Supporting Information

Blachly et al. 10.1073/pnas.1503587112

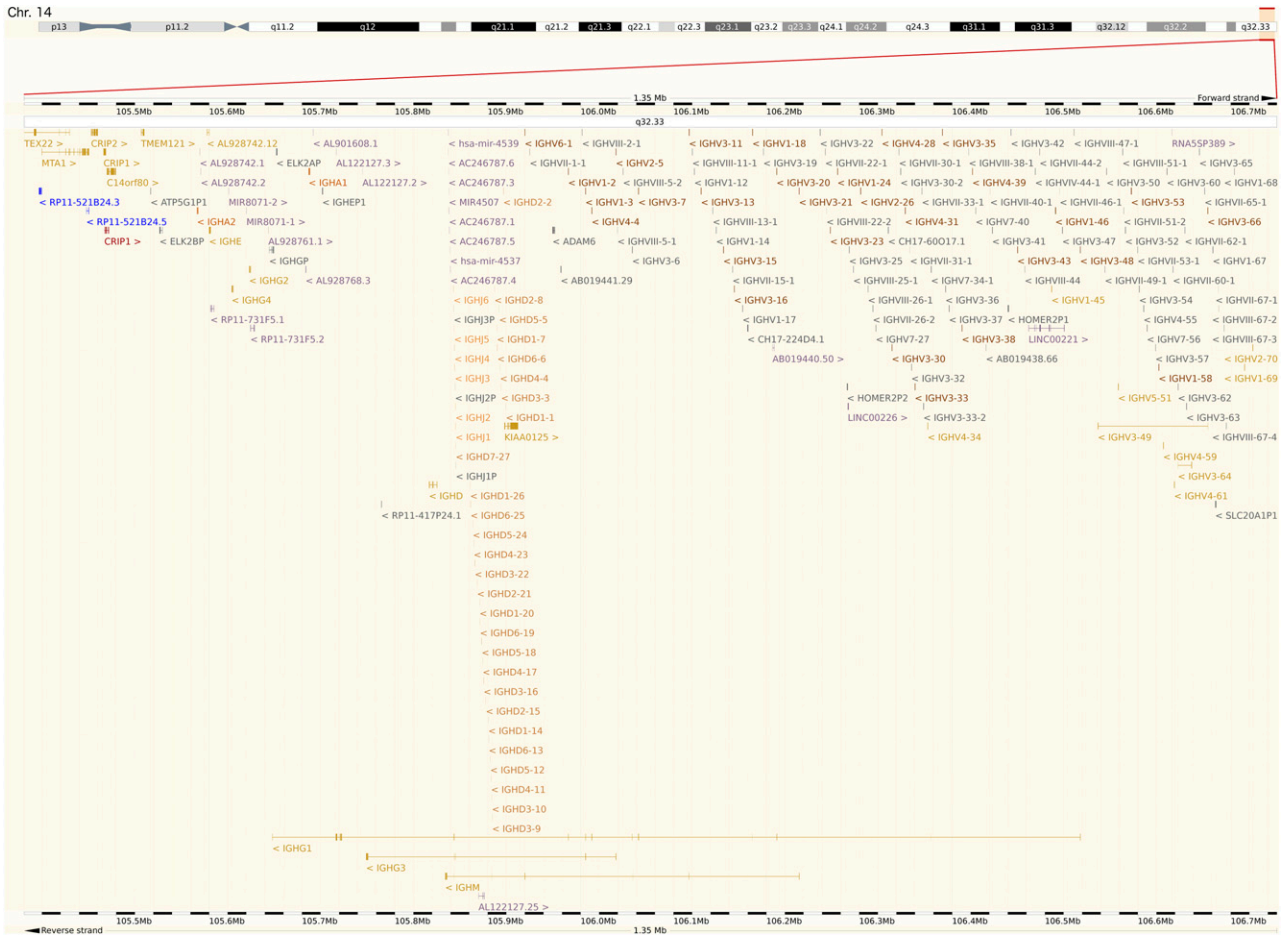

Fig. S1. Human chromosome 14 is illustrated with the portion of the 14q32.33 cytoband corresponding to the Ig heavy chain locus magnified. Overall, this region spans more than 1 megabase. As a result of V(D)J recombination, junctional diversity, and somatic hypermutation, a fully spliced Ig transcript (and by extension, the sequencing reads derived therefrom) bears far less resemblance to its originating genomic DNA than a typical transcript and spans large genomic distances, all problematic for conventional mapping software. (image from Ensembl release 77; ref. 1).

1. Flicek P, et al. (2014) Ensembl 2014. *Nucleic Acids Res* 42(Database issue, D1)D749–D755.

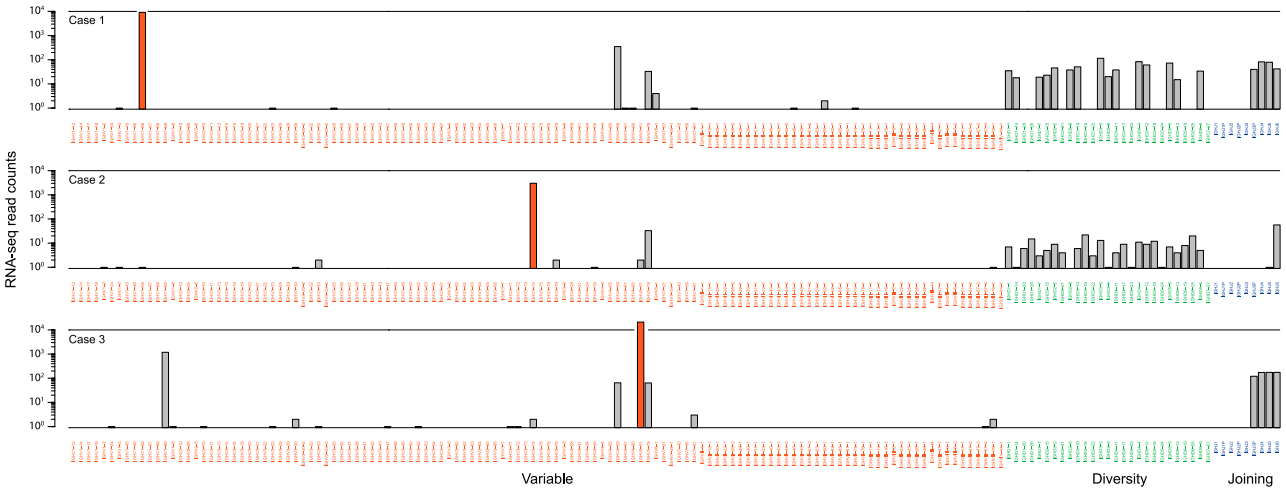

Fig. S2. Counts of RNA sequencing reads (*y* axis; log scale) mapping to the *IGH@* locus (x-axis) for three representative samples. Labels on the *x* axis are colored by type: Variable (V) genes are depicted in red; Diversity (D) genes are depicted in green; and Joining (J) genes are depicted in purple. A bar is present at each V, D, or J gene with height representing number of mapped reads; in each case, the V gene with the highest number of reads corresponds to the V gene detected by conventional analysis as well as the de novo reconstruction method and is colored in red. Mapping to and counting the D and J gene segments (green and purple, respectively) was problematic (no clear consensus in some cases; no mappings at all in others) and failed to predict the D and J segments.

Table S1. White blood cell count (WBC), absolute lymphocyte count (ALC), total number of mate reads, and total number of reads mapping to *IGH@* for each sample in the pilot set

RNA ID	WBC	ALC	Reads	IGH@
US-1422278	156.1	151.42	27183696	173466
US-1422282	91.8	89.96	24184258	419953
US-1422294	15.8	13.59	27490286	204751
US-1422302	30.3	25.72	27700310	282761
US-1422309	76	73.49	26376594	228077
US-1422311	12.2	10.74	25845072	246284
US-1422314	125	121.25	25951708	317911
US-1422321	22.5	18.2	27447262	253242
US-1422333	28.1	27.34	25845072	317009
US-1422335	22.7	21.63	26182804	265150
US-1422342	105.1	99.85	25845072	149119
US-1422350	25.1	24.92	26151792	323623
US-1422351	21.2	14.84	24376336	258669
US-1422352	144.6	140.26	23549136	292457
US-1422356	112.4	109.8	26972918	302169
US-1422366	12.6	6.94	27655638	182894
US-1422368	97.2	89.62	27554280	258087

Table S2. Summary of concordance

Paired-end Sequencing	N (%)
Read length, 91 nt*	
V gene match	16 (94)
V gene mismatch	1 (6)
Mutation status match	17 (100)
Mutation status mismatch	0
Mutation, MSE	0.23
Read length, 50 nt†	
V gene match	6 (100)
V gene mismatch	0
Mutation status match	6 (100)
Mutation status mismatch	0
Mutation, MSE	1.3

MSE, mean squared error.
*Ig-ID pipeline versus 17 clinical laboratory measures.
†Ig-ID pipeline versus 6 clinical laboratory measures.

Table S3. Determinations of *V* gene and percent *IGHV* mutation from the medical record (MR) and output from Ig-ID

Patient ID	RNA ID	*V* gene (MR)	*V* gene (Ig-ID)	Mutation, % (MR)	Mutation, % (Ig-ID)
2826	US-1422368	V3-53	IGHV3-74*03	6.1	8.8
2827	US-1422309	V3-74	IGHV3-53*01	8.8	6.1

Matching information (or in the case of patient ID and RNA ID, putatively matching) are highlighted in blue. Consecutively numbered patients 2826 and 2827 had sample tubes mislabeled. In the IMGT database, the total number of functional *IGHV* genes, including known alleles is 260. After multiplying by the total number of possible point mutations this increases exponentially. Because of this uniqueness, identification of specific patients can in some cases be made by comparing Ig-ID results with the medical record or other known source. Sample fingerprinting is an additional potentially important application of the Ig-ID procedure.