Characterizing haplotype diversity at the immunoglobulin heavy chain locus across human populations using novel long-read sequencing and assembly approaches

Corey T Watson[1*], Melissa Laird Smith[2*], William Gibson[2], Gintaras Deikus[2], Oscar Rodriguez[2], Maya Strahl[2], Matthew Pendleton[2], Phillip Comella[2], Lana Harshman[3], Wayne Marasco[4,5], Evan E. Eichler[3], Robert Sebra[2], Jonas Korlach[6], Andrew J. Sharp[2], Ali Bashir[2]

[1]Department of Biochemistry and Molecular Genetics, University of Louisville School of Medicine, Louisville, KY; [2]Icahn School of Medicine at Mount Sinai & Icahn Institute for Genomics and Multi-scale Biology, New York, NY; [3]Department of Genome Sciences, University of Washington, Seattle, WA; [4]Department of Cancer Immunology & AIDS, Dana-Farber Cancer Institute, Boston, MA; [5]Department of Medicine, Harvard Medical School, Boston, MA; [6]Pacific Biosciences, Menlo Park, CA; *Contributed Equally

Background

The human immunoglobulin (IG) gene regions are among the most structurally complex and polymorphic regions of the human genome.

→IG loci consist of duplicated variable (V), diversity (D), joining (J), and constant (C) genes that recombine in B cells to produce an individual's expressed antibody (Ab) repertoire **(Figure 1A**).

→The IG heavy chain (IGH) harbors ~50-60 IGHV, 23 IGHD, 6 IGHJ, and 9 IGHC functional/ORF genes, with >250 known coding alleles (and counting!) (**Figure 1A**).

→IGH is highly enriched for large complex structural and copy number variants (SVs; CNVs) up to 75 Kb in size, including insertions, deletions, and duplications (**Figure 1A**).

→Known coding single nucleotide polymorphisms (SNPs) and SVs/CNVs show considerable variation and evidence of selection between human populations (**Figure 1B,1C**).

→Extreme haplotype diversity has hindered the use of high-throughput genomic assays in the region (**Figure 1D**).

→However, in instances where IGH variants have been explicitly investigated in clinical cohorts, they associate with functional phenotypes (**Figure 1E**).

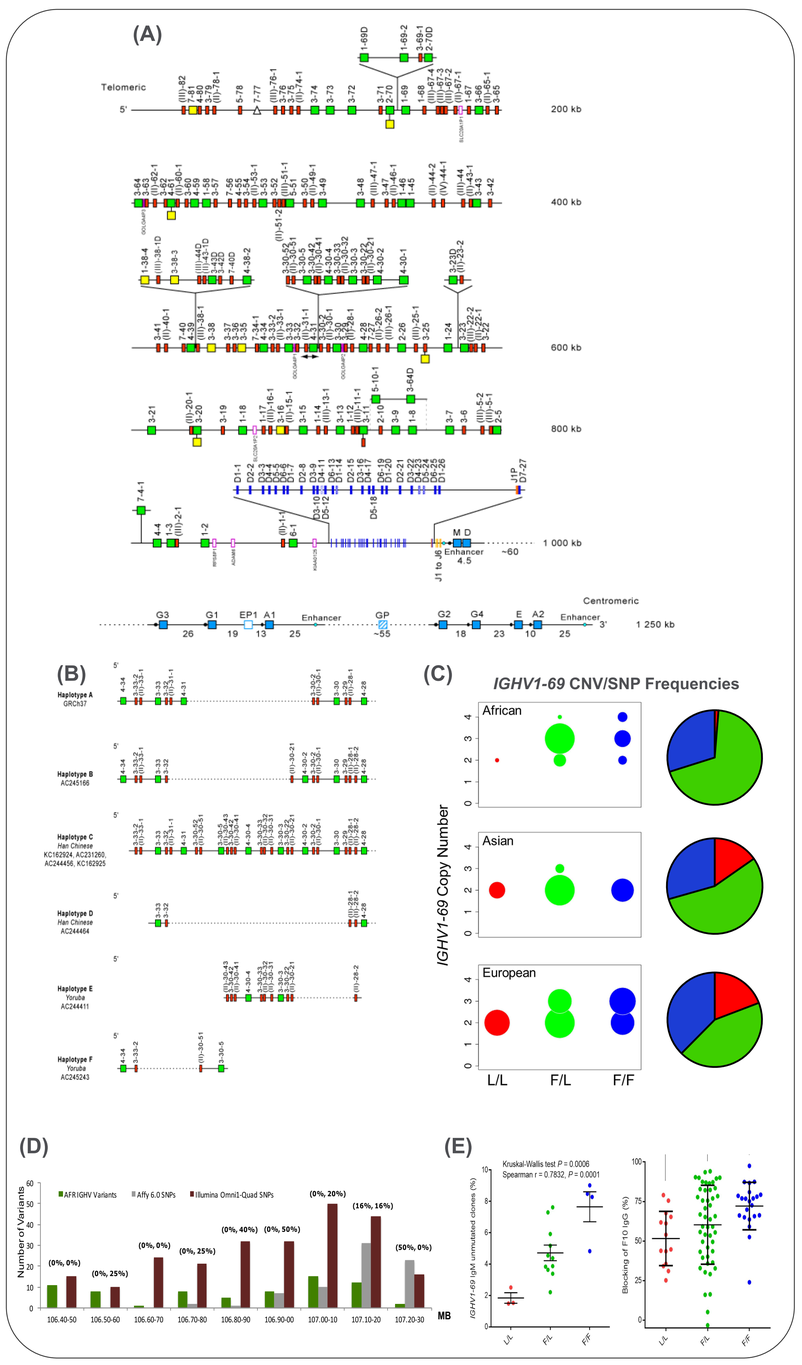


Figure 1. (A) IMGT map of IGH locus on chromosome 14[1], depicting IG functional, ORF, and pseudogenes, as well as alternate structural haplotypes. (B) Alternate haplotypes in the *IGHV3-30* region that contain large insertions and deletions[1,2]. (C) Inter-population variability observed for functional polymorphisms in the *IGHV1-69* region[2,3]. (D) GWAS arrays have low regional SNP density and poorly represent variants in the IGHV gene cluster[4]. (E) Germline polymorphism associates with *IGHV1-69* utilization in the naïve repertoire, and variability in the broadly neutralizing Ab anti-flu response[3].

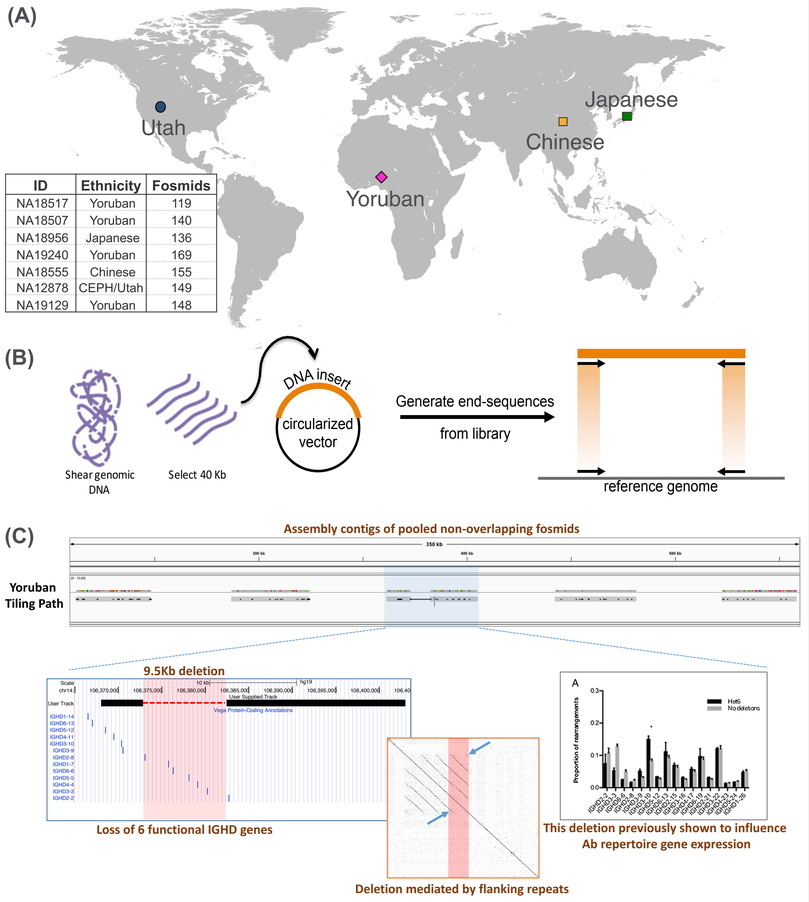
Building a Diverse Set of Reference Assemblies for the Human IGH locus

Problem: A major barrier to genetic & functional studies in IGH are due to the current paucity of genomic data in the region.

→The full ~1Mb IGH V, D, and J gene region (excluding IGHC) has only been sequenced two times[2,5].

→The current community IGH allele database, IMGT, is known to be incomplete, and ethnically biased[2,4,6,7,8].

Solution: Build a comprehensive map of sequence variation in IGH based on 14 complete IGH haplotypes assembled from 7 fosmid libraries of diverse ethnic origins (**Figure 2**).



ID	Ethnicity	Fosmids
NA18517	Yoruban	119
NA18507	Yoruban	140
NA18956	Japanese	136
NA19240	Yoruban	169
NA18555	Chinese	155
NA12878	CEPH/Utah	149
NA19129	Yoruban	148

Figure 2. (A) Geographic locations of the 7 individuals previously sampled for fosmid library construction[9]. The 1000 Genomes Project IDs of fosmid samples, their ethnicities, and number of clones processed per library are provided in the table (bottom left). (B) For fosmid library construction, genomic DNA from each individual was sheared and size selected; 40 kb fragments were cloned into fosmid vectors. Sanger sequences generated from the ends of ~1 million clones per library were mapped to the reference genome assembly[9], allowing for compilation of clone tiling paths across any locus of interest. We will utilize PacBio sequencing to generate a total of 14 ethnically diverse IGH reference assemblies from this fosmid resource. (C) Assemblies of initial fosmid tiling path in Yoruban NA18517 demonstrates utility of approach, and leads to the first complete description of a 9.5 Kb deletion, previously implicated in Ab repertoire gene usage variability[10].

Diploid Resolution of IGH (GIAB)

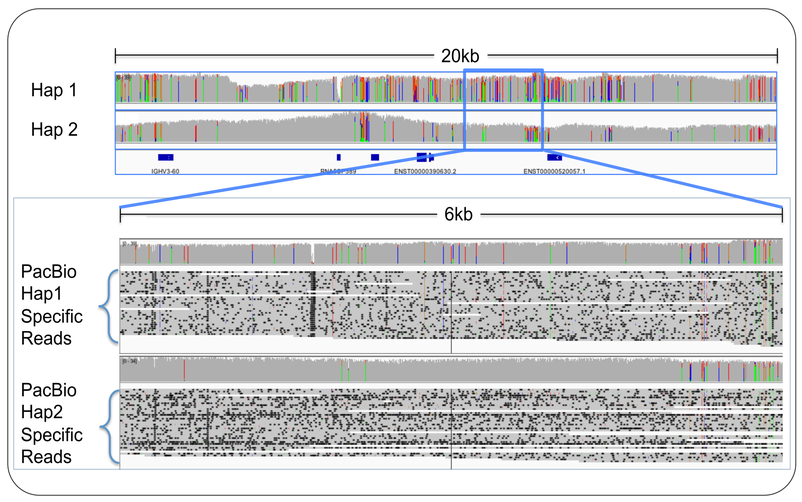


Figure 3: Diploid haplotype resolution of an Ashkenazi Jewish proband from the Genome in a Bottle (GIAB) Consortium. An Ashkenazi Jewish Trio was sequenced using multiple-technologies including short and long-reads. A sample 20 kb interval with SNPs within the *IGHV4-61* region is shown in the top panel. In the bottom panel, high-quality long-range phased SNPs allow PacBio reads two be partitioned into two distinct haplotypes via a novel algorithmic approach. These reads allow haplotype phasing of both SNPs and SVs within reads (*e.g.*, the small deletion shown in Hap1) and the potential for de novo assembled haplotypes spanning large regions/events.

Development of a novel long fragment capture and sequencing assay for IGH

Problem: Resolution of IGH complexity is challenging for standard genetic approaches.

→SNPs alone are unable to represent complex allelic and structural haplotypes.

→Short-read NGS data may allow for variant inference, but are often inaccurate. Phased assemblies are not possible.

Solution: Develop a robust approach for assaying IGH genetic variation locus-wide with nucleotide resolution (**Figure 4**) that leverages longer read lengths to improve assemblies and the characterization of novel haplotype variation.

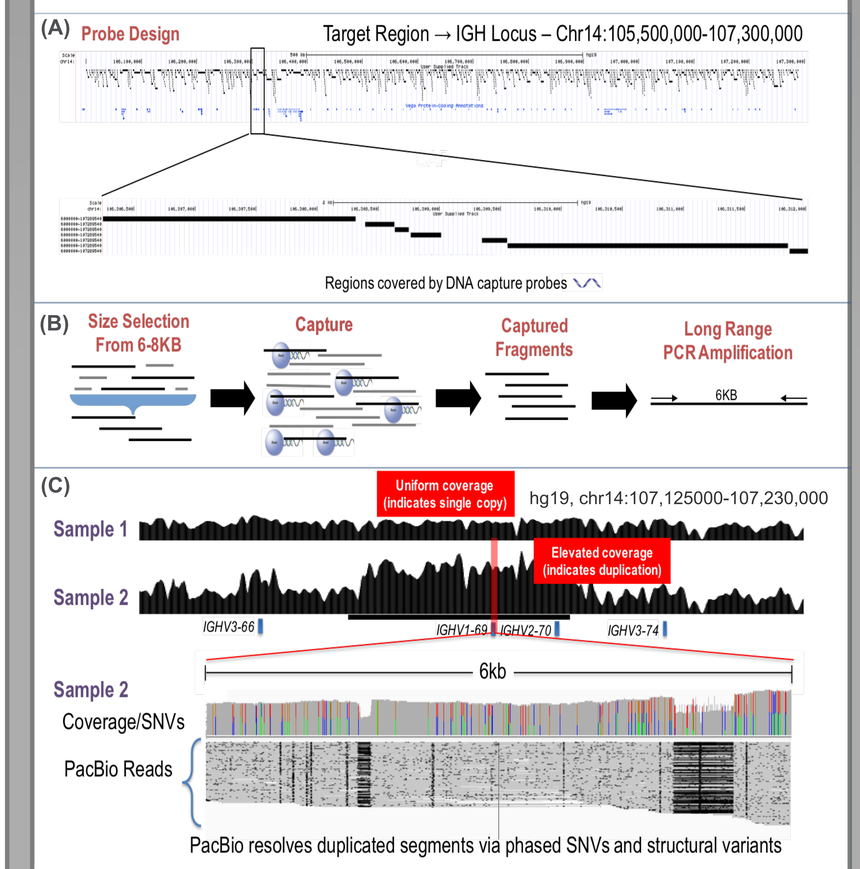


Figure 4. (A) Nimblegen SeqCap probes were designed across the entire IGH locus using all existing haplotype data, corresponding to ~1.4 Mb of unique sequence, with an estimated coverage of 94.6% of targeted bases. (B) Standard assay conditions were adapted to capture 6-8kb fragments from a haploid hydatidiform mole sample. (C) PacBio long-read sequencing allowed for more reliable reconstruction of large structural variants between haplotypes. A large tandem duplication variant overlapping *IGHV1-69* (black bar) is shown, which had been previously unresolvable with 300 bp MiSeq reads. This approach allowed for phasing of variants across the region and enabled the partitioning of reads into their respective tandem duplication blocks for improved assembly.

Outcomes & Future Directions

This work brings IGH into the modern genomics era, via:

→Resolving an expanded set of IGH haplotype maps and germline variants from a diverse set of human populations. These will allow for the discovery of novel genetic variation at this locus.

→Development of the beta design for the first locus-wide IGH genotyping platform. This will enable de novo, diploid resolution of IGH haplotypes, including: annotated germline IGH C, J, D, V allele calls; gene copy number; and a catalogue of non-coding SNPs and SVs.

We believe this will enable many lines of novel investigation. Most importantly, by further defining the full extent of IGH diversity, we can examine the impact of this on antibody response in disease.

Literature Cited

1.) Lefranc, M-P, Lefranc, G. 2001. The Immunoglobulin FactsBook. Academic Press, London.
2.) Watson, CT, *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.* 92:530-46.
3.) Avnir, Y, *et al.* 2016. *IGHV1-69* polymorphism modulates anti-influenza antibody repertoires, correlates with IGHV utilization shifts and varies by ethnicity. *Sci Rep.* srep20842.
4.) Watson, CT, Breden, F. 2012. The immunoglobulin heavy chain locus: genetic variation, missing data, and implications for human disease. *Genes Immun.* 13(5):363-73.
5.) Matsuda, F, *et al.* 1998. The complete nucleotide sequence of the human immunoglobulin heavy chain variable region locus. *J Exp Med.* 188:2151-62.
6.) Boyd, SD, *et al.* 2010. Individual variation in the germline Ig gene repertoire inferred from variable region gene rearrangements. *J Immunol.* 184(12):6986-92
7.) Scheepers, C, *et* al. 2015. Ability to develop broadly neutralizing HIV-1 antibodies is not restricted by the germline Ig gene repertoire. *J Immunol.* 194(9):4371-8.
8.) Gadala-Maria, D, et al. 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 U S A.* 112(8):E862-70.
9.) Kidd, JM, *et al.* 2008. Mapping and sequencing of structural variation from eight human genomes. *Nature.* 453: 56–64.
10.) Kidd, MJ, *et al.* 2016. DJ Pairing during VDJ Recombination Shows Positional Biases That Vary among Individuals with Differing IGHD Locus Immunogenotypes. *J Immunol.* 196(3):1158-64.

contact: corey.watson@louisville.edu; melissa.smith@mssm.edu; ali.bashir@mssm.edu