BIOINFORMATICS APPLICATIONS NOTE Vol. 26 no. 24 2010, pages 3129–3130
doi:10.1093/bioinformatics/btq604

Sequence analysis

Advance Access publication October 29, 2010

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]

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

Associate Editor: 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 analyzed 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 that the Stanford_S22 dataset can serve to benchmark the performance of IGH alignment utilities.

Results: We evaluated the performance of seven 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

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

Received on May 27, 2010; revised on October 5, 2010; accepted on October 21, 2010

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 germline 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 of around 370 nts to accumulate in excess of 30 point mutations. 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 11–37 nt.

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), JOINSOLVER (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 (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

*To whom correspondence should be addressed.

Table 1. The percentage of alignments from the Stanford_S22 dataset that were made, by various utilities, to IGHV, IGHD and IGHJ genes and allelic variants absent from the S22 genotype

Utility	IGHV (%)[a]	IGHD (%)[a]	IGHJ (%)[a]	Total (%)[b]
iHMMune-align	3.21 (0.21)	2.21 (1.27)	1.95 (0.0)	7.11
IMGT/VQ+JA	4.90 (0.22)	5.09 (2.81)	1.55 (0.0)	10.87
IgBLAST	3.84 (0.75)	3.96 (2.16)	0.85 (0.0)	8.39
Ab-origin	4.06 (0.22)	7.94 (5.53)	2.53 (0.0)	13.74
JOINSOLVER	6.17 (0.86)	6.93 (4.92)	1.24 (0.0)	7.89
SoDA	2.68 (0.29)	6.82 (6.63)	1.50 (0.0)	10.37
VDJSolver	6.87 (0.48)	1.96 (0.79)	0.71 (0.0)	9.09

[a]Errors involving an incorrect gene, rather than an incorrect allelic variant, are shown in brackets.
[b]Percentage of sequences that include an incorrect gene or allele for either the V, D or J.

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 the 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. These genotypes were found to be in agreement, with the exception of minor differences resulting from the presence of allelic variants in the S22 genotype that are not present in all repertoires, and misalignment of very infrequently rearranged genes. The inferred rearrangeable V, D and J genes of the S22 genome are available as Supplementary Tables S1, S2 and S3, respectively. S22 carries an apparent homozygous deletion of six contiguous D genes (Boyd *et al*., 2010) and is homozygous at the five other D loci where allelic variants exist. The IGHD6-25 gene, which we have previously suggested is non-functional (Lee *et al*., 2006), was also absent from the S22 genome. These features make the S22 dataset particularly useful for the evaluation of utility performance on D gene alignment. The alignment of D genes is one of the most difficult challenges in the partitioning of VDJ rearrangements and the nature of the challenge varies between D gene families.

The output from each utility was compared with the S22 genotype, and all errors were noted. The results for each utility are summarised as Table 1 and detailed as Supplementary Tables S4–S10. Overall, errors were made in the alignment of between 7.1% (iHMMune-align) and 13.7% (Ab-origin) of all sequences. Six unofficial V polymorphisms, previously inferred from sequence analysis (Boyd *et al*., 2010; Wang *et al*., 2008b), are seen in the S22 genome. These polymorphisms were present in 9.4% of all S22 VDJ genes, according to the iHMMune-align analysis. While failures to identify these polymorphisms were not scored as errors here, unless alignment utility repertoires incorporate these newly identified polymorphisms, an additional source of substantial error will remain.

4 CONCLUSION

Analysis using the Stanford_S22 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.

Conflict of Interest: none declared.

REFERENCES

Boyd,S.D. *et al*. (2010) Individual variation in the germline immunoglobulin gene repertoire inferred from VDJ rearrangements. *J. Immunol.*, **184**, 6986–6992.

Boyd,S.D. *et al*. (2009) Measurement and clinical monitoring of human lymphocyte clonality by massively parallel VDJ pyrosequencing. *Sci. Transl. Med.*, **1**, 12ra23.

Brakenhoff,R.H. *et al*. (1991) Chimeric cDNA clones: a novel PCR artifact. *Nucleic Acids Res.*, **19**, 1949.

Brochet,X. *et al*. (2008) IMGT/V-QUEST: the highly customized and integrated system for IG and TR standardized V-J and V-D-J sequence analysis. *Nucleic Acids Res*, **36**, W503–W508.

Corbett,S.J. *et al*. (1997) Sequence of the human immunoglobulin diversity (D) segment locus: a systematic analysis provides no evidence for the use of DIR segments, inverted D segments, 'minor' D segments or D-D recombination. *J. Mol. Biol.*, **270**, 587–597.

Gaëta,B. *et al*. (2007) iHMMune-align: Hidden Markov model-based alignment and identification of germline segments in immunoglobulin gene sequences. *Bioinformatics*, **23**, 1580–1587.

Jung,D. *et al*. (2006) Mechanism and control of V(D)J recombination at the immunoglobulin heavy chain locus. *Annu. Rev. Immunol.*, **24**, 541–570.

Lee,C.E.H. *et al*. (2006) Reconsidering the human heavy chain gene locus. 1.An evaluation of the expressed human IGHD gene repertoire. *Immunogenetics.*, **57**, 917–925.

Neuberger,M.S. (2008) Antibody diversification by somatic mutation: from Burnet onwards. *Immunol. Cell Biol.*, **86**, 124–132.

Ohm-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.

Souto-Carneiro,M.M. *et al*. (2004) Characterization of the human Ig heavy chain antigen binding complementarity determining region 3 using a newly developed software algorithm, JOINSOLVER. *J. Immunol.*, **172**, 6790–6802.

Volpe,J.M. *et al*. (2006) SoDA: implementation of a 3D alignment algorithm for inference of antigen receptor recombinations. *Bioinformatics*, **22**, 438–444.

Wang,X. *et al*. (2008a) Ab-origin: an enhanced tool to identify the sourcing gene segments in germline for rearranged antibodies. *BMC Bioinformatics*, **9** (Suppl. 12), S20.

Wang,Y. *et al*. (2008b) Many human immunoglobulin heavy-chain IGHV gene polymorphisms have been reported in error. Immunol. *Cell Biol.*, **86**, 111–115.