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