Genome Medicine

METHOD Open Access

Derivation of HLA types from shotgun sequence datasets

René L Warren[1], Gina Choe[1], Douglas J Freeman[1], Mauro Castellarin[1], Sarah Munro[1], Richard Moore[1] and Robert A Holt[1,2*]

Abstract

The human leukocyte antigen (HLA) is key to many aspects of human physiology and medicine. All current sequence-based HLA typing methodologies are targeted approaches requiring the amplification of specific HLA gene segments. Whole genome, exome and transcriptome shotgun sequencing can generate prodigious data but due to the complexity of HLA loci these data have not been immediately informative regarding HLA genotype. We describe HLAminer, a computational method for identifying HLA alleles directly from shotgun sequence datasets (http://www.bcgsc.ca/platform/bioinfo/software/hlaminer). This approach circumvents the additional time and cost of generating HLA-specific data and capitalizes on the increasing accessibility and affordability of massively parallel sequencing.

Background

Due to its central role in adaptive immunity, human leukocyte antigen (HLA) is implicated in wide ranging areas of medicine, from infectious disease and vaccinology to cancer, autoimmunity, aging and regenerative and transplantation medicine [1-7]. The HLA locus is the most polymorphic region of the genome with over 5,000 variant HLA-class I allelic sequences catalogued to date. This genetic heterogeneity is the principal challenge to HLA typing methodologies, and it is the reason why this region has remained largely opaque to analysis by next-generation sequencing (NGS) platforms. Conventional sequence-based HLA typing approaches, the most recent of which exploits the sequence throughput of the Illumina MiSeq [8] and relatively long sequence reads of the 454 NGS platform [9], are targeted assays that rely on amplification of hypervariable sub-regions of these loci and variant detection within these amplicons. As such, HLA calls are based on sequence information that is not as comprehensive as for shotgun datasets, and must be generated *de novo* for each subject. The widespread uptake of large-scale genome, exome and transcriptome shotgun sequencing approaches for biomedical research, and now for clinical use, prompted us to explore the utility of these types of NGS data sets for HLA typing. The need has been for a solution to the problem of managing the many millions of short sequence reads NGS technologies produce, managing the many thousands of reference allele sequences, and integrating all of these data in a manner that maximally informs HLA content. Here we present a method for HLA allele prediction from next-generation shotgun sequence datasets. We focus on data generated from the Illumina platform, from which most sequence data are currently derived worldwide. Importantly, HLA allele assignments from shotgun datasets can not be derived from standard alignment-based interpretive methods for the simple reason that the extant genome reference sequences [10,11] on which these methods rely do not provide any useful representation of HLA allelic diversity. Therefore, we have developed a computational pipeline that derives HLA allele predictions by targeted assembly of shotgun sequence data and comparison to a database of reference allele sequences. Our solution allows, for the first time, application of the power of NGS to the interrogation of one of the most important and complex sets of human genes. Our method is scalable, such that it will provide utility in extracting HLA information even from very large sequence data sets, such as those currently being compiled by various international consortia [12-15].

Materials and methods

Library construction and sequencing

Written informed consent was obtained from all donors and samples were collected following assessment of tissue specimens by a pathologist according to standardized

* Correspondence: rholt@bcgsc.ca
[1]BC Cancer Agency, Michael Smith Genome Sciences Centre, Vancouver, British Columbia V5Z 1L3, Canada
Full list of author information is available at the end of the article

operating procedures, immediately following surgical resection. Library construction and Illumina sequencing were performed as previously described for RNA-Seq [16] and whole genome shotgun (WGS) [17]. For the colorectal cancer (CRC) RNA-Seq study, four lanes of 100-nucleotide paired-end sequences were obtained for each of the two pools, providing an average of 5 million paired reads per sample. For WGS, approximately 430 million paired 100-nucleotide WGS reads (approximately 30× depth coverage human genome) from normal and tumor samples from four diffuse large B cell lymphoma patients were processed [17]. The sequencing data from the CRC study have been submitted to the NCBI Sequence Read Archive [18] under accession number SRP010181. A file describing the sample libraries is available at [19].

Exome capture libraries were prepared using the SureSelect system (Agilent) according to the manufacturer's instructions. Approximately 30 million (normal samples) and 120 million (normal plus tumor samples) 100-nucleotide exon capture paired-end sequence reads were generated from three ovarian cancer patients whose HLA alleles were verified by PCR-based methods. Verification of HLA allele predictions was accomplished by PCR amplification of exons 2 and 3 from HLA-I A, B and C, followed by capillary sequencing as previously described [20].

IMGT/HLA sequences

HLA coding DNA sequence (CDS) and genomic sequence databases from release 3.3.0 and 3.4.0 were obtained, respectively, from [21]. HLA-I exon 2 and 3 concatenated sequence FASTA files were prepared using exon coordinates available from the flat file database (EMBL format) released by IMGT [22]. For HLA allele predictions from RNA-Seq data, we used concatenated exons 2 and 3 as sequence targets for assembly using the TASR assembly tool [23]. For predictions from genome and exome NGS data, we used HLA-I genomic sequences from major genes A, B and C.

Computational HLA allele predictions by targeted read assembly

HLA CDS or genomic sequences from IMGT/HLA (sequence targets) are read by TASR (default options used with -i 1), creating a hash table of every possible 15-nucleotide word (k-mers) encountered. NGS data sets are interrogated for the presence of one of these k-mers in 5' (on either strand) and candidate reads recruited. Recruited reads seed the assembly in a manner analogous to that of SSAKE [24]. Only sequence contigs equal to or larger than a user-determined length (200 nucleotides chosen for this study) are considered for further analysis. Reciprocal BLAST [25] (v.2.2.22 with options -a 8 -F F -p blastn -m 7) alignments are performed between the contig and HLA CDS or genomic sequence databases depending on the read source (RNA-Seq or WGS and exon capture), parsed at runtime using PERL Bio::SearchIO modules and summarized. HLAminer parses these alignment files and generates a score and probability for each putative HLA coding variant identified from sequence contigs. Briefly, for each assembled contig, best BLAST HLA alignments are reported, tracking the sequence identity over the alignment portion, as well as over the length of the contig. Contigs are organized by increasing number of HLA sequences they co-characterize best, listing all possible *ex-equo* best hits and tracking HLA sequences that, reciprocally, best identify each contig. For each putative HLA, a score S_{HLA} is calculated as the sum of score computed for each contig aligning to it. Individual contig scores factor in the contig depth of coverage, length and percent sequence identity, such that a score reflects the number of bases aligned to a particular HLA allele. A reciprocal best hit where a given HLA aligns best to a given contig doubles the score for the identified HLA sequence:

$$S_{HLA} = \sum_{Contig=1}^{n} Score_{Contig} = size^{*}depth^{*}\%sequence\_identity$$

For any given contig, the probability of characterizing a single HLA allele by chance is equal to the inverse proportion of HLA sequences in the sequence database. And since shorter contigs may not capture sufficient bases to characterize any one type unambiguously, the probability P that a contig characterizes one or another HLA type is mutually exclusive such that:

$$P_{Contig_1\_is\_HLA_x} = \sum P_{HLA}$$

The expect value (Eval) of each computationally determined HLAx, Eval$_{\mathrm{HLA}x}$, is calculated as:

$$Eval_{HLA_x} = (P_{Contig_1 \text{ is } HLA_x} * P_{HLA_x \text{ is } Contig_1}) * (P_{Contig_2 \text{ is } HLAx} * P_{HLA_x \text{ is } Contig_2}) * \ldots * (P_{Contig_n \text{ is } HLA_x} * P_{HLA_x \text{ is } Contig_n})$$

since individual contig probabilities and reciprocal best hits are independent events. A short list of HLA allele groups (for example, A*02) and protein coding alleles (for example, A*02:01), sorted by decreasing score, are catalogued for each major HLA gene. When separate contigs characterize the same types, only the types that overlap are reported, unless the non-overlapping ones are characterized by additional, distinct contig (s). In addition, we summarize ambiguous HLA alleles using the P designation, when applicable.

Simulated data sets

In separate experiments, we removed HLA CDS, exonic regions and genes from 15K randomly selected Ensembl [26] transcripts, approximately 220K exon regions [27] (SureSelect Target Enrichment, Agilent Technologies, Inc. [28] and the HuRef genome [11]). For each data

type, we randomly generated 20 sets of six (2 × A, 2 × B, 2 × C) HLA-I alleles (total of 120 alleles). In triplicate experiments, we merged each set of six sequences with HLA-less CDS, exon regions or HuRef, respectively, and simulated at various depth of coverage 50-, 75-, 100- or 150-nucleotide paired-end reads with 0.5, 1, 2 or 3 error using SAMtools [29] wgsim, and ran TASR and BLASTN as described above. For the simulation from direct read pair alignments, we used the simulated reads described above and ran BWA [30] with defaults and generated HLAminer predictions from SAM files.

Assessment metrics

We define the sensitivity as a proportion, that is, the number of HLA allele groups or protein coding alleles detected over the sum of distinct groups or protein coding alleles randomly chosen for the simulation or predicted by PCR, when applicable. The ambiguity rate is the proportion of all ambiguous predictions per total allele groups or protein coding alleles predicted. Ambiguous predictions arise when HLA allele groups or protein coding gene names differ despite having an identical score and probability. The specificity is defined as a proportion of number of groups or alleles predicted accurately divided by the total number of groups or alleles detected, respectively.

HLA typing

HLA class I alleles were predicted directly from the RNA-Seq data as described [20]. Briefly, genomic DNA was extracted from patient granulocytes, and exons two and three from HLA class I genes (A, B, and Cw) were amplified by PCR [31]. PCR amplicons were cloned and sequenced using an ABI 3730XL instrument, according to standard procedures. Clone sequences were assembled using Phred/Phrap/Consed [32]. The resulting sequence data were aligned against all available exon 2 and 3 nucleotide sequences from the 3.1.0 release of the IMGT/HLA database [22] using ClustalW [33]. Protein coding allele assignments [34] were based on high-quality exact or synonymous matches at informative nucleotide positions.

Results and discussion

To maximize the performance of HLAminer with short read data, we implemented stringent, localized, *de novo* assembly of sequence reads prior to alignment (Figure 1, left). Direct alignment of reads to reference alleles is also supported (Figure 1, right), but at the present time we find this modality has modest utility due to current limitations on read length. HLAminer predictions are arranged by HLA gene (for example, HLA class I A, B and C) and for each, putative alleles are ranked by highest scoring HLA protein coding alleles. A confidence value reflects the likelihood of each prediction (expect value) on a log10 scale. A sample output from HLAminer is shown (Table 1).

For initial evaluation of HLAminer we relied on simulated data sets, which allowed us to determine the influence on performance of sequencing parameters such as depth of coverage, sequence read length, and sequence error. We produced simulated data sets for each of the three formats, RNA-Seq, WGS and exome, by taking reference sequences (Ensembl transcripts, the HuRef genome, and hg19 exon capture regions, respectively) and substituting HLA-I A, B and C sequences with two randomly chosen alleles of HLA-I A, B and C. From these modified references we generated faux sequence reads. For each sequence format, 20 such data sets were generated and these were queried in triplicate, yielding a total of 360 allele predictions per condition tested. The sensitivity, specificity and ambiguity of HLA class I allele prediction was evaluated by comparing the highest-scoring HLAminer predictions to the randomly selected alleles. By ambiguity we mean the prediction of multiple, equally probably alleles.

HLA nomenclature (for example, HLA A*02:01) defines the digits immediately following the asterisk as the allele group (two-digit resolution, formerly referred to as supertype) and the next set of digits (those following the semicolon, often referred to as four-digit resolution) as the individual protein coding allele [34]. Further separators and digits are sometimes used to describe allelic variants that contain silent nucleotide differences. Using simulated data, we found that at the level of HLA allele groups, RNA-Seq data provided high sensitivity and specificity (each >95.7%) with a low ambiguity (<4.5%), even at relatively low coverage (<5 million total read pairs) (Figure 2; Additional file 1). Likewise, WGS provided high sensitivity and specificity (each >97.3%) and no observable ambiguity (0.0%) for prediction of allele groups, but required substantially higher sequence depth, on the order of 400 million paired reads, to achieve this (Figure 2; Additional file 1). This is the equivalent of approximately 30× genome coverage with 100-nucleotide reads. For both RNA-Seq and WGS data, predictions at the level of individual protein coding alleles showed very similar sensitivity and specificity to that observed for allele group predictions, but ambiguity levels increased to approximately 30% (Figure 2; Additional file 1). Our expectations for HLA allele prediction from exome data were low, because allelic diversity of HLA coding sequence tends to have limited representation in standard capture reagents. For example, the Agilent SureSelect system that we use at our center contains 36 120-nucleotide RNA probes targeting the HLA class I region of hg19. Still, we included evaluation of this data type for the purpose of completeness, and

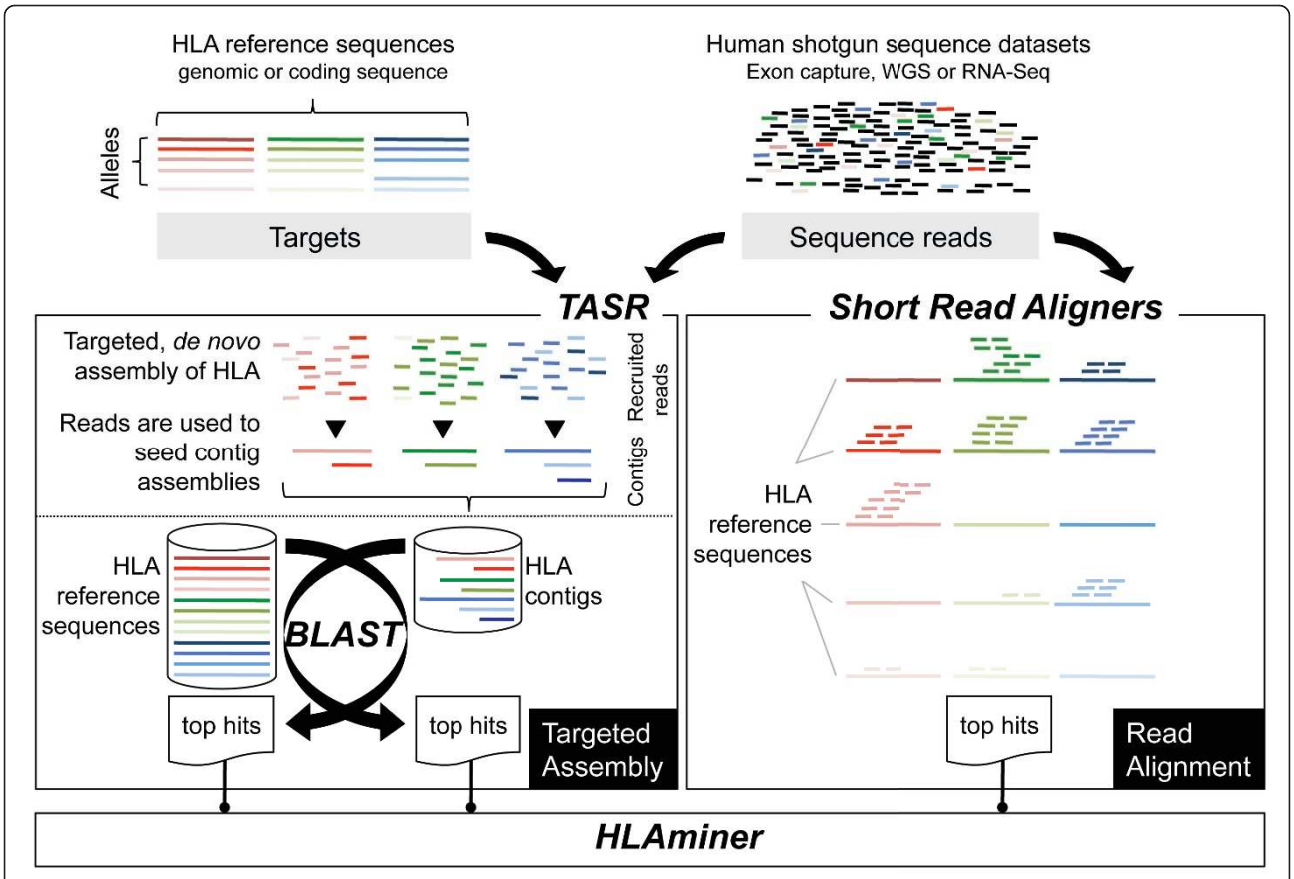

Figure 1 Computational predictions of HLA-I from shotgun data by targeted assembly (left) or read alignment (right). For targeted assembly, NGS reads having their first fifteen 5′ bases matching one of HLA CDS (RNA-Seq) or genomic (WGS/exon capture) sequences are recruited and assembled *de novo* with TASR. Resulting sequence contigs are aligned against a database sequence of all predicted HLA CDS (RNA-Seq) or genomic sequences (WGS/exon capture), tracking best HLA hit(s). Reciprocal best alignments are considered in the same manner. Putative allele assignments from shotgun datasets (HLAminer) are informed by contig length, depth of coverage and similarity to reference sequences, when applicable. The probability of each prediction being correct is estimated by determining the probability of that prediction being observed by chance.

with the understanding that a variety of HLA alleles could possibly be captured by imperfectly matching probes of this length. Our simulations revealed that exome data did in fact show some modest utility for HLA prediction, at least at the allele group level. For allele group prediction, high specificity (92.8%) and low ambiguity (4.7%) could be achieved at low coverage (40 million read pairs); however, considerably higher coverage was necessary to increase sensitivity, and even at very high exome coverage (240 million read pairs) sensitivity never approached that observed for the other data types. By comparison, for RNA-Seq, 5 million and 3 million 100-nucleotide RNA-Seq read pairs are required for 95% sensitivity and specificity, respectively. For WGS, 427 million and 57 million 100-nucleotide read pairs are needed for 95% sensitivity and specificity, respectively. Under the conditions tested, exome data did not provide such high levels of detection and prediction accuracy at any read depth and performance for predicting individual protein coding alleles from exome data was uniformly poor (Figure 2).

Overall, from simulation, RNA-Seq datasets provided the greatest utility for HLA prediction. This may be due, in part, to lower representation in RNA-Seq data of off-target regions, such as the minor HLA class I genes, pseudogenes, and HLA class II genes, compared to genome or exome data, where these regions would be expected to have approximately equal representation as the class I alleles of interest, A, B and C. The stark contrast in HLAminer predictions derived from RNA-Seq compared to WGS or exome capture highlights intrinsic properties of these datasets and their value for computational HLA predictions. Functional HLA-I alleles are expressed on all nucleated cells, and despite possible amplification biases in the RNA-Seq library construction protocol, the high abundance of HLA-I transcripts is such that relatively low

Table 1 Output from HLAminer HLA class I predictions from a CRC patient 100-nucleotide RNA-Seq sample

Allele[a]	Score[b]	Expect value (Eval)	Confidence (-10 × log10(Eval))
HLA-A[c]			
Prediction[d] 1 - A*02			
A*02:01P	64038.03	1.63E-06	57.9
Prediction 2 - A*11			
A*11:01P	5463.99	5.30E-09	82.8
HLA-B			
Prediction 1 - B*27			
B*27:05P	64579.61	2.67E-18	175.7
Prediction 2 - B*07			
B*07:02P	56662.08	6.63E-12	111.8
HLA-C			
Prediction 1 - C*07			
C*07:02P	49419.33	5.23E-08	72.8
Prediction 2 - C*02			
C*02:02P[e]	20466.00	6.64E-16	151.8
C*02:21[e]	20466.00	6.64E-16	151.8

[a]HLA protein coding alleles validated by PCR are shown in bold face. [b]The protein coding allele predictions are arranged by decreasing score from most to less likely. [c]Most likely HLA class I allele groups and protein coding alleles (Confidence (-10 × log10(Eval)) ≥ 20 Score ≥ 1,000) for each gene. [d]The prediction rank factors in the maximum score for each predicted allele. [e]Ambiguity arises when two or more HLA allele group or protein coding alleles have the same score (for example, C*02:02P and C*02:21).

depth of sequencing is needed for robust predictions (approximately 5 million). On the other hand, non-functional (null) HLA alleles that are present in the genome (but transcriptionally silent) can confound HLA prediction from WGS or exome capture data, since the functional alleles and null alleles are equally represented in these data types. HLAminer has the functionality to report predictions from null alleles, if desired.

We explored further the effects of read length (up to 150 nucleotides) and sequencing errors (up to 3%) with RNA-Seq data. Not unexpectedly, performance improved with increasing read length and decreasing base error (Table 2). Reads with length less than 75 nucleotides and error rates higher than 1% significantly impacted performance for prediction of individual protein coding alleles, but prediction of allele groups remained robust (Table 2).

Next, we evaluated the performance of HLAminer with real shotgun datasets, including RNA-Seq data from 16 CRC libraries (RL Warren, DJ Freeman, P Watson, RA Moore, EA Allen-Vercoe, RA Holt, manuscript submitted), WGS and RNA-Seq from four lymphoma libraries [17] and exon capture data from three ovarian cancer libraries (Figure 2; Additional file 1). HLA predictions were compared to results from these same subjects obtained from standard PCR and capillary sequence-based typing [20]. Results mirrored those obtained from simulated data. For all data types, prediction of allele groups was more reliable than prediction of individual protein coding alleles. For prediction of allele groups, the CRC RNA-Seq data yielded predictions with highest sensitivity and specificity (>96.5%) and low ambiguity (<2.4%), even at low sequence depth (approximately 5 million pairs per sample). From a total of 81 HLA allele groups predicted by HLAminer on the CRC cohort shotgun data only a single allele group prediction conflicted with PCR-based typing results (Additional file 2). For WGS and exome data, high sensitivity and specificity could also be achieved, but only at much higher depth of coverage. For all data types, the ambiguity associated with prediction of individual protein coding alleles were higher than for prediction of allele groups, with predictions from exome data sets more significantly impacted than predictions from WGS or RNA-Seq data sets. HLAminer predictions were also benchmarked on low-coverage 100-nucleotide WGS data from 20 individuals of the 1000 Genomes cohort [15]. HLA class I allele predictions obtained from these same HapMap samples by the targeted PCR method of Erlich and colleagues have been previously published [9]. Applying HLAminer to this data set, allele group sensitivity and specificity of 86.7 ± 15.9% and 86.3 ± 16.1% were achieved (Additional file 3), despite the relatively low number of genome shotgun reads processed (mean ± standard deviation of 361.2 ± 80.9 million. Further, our results from these 1000 Genomes samples are consistent with those we obtained from the diffuse large B cell lymphoma control normal tissue (sensitivity and specificity of 68.8 ± 31.5% and 71.3 ± 21.8%) and tumor tissue (sensitivity and specificity of 75.0 ± 20.4% and 87.5 ± 14.4%) WGS datasets, for which substantially higher sequence coverage was available (approximately 1.1 billion reads per sample). As discussed, the data type availability (WGS) and the lower depth of

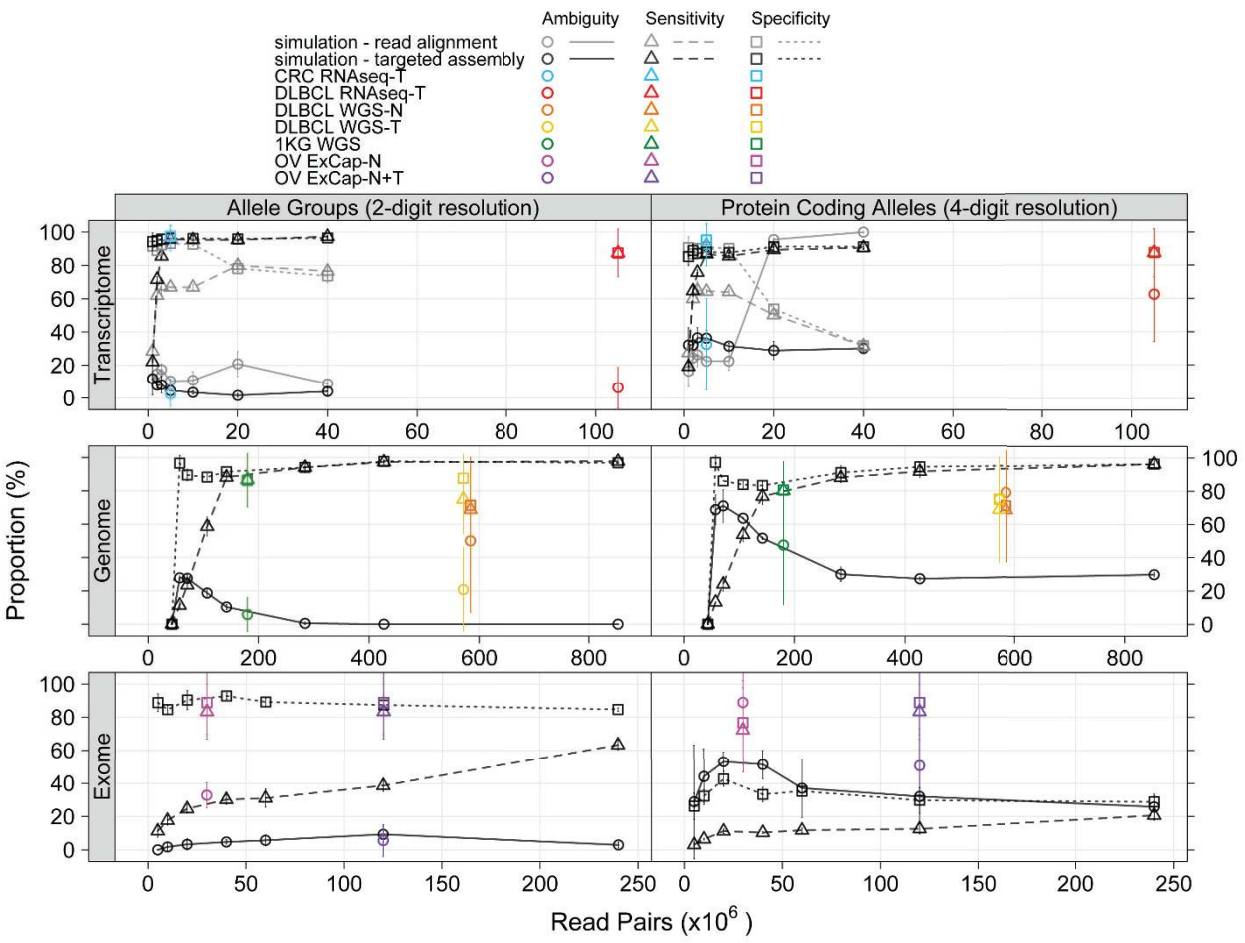

Figure 2 HLAminer performance. HLA allele group and protein coding allele predictions derived from targeted read assembly (black symbols) or direct read alignment (grey symbols) of simulated 100-nucleotide RNA-Seq, WGS and exon capture (ExCap) datasets were compared to original, spiked-in, HLA sequences and performance metrics evaluated (ambiguity, sensitivity and specificity represented by circle, triangle and square symbols, respectively). HLAminer predictions were also obtained from targeted assembly of colorectal cancer (CRC; blue symbols), lymphoma (DLBCL; red, orange and yellow symbols), 1000 Genomes (1KG; green symbols) and ovarian cancer (OV; violet and magenta symbols) patient tumor (T) and/or matched normal (N) shotgun datasets and compared to PCR-based HLA types to calculate performance metrics.

coverage (10- to 20-fold) are both limiting factors for HLAminer predictions.

HLAminer can evaluate reads by direct alignment (Figure 1, right). However, with Illumina read lengths currently ranging from 100 to 150 nucleotides, this approach has limited utility at the present time. At best we observed 80.0 ± 3.5% sensitivity and 78.2 ± 2.8% specificity (mean ± standard deviation; Figure 2, top panel; Additional file 1).

Regardless of input data, HLAminer predictions for HLA allele groups (two-digit resolution) are more robust than for HLA protein-coding alleles (four-digit resolution) (Figure 2; Additional file 1). Both the sensitivity and specificity of four-digit allele predictions are reduced relative to their two-digit counterparts, but changes to the ambiguity of predictions are more pronounced. For example, with 5 million × 100-nucleotide simulated RNA-Seq read pairs, four-digit predictions show a 8.9% decrease in sensitivity, 8.2% decrease in specificity, and a 31.9% increase in ambiguity, compared to two-digit predictions. This is because HLA coding alleles often differ by only a single base. In contrast to conventional HLA genotyping methods where sequence analysis is restricted to HLA amplicons, a target of reduced complexity compared to shotgun sequence data, HLAminer interrogates the full diversity of sequence information in whole transcriptome, whole genome or whole exome datasets. Here, single base differences can be more easily missed due to factors such as low or unequally distributed sequence coverage and base errors. Thus, the performance of HLAminer for robust four-digit HLA allele calls is a limitation of the current data sets and performance is expected to improve as sequencing

Table 2 Effect of read length and base error on HLAminer predictions from targeted assembly of simulated RNA-Seq data[a]

HLA allele resolution	Base error (%)	Read length (nucleotides)	Sensitivity (mean ± SD%)	Specificity (mean ± SD%)	Ambiguity(mean ± SD%)
Two-digit	1.0	50	13.62 ± 2.80	92.86 ± 10.10	19.06 ± 16.53
		75	62.32 ± 3.62	90.27 ± 3.28	8.93 ± 2.67
		100	95.72 ± 0.53	96.31 ± 0.02	4.46 ± 3.84
		150	97.97 ± 2.80	95.73 ± 3.63	0.00 ± 0.00
Two-digit	0.5	100	93.04 ± 4.60	96.39 ± 1.44	2.91 ± 1.69
	1.0		95.72 ± 0.53	96.31 ± 0.02	4.46 ± 3.84
	2.0		64.64 ± 4.79	96.13 ± 1.43	13.40 ± 3.02
	3.0		6.67 ± 2.51	100.00 ± 0.00	8.59 ± 8.34
Four-digit	1.0	50	7.78 ± 1.92	60.51 ± 20.02	27.78 ± 4.81
		75	51.94 ± 2.93	77.36 ± 5.44	37.38 ± 11.16
		100	86.84 ± 1.75	88.13 ± 1.41	36.32 ± 4.76
		150	93.33 ± 3.63	93.07 ± 2.91	22.87 ± 2.40
Four-digit	0.5	100	84.72 ± 5.42	89.65 ± 2.25	24.03 ± 3.00
	1.0		86.84 ± 1.75	88.13 ± 1.41	36.32 ± 4.76
	2.0		56.94 ± 1.73	87.49 ± 4.77	39.14 ± 5.73
	3.0		4.44 ± 2.10	68.69 ± 17.03	37.22 ± 25.62

[a]In triplicate experiments, 5 million read pairs 50, 75, 100 or 150 nucleotides in length (top) and 100-nucleotide read pairs having 0.5, 1, 2 or 3% errors (bottom) were randomly generated from 20 sets of transcripts, each containing 6 randomly chosen reference HLA alleles. HLAminer predictions derived from targeted read assembly were compared to each reference set and the performance of HLAminer was assessed by measuring the specificity, sensitivity and ambiguity. SD, standard deviation.

technologies evolve to offer greater accuracy and read length at reduced cost.

Conclusions

HLAminer is the implementation of a strategy for automated HLA typing directly from NGS data sets. It is a fundamentally different approach compared to conventional methods that all rely on first amplifying HLA genes. The identification of allelic variants from an individual NGS data set by simple sequence search or alignment is precluded by the complexity of the locus, the massive allelic diversity in the population and limitations of short sequence reads to adequately capture these variations. The option of typing, retrospectively, existing cohorts for which NGS data have already been generated is enabling, particularly for large community resource projects [12-15]. In this context, the HLA info is value-added, as no additional cost is necessary to generate further HLA specific data from an existing data set. The method can also be applied prospectively. In fact, it may turn out to be the case that it is most efficient to do all HLA typing by shotgun sequencing, since these types of data sets are maximally informative and are becoming routine to generate.

It is recognized that certain HLA allele combinations are common in certain populations, presumably due to linkage disequilibrium [35]. For example the combination of HLA-I A*01; C*07; B*08 is common in some western European populations. These conserved extended haplotypes are not yet well represented in HLA databases, but in future we will explore the possibility of using this type of information to further improve computational HLA typing. We also expect to extend our approach to prediction of HLA class II alleles. HLAminer is available for public use [36].

Abbreviations

CDS: coding DNA sequence; CRC: colorectal cancer; HLA: human leukocyte antigen; NGS: next-generation sequencing; PCR: polymerase chain reaction; WGS: whole genome shotgun.

Acknowledgements

This work was supported by the Canadian Institutes of Health Research and Genome British Columbia.

Author details

[1]BC Cancer Agency, Michael Smith Genome Sciences Centre, Vancouver, British Columbia V5Z 1L3, Canada. [2]Department of Molecular Biology and Biochemistry, Simon Fraser University, Burnaby, British Columbia V5A 1S6, Canada.

Authors' contributions

RLW and RAH designed the research; DJF extracted RNA and constructed the sequencing libraries; SM and GC verified predicted HLA calls by conventional sequencing; MC provided ovarian cancer exon capture sequence datasets; RM coordinated the sequencing; RLW and RAH analyzed the data and made the figures; RLW and RAH wrote the manuscript. All authors read and approved the final manuscript.

Competing interests

The authors declare that they have no competing interests.

Received: 8 June 2012 Revised: 12 October 2012
Accepted: 10 December 2012 Published: 10 December 2012

References

1. Carrington M, O'Brien SJ: **The influence of HLA genotype on AIDS.** *Annu Rev Med* 2003, **54**:535-551.
2. Dawson DV, Ozgur M, Sari K, Ghanayem M, Kostyu DD: **Ramifications of HLA class I polymorphism and population genetics for vaccine development.** *Genet Epidemiol* 2001, **20**:87-106.
3. Fernando MM, Stevens CR, Walsh EC, De Jager PL, Goyette P, Plenge RM, Vyse TJ, Rioux JD: **Defining the role of the MHC in autoimmunity: a review and pooled analysis.** *PLoS Genet* 2008, **4**:e1000024.
4. Mizuki N, Meguro A, Ota M, Ohno S, Shiota T, Kawagoe T, Ito N, Kera J, Okada E, Yatsu K, Song YW, Lee EB, Kitaichi N, Namba K, Horie Y, Takeno M, Sugita S, Mochizuki M, Bahram S, Ishigatsubo Y, Inoko H: **Genome-wide association studies identify IL23R-IL12RB2 and IL10 as Behçet's disease susceptibility loci.** *Nat Genet* 2010, **42**:703-706.
5. Rioux JD, Goyette P, Vyse TJ, Hammarström L, Fernando MM, Green T, De Jager PL, Foisy S, Wang J, de Bakker PI, Leslie S, McVean G, Padyukov L, Alfredsson L, Annese V, Hafler DA, Pan-Hammarström Q, Matell R, Sawcer SJ, Compston AD, Cree BA, Mirel DB, Daly MJ, Behrens TW, Klareskog L, Gregersen PK, Oksenberg JR, Hauser SL: **Mapping of multiple susceptibility variants within the MHC region for 7 immune-mediated diseases.** *Proc Natl Acad Sci USA* 2009, **106**:18680-18685.
6. Ryder LP, Svejgaard A, Dausset J: **Genetics of HLA disease association.** *Annu Rev Genet* 1981, **15**:169-187.
7. Shugart YY, Wang Y, Jia WH, Zeng YX: **GWAS signals across the HLA regions: revealing a clue for common etiology underlying infectious tumors and other immunity diseases.** *Chin J Cancer* 2011, **30**:226-230.
8. Wang C, Krishnakumar S, Wilhelmy J, Babrzadeh F, Stepanyan L, Su LF, Levinson D, Fernandez-Viña MA, Davis RW, Davis MM, Mindrinos MN: **High-throughput, high-fidelity HLA genotyping with deep sequencing.** *Proc Natl Acad Sci USA* 2012, **109**:8676-8681.
9. Erlich RL, Jia X, Anderson S, Banks E, Gao X, Carrington M, Gupta N, DePristo MA, Henn MR, Lennon NJ, de Bakker PI: **Next-generation sequencing for HLA typing of class I loci.** *BMC Genomics* 2011, **12**:42.
10. International Human Genome Sequencing Consortium: **Finishing the euchromatic sequence of the human genome.** *Nature* 2004, **431**:931-945.
11. Levy S, Sutton G, Ng PC, Feuk L, Halpern AL, Walenz BP, Axelrod N, Huang J, Kirkness EF, Denisov G, Lin Y, MacDonald JR, Pang AW, Shago M, Stockwell TB, Tsiamouri A, Bafna V, Bansal V, Kravitz SA, Busam DA, Beeson KY, McIntosh TC, Remington KA, Abril JF, Gill J, Borman J, Rogers YH, Frazier ME, Scherer SW, Strausberg RL, Venter JC: **The Diploid Genome Sequence of an Individual Human.** *PLoS Biol* 2007, **5**:e254.
12. Cancer Genome Atlas Research Network, McLendon R, Friedman A, Bigner D, Van Meir EG, Brat DJ, Mastrogianakis GM, Olson JJ, Mikkelsen T, Lehman N, Aldape K, Yung WK, Bogler O, Weinstein JN, VandenBerg S, Berger M, Prados M, Muzny D, Morgan M, Scherer S, Sabo A, Nazareth L, Lewis L, Hall O, Zhu Y, Ren Y, Alvi O, Yao J, Hawes A, Jhangiani S, *et al*: **Comprehensive genomic characterization defines human glioblastoma genes and core pathways.** *Nature* 2008, **455**:1061-1068.
13. Human Microbiome Jumpstart Reference Strains Consortium, Nelson KE, Weinstock GM, Highlander SK, Worley KC, Creasy HH, Wortman JR, Rusch DB, Mitreva M, Sodergren E, Chinwalla AT, Feldgarden M, Gevers D, Haas BJ, Madupu R, Ward DV, Birren BW, Gibbs RA, Methe B, Petrosino JF, Strausberg RL, Sutton GG, White OR, Wilson RK, Durkin S, Giglio MG, Gujja S, Howarth C, Kodira CD, Kyrpides N, *et al*: **A catalog of reference genomes from the human microbiome.** *Science* 2010, **328**:994-999.
14. International Cancer Genome Consortium, Hudson TJ, Anderson W, Artez A, Barker AD, Bell C, Bernabé RR, Bhan MK, Calvo F, Eerola I, Gerhard DS, Guttmacher A, Guyer M, Hemsley FM, Jennings JL, Kerr D, Klatt P, Kolar P, Kusada J, Lane DP, Laplace F, Youyong L, Nettekoven G, Ozenberger B, Peterson J, Rao TS, Remacle J, Schafer AJ, Shibata T, Stratton MR, *et al*: **International network of cancer genome projects.** *Nature* 2010, **464**:993-998.
15. 1000 Genomes Project Consortium, Altshuler D, Durbin RM, Abecasis GR, Bentley DR, Chakravarti A, Clark AG, Collins FS, De La Vega FM, Donnelly P, Egholm M, Flicek P, Gabriel SB, Gibbs RA, Knoppers BM, Lander ES, Lehrach H, Mardis ER, McVean GA, Nickerson DA, Peltonen L, Schafer AJ, Sherry ST, Wang J, Wilson R, Gibbs RA, Deiros D, Metzker M, Muzny D, Reid J, *et al*: **A map of human genome variation from population-scale sequencing.** *Nature* 2010, **467**:1061-1073.
16. Castellarin M, Warren RL, Freeman JD, Dreolini L, Krzywinski M, Strauss J, Barnes R, Watson P, Allen-Vercoe E, Moore RA, Holt RA: **Fusobacterium nucleatum infection is prevalent in human colorectal carcinoma.** *Genome Res* 2012, **22**:299-306.
17. Morin RD, Mendez-Lago M, Mungall AJ, Goya R, Mungall KL, Corbett RD, Johnson NA, Severson TM, Chiu R, Field M, Jackman S, Krzywinski M, Scott DW, Trinh DL, Tamura-Wells J, Li S, Firme MR, Rogic S, Griffith M, Chan S, Yakovenko O, Meyer IM, Zhao EY, Smailus D, Moksa M, Chittaranjan S, Rimsza L, Brooks-Wilson A, Spinelli JJ, Ben-Neriah S, *et al*: **Frequent mutation of histone-modifying genes in non-Hodgkin lymphoma.** *Nature* 2011, **476**:298-303.
18. **NCBI Sequence Read Archive..** [http://www.ncbi.nlm.nih.gov/Traces/sra/sra.cgi].
19. **Sample libraries..** [ftp://ftp.bcgsc.ca/supplementary/CRC2012/].
20. Warren RL, Freeman JD, Zeng T, Choe G, Munro S, Moore R, Webb JR, Holt RA: **Exhaustive T-cell repertoire sequencing of human peripheral blood samples reveals signatures of antigen selection and a directly measured repertoire size of at least 1 million clonotypes.** *Genome Res* 2011, **21**:790-797.
21. **HLA CDS and genomic sequences..** [ftp://ftp.ebi.ac.uk/pub/databases/imgt/mhc/hla/].
22. Robinson J, Waller MJ, Fail SC, McWilliam H, Lopez R, Parham P, Marsh SG: **The IMGT/HLA database.** *Nucleic Acids Res* 2009, **37**:D1013-D1017.
23. Warren RL, Holt RA: **Targeted assembly of short sequence reads.** *PLoS ONE* 2011, **6**:e19816.
24. Warren RL, Sutton GG, Jones SJ, Holt RA: **Assembling millions of short DNA sequences using SSAKE.** *Bioinformatics* 2007, **23**:500-501.
25. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: **Basic local alignment search tool.** *J Mol Biol* 1990, **215**:403-410.
26. Flicek P, Amode MR, Barrell D, Beal K, Brent S, Chen Y, Clapham P, Coates G, Fairley S, Fitzgerald S, Gordon L, Hendrix M, Hourlier T, Johnson N, Kähäri A, Keefe D, Keenan S, Kinsella R, Kokocinski F, Kulesha E, Larsson P, Longden I, McLaren W, Overduin B, Pritchard B, Riat HS, Rios D, Ritchie GR, Ruffier M, Schuster M, *et al*: **Ensembl 2011.** *Nucleic Acids Res* 2011, **39**:D800-D806.
27. Gnirke A, Melnikov A, Maguire J, Rogov P, LeProust EM, Brockman W, Fennell T, Giannoukos G, Fisher S, Russ C, Gabriel S, Jaffe DB, Lander ES, Nusbaum C: **Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing.** *Nat Biotechnol* 2009, **27**:182-189.
28. **SureSelect Target Enrichment..** [https://earray.chem.agilent.com/earray/].
29. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R: **The Sequence Alignment/Map format and SAMtools.** *Bioinformatics* 2009, **25**:2078-2079.
30. Li H, Durbin R: **Fast and accurate short read alignment with Burrows-Wheeler transform.** *Bioinformatics* 2009, **25**:1754-1760.
31. Cereb N, Maye P, Lee S, Kong Y, Yang SY: **Locus-specific amplification of HLA class I genes from genomic DNA: Locus-specific sequences in the first and third introns of HLA-A, -B, and -C alleles.** *Tissue Antigens* 1995, **45**:1-11.
32. **Phred/Phrap/Consed..** [http://www.phrap.org].
33. Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, Valentin F, Wallace IM, Wilm A, Lopez R, Thompson JD, Gibson TJ, Higgins DG: **Clustal W and Clustal × version 2.0.** *Bioinformatics* 2007, **23**:2947-2948.
34. Marsh SG, Albert ED, Bodmer WF, Bontrop RE, Dupont B, Erlich HA, Fernández-Viña M, Geraghty DE, Holdsworth R, Hurley CK, Lau M, Lee KW, Mach B, Maiers M, Mayr WR, Müller CR, Parham P, Petersdorf EW, Sasazuki T, Strominger JL, Svejgaard A, Terasaki PI, Tiercy JM, Trowsdale J: **Nomenclature for factors of the HLA system, 2010.** *Tissue Antigens* 2010, **75**:291-455.
35. de Bakker PI, McVean G, Sabeti PC, Miretti MM, Green T, Marchini J, Ke X, Monsuur AJ, Whittaker P, Delgado M, Morrison J, Richardson A, Walsh EC, Gao X, Galver L, Hart J, Hafler DA, Pericak-Vance M, Todd JA, Daly MJ, Trowsdale J, Wijmenga C, Vyse TJ, Beck S, Murray SS, Carrington M, Gregory S, Deloukas P, Rioux JD: **A high-resolution HLA and SNP haplotype map for disease association studies in the extended human MHC.** *Nat Genet* 2006, **38**:1166-1172.
36. **HLAminer..** [http://www.bcgsc.ca/platform/bioinfo/software/hlaminer/].

doi:10.1186/gm396
Cite this article as: Warren *et al.*: Derivation of HLA types from shotgun sequence datasets. *Genome Medicine* 2012 **4**:95.