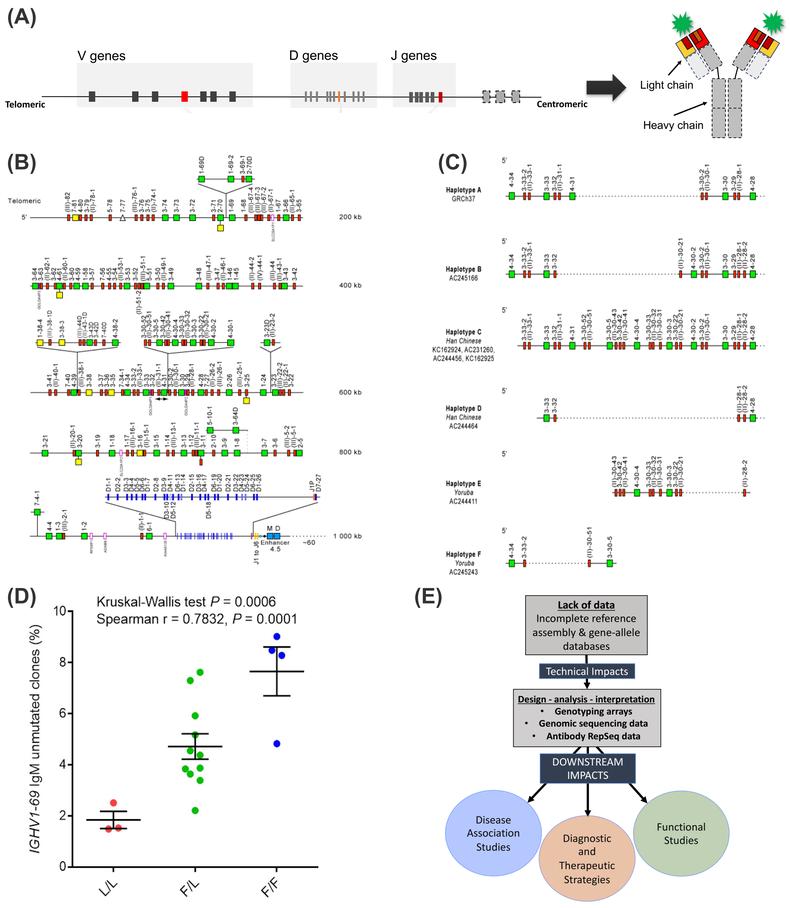
Genomic characterization of the immunoglobulin heavy chain variable gene locus in individuals of African, Asian, and European descent reveals elevated haplotype diversity

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

[1]Icahn School of Medicine at Mount Sinai & Icahn Institute for Genomics and Multi-scale Biology, New York, NY; [2]Department of Biochemistry and Molecular Genetics, University of Louisville School of Medicine, Louisville, KY; [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

Background



The human immunoglobulin (IG) gene regions are among the most structurally complex and polymorphic regions of the genome. IG genes recombine in B cells to produce antibodies. (**Figure A**). The IG heavy chain (IGH) harbors ~50-60 IGHV, 23 IGHD, 6 IGHJ, and 9 IGHC functional/ORF genes, with >250 known coding alleles (**Figure B**). It is highly enriched for large complex structural variations up to 75 Kb in size (**Figure B and C**). Extreme haplotype diversity (**Figure C)** has hindered the use of high-throughput genomic assays in the region and the lack of the data has had impacts in disease association studies and functional studies (**Figure E**). However, targeted studies of IGH variants have been shown to associate with antibody expression and function, and higher –level phenotypes in clinical cohorts[3]. (**Figure D)**.

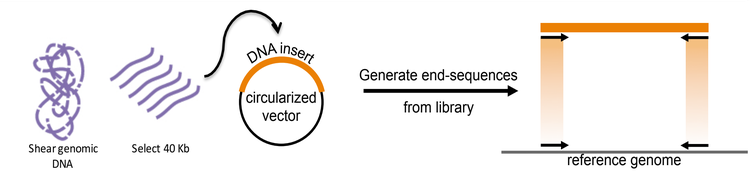
Building a Diverse Set of Reference Assemblies for the Human IGH locus

Problem: A major barrier to genetic & functional studies in IGH is 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].

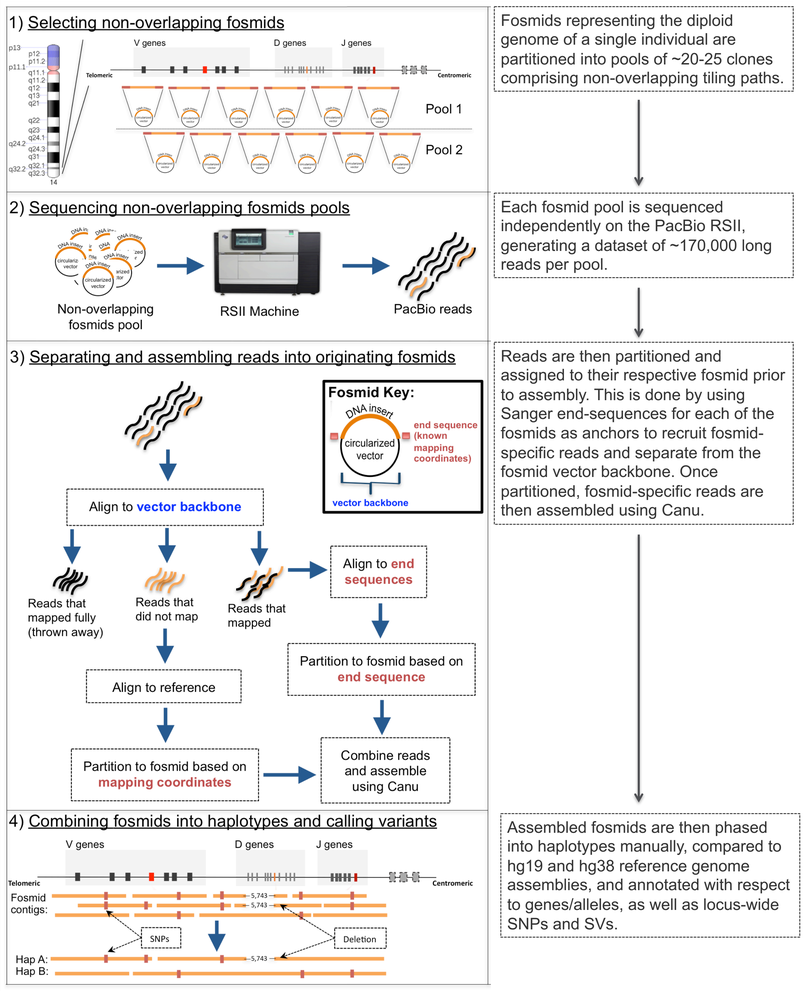
→The current community IGH allele database, IMGT, is known to be incomplete, and ethnically biased[2,4,6,7,8].

Solution: Build a comprehensive map of sequence variation in IGH based on 16 complete IGH haplotypes assembled from 8 fosmid libraries of diverse ethnic origins



- For fosmid library construction, genomic DNA from each individual was sheared and size selected; 40 kb fragments were cloned into fosmid vectors[9]. Sanger sequences generated from the ends of ~1 million clones per library were mapped to the reference genome assembly, allowing for compilation of clone tiling paths across IGH. Fosmids are then sequenced using Pacbio to generate a total of 16 ethnically diverse IGH reference assemblies from this fosmid resource. Geographic origins of the 8 individuals previously sampled for fosmid library construction are shown.

Methods



Results

Sample	Population	Number of fosmids	Average fosmid coverage	Number of subreads	Average read length	Min read length	Max read length
ABC10	African	78	2.5	665,435	7,533	50	47,195
ABC11	Asian	89	2.61	700,484	6,381	50	43,554
ABC12	European	89	2.25	503,528	7,352	50	42,141

Table 1. Statistics for fosmids for ABC10, ABC11 and ABC12 libraries

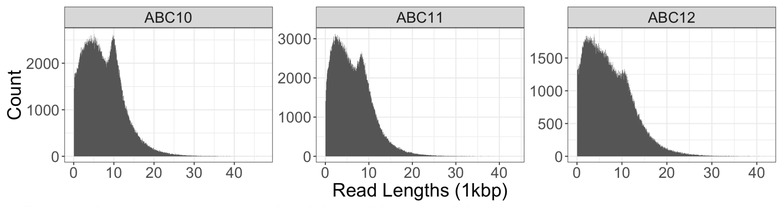


Figure 1. Read length distributions for ABC10, ABC11 and ABC12.

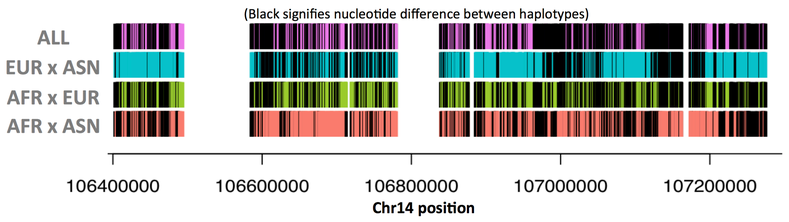


Figure 2. Comparing haplotype sequence across the three samples. Colored regions represent identical tracks of sequences. Across all the samples ("ALL"), the mean length of identical haplotype sequences is 335 bp and the max length is 8,861 bps. Between the European and Asian sample ("EUR x ASN"), the mean length is 610 bp and the max length is 37,344 bp. Between the African and European sample ("AFR x EUR"), the mean length is 376 np and the max length is 8,681 bp, and between the African and Asian sample ("AFR x ASN"), the mean length is 591 bp and the max length is 35,597 bp. The relatively small blocks of shared haplotypes suggests high levels of haplotype diversity between individuals

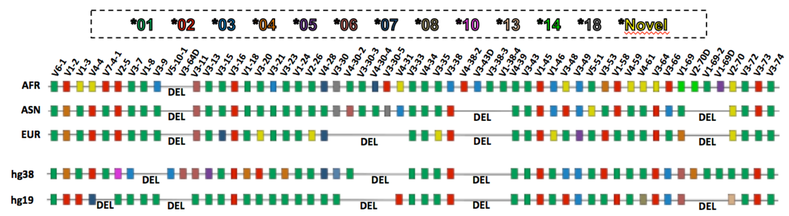


Figure 3. Annotation of genes in the IGHV haplotypes across all samples and reference genomes, hg38 and hg19. Each column is the gene and it is color coded by the allele type. The African haplotype (AFR) has 55 genes, the European haplotype (EUR) has 42 genes, and the Asian haplotype (ASN) has 48 genes. There were 8, 2, and 5 novel alleles in African, Asian, and European haplotypes, respectively.

Results (Cont'd)

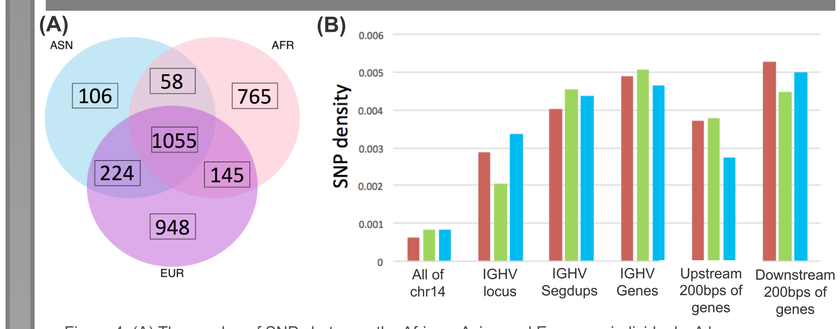


Figure 4. (A) The overlap of SNPs between the African, Asian and European individuals. A large proportion of SNPs are unique to each individual. (B) The SNP density in different regions of the IGHV locus and chromosome 14, revealing a greater density of SNPs in IGHV, particularly in segmental duplications and genic regions.

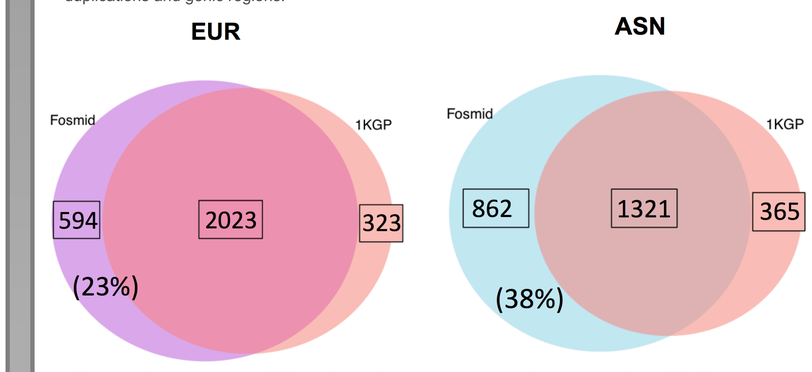


Figure 5. Overlap between SNPs called in fosmids and those called as part of the 1000 Genomes Project reveals low concordance between datasets.

Future Directions

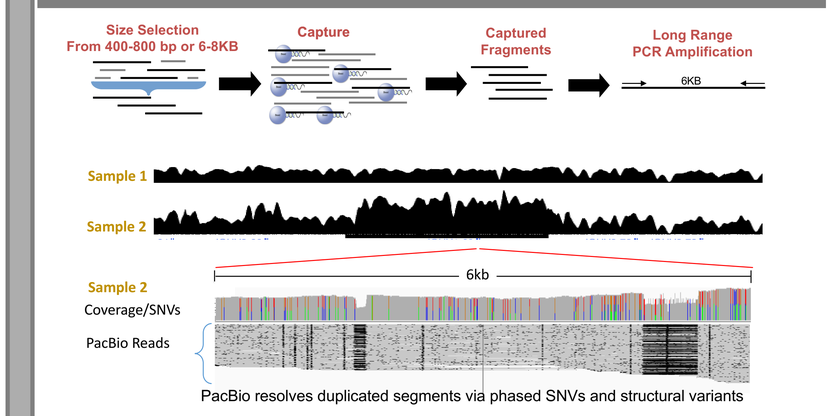


Figure 1. Using the sequence haplotypes to build a more effective genotyping assay. Current assays are inadequate for comprehensively genotyping IGH variants (SNPs and SVs) in a high-throughput manner. By having more sequence resolved haplotypes, we can build better tools and analysis pipelines for accurately genotyping the locus. We developed a pilot capture-PacBio sequencing protocol that uses the NimbleGen SeqCap EZ system combined with PacBio sequencing.

Figure 2. We tested our capture panel on a haploid hydatidiform mole sample, CHM1, the same individual we previously used to reconstruct the first IGHV locus haplotype from a single chromosome; this now serves as the representation for IGHV in the hg38 reference assembly. Reads CHM1 sequenced for this experiment were assembled and mapped back to hg38, identifying only 42 SNPs and 4 large INDELs.

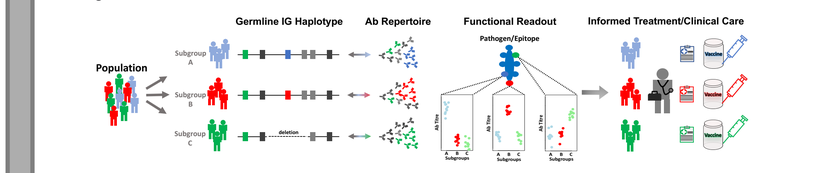


Figure 3. Using the assembled haplotypes in combination with capture data from clinical cohorts, we can start to link genetics to antibody expression/function and disease.

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: oscar.rodriguez@icahn.mssm.edu; corey.watson@louisville.edu; melissa.smith@mssm.edu; ali.bashir@mssm.edu

Pacific Biosciences, PacBio, SMRT, SMRTbell and the Pacific Biosciences logo are trademarks of Pacific Biosciences of California, Inc. All other trademarks are the property of their respective owners. © 2012 Pacific Biosciences of California, Inc. All rights reserved.