Benchmarking the performance of human antibody gene alignment utilities using a 454 sequence dataset

Katherine J.L. Jackson 1,2,*, Scott Boyd 3, Bruno A. Gaëta 1 and Andrew M. Collins 2

Schools of 1Computer Science and Engineering, and 2 Biotechnology and Biomolecular Sciences, University of New South Wales, Sydney, NSW, 2052, Australia.
3Department of Pathology, Stanford University, Stanford, CA, 94305, USA.

Associate Editor: Prof. John Quackenbush

ABSTRACT

Motivation: Immunoglobulin heavy chain genes are formed by recombination of genes randomly selected from sets of IGHV, IGHD and IGHJ genes. Utilities have been developed to identify genes that contribute to observed VDJ rearrangements, but in the absence of datasets of known rearrangements, the evaluation of these utilities is problematic. We have analysed thousands of VDJ rearrangements from an individual (S22) whose IGHV, IGHD and IGHJ genotype can be inferred from the dataset. Knowledge of this genotype means the Stanford_S22 dataset can serve to benchmark the performance of IGH alignment utilities.

Results: We evaluated the performance of 7 utilities. Failure to partition a sequence into genes present in the S22 genome was considered an error, and error rates for different utilities ranged from 7.1% to 13.7%.

Availability: Supplementary data includes the S22 genotypes and alignments. The Stanford_S22 dataset and an evaluation tool is available at http://www.emi.unsw.edu.au/~ihmmune/IGHUtilityEval/.

Contact: katherine.jackson@unsw.edu.au

1 INTRODUCTION

Effective responses to the antigenic diversity of the microbial world require the human immune system to produce antibodies with specificities of similar diversity. The genetic diversity necessary to encode such an antibody repertoire is generated by recombination, for immunoglobulin genes are created by the joining of a number of genes. As each B cell develops within the bone marrow, the rearranged immunoglobulin heavy chain (IGH) variable region genes are created by essentially random recombination of single IGHV (V), IGHD (D), and IGHJ (J) genes, each selected from sets of germ-line genes (Jung et al., 2006). During the recombination process, the recombining gene ends can be trimmed, and variable numbers of non-template encoded nucleotides may be inserted between the recombining genes. Additional diversity is generated during an immune response by the process of somatic hypermutation (SHM) (Neuberger, 2008), which can lead a rearranged VDJ gene to accumulate in excess of 30 mutations within its ~370 nucleotides. The processes of VDJ recombination and SHM, when combined with the high similarity of many of the genes, lead to difficulty in identifying the germline contributions to IGH rearrangements. This is particularly true of the D genes, which range in length from just 11 to 37 nucleotides.

A number of utilities have been developed to identify the germline origins of VDJ rearrangements. We have previously used sets of clonally-related sequences to evaluate such utilities (Gaëta et al., 2007). This provided a useful test, but as rearrangements with differing characteristics (long D, short D, highly mutated, etc.) present different challenges for the utilities, larger and more varied datasets are required for proper evaluation.

Here we present an evaluation of IGH alignment utilities based on alignment of large 454 sequencing datasets from single individuals. The method is explored using the Stanford_S22 dataset. The many thousands of VDJ rearrangements in this dataset allow the individual genotype at the IGH variable gene loci to be inferred (Boyd et al., 2010). By assessing the number of alignments made to genes that are absent from the individual genotype using our evaluation tool, such datasets of rearrangements can be used for evaluation and improvement of IGH alignment utilities.

2 METHODS

The Stanford_S22 dataset of VDJ sequences was produced by 454 pyrosequencing of genomic DNA derived from peripheral blood mononuclear cells of a single individual, as previously described (Boyd et al., 2009). Apparent chimeric sequences (Brakenhoff et al., 1991) and those containing base-pair insertions and deletions were identified and excluded. Chimeric sequences were identified by BLAST alignments against the germline V repertoire. Chimeric sequences had disparate 5‘ and 3‘ portions of their rearranged V genes. Clonally-related sets were also identified and a single representative sequence from each set was retained in the dataset.

The dataset of 13,153 sequences was aligned using iHMMune-align (Gaëta et al., 2007), IMGT/V-QUEST+JCTA (Brochet et al., 2008), JOIN-SOLVER (Souto-Carneiro et al., 2004), SoDA (Volpe et al., 2006), VDJsolver (Ohm-Laursen et al., 2006), Ab-origin (Wang et al., 2008a) and NCBI’s IgBLAST. All alignments were completed using each utility’s default parameters and germline repertoires. The iHMMune-align repertoire includes a number of putative polymorphisms not present in the repertoires of other utilities. The S22 genotype was then determined as previously described.

*to whom correspondence should be addressed

© The Author (2010). Published by Oxford University Press. All rights reserved. For Permissions, please email: journals.permissions@oxfordjournals.org

Errors involving an incorrect gene, rather than an incorrect allelic variant, are shown in Table 1.

Table 1. The percentage of alignments from the StanfordS22 dataset that were made, by various utilities, to IGHV, IGHD and IGHJ genes and allelic variants absent from the S22 genotype.

<table>
<thead>
<tr>
<th>Utility</th>
<th>IGHV (%)</th>
<th>IGHD (%)</th>
<th>IGHJ (%)</th>
<th>Total (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>iHMMune-align</td>
<td>3.21 (0.21)</td>
<td>2.21 (1.27)</td>
<td>1.95 (0.0)</td>
<td>7.11</td>
</tr>
<tr>
<td>IMGT</td>
<td>4.90 (0.22)</td>
<td>5.09 (2.81)</td>
<td>1.55 (0.0)</td>
<td>10.87</td>
</tr>
<tr>
<td>IgBLAST</td>
<td>3.84 (0.75)</td>
<td>3.92 (2.16)</td>
<td>0.85 (0.0)</td>
<td>8.39</td>
</tr>
<tr>
<td>Ab-origin</td>
<td>4.06 (0.22)</td>
<td>7.94 (5.53)</td>
<td>2.53 (0.0)</td>
<td>13.74</td>
</tr>
<tr>
<td>JOINSOLVER</td>
<td>6.17 (0.86)</td>
<td>6.59 (4.92)</td>
<td>1.24 (0.0)</td>
<td>8.25</td>
</tr>
<tr>
<td>SoDA</td>
<td>2.68 (0.29)</td>
<td>6.82 (6.63)</td>
<td>1.50 (0.0)</td>
<td>10.37</td>
</tr>
<tr>
<td>VDJSolver</td>
<td>6.87 (0.48)</td>
<td>1.96 (0.79)</td>
<td>0.71 (0.0)</td>
<td>9.09</td>
</tr>
</tbody>
</table>

* Percentage of sequences that include an incorrect gene or allele for either the V, D or J (Boyd et al., 2010), by independent analysis of results from iHMMune-align, IMGT/V-QUEST+JCTA and IgBLAST.

The percentages of V, D and J alignments made to genes and alleles absent from the S22 genotype were then determined for each utility. As the S22 genotype includes putative polymorphisms that are only found in the iHMMune-align repertoire, alignments of sequences containing such polymorphisms were considered correct if a utility identified the most similar allele within its repertoire. Although IGHD5-5 may be missing from the S22 genome, alignments to IGHD5-5 were accepted as correct, as IGHD5-5 and IGHD5-18 have identical coding regions. For utilities that report a number of possible V, D and J genes, only the highest scoring genes were included in the analysis. If a number of genes achieved the same high score, and these equally high scoring genes included an S22 gene, then the alignments were accepted as correct. For utilities that reported a single possible gene, or where there was only a single high scoring gene, if this gene was absent from the S22 genome the alignment was considered incorrect. Alignments to inverse D genes were considered incorrect given evidence challenging their involvement in the generation of diversity (Corbett et al., 1997), and given that review of the SoDA inverse D alignments demonstrated them to be invariably short and improbable (data not shown).

3 DISCUSSION

No VDJ rearrangement can be partitioned with absolute certainty, but any alignment made to genes that are absent from the genome of the individual under study must be in error. This study did not attempt to identify all incorrect alignments, but rather it detected errors by reference to an individual’s inferred genotype, providing a measure of the minimum frequency of misalignments for each utility. The S22 genotype was independently determined using three utilities to report a number of possible V, D or J assignments of the IGH genes for iHMMune-align, JOINSOLVER and IgBLAST. This approach should now allow the performance of all utilities to be further refined.

4 CONCLUSION

Analysis using the StanfordS22 dataset demonstrates how 454 sequence datasets can be used to evaluate the performance of IGH alignment utilities. Analysis found the fewest definitive misassignments of the IGH genes for iHMMune-align, JOINSOLVER and IgBLAST. This approach should now allow the performance of all utilities to be further refined.

REFERENCES

Ohn-Laursen, L., et al. (2006) No evidence for the use of DIR, D-D fusions, chromosome 15 open reading frames or VH replacement in the peripheral repertoire was found on application of an improved algorithm, JointML, to 6329 human immunoglobulin H rearrangements, *Immunology*, **119**, 265-277.