequencer with 101 bp long reads.

Detailed sample and run identifiers are listed in Supplementary Table S11. The binary hit matrices of the iVacALL samples can be found in Supplementary Material S13.

2.6 Coverage depth simulation

To investigate the influence of coverage depth on prediction accuracy, artificial data were generated from all 1000 Genomes Project exome sequencing benchmark samples using randomized subsets of decreasing size drawn from the original reads to simulate different coverage depth conditions, until the number of remaining reads amounted to as little as ~0.2× fold coverage on HLA-I loci.

2.7 Performance measure

We based our comparison on the percentage of correctly predicted HLA alleles (two-digit and four-digit) per sample. This measure was used by Boegel *et al.* and is similar to the definition of sensitivity used by Warren *et al.* and accuracy by Kim *et al.* Correctness of zygosity prediction was used as a second, independent performance measure, where the zygosity of a locus was considered to be correct if the predicted zygosity matched the experimentally determined zygosity without considering the correctness of the typed alleles.

2.8 Implementation and availability

The NGS analysis pipeline was implemented in Python 2.7 using the Pandas 0.12 (pandas.pydata.org) module with HDF5 1.8.11 (www.hdfgroup.org/HDF5) data persistence support. Read mapping was performed using RazerS 3.1 (Weese *et al.*, 2012) and Bowtie 2 (Langmead and Salzberg, 2012).

The ILP was formulated with the Python package Pyomo, which is part of Coopr 3.3 (software.sandia.gov/trac/coopr), and solved with ILOG CPLEX 12.5 (www.ilog.com). CPLEX is free of charge for academic use, but open-source ILP solvers like GLPK can be used as well, with a single configuration option. Statistical analysis was conducted with R 3.0.2. Bootstrapping with 100 000 repetitions was used to calculate 95% confidence intervals. OptiType is available under a BSD open-source license. The complete source code can be downloaded from github.com/FRED-2/OptiType.

3 RESULTS

HLA typing with four-digit-level accuracy is essential for clinical applications like the development of individualized patient-specific vaccines and transplantation. Therefore, OptiType has been optimized to yield correct HLA typing results on four-digit resolution (i.e. on the protein-coding level) for distinct read lengths and different types of sequencing technologies. Performance of OptiType has been evaluated on exome sequencing, WGS and RNA-Seq data.

3.1 Overall performance

OptiType was benchmarked on all datasets that were used by other *in silico* methods, including HLAminer by Warren *et al.*, ATHLATES by Liu *et al.*, seq2HLA by Boegel *et al.*, HLAforest by Kim *et al.* and the most recent HLA typing method by Major *et al.* On the 361 benchmark samples, OptiType achieved an accuracy of 97.1% (CI_{95}: 96.1–97.80%) on four-digit level and 99.3% (CI_{95}: 98.7–99.7%) on two-digit level, correctly predicting 939 of 950 heterozygous loci and 127 of 133 homozygous loci (Supplementary Table S10). Because two-digit typing has little relevance to clinical applications, we present only four-digit performance in the comparison.

OptiType outperforms comparable methods on all datasets by 4 to 15% accuracy, corresponding to a 65 to 83% lower rate of incorrect allele predictions (Fig. 2, Supplementary Table S1). Statistical significance was confirmed in each case by a sign test at an α-level of 0.05. Only ATHLATES showed comparable performance on their benchmark dataset consisting of 11 samples.

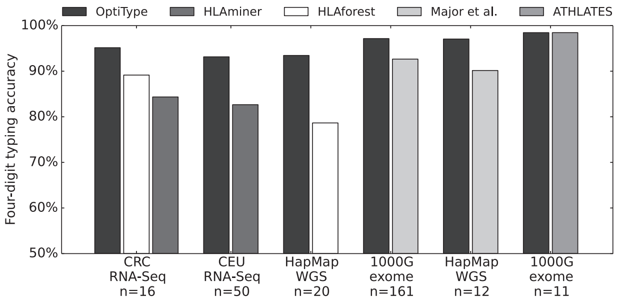

Fig. 2. Performance comparison of HLA typing algorithms. OptiType's average prediction accuracy for major HLA-I loci was compared with four other published HLA typing methods capable of four-digit typing on publicly available datasets previously used to evaluate these methods

Applying OptiType on all 253 paired-end Illumina exome sequencing runs of the 1000 Genomes Project yielded an average accuracy of 97.6% (CI_{95}: 96.7–98.4%). Detailed prediction information can be found in Supplementary Table S2. Heterozygosity was correctly predicted for 667 of 676 (98.7%), and homozygosity for 80 of 83 loci (96.4%).

Performance of OptiType has also been benchmarked on an in-house exome sequencing dataset of 10 ALL patient samples, which have been gathered as part of the iVacALL project (Kyzirakos *et al.*, 2013), yielding an accuracy of 96.7% (CI_{95}: 91.7–100%). All heterozygous and homozygous cases were detected correctly. Detailed prediction results can be found in Supplementary Tables S1 to S10.

3.2 Influence of intronic reconstruction

To analyze the influence of intron sequence reconstruction for DNA sequencing data, a modified version of OptiType was tested on the 1000 Genomes Project dataset using only exon 2 and 3 sequences as reference. Reads were mapped with Bowtie 2's local alignment (soft clipping) setting to avoid losing reads at the exon boundaries. Mismatch tolerance was similar to that of the RazerS3 mapping settings.

As exons 2 and 3 are ~270 bp long each, a significant amount of paired reads could be mapped with just one mate, effectively turning them into single-ended hits. Therefore, we evaluated prediction performance with two different hit matrix construction rules: once with allowing mapping pairs only and once with including mapped reads without mapped mates as well.

OptiType yielded an accuracy of 93.5% (CI_{95}: 91.8–95.1%) with the strict mapping pair approach and 90.6% (CI_{95}: 89.0–92.3%) with the hybrid approach allowing single-end hits as well, showing a 2.7- to 3.9-fold increase in error compared with the 97.6% accuracy of the default OptiType pipeline using intronic sequences.

3.3 Influence of HLA enrichment and coverage depth

To determine the effects of specific HLA enrichment on prediction accuracy, we investigated a sample with an average coverage

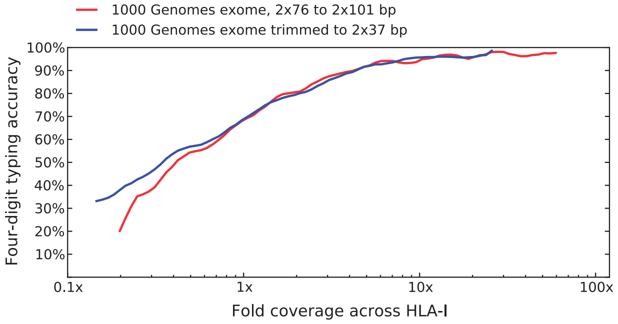

Fig. 3. Coverage and read length dependence of prediction accuracy. To determine the influence of coverage depth on HLA typing accuracy, reads of 253 exome sequencing runs of the 1000 Genomes Project were subsampled >4000 times to simulate different coverage depth conditions. To investigate the impact of read length on performance, original reads were trimmed to 37 bp and evaluated with the same subsampling procedure. Read length alone shows little effect on prediction accuracy, and an average coverage depth greater than 10× over the HLA-I loci was already found to yield maximal accuracy

depth of ~4100× on HLA-I loci, from which a decreasing number of reads were randomly extracted, simulating decreasing coverage depth. Based on the extracted subsets of reads, the HLA genotype was predicted and compared against the experimentally derived genotype to determine accuracy. A fully correct genotype prediction could be consistently achieved using as little as ~0.3% of the total amount of reads (~12 × coverage), corresponding to ~15% of reads of the non-HLA specific exome-enriched sample of the same subject.

To investigate the dependence of prediction accuracy on coverage depth on a broader set of samples, a simulation experiment was carried out using all 1000 Genomes Project exome sequencing samples by similarly resampling runs with restricted number of reads to examine different coverage depth conditions. A total of 253 individual samples were resampled >4000 times. An accuracy of 95% was shown to be achievable at 10× average coverage depth on HLA-I loci (Fig. 3).

To assess the influence of read length on performance, reads were trimmed to 2 × 37 bp and evaluated similarly. No detrimental effects have been observed, suggesting that the method is just as reliable with short reads as with longer reads.

4 CONCLUSION

The presented *in silico* HLA typing pipeline OptiType performs fully automated HLA typing with four-digit resolution on NGS data from RNA-Seq, exome sequencing and WGS technologies. Performance of OptiType was benchmarked on datasets of the above sequencing technologies with read lengths ranging from 2 × 37 bp to 2 × 101 bp and showed an accuracy of 99.3% (CI_{95}: 98.7–99.7%) on two-digit-level and of 97.1% (CI_{95}: 96.1–97.80%) on four-digit-level typing. In terms of zygosity prediction, OptiType achieved an accuracy of 98.4% (CI_{95}: 97.5–99.1%) on 361 benchmarked runs, correctly predicting 939 of 950 heterozygous loci and 127 of 133 homozygous loci. OptiType is applicable to NGS data of different sources and outperforms previously published *in silico* HLA typing approaches on both two- and four-digit resolution. The latter is especially important in clinical applications like individualized vaccine design, prevention of graft-versus-host disease and treatment of autoimmune diseases. Additionally, OptiType, as an *in silico* approach, provides the benefits of great cost reduction and a decrease of turnaround time in comparison with state-of-the-art experimental HLA typing methods. Runtimes are typically on the order of minutes per sample (including read mapping) and thus permit an efficient integration into existing NGS analysis pipelines.

In general, coverage depth, as seen in the enrichment and simulation studies, does not play a major role above a certain level. As previously observed by Major *et al.*, the number of covered bases has a stronger influence on the prediction outcome than coverage depth. Short reads, while increasing the complexity of the problem because of higher mapping ambiguities, did not have a negative effect on our method's performance.

Incorrect predictions were mostly found to be caused by three distinct issues. First, sequence stretches not covered by any reads can make it impossible to resolve the ambiguity between the correct allele and alleles differing only on the uncovered segments.

Second, zygosity detection occasionally fails in cases where alleles with high sequence similarity constitute a heterozygous locus. In such cases including both alleles in the solution has little impact on the total number of explained reads compared with including just one of them; therefore, OptiType favors the homozygous solution. This problem is normally encountered if the two alleles' distiguishing segments have considerably lower coverage than the rest of their sequence. Third, while typing minor loci generally helps with finding the actual source of reads mapping to both minor and major loci, it is not able to resolve all ambiguities for every genotype. Additionally, experimental typings of the benchmark datasets were sometimes found to be inaccurate, as also observed for the 1000 Genomes Project samples (Erlich *et al.*, 2011). This limits the accuracy that can be achieved on these datasets.

It is important to ensure an equal a priori chance for every allele to be identified by minimizing the disadvantage of alleles with only partial sequence information. Therefore, only exons 2 and 3 and their flanking intron sequences were used as reference, reconstructing unknown intron sequences with a phylogeny-based approach for incomplete alleles. Including intron sequences not only helped retain more read pairs, but information from intronic hits was found to be beneficial to performance. Furthermore, with an increasing number of completely sequenced HLA alleles, the used reference sequences could be extended beyond regions surrounding exons 2 and 3, reducing ambiguities and increasing prediction accuracy of OptiType.

To summarize, OptiType is a fast and accurate HLA typing method based on NGS data, which provides an alternative approach to common HLA genotyping methods. It can be easily adapted to predict genotypes for loci other than HLA-I such as HLA-II and transporter associated with antigen processing. Nevertheless, the predictions are restricted to the used reference and, therefore, can predict only known alleles.

ACKNOWLEDGEMENTS

The authors would like to thank Michael Wittig (Institute of Clinical Molecular Biology, Christian-Albrechts-University of Kiel, Germany) and Agilent Technologies Deutschland GmbH (Böblingen, Germany) for kindly providing the custom HLA enrichment kit.

Funding: This study was partially funded by Deutsche Forschungsgemeinschaft (SFB 685/B1, KO 2313/6-1) and Bundesministerium für Bildung und Forschung (01GU1106).

Conflict of interest: none declared.

REFERENCES

Bentley,G. *et al.* (2009) High-resolution, high-throughput HLA genotyping by next-generation sequencing. *Tissue Antigens*, **74**, 393–403.

Blasczyk,R. *et al.* (1997) The nature of polymorphism of the HLA class I non-coding regions and their contribution to the diversification of HLA. *Hereditas*, **127**, 7–9.

Boegel,S. *et al.* (2013) HLA typing from RNA-Seq sequence reads. *Genome Med.*, **4**, 102.

Bradley,B. (1991) The role of HLA matching in transplantation. *Immunol. Lett.*, **29**, 55–59.

Danzer,M. *et al.* (2013) Rapid, scalable and highly automated HLA genotyping using next-generation sequencing: a transition from research to diagnostics. *BMC Genomics*, **14**, 221.

Döring,A. *et al.* (2008) SeqAn an efficient, generic C++ library for sequence analysis. *BMC Bioinformatics*, **9**, 11.

Erlich,R.L. *et al.* (2011) Next-generation sequencing for HLA typing of class I loci. *BMC Genomics*, **12**, 42.

Gabriel,C. *et al.* (2009) Rapid high-throughput human leukocyte antigen typing by massively parallel pyrosequencing for high-resolution allele identification. *Hum. Immunol.*, **70**, 960–964.

Gonzalez-Galarza,F.F. *et al.* (2011) Allele frequency net: a database and online repository for immune gene frequencies in worldwide populations. *Nucleic Acids Res.*, **39** (**Suppl. 1**), D913–D919.

Haralambieva,I.H. *et al.* (2013) The genetic basis for interindividual immune response variation to measles vaccine: new understanding and new vaccine approaches. *Expert Rev. Vaccines*, **12**, 57–70.

Karp,R.M. (1972) Reducibility among combinatorial problems. In: *Complexity of Computer Computations*. Plenum Press, New York, NY, pp. 85–103.

Kim,H.J. and Pourmand,N. (2013) HLA Haplotyping from RNA-seq data using hierarchical read weighting. *PloS One*, **8**, e67885.

Kyzirakos,C. *et al.* (2013) iVacALL: utilizing next-generation sequencing for the establishment of an individual peptide vaccination approach for paediatric acute lymphoblastic leukaemia. *Bone Marrow Transplant.*, **48**, S401.

Langmead,B. and Salzberg,S.L. (2012) Fast gapped-read alignment with Bowtie 2. *Nat. Methods*, **9**, 357–359.

Lank,S.M. *et al.* (2012) Ultra-high resolution HLA genotyping and allele discovery by highly multiplexed cDNA amplicon pyrosequencing. *BMC Genomics*, **13**, 378.

Lank,S.M. *et al.* (2010) A novel single cDNA amplicon pyrosequencing method for high-throughput, cost-effective sequence-based HLA class I genotyping. *Hum. Immunol.*, **71**, 1011–1017.

Leinonen,R. *et al.* (2011) The European nucleotide archive. *Nucleic Acids Res.*, **39** (**Suppl. 1**), D28–D31.

Liu,C. *et al.* (2013) ATHLATES: accurate typing of human leukocyte antigen through exome sequencing. *Nucleic Acids Res.*, **41**, e142.

Major,E. *et al.* (2013) HLA typing from 1000 genomes whole genome and whole exome illumina data. *PloS One*, **8**, e78410.

Montgomery,S.B. *et al.* (2010) Transcriptome genetics using second generation sequencing in a *Caucasian* population. *Nature*, **464**, 773–777.

Moonsamy,P.V. *et al.* (2013) High throughput HLA genotyping using 454 sequencing and the Fluidigm Access Array System for simplified amplicon library preparation. *Tissue Antigens*, **81**, 141–149.

NCBI Resource Coordinators. (2013) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, **41**, D8–D20.

Opelz,G. *et al.* (1999) HLA compatibility and organ transplant survival. Collaborative transplant study. *Rev. Immunogenet.*, **1**, 334.

Ovsyannikova,I.G. and Poland,G.A. (2011) Vaccinomics: current findings, challenges and novel approaches for vaccine development. *AAPS J.*, **13**, 438–444.

Robinson,J. *et al.* (2013) The IMGT/HLA database. *Nucleic Acids Res.*, **41**, D1222–D1227.

Schrijver,A. (1998) *Theory of Linear and Integer Programming*. John Wiley & Sons, Chichester, West Sussex, UK.

Shiina,T. *et al.* (2012) Super high resolution for single molecule-sequence-based typing of classical HLA loci at the 8-digit level using next generation sequencers. *Tissue Antigens*, **80**, 305–316.

Sievers,F. *et al.* (2011) Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol. Syst. Biol.*, **7**, 539.

The 1000 Genomes Project Consortium. (2012) An integrated map of genetic variation from 1,092 human genomes. *Nature*, **491**, 1.

The International HapMap Consortium. (2005) A haplotype map of the human genome. *Nature*, **437**, 1299–1320.

Thorsby,E. and Lie,B.A. (2005) HLA associated genetic predisposition to autoimmune diseases: genes involved and possible mechanisms. *Trans. Immunol.*, **14**, 175–182.

Undlien,D.E. *et al.* (2001) HLA complex genes in type 1 diabetes and other autoimmune diseases. Which genes are involved? *Trends Genet.*, **17**, 93–100.

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

Weese,D. *et al.* (2012) RazerS 3: faster, fully sensitive read mapping. *Bioinformatics*, **28**, 2592–2599.