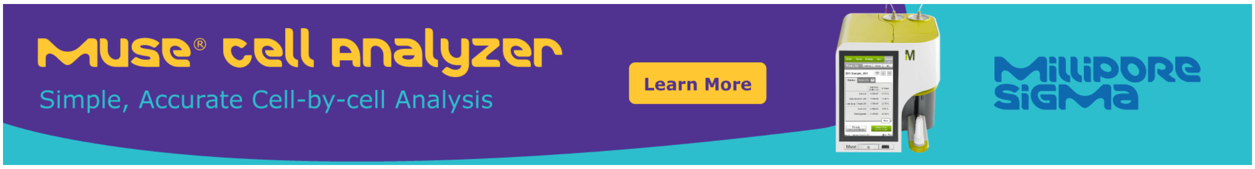

The Journal of Immunology

Comment on ''A Database of Human Immune Receptor Alleles Recovered from Population Sequencing Data''

Corey T. Watson, Frederick A. Matsen IV, Katherine J. L. Jackson, Ali Bashir, Melissa Laird Smith, Jacob Glanville, Felix Breden, Steven H. Kleinstein, Andrew M. Collins and Christian E. Busse

J Immunol 2017; 198:3371-3373; ;
doi: 10.4049/jimmunol.1700306
http://www.jimmunol.org/content/198/9/3371

The Journal of Immunology is published twice each month by
The American Association of Immunologists, Inc.,
1451 Rockville Pike, Suite 650, Rockville, MD 20852

Print ISSN: 0022-1767 Online ISSN: 1550-6606.

The Journal of Immunology

Comment on "A Database of Human Immune Receptor Alleles Recovered from Population Sequencing Data"

It was with great interest that we read the recently published article by Yu et al. (1), which proposes a solution to the problem of building complete and accurate databases of germline Ig (IG) and TCR genes and alleles. This highlights one of the most formidable challenges in the immunogenetics field, as it has become apparent in recent years that existing germline databases (GLDB) are neither complete (i.e., lack existing alleles) (2–7) nor accurate in some cases (i.e., contain nonexisting alleles) (8). The impacts of this problem have most prominently come into focus in the context of IG/TCR expressed repertoire sequence datasets, the analysis and interpretation of which critically depend on the use of accurate GLDBs. Indeed, recent GLDB improvements via the inclusion of previously undetected IG alleles in repertoire sequence analysis have demonstrated the potential for direct consequences on human health research (9). In this comment, we enumerate some difficulties inherent in employing the data used by Yu et al. (1) to build a GLDB, and argue that a broad-based collaborative effort using a variety of data types is needed to achieve the goal of a complete yet reliable GLDB.

In their article (1), the authors develop a pipeline for identifying novel IG/TCR alleles from single nucleotide polymorphism (SNP) genotype data, and apply it to diverse population samples of the 1000 Genomes Project (G1K) (10–12) to build "Lym1K," a GLDB covering the human TCR β (TRB), TCR α (TRA), IGH, and IGK and IGL, summarized as "IGL" chain loci. Across the variable (V), diversity (D) and joining (J) genes in these loci, the authors report the discovery of 8750 germline alleles not currently curated in the international ImMunoGeneTics information system (13, 14). At face value, this finding is profound, and suggests potential for augmenting IG/TCR GLDBs using existing or newly generated genotype data. However, we are concerned about the accuracy of the underlying data, and the fact that erroneous genotypes/haplotypes used as input will result in incorrectly inferred IG/TCR alleles. We argue that users of such an approach should exercise due caution.

There are at least three potential caveats concerning the use of G1K data and similar short-read sequencing datasets for variant discovery, genotyping, and the downstream inference of novel IG/TCR alleles. These include:

1) the repetitive nature and structural complexity of IG and TCR loci;
2) the unknown extent of haplotype diversity and prevalence of large copy number variants (CNVs) involving genes in these regions; and
3) the use of source material derived from immortalized B cell lines.

The variant calls produced by G1K are only as reliable as the underlying short-read sequencing technologies used. Mapping of short-read data (for G1K Phase 3, reads can be as short as 70 bp) can be confounded in complex genomic loci (15–17), such as IG and TCR, which are characterized by a highly repetitive sequence architecture and extensive haplotype diversity (5, 18–23). Each of the IG and TCR loci consist of ∼40 or more phylogenetically related functional/open reading frame V, D (in IGH and TRB), and J genes, which exhibit high sequence homology that in some cases can reach 100% (e.g., for alleles at *IGHV3-30* and related paralogs) (13, 14). Importantly, due to the fact that germline allele databases are incomplete, the degree of "allele sharing" between genes within IG/TCR loci is not fully understood. This would be expected to be a serious issue in the IGK locus, as nearly every V gene resides within two tandem duplication blocks, between which direct gene conversion events have been described (18). The repetitive nature of sequences in these loci creates the potential for mismapping of reads and ultimate assignment of variants to the incorrect genes.

A second critical consideration is that variant calls made using standard short-read data and bioinformatics pipelines are restricted to loci present in the genome reference assembly used for read mapping. This is important, as haplotype variability, in the form of large CNVs and SNPs is common in the IG and TCR loci (5, 18–23). Therefore, a single reference assembly poorly represents standing haplotype variation in any given population being screened. As noted by Yu et al. (1), there are in fact many genes missing from the current reference assemblies (e.g., GRCh37 and GRCh38), and thus by definition, it is impossible to make reliable genotype and allele calls for these genes. In total, using IGH as an example, there are at least 16 known functional/open reading frame V genes and >220 kbp of genomic sequence present in haplotypes in the human population that are not represented in GRCh37 (the assembly used by G1K for Phase 3 read mapping and variant calling) (5). In some populations, these alternate haplotypes represent the major allele, indicating that the majority of samples screened would carry a sequence absent from GRCh37 (5). The technical effects of this "missing sequence" are not known, but it would be expected that reads representing alternate haplotypes and genes in any given individual would have the potential to be incorrectly mapped to off-target genes that are present in the reference. Directly related to this, the presence of CNVs in a sample can cause other problems for short-read mapping and downstream genotype inference. For example, heterozygous gene deletions (hemizygotes) can masquerade as homozygotes for a given SNP or coding allele, whereas paralogous sequence variants between close gene duplicates can result in artifactual heterozygote calls (24, 25).

Furthermore, it is important to take the source of genomic DNA used by a study into account. In the case of G1K, DNA was extracted from lymphoblastoid cell lines, i.e., B cells immortalized by EBV. Therefore, a fraction of the IG loci in these lines has undergone V(D)J recombination, which can lead to a reduction or complete loss of reads (lower read depth) overlapping proximal V, D, and distal J genes within a given sequencing library (5, 26). Low read coverage can directly impact the reliability of variant discovery and genotype calling (11, 12). Additionally, hypermutated memory B cells can be the target of EBV transformation (27), which will result in the presence of non-germline IG gene mutations in DNA isolated from lymphoblastoid cell lines, resulting in potential false-positive allele calls; evidence of somatic hypermutation at IG genes that have undergone V(D)J recombination has been directly observed in G1K samples (5). Although not directly applicable to G1K data, it should also be noted that similar issues concerning the reliability of genotyping due to somatic rearrangements have been reported for TCR loci using DNA isolated from blood (28). Requiring variants to be present in multiple individuals or conducting analyses in family-based datasets could potentially help mitigate this issue, but the reliability of such an approach would need to be demonstrated, as somatic mutations at hot-spots likely recur.

Unfortunately, because population and genomic resources in the IG and TCR gene regions remain limited, the true impacts of the potential caveats laid out above remain difficult to assess. However, as part of the Phase 3 data release, G1K has used quality control metrics from low-coverage data across >2,600 human samples to directly assess the “accessibility” of every base in the genome to sequencing technologies used currently by the consortium [see Refs. (12) and (29)]. Using this approach, certain bases have been masked as having potentially higher false-positive and -negative variant call rates. Using IGH as an example, >25% of bases within the coding exons of 62/83 IGHV, D, and J genes in GRCh38 fall within this category, even when using the least stringent (“pilot”) criteria established by G1K. Although this does not by definition mean calls made in these regions are incorrect, we would argue it implies that their reliability is difficult to assess at this time. Indeed, G1K found that variant calls at these masked bases also had higher failure rates using alternative variant discovery/genotyping methods (12).

Taken together, the caveats discussed above suggest that databases constructed from alleles inferred from short-read genomic data should be carefully vetted, bearing in mind that even a single incorrect genotype within an IG/TCR gene can impact the reliability of haplotype phasing and allele inference for that gene. Therefore, we urge users to critically examine and consider both the features of the data underlying a given allele call, such as read lengths, coverage depth, library construction methods, cohort sample size, and the source of DNA, as well as the bioinformatics methods and the genic and sequence content of the genome reference assembly used for read mapping and variant calling. It is likely that all of these will impact the reliability of the allele database constructed, and importantly, may be more or less critical depending on the locus or gene/allele in question.

Finally, in addition to understanding factors related to the underlying data used, systems for thorough validation and benchmarking should be implemented to ensure low error rates. Such efforts have proven critical for the development of allele calling and genotyping methods using short-read data in other immune loci of comparable complexity (e.g., KIR and HLA) (30, 31). A basic cross-referencing of variant calls to other databases may be a useful strategy in certain circumstances, but would be expected to be problematic if variants in that database are not mutually exclusive from the variant call set used for allele inference. For example, dbSNP (32), used by Yu et al. (1) for filtering of calls from their pipeline, contains SNPs directly submitted by G1K, and thus an overlap of G1K IG/TCR variants and dbSNP would be expected, not offering an unbiased form of validation. Furthermore, if a database cross-referencing approach is used, the secondary database must be reliable, and may itself require careful filtering. For example, there are 62 G1K SNPs across 24/44 IGHV genes (GRCh37) that are cataloged by dbSNP, but are flagged as “suspect” variants potentially representing false-positives.

We hope that this debate can motivate a concerted effort on the part of our community to find sustainable strategies to improve and complement the current IG/TCR GLDBs. Over the coming years, in addition to population genome sequencing efforts by short-read platforms, data from long-read technologies and inferred alleles from expressed repertoire sequencing efforts will become generally available. It is clear that a multitude of approaches can and will be taken to create reference GLDBs from these data, but we should recognize that the quality of a GLDB cannot be measured by its allele count. Instead, we consider it to be the most productive path to set our current focus on the creation of GLDBs containing high-confidence, independently confirmed genes/alleles, even if stringent confirmation requirements result in the exclusion of rare alleles. In addition, the community should strive to develop statistics that describe the uncertainty associated with an individual allele to provide a transparent measure for users. Ultimately, however, it is worth considering that studies requiring the precise germline sequence of a specific donor may necessitate direct sequencing of the individual, instead of relying on a reference database. Ideally, in line with the principles of the Reproducible Research Standard (33), both databases and their underlying datasets should be available under a free and open licensing scheme to facilitate further development. Finally, it is important to note that the issues discussed here are not limited to human GLDBs, and will apply to other species, including murine and nonhuman primate models (7, 34). We are convinced that a community effort toward achieving these goals has the potential to greatly enhance the analysis of repertoire sequencing studies across the field and provide more detailed and reliable insights into adaptive immune responses in the context of infection, autoimmunity, and malignancies.

Corey T. Watson,[*] Frederick A. Matsen, IV,[†]
Katherine J. L. Jackson,[‡] Ali Bashir,[§,¶] Melissa
Laird Smith,[§,¶] Jacob Glanville,[‖,#] Felix Breden,[**]
Steven H. Kleinstein,[††,‡‡] Andrew M. Collins,[§§] and
Christian E. Busse[¶¶]

[*]Department of Biochemistry and Molecular Genetics, University of Louisville School of Medicine, Louisville, KY 40202; [†]Fred Hutchinson Cancer Research Center, Seattle, WA 98109; [‡]Immunogenomics Laboratory, Immunology Division, Garvan Institute of Medical Research, Sydney, 2010 New South Wales, Australia; [§]Icahn Institute for Genomics and Multiscale Biology, Icahn School of Medicine at Mount Sinai, New York, NY 10029;

[¶]Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, NY 10029; [‖]Institute for Immunity, Transplantation, and Immunology, Stanford University School of Medicine, Stanford, CA 94305; [#]Computational and Systems Immunology, Stanford University School of Medicine, Stanford, CA 94305; [**]Department of Biological Sciences, Simon Fraser University, Burnaby, British Columbia, Canada V5A 1S6; [††]Interdepartmental Program in Computational Biology and Bioinformatics, Yale University, New Haven, CT 06511; [‡‡]Department of Pathology, Yale School of Medicine, New Haven, CT 06520; [§§]School of Biotechnology and Biomolecular Sciences, University of New South Wales, Sydney, 2052 New South Wales, Australia; and [¶¶]Division of B Cell Immunology, German Cancer Research Center, 69120 Heidelberg, Germany

ORCIDs: 0000-0003-0607-6025 (F.A.M.); 0000-0002-9952-7069 (K.J.L.J.); 0000-0001-7151-3207 (M.L.S.); 0000-0001-9762-6314 (F.B.); 0000-0003-4957-1544 (S.H.K.); 0000-0001-7553-905X (C.E.B.).

Address correspondence and reprint requests to Dr. Corey T. Watson or Dr. Christian E. Busse, University of Louisville School of Medicine, 580 S. Preston Street, Baxter II, Room 221A, Louisville, KY 40202 (C.T.W.) or Division of B Cell Immunology, German Cancer Research Center, Im Neuenheimer Feld 280, 69120 Heidelberg, Germany (C.E.B.). E-mail addresses: corey.watson@louisville.edu (C.T.W.) or christian.busse@dkfz-heidelberg.de (C.E.B.)

Abbreviations used in this article: CNV, copy number variants; G1K, 1000 Genomes Project; GLDB, germline database; SNP, single nucleotide polymorphism.

References

1. Yu, Y., R. Ceredig, and C. Seoighe. 2017. A database of human immune receptor alleles recovered from population sequencing data. *J. Immunol.* 198: 2202–2210.
2. Wang, Y., K. J. Jackson, B. Gäeta, W. Pomat, P. Siba, W. A. Sewell, and A. M. Collins. 2011. Genomic screening by 454 pyrosequencing identifies a new human IGHV gene and sixteen other new IGHV allelic variants. *Immunogenetics* 63: 259–265.
3. Boyd, S. D., B. A. Gaëta, K. J. Jackson, A. Z. Fire, E. L. Marshall, J. D. Merker, J. M. Maniar, L. N. Zhang, B. Sahaf, C. D. Jones, et al. Individual variation in the germline Ig gene repertoire inferred from variable region gene rearrangements. *J. Immunol.* 2010. 184: 6986–6992.
4. Scheepers, C., R. K. Shrestha, B. E. Lambson, K. J. Jackson, I. A. Wright, D. Naicker, M. Goosen, L. Berrie, A. Ismail, N. Garrett, et al. 2015. Ability to develop broadly neutralizing HIV-1 antibodies is not restricted by the germline Ig gene repertoire. *J. Immunol.* 194: 4371–4378.
5. Watson, C. T., K. M. Steinberg, J. Huddleston, R. L. Warren, M. Malig, J. Schein, A. J. Willsey, J. B. Joy, J. K. Scott, T. A. Graves, et al. 2013. Complete haplotype sequence of the human immunoglobulin heavy-chain variable, diversity, and joining genes and characterization of allelic and copy-number variation. *Am. J. Hum. Genet.* 92: 530–546.
6. Gadala-Maria, D., G. Yaari, M. Uduman, and S. H. Kleinstein. 2015. Automated analysis of high-throughput B-cell sequencing data reveals a high frequency of novel immunoglobulin V gene segment alleles. *Proc. Natl. Acad. Sci. USA* 112: E862–E870.
7. Corcoran, M. M., G. E. Phad, V. B. Néstor, C. Stahl-Hennig, N. Sumida, M. A. A. Persson, M. Martin, and G. B. Karlsson Hedestam. 2016. Production of individualized V gene databases reveals high levels of immunoglobulin genetic diversity. *Nat. Commun.* 7: 13642.
8. Wang, Y., K. J. Jackson, W. A. Sewell, and A. M. Collins. 2008. Many human immunoglobulin heavy-chain IGHV gene polymorphisms have been reported in error. *Immunol. Cell Biol.* 86: 111–115.
9. Xochelli, A., A. Agathangelidis, I. Kavakiotis, E. Minga, L. A. Sutton, P. Baliakas, I. Chouvarda, V. Giudicelli, I. Vlahavas, N. Maglaveras, et al. 2015. Immunoglobulin heavy variable (IGHV) genes and alleles: new entities, new names and implications for research and prognostication in chronic lymphocytic leukaemia. *Immunogenetics* 67: 61–66.
10. 1000 Genomes Project Consortium, Abecasis, G. R., D. Altshuler, A. Auton, L. D. Brooks, R. M. Durbin, R. A. Gibbs, M. E. Hurles, and G. A. McVean. 2010. A map of human genome variation from population-scale sequencing. [Published erratum appears in 2011 *Nature* 473: 544.] *Nature* 467: 1061–1073.
11. 1000 Genomes Project Consortium, Abecasis, G. R., A. Auton, L. D. Brooks, M. A. DePristo, R. M. Durbin, R. E. Handsaker, H. M. Kang, G. T. Marth, and G. A. McVean. 2012. An integrated map of genetic variation from 1,092 human genomes. *Nature* 491: 56–65.
12. 1000 Genomes Project Consortium, Auton, A., L. D. Brooks, R. M. Durbin, E. P. Garrison, H. M. Kang, J. O. Korbel, J. L. Marchini, S. McCarthy, G. A. McVean, and G. R. Abecasis. 2015. A global reference for human genetic variation. *Nature* 526: 68–74.
13. Lefranc, M.-P. L. G. 2001. *The Immunoglobulin factbook*, Vol. 262. Academic Press, London.
14. Giudicelli, V., D. Chaume, and M.-P. Lefranc. 2005. IMGT/GENE-DB: a comprehensive database for human and mouse immunoglobulin and T cell receptor genes. *Nucleic Acids Res.* 33: D256–D261.
15. Musumeci, L., J. W. Arthur, F. S. G. Cheung, A. Hoque, S. Lippman, and J. K. V. Reichardt. 2010. Single nucleotide differences (SNDs) in the dbSNP database may lead to errors in genotyping and haplotyping studies. *Hum. Mutat.* 31: 67–73.
16. Li, H. 2014. Toward better understanding of artifacts in variant calling from high-coverage samples. *Bioinformatics* 30: 2843–2851.
17. Reumers, J., P. De Rijk, H. Zhao, A. Liekens, D. Smeets, J. Cleary, P. Van Loo, M. Van Den Bossche, K. Catthoor, B. Sabbe, et al. 2011. Optimized filtering reduces the error rate in detecting genomic variants by short-read sequencing. *Nat. Biotechnol.* 30: 61–68.
18. Watson, C. T., K. M. Steinberg, T. A. Graves, R. L. Warren, M. Malig, J. Schein, R. K. Wilson, R. A. Holt, E. E. Eichler, and F. Breden. 2015. Sequencing of the human IG light chain loci from a hydatidiform mole BAC library reveals locus-specific signatures of genetic diversity. *Genes Immun.* 16: 24–34.
19. Mackelprang, R., C. S. Carlson, L. Subrahmanyan, R. J. Livingston, M. A. Eberle, and D. A. Nickerson. 2002. Sequence variation in the human T-cell receptor loci. *Immunol. Rev.* 190: 26–39.
20. Mackelprang, R., R. J. Livingston, M. A. Eberle, C. S. Carlson, Q. Yi, J. M. Akey, and D. A. Nickerson. 2006. Sequence diversity, natural selection and linkage disequilibrium in the human T cell receptor alpha/delta locus. *Hum. Genet.* 119: 255–266.
21. Li, H., X. Cui, S. Pramanik, and N.-O. Chimge. 2002. Genetic diversity of the human immunoglobulin heavy chain VH region. *Immunol. Rev.* 190: 53–68.
22. Chimge, N.-O., S. Pramanik, G. Hu, Y. Lin, R. Gao, L. Shen, and H. Li. 2005. Determination of gene organization in the human IGHV region on single chromosomes. *Genes Immun.* 6: 186–193.
23. Kidd, M. J., Z. Chen, Y. Wang, K. J. Jackson, L. Zhang, S. D. Boyd, A. Z. Fire, M. M. Tanaka, B. A. Gaëta, and A. M. Collins. 2012. The inference of phased haplotypes for the immunoglobulin H chain V region gene loci by analysis of VDJ gene rearrangements. *J. Immunol.* 188: 1333–1340.
24. Estivill, X., J. Cheung, M. A. Pujana, K. Nakabayashi, S. W. Scherer, and L.-C. Tsui. 2002. Chromosomal regions containing high-density and ambiguously mapped putative single nucleotide polymorphisms (SNPs) correlate with segmental duplications in the human genome. *Hum. Mol. Genet.* 11: 1987–1995.
25. Ho, M. R., K. W. Tsai, C. H. Chen, and W. C. Lin. 2011. dbDNV: a resource of duplicated gene nucleotide variants in human genome. *Nucleic Acids Res.* 39: D920–D925.
26. Luo, S., J. A. Yu, and Y. S. Song. 2016. Estimating Copy Number and Allelic Variation at the Immunoglobulin Heavy Chain Locus Using Short Reads. *PLOS Comput. Biol.* 12: e1005117.
27. Kozbor, D., and J. C. Roder. 1981. Requirements for the establishment of high-titered human monoclonal antibodies against tetanus toxoid using the Epstein-Barr virus technique. *J. Immunol.* 127: 1275–1280.
28. Schwienbacher, C., A. De Grandi, C. Fuchsberger, M. F. Facheris, M. Svaldi, M. Wjst, P. P. Pramstaller, and A. A. Hicks. 2010. Copy number variation and association over T-cell receptor genes–influence of DNA source. *Immunogenetics* 62: 561–567.
29. 1000 Genomes Project. webpage: http://www.internationalgenome.org/faq/why-only-85-genome-assayable/
30. Dilthey, A., C. Cox, Z. Iqbal, M. R. Nelson, and G. McVean. 2015. Improved genome inference in the MHC using a population reference graph. *Nat. Genet.* 47: 682–688.
31. Norman, P. J., J. A. Hollenbach, N. Nemat-Gorgani, W. M. Marin, S. J. Norberg, E. Ashouri, J. Jayaraman, E. E. Wroblewski, J. Trowsdale, R. Rajalingam, et al. 2016. Defining KIR and HLA Class I Genotypes at Highest Resolution via High-Throughput Sequencing. *Am. J. Hum. Genet.* 99: 375–391.
32. Sherry, S. T., M. H. Ward, M. Kholodov, J. Baker, L. Phan, E. M. Smigielski, and K. Sirotkin. 2001. dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.* 29: 308–311.
33. Stodden, V. 2009. The legal framework for reproducible scientific research: licensing and copyright. *Comput. Sci. Eng.* 11: 35–40.
34. Collins, A. M., Y. Wang, K. M. Roskin, C. P. Marquis, and K. J. Jackson. 2015. The mouse antibody heavy chain repertoire is germline-focused and highly variable between inbred strains. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 370: 20140236.

www.jimmunol.org/cgi/doi/10.4049/jimmunol.1700306

Response to Comment on "A Database of Human Immune Receptor Alleles Recovered from Population Sequencing Data"

We share the view of Dr. Watson and colleagues on the importance of germline databases (GLDB) of immune receptor genes and thank them for their interest in our work (1). Whereas we fully agree with them on many of the caveats that apply to alleles inferred from genome sequence data (discussed further below), we feel that it is important for us first to clarify the objectives of our article, to avoid any misinterpretation. Watson et al. state that we propose a solution to the incompleteness and inaccuracy of existing