Title

The Antibody Repertoire of Colorectal Cancer.

Permalink

https://escholarship.org/uc/item/21n171bt

Journal

Molecular & Cellular Proteomics, 16(12)

Authors

Cha, Seong

Bonissone, Stefano

Na, Seungjin

et al.

Publication Date

2017-12-01

DOI

10.1074/mcp.RA117.000397

Peer reviewed

The Antibody Repertoire of Colorectal Cancer*

Seong Won Cha‡, Stefano Bonissone§, Seungjin Na¶, Pavel A. Pevzner¶, and Vineet Bafna¶||

Immunotherapy is becoming increasingly important in the fight against cancers, using and manipulating the body's immune response to treat tumors. Understanding the immune repertoire—the collection of immunological proteins—of treated and untreated cells is possible at the genomic, but technically difficult at the protein level. Standard protein databases do not include the highly divergent sequences of somatic rearranged immunoglobulin genes, and may lead to miss identifications in a mass spectrometry search. We introduce a novel proteogenomic approach, AbScan, to identify these highly variable antibody peptides, by developing a customized antibody database construction method using RNA-seq reads aligned to immunoglobulin (Ig) genes.

AbScan starts by filtering transcript (RNA-seq) reads that match the template for Ig genes. The retained reads are used to construct a repertoire graph using the "split" de Bruijn graph: a graph structure that improves on the standard de Bruijn graph to capture the high diversity of Ig genes in a compact manner. AbScan corrects for sequencing errors, and converts the graph to a format suitable for searching with MS/MS search tools. We used AbScan to create an antibody database from 90 RNA-seq colorectal tumor samples. Next, we used proteogenomic analysis to search MS/MS spectra of matched colorectal samples from the Clinical Proteomic Tumor Analysis Consortium (CPTAC) against the AbScan generated database. AbScan identified 1,940 distinct antibody peptides. Correlating with previously identified Single Amino-Acid Variants (SAAVs) in the tumor samples, we identified 163 pairs (antibody peptide, SAAV) with significant cooccurrence pattern in the 90 samples. The presence of coexpressed antibody and mutated peptides was correlated with survival time of the individuals. Our results suggest that AbScan (https://github.com/csw407/AbScan.git) is an effective tool for a proteomic exploration of the immune response in cancers. ***Molecular & Cellular Proteomics 16: 10.1074/mcp.RA117.000397, 2111–2124, 2017.***

Cancer immunotherapy, which attempts to tackle cancer using the body's own immune response, has been very successful in boosting the survival rates of patients with leukemia and other blood cancers (1–3). This field of research is expanding rapidly, and has been extended to include other cancer subtypes, including solid tumors (4–6).

Immunotherapy is more specific than generic typical cancer treatments targeting fast-growing cells directly. It can take the form of cancer vaccines (neoantigens that stimulate an immune response) (7, 8), monoclonal antibodies, which target cancer cells expressing specific (neoantigenic) proteins (9) or immune checkpoint inhibitors that activate suppressed immune cells (10–12). The development of new forms of cancer immunotherapy could be greatly helped by knowledge of the cancer specific immune response, especially in understanding the antibodies and neoantigens specific to cancer.

This is a challenge because of the millions of distinct antibodies that are circulating in the blood. We still have only limited knowledge of the antibody responses that target individual disease-related antigens and epitopes. There are only a few known examples in infectious disease (13) and autoimmune disease (14). On top of that, recent methods that characterize the antibody repertoire use serum or plasma samples as their source for antibody analysis. However, the antibodies in these samples include the pool of all antibodies binding to multiple antigens, as well as the antibodies produced by numerous previous immune responses (15–19). Screening the antibodies based on their binding to preselected antigens may also not work, as all possible neoantigens existing in a sample cannot be known, and some important antigens may be post-translationally modified (20, 21) or cleaved (22).

Another approach to understanding the antibody repertoire is by isolating the B-cells that respond to a target immunogenic antigen. Plasmablasts (23, 20), memory B cells (24–26), and tissue infiltrating B cells (27–29) have been used to characterize the functional antibody repertoire (30, 31). The method works, but it requires a dedicated workflow to isolate the B-cells and sequence the antibody clones. Here, we propose a more direct method for discovering antibody peptides in tumor samples.

From the ‡Department of Electrical and Computer Engineering, University of California, San Diego, La Jolla, California; §Digital Proteomics, LLC, La Jolla, California; ¶Department of Computer Science and Engineering, University of California, San Diego, La Jolla, California 92037

Received October 11, 2017
Published, MCP Papers in Press, October 18, 2017, DOI 10.1074/mcp.RA117.000397
Author contributions: S.W.C., S.B., P.A.P., and V.B. designed research; S.W.C. and V.B. performed research; S.W.C., S.N., and V.B. analyzed data; S.W.C. and V.B. wrote the paper.

Recently, we and others have developed pipelines for identifying mutated peptides expressed specifically in cancer (32–34). In our approach, we mine a general transcript resource (such as The Cancer Genome Atlas Project) to extract transcript sequences, identify novel mutations, and junctions, and then encode them into a complex database. This database is then searched via a proteogenomic approach, to identify peptides that are seen only in tumor proteome samples. Interestingly, our initial search of the Clinical Proteomic Tumor Analysis Consortium (CPTAC)[1] colorectal tumor samples identified a number of antibody peptide sequences (32). supplemental Fig. S1 shows the example of some antibody peptides identified in the search. At the time, there were questions regarding the provenance of the discovery, as we did not expect to find antibody peptides in colon tissue. They could be antibodies from tumor infiltrating lymphocytes (TIL), circulating antibodies from blood contamination, encoding general proteome variation, or even mis-identifications. Moreover, our databases were not specifically designed to capture Ig regions, so we were only identifying peptides from some of the annotated Ig genes on the human reference.

AbScan is a new tool for identifying all antibody (Ig) peptides in a sample by searching mass spectral data sets against RNA-seq data sets. AbScan is a proteogenomic tool that scans transcript and genomic data, preferably, but not exclusively from the same samples as the proteomic data; it creates specialized antibody sequence databases that can search tandem mass spectra. As the antibody sequences are hypervariable, identifying and characterizing transcripts encoding Ig genes is a challenging endeavor. We devised a special construct called the "split" de Bruijn (SdB) graph to encode all Ig transcripts in a compact fashion, then show the power of this approach compare with other methods. AbScan also uses a customized pipeline to search these antibody databases and identify expressed antibody peptides, while controlling for false discoveries. We evaluated sensitivity and specificity of AbScan by benchmarking it on simulated data sets, pure antibody mixtures, normal colon tissues, and colorectal cell-lines. We further applied AbScan to 90 colorectal samples from the CPTAC project and demonstrated that the antibody repertoire was characterized by significant cooccurrence pattern in 163 pairs of antibody peptides and Single Amino-Acid Variant (SAAV) pairs, and the cooccurring pairs were correlated with patient survival.

5. EXPERIMENTAL PROCEDURES

Experimental MS Data Sets, Sequence Databases, and Search Parameters—We analyzed four spectral data sets, which have been described in previous work.

- 90 colorectal tumor samples (https://cptac-data-portal.georgetown.edu/cptac/s/S022) (35)
- 30 normal colon biopsies (https://cptac-data-portal.georgetown.edu/cptac/s/S019) (35)
- Colon cancer cell-lines LIM1215, LIM1899, and LIM2405 (http://proteomecentral.proteomexchange.org/cgi/GetDataset?ID=PXD000120, PXD000120) (36), and
- Purified polyclonal antibody mixture (ftp://massive.ucsd.edu/MSV000081401, MSV000081401) (37)

We searched each tandem mass spectra against three different databases. These included

- Ensembl database version GRCh38 (38)
- The "split" de Bruijn (SdB) graph based database driven by the method described in this paper, and
- A de Bruijn (dB) graph based database (34)

We used MS-GF+ (version 1.1.0) (39) with the following parameters: parent mass tolerance of 20 ppm, and allowed post-translational modifications of fixed carbamidomethyl C and optional oxidized Methionine. Common contaminants were excluded. ProteoWizard (v3.0.3827) (40), and ReAdW (v1.1 and v4.3.1) (41) were used for the peaklist-generating software. Number of missed and nonspecific cleavages permitted was 1. Trypsin was used to generate peptides for three colorectal data sets, and Trypsin, Asp-N, Chymotrypsin, Elastase were used for the purified polyclonal antibody mixture data set. A multistage FDR (See Experimental Procedures–"Multistage search") was applied to identify the PSMs from the SdB and dB driven databases (See Supplemental Table 1).

Database Construction Using "split" de Bruijn (SdB) and de Bruijn (dB) Graphs—AbScan constructs the "split" de Bruijn (SdB) graphs for multiple RNA-seq data sets from tumors. A de Bruijn (dB) is constructed for a fair comparison. Followings are the steps to generate a custom MS searchable database:

1 **Read filtering.** Filter out all RNA-seq reads not sampling an Ig gene

2 **SdB graph construction.** Create a SdB graph based database from filtered reads

3 **Error correction.** Identify and eliminate sequencing errors

4 **FASTA database construction.** The SdB graph is used to generate an MS searchable FASTA formatted database, as well as scripts to identify the context of the peptide on the antibody sequence.

For comparing the performance of SdB graphs to dB graphs, we used an implementation of dB graphs customized for the discovery of antibody peptides (34).

Read Filter—All antibodies are a combination of relatively fixed framework (FR), and hyper-variable complementarity determining regions (CDR), with the order given by "FR1, CDR1, FR2, CDR2, FR3, CDR3, FR4". The typical lengths of CDRs in human are 15 to 30 nt for CDR1 and CDR2, and 24 to 36 nt for CDR3 (42). On the other hand, lengths of RNA-seq read in our data sets varied from 76 to 100 nt. Therefore, we expect most RNA-seq reads to cover some part of a framework region, and could use this to filter RNA-seq reads from Ig genes. In addition, we employed keyword matching to recover non-mapped Ig gene encoding reads by creating a list of k-mer sequences from all Ig genes in the IMGT reference (43), and selecting all reads that matched one of the k-mers.

An appropriate value of parameter k was determined by comparison with decoy data obtained by reversing the IMGT reference sequences. As k is made smaller, we can quantify the false matches by the number of reads that match decoy k-mers. For any value of k, the false discovery rate is given by

[1] The abbreviations used are: TCGA, The Cancer Genome Atlas Project; CPTAC, Clinical Proteomic Tumor Analysis Consortium; TIL, Tumor Infiltrating Lymphocyte; Ig, Immunoglobulin; Sdb, "Split" de Bruijn; dB, de Bruijn; FR, Framework Region; CDR, Complementarity Determining Region; IMGT, The international ImMunoGeneTics information system; COAD, Colon adenocarcinoma; DP, Digital Proteomics; SAAV, Single Amino-Acid Variant.

$$FDR(k) = \frac{Number\ of\ reads\ matching\ decoy\ k - mer}{Number\ of\ reads\ matching\ target\ k - mer}$$

We selected the smallest value of k that resulted in a FDR below 1%, $k = 19$ was used for filtering (supplemental Fig. S2). Quality filtering was applied additionally to trim the part of the poor quality reads. We trimmed the 3′ end of the reads if their quality threshold were less than the threshold value (10). We excluded the read if the trimmed part was longer than $\frac{2}{3}$ of the read length or the overall quality of the reads were below than the threshold value (25).

SdB Graph Construction—Typical de Bruijn (dB) graph construction is as follows: Given a set of reads, the dB graph for this set is defined as follows: each k-mer from reads is a node in the graph. Nodes u and v are connected by an edge, if there exists a $(k + 1)$-mer in reads whose k-suffix is u and whose k-prefix is v. dB graphs are a powerful construct because they help remove redundancy in read coverage, and can be efficiently constructed without the need to compare all pairs of reads to test for overlap (44–46). In the ideal case, each of the Ig genes is a path in the graph, and each path in the graph corresponds to a putative Ig gene. Errors can arise in dB graph construction if two unrelated reads share the same k-mer (we denote these as "false-positive" overlaps), or if reads from the same molecule do not share a k-mer because of sequencing errors (false-negatives).

False edges in the dB graph can also arise because of repeated k-mers. Specifically, a repeated substring of size greater than or equal to k will lead to false edges in a k-mer based dB graph. The error could be controlled by selecting larger values of k, but that would result in a higher false-negative rate. We reasoned that the exact match requirement in k-mer dB graph is restrictive. For example, consider two reads that overlap over 40 bp. The probability that this overlap contains $k = 30$ consecutive nucleotides with no error in both reads is 65.5%. On the other hand, the probability that this overlap contains 30 consecutive nucleotides with at most one error is 93.0%. Therefore, allowing for an approximate match improves sensitivity from 65.5% to 93.0%. See supplemental Method - "Analytical comparison of SdB and dB graphs" for a rigorous analysis.

An alternative approach is to do an error-correction before matching. BayesHammer uses a Bayesian approach on 1-neighborhoods of k-mers to correct reads, before constructing a k-mer dB graph (47). In Ig genes, however, we use RNA data to identify variation, and the variable coverage makes it difficult to distinguish sequencing errors from true genetic variation. The SdB graph handles this problem of nonuniform coverage through correction on local nodes like the IGdb, Trinity and IDBA-tran (48, 49, 34).

On top of that, SdB graph applies a binning technique to solve the approximate matching problem efficiently. To obtain 1-neighborhoods of k-mers, we need the pairwise distance of every existing k-mers observed from the reads, which may increase the computation time for large value of k. We divided the k-mers into two parts: one (r-mer) for binning, and the other (l-mer) for 1-neighborhood testing. The size of the bin decides the average number of nodes required the pairwise distance computation. Note that the SdB graph is a generalization of prior approaches with bin size of k (respectively, 0) corresponding to a standard dB graph (respectively, BayesHammer like graph). In our tests, we did some empirical tests to choose r and l and found the performance to be robust to difference choices. Therefore, we worked with $r = 10$, $l = 20$, leaving the optimization of parameters to future work. However, to allow for fair comparisons, we tested the SdB graphs against dB graphs using a range of values of k.

Given r, l, we build a SdB graph as follows

1 Each node initially corresponds to a distinct $(r + l)$-mer from the read. Node $u = (x, y)$, in which x is a length r of prefix of the node, and y is a length l of suffix of the node.

2 Consider nodes $u = (x, y)$, and $v = (x', y')$. We connect u and v by an edge, if the $r + l$ - 1 suffix of u matches the prefix of v, and a read matches the combined sequence. The weight of an edge (u, v) is the number of reads that contain the combined sequence. The weight of node u is the maximum of the sum of incoming or outgoing edge-weights. This operation mimics a standard dB graph construction.

3 Consider nodes in order of decreasing weights, and repeat the following until no node is left

(a) Pick node $u = (x, y)$. For all nodes $u' = (x', y')$, merge u with u' (and remove from further consideration) if $d_h(x, x') = 0$, $d_h(y, y') \leq 1$ and u is the heaviest, in which $d_h(x, x')$ is hamming distance between x and x'. Merge any multiedges into a single edge of weight equal to the sum of the weights of the merged edges. Note that the actual implementation speeds this computation by hashing on the prefix strings.

The construction of a SdB is illustrated in supplemental Fig. S3. A (3, 3) SdB graph successfully compacted the data with no false-positives or false-negatives except because of sequencing error. As supplemental Fig. S3 shows, there are two distinct paths, corresponding to the two genes. However, because of sequencing error, we see a small branching in gene 1. This can be controlled by an error correction procedure, described in the next section. In contrast, supplemental Fig. S4 shows examples of dB graphs with the choice of $k = 4$ and $k = 5$ using same reads. A 4-mer dB graph connected false edges at node "*GAAT*", producing false paths combining gene 1 and 2. On the other hand, a 5-mer dB graph failed to connect edges in both genes, and neither gene could be represented by a single path. In the Results section, we systematically compare the performance of SdB graphs and dB graphs.

Error Correction—Sequencing errors also result in false overlaps. An error toward the end of the read (within k nucleotides) leads to a "tip" in the dB graph, whereas an error in the middle of the read leads to a "bulge." After its construction, the SdB graph can be viewed as a regular $(r + l)$-mer graph; graph simplification methods, such as tip clipping and bulge removing can be applied. For transcript assembly, uniform coverage pruning may delete some true sequences, so we use a proportional approach to rescue lower-abundance transcripts similar to the one used by IGdb, Trinity and IDBA-tran (48, 49, 34).

Assuming for simplicity that sequencing errors are independent and identically distributed with ε_s denoting the nucleotide error probability. The number of reads matching a specific k-mer is proportional to $(1 - \varepsilon_s)^k$. On the other hand, the number of reads matching a k-mer with a mismatch at a specific position is proportional to $\frac{1}{3} \cdot \varepsilon_s \cdot (1 - \varepsilon_s)^{k-1}$. The expected ratio of read depths of the true edge to any false edge is given by

$$\frac{(1 - \varepsilon_s)}{\frac{1}{3} \cdot \varepsilon_s}$$

The expression is usually $\gg 1$, for typical values of $\varepsilon_s \sim 1\%$. Therefore, sequencing errors can be overcome if the sequence coverage is high enough.

AbScan differentiates true mutations from sequencing errors using the same idea. In ideal case, any genes conveying mutations are regarded as separate genes and the graph maintains separate paths. However, if two genes are separated only by a few polymorphisms, then the graph may merge some nodes in paths. For SdB graphs, an exact match requirement for the r-mer would result in a bulge where one collection of r-mers carry the mutation, and the other collection contain r-mers carrying the reference nucleotide. In the case of a true mutation, these bulge would be well-supported by reads, and not removed during error correction.

FASTA Conversion—To use the SdB graph to construct a FASTA database, we associated a sequence with each node. The sequence of the source is the *r*-mer; the sequence associated with the sink is the last nucleotide of its *r*-mer concatenated with its *l*-mer. For all nonsource, nonsink nodes, the associated sequence is simply the last nucleotide of the *r*-mer. The sequence of a path in the graph is the concatenation of sequences associated with nodes on the path. A compact FASTA database is constructed from the SdB graph by enumerating the paths as described. The sequences in the path were converted to the amino acid FASTA format to generate a database for the MS/MS database search tools, using the SpliceDB tool (32) for this conversion. 69.3MB of FASTA form amino acid database was created, concentrating on the antibody sequence generated from 162.7GB of RNA-seq bam files.

Multistage Search—The antibody database adds some noise to the search and it is possible that a PSM to a known peptide has a better score against an antibody peptide, leading to false identification. As a conservative strategy to avoid false identifications, we use a modified multistage search (33). We first searched all spectra using MS-GF+ against a known protein database (Ensembl version GRCh38) (38). All PSMs identified as a non-Ig known peptide from the spectrum level 1% FDR search of the Ensembl database were excluded from the second search. Spectra that could not be matched were searched using MS-GF+ against the antibody database, using a target-decoy strategy with 1% spectrum level FDR.

Comparison to rnaSPAdes—As transcriptome assembly is a well-established research area, we built database using a popular transcriptome assembly tool, rnaSPAdes (50) to compare with AbScan. To make a fair comparison, we applied the identical read set for assembly. SPAdes version 3.9.0 was used with options "-only-assembler" and "-rna". The output nucleotide sequence translated to FASTA form amino acid sequence for MS/MS search.

Identifying Antibody Peptide Location—For all identified antibody PSMs, we found the most likely position in the antibody structure. To do this, we recovered the nucleotide sequence of the peptide from the SdB graph, then compared it to sequences with the IMGT sequence to find the best matched position of each PSM to IMGT reference sequences. Finally, we incorporated gaps to the position using IMGT multiple sequence alignments to get a normalized position. Fig. 1 shows the expected position of the peptides we identified from the colorectal tumor MS/MS data and polyclonal antibody MS/MS data. Each horizontal black line represents the distinct peptide sequence. Peptides that do not map to IMGT reference sequences are not displayed.

Statistical Test for Antibody Enrichment—The two-stage search resulted in PSM identifications with spectra matching "known peptides" and antibody peptides (SdB database). If in some sample, the MS/MS data was known to not contain any antibody peptide (*e.g.* cell-line), then any PSM in the SdB database corresponds to a false identification. The number of false identifications is expected to grow linearly with the number of known peptide identifications. Therefore, we considered the fraction

$$\frac{\#\ \textit{of PSMs in SdB database}}{\#\ \textit{of PSMs in known peptide database}}$$

for all MS/MS data, and considered the Null hypothesis that this fraction was constant in all cases, colorectal tumor, colorectal normal, and colorectal cell-lines. To calculate the *p* values, we applied Pearson's χ^2 test in a 2 × 2 contingency table.

$$\chi^2 = \sum_{i=1}^{4} \frac{(O_i - E_i)^2}{E_i}$$

in which

χ^2 = Pearson's cumulative test statistic,
O_i = Number of observation of type i
E_i = Theoretical frequency of type i

The *p* value was calculated from the χ^2 distribution table.

Antibody and SAAV Peptides Correlation Test—Consider a table, where columns correspond to samples (each column is a different sample), and rows correspond either to SAAV peptides (possible antigens) or to antibody peptides. The cells mark the presence or absence of the peptides in the specific sample. For any pair of antibody and SAAV peptides, we used the Fisher's exact test to measure correlation of occurrence. As many pairs were used, we used a target-decoy based approach to compute the false discovery rate for any *p* value cut-off.

For each row, the columns were permuted independently so that any correlation between two rows (an antibody peptide, SAAV peptide pair) was just by chance, and a Fisher exact test was used to compute the correlation among all pairs. Highly correlated pairs of antibody and SAAV peptides were identified by applying a 5% FDR threshold.

Measuring Immune Response—Measurement of the immune response for each individual was accomplished by counting the number of antibody PSMs. As the total number of spectra and their quality were not identical for every sample, the antibody PSMs were normalized by the total number of PSMs to the "known database" search. Fig. 2(c) shows the distribution of the normalized immune response of each individual in both tumor and normal samples. We simply took the top 45% and bottom 45% group of individuals in terms of their normalized immune response.

Survival Rate Comparison—We designed a method that takes a collection of peptides, and samples, groups samples based on cooccurrence of peptides, and tests if the individual groups have different survival times. Specifically, we used the following strategy:

1 Represent each peptide *p* as a binary vector **p** over all samples with $\mathbf{p}_i = 1$ (respectively, $\mathbf{p}_i = 0$) indicating the presence (respectively, absence) of peptide *p* in sample *i*.

2 Cluster the peptide vectors into two groups (arbitrarily labeled +, −) using 2-means clustering.

3 Assign score S_i to each sample i using.

S_i = (Number of "+" assigned peptides in sample *i*) − (Number of "−" assigned peptides in sample *i*).

4 Pick two sets of samples: Bottom (45%) of all samples with the lowest score, and top (45%) with the highest score.

5 Perform the Kaplan Meier log-rank test on the two groups of samples to test for correlation with clinical outcome.

We performed this test using all antibody and mutated peptides that significantly cooccurred in the samples exceeding a 5% FDR threshold of correlation test. The test statistic could include some unknown bias, and it wasn't clear if they followed the χ^2 distribution used to compute a *p* value. To test this, we set two groups of patients in which each group included 45% of random samples without replacement, and then we calculated the test statistics of two groups of random patients by log-rank test. We repeated the process 10,000 times to create the distribution of test statistics (supplemental Fig. S5).

RESULTS

Analytical Comparison of SdB and dB Graphs—We compared the performance of SdB graphs *versus* dB graphs using both analytical methods as well as empirical data from simulations. Let p_s denote the probability that a randomly chosen pair of nucleotides is identical. Thus, the probability of a false *k*-mer match is ${p_s}^k$. To allow for fair comparisons, parameters *r*, *l*, *k* were selected so that the probability of an (*r*, *l*) match among unrelated reads in a SdB graph is the same as the probability of a k-mer match in dB graph. Specifically (Supplementary Methods—"Analytical comparison of SdB and dB graphs"),

$$p_s^k \geq p_s^{(k+l)} \cdot \left(1 + l\,\frac{1-p_s}{p_s}\right) \qquad \text{(Eq. 1)}$$

$$k \leq r + l + \log_{p_s}\left(1 + l\,\frac{1-p_s}{p_s}\right) \qquad \text{(Eq. 2)}$$

For any r, l, we chose k to be the largest value satisfying constraint 2. We also computed the probability of false overlaps. (See Supplemental Methods–"Comparison between the SdB and dB graph mathematically"). Using these calculations, we can show that SdB graphs have significantly lower false negative rates compared with dB graphs. For example, let $p_s = \frac{1}{4}$. When $r = 10$, $l = 20$, the choice of $k = 27$ equated the false overlap rates for both methods at $5.55 \cdot 10^{-17}$. However, for an overlap of 40 bp, $\varepsilon = 0.01$, we computed a false negative rate of 13.9% for the dB graphs *versus* a rate of 2.2% for the SdB graphs.

To test these theoretical results, data was simulated by generating the 100, 000 overlapping regions, of length from 30 to 100, with uniform sequencing error rate ε. Reads were connected by a path in the dB graph, if there was at least one k-mer consecutive sequence without an error. Similarly, they were connected by a path in a SdB graph, if there was an $(r + l)$-mer in which the first r nucleotides had no error and the following l-mer had at most one mismatch. supplemental Fig. S6 showed a complete concordance between theoretical and simulated results. The sensitivity for all methods increases with length of overlap and decreases with higher ε. SdB graphs consistently outperform dB graphs.

Comparison on Simulated Antibody Reads—To provide a more direct comparison of the performance of SdB graphs and dB graphs on Ig sequences, we employed a second simulation, starting with a single IMGT reference antibody sequence denoted by A. Note that an antibody (supplemental Fig. S7) is a "Y" shape protein and consists of a variable region and constant region. The variable region is formed by selecting a gene from each of 3 sets V, D, and J which are brought together by recombination and splicing. The combined variable region itself can be divided into a framework (FR) which is relatively constant, and three hypervariable complementarity determining regions (CDRs; supplemental Fig. S7) (42). In the simulation, A was created by joining known V, D, and J regions (IGHV1-18*01, IGHD1-1*01, IGHJ1*01; (51); supplemental Fig. S8). We generated a collection D of decoy sequences in which each nucleotide was chosen uniformly at random, except for the insertion (at a random position) of a single substring of A. The insertions were of varying lengths ranging from 20 to 26. The antibody reference A and decoy gene sequence collection D were used as a template to simulate reads, using the tool *wgsim* (https://github.com/lh3/wgsim), with sequence error rate set at $\varepsilon = 1\%$. A dB graph and a SdB graph was built using these reads to measure the false positive and false negative results from these graphs.

Let $G = (V, E)$ denote the dB or SdB graph, depending on context, whereas the graph $G_A = (V_A, E_A)$ is constructed solely using A. In the ideal scenario, G and G_A should be identical. Therefore (supplementary Methods), the false negative rate (denoted by F) can be estimated using

$$F = \frac{|E - E_A|}{|E_A|}$$

The false positive rate was measured indirectly, using divergence (denoted by D)

$$D = \frac{\sum_{n^{\varepsilon}V_A}((n_i - 1) + (n_o - 1))}{|E_A|}$$

in which n_i is the in-degree and n_o is the out-degree of node $n^{\varepsilon}V_A$. The divergence provides a measure of false connections for the antibody sequence A.

Note that false positive edges can also arise because of sequencing errors. However, most dB construction corrects for such errors by choosing an appropriate threshold for coverage, along with other methods (48, 49, 34). However, the appropriate threshold is different for each value of coverage. We chose a principled method for choosing coverage to remove false positive edges because of sequencing errors for both dB, and SdB. After coverage filtering, the false negative rate and divergence was measured as a function of increased coverage (Fig. S9). Supplemental Fig. S9 shows an explicit tradeoff among false negatives, and divergence (false positives) for dB graph methods. At any specific fixed coverage parameter (*e.g.* 10×), the false negative rate of the dB graph increases with increasing values of k, even as divergence decreases, making it difficult to simultaneously improve both metrics. In contrast, SdB graphs show consistently lower divergence and false negatives for all coverage values.

Read Filtering—Before we construct the split de Bruijn graph, we need to collect the reads that encode Ig gene transcripts (See Experimental Procedures - "Read filter"). We tested the quality of read filtering by a partial alignment of filtered reads to the reference antibody sequences. A virtual antibody reference was set to represent the variable regions of all antibodies. We adjusted the gap between this virtual antibody and each individual antibody using the IMGT antibody reference with gap. The matching k-mer was used to anchor the alignment, and the extent of the alignment was determined simply by the length of read on each side of the k-mer. The anchored position of the read was transferred to virtual antibody position and used to estimate the overall coverage. We counted the number of reads passing through each unique position of the virtual antibody. supplemental Fig. S10 describes a coverage because of the partial alignment of

TABLE I
Number of identified PSM (peptides)

Data set	Sdb Graph DB	DB graph DB	Ensembl DB
Cell line	0 (0)	0 (0)	117,679 (14,527)
Normal	711 (113)	700 (96)	1,705,785 (85,956)
Tumor	54,909 (1940)	16,364 (1088)	5,573,094 (129,886)
IG purified	16,404 (3029)	9,576 (2338)	989 (246)

all filtered reads, and shows that the reads are filtered without apparent bias except at the very end of the sequence.

MS-MS Based Discovery of Antibody Peptides—We used four mass spectrometry data-sets. To test the algorithms, a data-set of spectra acquired from a purified polyclonal antibody mixture (*antibody purified*) was used (37). To test for antibody peptides in tumor samples, we used a collection of MS/MS spectra from 90 distinct colorectal tumor samples from the CPTAC project (35) (*colorectal tumor*). As negative control, we used spectra acquired from 30 normal colon biopsies (35) (*colorectal normal*). As a second control, we used spectra from colon cancer cell-lines LIM1215, LIM1899, and LIM2405 (denoted as *colon cell-lines*) (36).

SdB and dB graphs were designed and implemented, using 162.7GB RNA-seq reads of 90 individuals downloaded from The Cancer Genome Atlas (TCGA) repository (52). The two approaches resulted in a 69.3MB and 107.8MB FASTA-formatted amino acid database. A multistage search (See Experimental Procedures Multistage Search) using known proteins and SdB graphs (respectively, known proteins and dB graphs) was conducted to identify peptide spectrum matches (PSMs). A summary of the results of those searches is presented in Table I. The list of identified spectra and other details are presented in supplemental Table S1 - "Link to the list of PSM and spectrum image."

Note that the antibody-purified data set presents an interesting challenge, as the SdB graph was constructed from RNA of completely different individuals. Even so, our search identified 16,404 antibody PSMs (3167 peptides) out of 116,018 total spectra (PSM identification rate 14%). Fig. 1*A* shows that the identified peptides cover the entire space of the antibody. Table I also allows for a comparison of the SdB graph and dB graph databases, as both use the same read set as their inputs. At identical FDR cut-off (1%), SdB graphs identify 3.3× as many PSMs as the dB for the colorectal tumor, and 1.7× as many PSMs for antibody purified dataset, consistent with simulation results. On the other hand, the number of PSMs identified in the colorectal normal, and cell-line colorectal samples are similar, validating the proposition that SdB graphs can filter out erroneous PSMs at the same rate as dB graphs. Therefore, SdB graphs reduce both false positives and false negatives in the real data, identifying more true PSMs without increasing false PSMs.

In the sample matched colorectal tumor spectra, 54,909 PSMs (1,940 peptides) were identified. We asked if these large numbers of antibody peptides originated from tumor infiltrating lymphocytes, or from other sources. For example, these immunoglobulin identifications could simply correspond to floating antibodies from blood contamination, or they could be misidentified (modified) peptides. In the first case, we would expect to see similar numbers of antibody peptides in colorectal tumor and colorectal normal data set. In the second case, we would expect to see similar numbers of antibody peptides in colorectal cancer, and colon cell-lines (Fig. 2*A*).

We normalized PSM counts to the number of PSMs in the Known DB before comparing across samples (Fig. 2*B*). The normalized PSM count in the colorectal normal data-set was only 4.69% of the colorectal tumor counts (p value < 0.0001; See Experimental Procedures, Statistical Test for Antibody Enrichment). The normalized PSM count in colon cell-lines was 0 consistent with the observation that TILs were the source of the antibody peptides observed in the colorectal tumor. Although it is likely that the actual numbers would depend on experimental handling of tumor interstitial fluid (TIF), the tumor and normal cells were processed in an identical fashion, and would have similar biases in terms of TIF handling. We additionally tested the samples for presence of biomarkers, and identified 113 PSMs matching CD38 in tumor samples compared with 1 PSM in normal samples. Similarly, we observed CD74 predominantly in tumor samples (684 PSMs *versus* 34 PSMs). These represent significant enrichment even after accounting for the 3x larger number of tumor samples. CD38 is a glycoprotein found on the surface of many immune cells including CD4+, CD8+, B-lymphocytes and natural killer cells (53, 54), whereas CD74 has been reported to possibly reflect an intratumoral immune response with TIL association (55).

The assembly of RNA-seq reads is a well-established research area of genomics (44, 48, 49). However, genome assembly tools are designed to be general, and may not do a good job of assembling Ig genes. As the reconstruction of Ig genes from the RNA-seq reads was a key part of our pipeline, we asked if the use of the RNA assembly tools could provide better results. To test this, a popular transcriptome assembly tool, rnaSPAdes (50), was used to assemble RNA-seq reads from one colorectal tumor sample. We searched the MS/MS data from the same sample against databases constructed using rnaSPAdes and SdB graphs. The number of spectra identified using the SdB graph method was 2450, compared with 528 using rnaSPAdes, suggesting that general purpose transcript assembly tools were not suitable for studying the antibody repertoire at the protein level.

SAAV Discovery—We used the SAAV peptides from the results of previous studies (33, 34), but with additional filtering. We remove all peptides in which the mutation has a mass difference of one. We also enumerate all reference peptides with common modifications that shared some sequence tag with the mutated peptide and scored them to see if a refer-

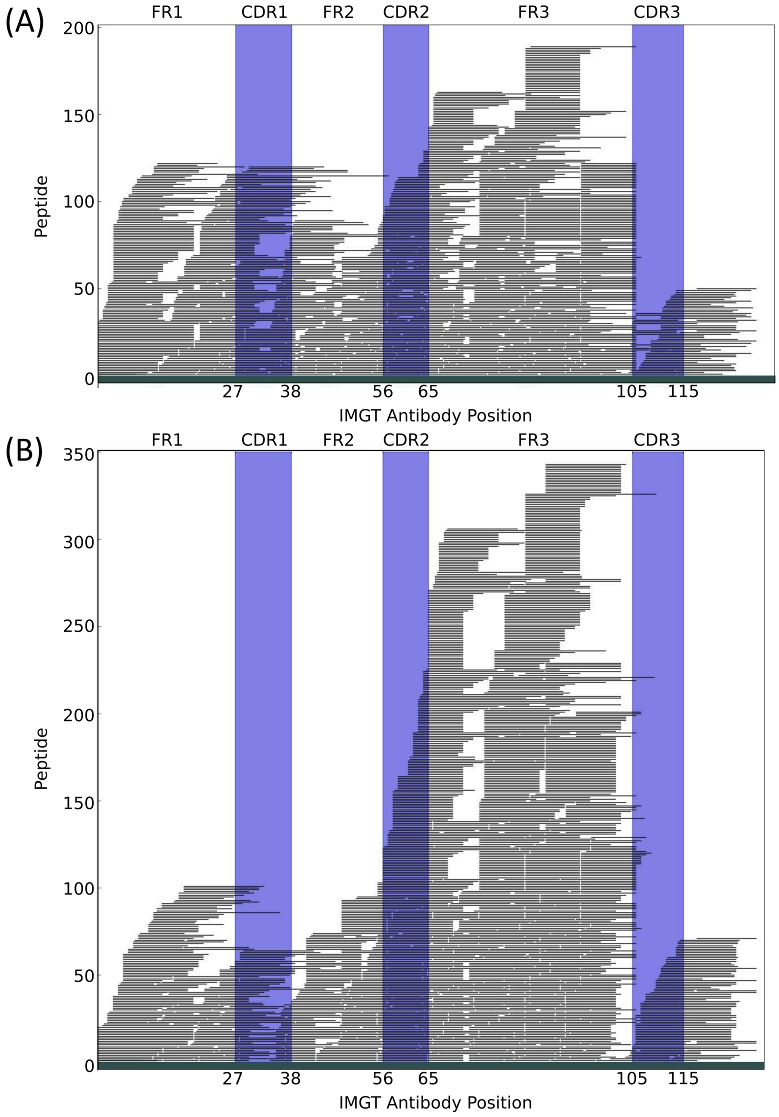


FIG. 1. **Relative locations of identified antibody peptides.** Each horizontal black line represents a distinct peptide sequence. Trypsin was applied for the colorectal tumor MS/MS spectra assessment, and four different enzymes were applied for polyclonal antibody MS/MS spectra assessment. Both spectra sets were searched against the same antibody database constructed using tumor RNA-seq reads driven by TCGA. *A,* Antibody PSMs from colorectal tumor MS/MS data. *B,* Antibody PSMs from polyclonal antibody MS/MS data.

ence peptide could better explain the data. The final list of mutated peptides is presented in supplemental Table S3, and the annotated mass spectra are in MassIVE, and link is provided in supplemental Table S1. The filtered list contains 677 SAAV peptides.

Antibody Peptide-SAAV Peptide Correlation—We asked if the antibody peptides discovered in the colorectal tumor data set could be targeting specific neo-antigens. The neo-antigens are possibly mutated peptides that are recognized by TILs and antibodies. Many somatically mutated peptides had been detected in the colorectal tumor data in the original seminal study (35) and our own group's re-analysis (34). supplemental Table S4 shows the occurrence of mutated nonreference peptides, and all antibody peptides in each of the 90 samples. As peptides that are polymorphic in the population could still be somatic in individuals, and some polymorphisms are known to be functionally deleterious, we used all mutated nonreference peptides.

For every antibody peptide-SAAV peptide pair in this table, a calculation was made to determine the significance of cooccurrence using the Fisher exact test. Because many pairs were to be tested, a target-decoy approach was used to compute the false discovery rate for significant pairs. The decoy statistics were computed by permuting the occurrence of each peptide in the sample. Fig. 3 shows the distribution of *p* values computed from the target and the decoy table. At a nominal *p* value threshold of 0.00025, we see 163 pairs that

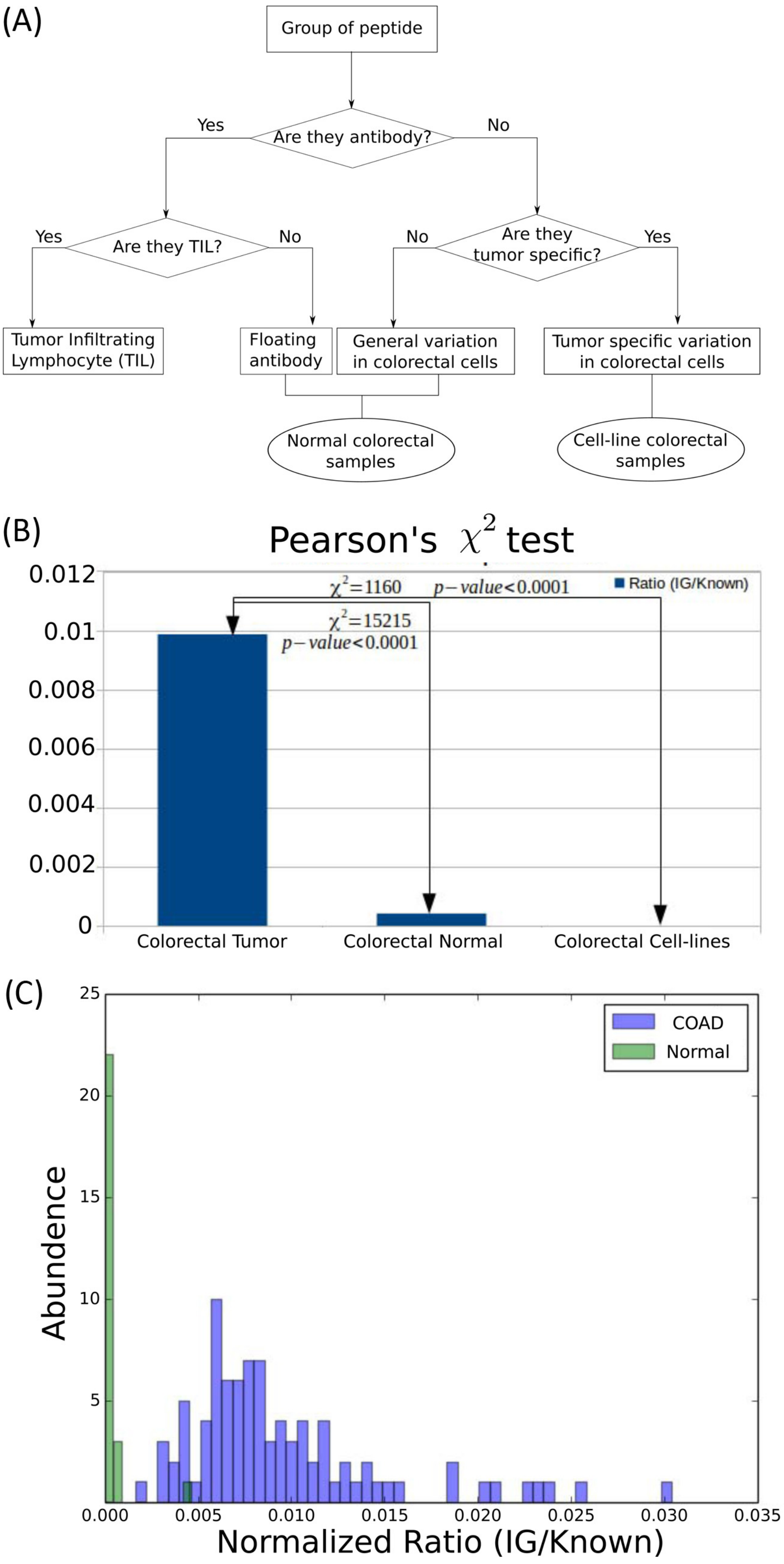
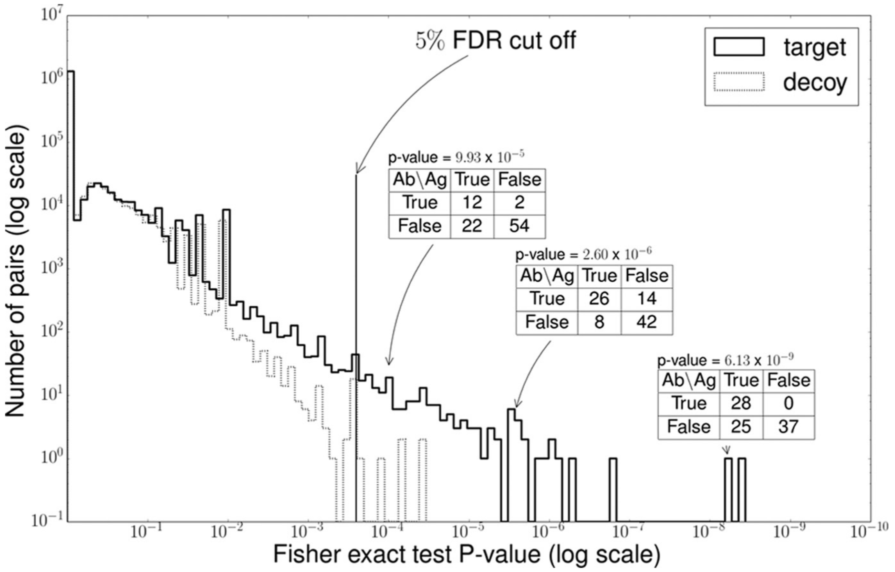

(A)
Group of peptide
Are they antibody?
Yes
No
Are they TIL?
Yes
No
Are they tumor specific?
No
Yes
Tumor Infiltrating Lymphocyte (TIL)
Floating antibody
General variation in colorectal cells
Tumor specific variation in colorectal cells
Normal colorectal samples
Cell-line colorectal samples
(B)
Pearson's χ^2 test
$\chi^2=1160$ $p-value<0.0001$
$\chi^2=15215$ $p-value<0.0001$
Ratio (IG/Known)
0.012
0.01
0.008
0.006
0.004
0.002
0
Colorectal Tumor
Colorectal Normal
Colorectal Cell-lines
(C)
COAD
Normal
25
20
15
10
5
Abundence
0.000
0.005
0.010
0.015
0.020
0.025
0.030
0.035
Normalized Ratio (IG/Known)

FIG. 3. **Peptide correlation test.** We tested the correlation between the antibody peptides and mutated peptides. For every pair of peptides, we counted the number of samples cooccurring with these peptides and then we applied Fisher exact test to calculate the *p* value. For example, the peptide pairs of NTLYLQMDSLR (antibody) and AAQAQGQSCEYSLMVGYQCGQVF(Q→R) (SAAV peptide) cooccurred in 26 samples, and there was a coabsence in 42 samples. It was revealed that 68 of the 90 samples shared the cooccurrence of this pair with a *p* value of 2.60 × 10^{-6}. We drew the histogram of *p* values of all pairs in supplemental Table S4. We also drew the histogram of the *p* values from the decoy table generated by the random permutation of values. A 5% FDR threshold was applied to collect the high correlated pairs.



Ab\Ag	True	False
True	12	2
False	22	54

Ab\Ag	True	False
True	26	14
False	8	42

Ab\Ag	True	False
True	28	0
False	25	37

exceed this threshold, *versus* 5 decoy pairs, suggesting a small false discovery rate of ≤ 5%.

One example of these cooccurring pairs is the the antibody peptide NTLYLQMDSLR, and SAAV peptide AAQAQGQSCEYSLMVGYQCGQVF(Q→R). The antibody peptide NTLYLQMDSLR belongs to variable region of IGHV3–64D*06 and the mutated peptide *pep* = AAQAQGQSCEYSLMVGYQCGQVF(Q→R) belongs to the gene FBLN1 reported to be downregulated in colorectal cancer cells (56). Among 90 samples, both peptides are expressed in 26 samples, and neither is found in 42 samples, giving a Fisher exact test *p* value of 2.59 × 10^{-6}. supplemental Fig. S11 shows examples of peptide spectrum matches of these peptides. The mutation in peptide *pep* is a known polymorphism (dbSNP rsID136730). However, the mutation is very low frequency in normal population surveys 0.14% in ExAC, and 0.04% in 1000 Genomes project (57) compared with its occurrence in 34 out of 90 samples. It is also known that nonsomatic, self-peptides can elicit an immune response against tumor cells (58). Therefore, the functional relevance of *pep* cannot be rejected based solely on its classification as (non)somatic. Finally, it is important to assert that cooccurrence does not indicate cooccurrence only between the specific antibody peptide and the SAAV, but rather between the antibody carrying NTLYLQMDSLR and some peptide in the mutated version of FBLN1 product. In fact, we see another antibody peptide LSCAASGFSFR in the FR1/CDR1 region that also cooccurs with *pep* (*p* value: 9.93 × 10^{-5}). Therefore, we did not filter antibody peptides by their location (CDR/FR) before testing for correlation.

We also used both cooccurrence and coabsence to test correlation. Although “absence” of a peptide may be because of experimental protocol, coabsence is also indicative of a correlation. As an extreme example, the pair of antibody peptides NGPSVFPLAPSSK and mutated peptide AGRPVICATQMLESMIK were observed in 29 samples and 28 samples, respectively. If they had no correlation, then over the 90 samples, we would expect 9 samples to carry them both, just by chance. Instead we see zero (*p* value: 1.44 · 10^{-6}). This suggests that the existence of this mutated peptide (perhaps indirectly) reduced the affinity to this specific antibody leading to a negative correlation, and the effect was independent of the event that we missed identifying the peptides.

Correlation Between Antibody Expression and Survival Status—The antibody peptide repertoire might provide a snapshot of the immune response to cancer. We anticipated that the patients with higher immune response could have a different clinical outcome than those with lower immune response because of the role of TILs in mediating response to cancer (59–61).

We first measured the immune response of an individual as the fraction of identified peptides that came from the antibody

FIG. 2. **Comparison of identified antibody PSMs per experiment and sample.** *A,* The source of antibody peptides in different samples. PSMs that match nonreference peptides are either mutations or antibody peptides. Antibody peptides should not be observed in cell-lines. However, floating antibodies could be observed in normal colorectal samples. Antibodies from Tumor infiltrating lymphocytes should only be observed in tumor samples. *B,* Occurrence of antibody peptides in tumor, normal, and tumor derived cell-lines are significantly different for MS/MS spectra of tumor, normal, and cell-line colorectal samples. Each spectra set was searched against the Ensembl GRCh38 protein database (38) and a custom antibody database. The number of PSMs identified as antibody peptides were 54K (*colorectal tumor*), 711 (*colorectal normal*), and 0 (*Cell-lines*). The PSM counts were normalized against the number of PSMs to known peptides (5.5 M in *colorectal tumor*, 1.7 M in *colorectal normal*, and 0.1 M in *Cell-lines*). The normalized ratios suggest that a significantly larger fraction of the colorectal tumor PSMs are antibody peptides, compared with the other two data-sets (Pearson's χ^2 *p* value < 10^{-4}). *C,* The distribution of the number of samples carrying a normalized fraction of antibody peptides. COAD samples carry a higher fraction of antibody peptides.

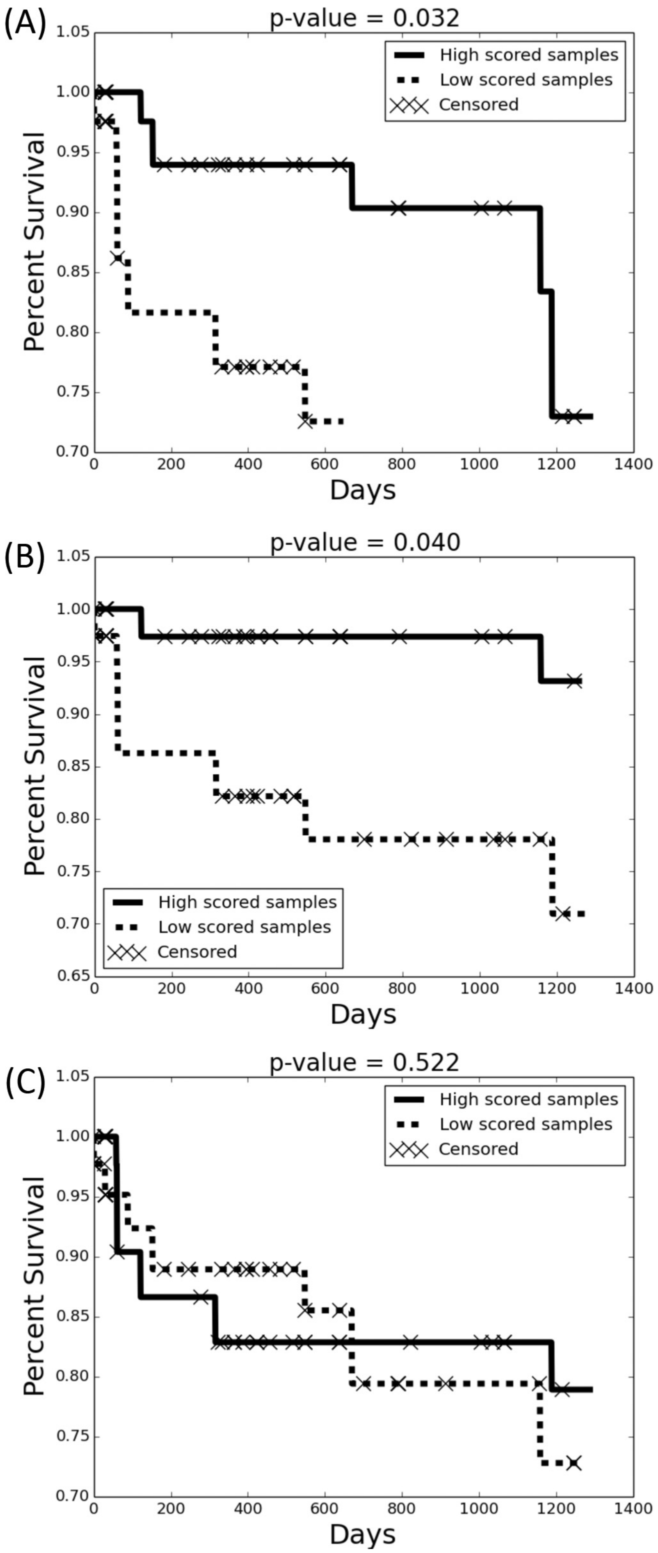


FIG. 4. **Kaplan-Meier survival estimator.** For any subset of peptides, we bi-partioned peptides based on coexpression in samples. Next, we scored each sample based on the homogeneity of peptides from a single partition in that sample (Methods). The highest and lowest scoring samples (45% each) were grouped, and were tested to determine the clinical outcome. The Kaplan-Meier survival estimator and log-rank test were applied to test the difference of the clinical outcome of two groups. When testing with cooccurring mutated peptide/antibody peptide pairs, we observed a significant correlation with survival (Plot (*A*): *p* value = 0.032). In contrast, the correlation was reduced when testing with only antibody peptides (Plot (*B*): *p* value = 0.040), and there was no-correlation when testing with mutated peptides. (Plot (*C*): *p* value = 0.522).

repertoire, and identified a subset of individuals as high-responders and low-responders (See Experimental Procedures, Measuring immune response, and Fig. 2*C*). We used the days-to-death values to get the Kaplan-Meier survival estimator for the two groups. Next, we used a log-rank test to compute a *p* value for the difference between the two curves. The *p* value was 0.75, indicating that we could not reject the Null hypothesis (supplemental Fig. S13).

We also considered the possibility that some, but not all peptides mediate a positive clinical outcome. Further, these peptides would be expressed in multiple individuals with similar outcomes. To test this hypothesis, we designed a method that takes any group of peptides, and clusters samples based on coexpression, but without knowledge of the clinical outcome in the individuals (See Experimental Procedures, Survival Rate Comparison). For a given collection of peptides, we tested the null hypothesis that there is no correlation among sample grouping and the clinical outcome.

We computed an empirical null distribution by choosing random subsets of individuals, and performing the log-rank test against clinical outcome. supplemental Fig. S5 shows that the test statistic under null hypothesis closely follows the theoretical χ^2 distribution.

In contrast, when we tested sample grouping using the correlated antibody, SAAV peptide pairs (See Experimental Procedures, Antibody and SAAV Peptides Correlation Test), we observed a significant differential response with *p* value: 0.032 (Fig. 4*A*). We also tested this method using two other groups of peptides. When we used all antibody peptides we also obtained a differential response with *p* value 0.040 (Fig. 4*B*). However, testing with all mutated peptides, we did not observe significant differential response, obtaining a *p* value of 0.522 (Fig. 4*C*). The small number of samples implies that our study is not fully powered and the results need to be replicated in larger cohorts. Nevertheless, they do show that antibody expression could be correlated with the clinical outcomes.

Discussion and Future Study—Understanding the immune response to cancer is key to cancer immunotherapy. Current approaches use serum or plasma samples and specifically focus on isolating differentiated B cells for analyzing antibodies. However, the serum antibody repertoire may contain a larger pool of antibody sequences, not just the ones responding to tumor neo-antigens. In this paper, we mined spectra acquired from isolated (colorectal) tumor cells, and identified a large number of antibody peptides. Our results suggest that infiltrating lymphocytes in the tumors generate antibodies in response to the tumor. They also suggest that somatic coding mutations in the tumor genome act as neoantigens triggering antibody generation. We observed recurrence of antibody and mutated peptide sequences that cannot be explained as chance events, and showed a positive association between clinical outcome (survival time), and the antibody response. Together, the results underscore the need for systematic analysis of the tumor antibody repertoire.

The identification of antibody peptides using tandem mass spectrometry is technically challenging. In the ideal case, the spectra should be searched against transcript data from differentiated B-cells from the same individual. However, that data may not always be available. Moreover, it is not known if circulating B cells have the same antibody repertoire as the tumor infiltrating lymphocytes. In this paper, we used RNA-seq data, not from isolated B cells, but from the same tissue that the proteome was extracted. Nevertheless, we managed to get significant coverage of antibody peptides. We identified a large number of peptides even when we used MS data from unmatched samples. Future research will focus on the differences among different sequencing approaches, such as IG-seq, and RNA-seq.

The hyper-variability of antibody sequences makes it challenging to construct databases that can be searched with MS spectra. We proposed a new structure, called the SdB graph, and showed improved performance in compressing and creating MS-searchable databases relative the dB graphs. The SdB graphs are later converted into Fasta formatted databases that can be used for search with any tool. The software for developing SdB graph should be generally applicable for any hypervariable region, and is available for download. These techniques described here can be further improved and those will be the focus of future research.

We found that the SdB graph database generated from RNA-seq of TCGA tumor samples was also helpful in identifying antibodies from completely different samples. This raises the possibility that multiple RNA-seq samples from a specific tumor type could be used as a universal database, reducing the need for matched RNA and protein samples for decoding the immune repertoire. This will be explored in future work. At the end, we also hope that our preliminary results spurs a further investigation of the clinical outcome based on immune system response, and the development of diagnostic tools and therapies that can emerge from an analysis of the tumor immune repertoire.

Acknowledgments—We thank Dr. Alan Carter Covell for his assistance with proofreading the manuscript.

DATA AVAILABILITY

All datasets are available via urls below. 1. 90 colorectal tumor samples (https://cptac-data-portal.georgetown.edu/cptac/s/S022); 2. 30 normal colon biopsies (https://cptac-data-portal.georgetown.edu/cptac/s/S019); 3. Colon cancer cell-lines LIM1215, LIM1899, and LIM2405 (http://proteomecentral.proteomexchange.org/cgi/GetDataset?ID=PXD000120, PXD000120); 4. Purified polyclonal antibody mixture (ftp://massive.ucsd.edu/MSV000081401, MSV000081401).

* Dr. Vineet Bafna, Dr. Pavel A. Pevzner, and Seong Won Cha were supported by a grant from the NIH (P41RR24851, 1R01GM114362). Dr. Seungjin Na were supported by a grant from the NIH (5P41GM103484–07). Dr. Vineet Bafna is a cofounder, has an equity interest, and receives income from Digital Proteomics, LLC. The terms of this arrangement have been reviewed and approved by the University of California, San Diego in accordance with its conflict of interest policies.

S This article contains supplemental material.

‖ To whom correspondence should be addressed: Department of Computer Science and Engineering, University of California, San Diego, 9500 Gilman Drive, Mail Code 0404, La Jolla, CA 92093-0404. Tel.: 1-858-822-4978; E-mail: vbafna@cs.ucsd.edu.

REFERENCES

1. Curran, E., Corrales, L., and Kline, J. (2015) Targeting the innate immune system as immunotherapy for acute myeloid leukemia. *Front. Oncol.* **5,** 83
2. Riches, J. C., and Gribben, J. G. (2014) Immunomodulation and immune reconstitution in chronic lymphocytic leukemia. *Semin. Hematol.* **51,** 228–234
3. Vincent, K., Roy, D. C., and Perreault, C. (2011) Next-generation leukemia immunotherapy. *Blood* **118,** 2951–2959
4. Jimenez-Luna, C., Prados, J., Ortiz, R., Melguizo, C., Torres, C., and Caba, O. (2016) Current status of immunotherapy treatments for pancreatic cancer. *J. Clin. Gastroenterol.* **50,** 836–848
5. Kottschade, L., Brys, A., Peikert, T., Ryder, M., Raffals, L., Brewer, J., Mosca, P., and Markovic, S. (2016) A multidisciplinary approach to toxicity management of modern immune checkpoint inhibitors in cancer therapy. *Melanoma Res.* **26,** 469–480
6. De Vries, J., and Figdor, C. (2016) Immunotherapy: Cancer vaccine triggers antiviral-type defences. *Nature* **534,** 329–331
7. Lollini, P. L., Cavallo, F., Nanni, P., and Forni, G. (2006) Vaccines for tumour prevention. *Nat. Rev. Cancer* **6,** 204–216
8. Rosenberg, S. A., Yang, J. C., and Restifo, N. P. (2004) Cancer immunotherapy: moving beyond current vaccines. *Nat. Med,* **10,** 909–915
9. Sampson, J. H., Archer, G. E., Mitchell, D. A., Heimberger, A. B., and Bigner, D. D. (2008) Tumor- specific immunotherapy targeting the EGFRvIII mutation in patients with malignant glioma. *Semin. Immunol.* **20,** 267–275
10. Boutros, C., Tarhini, A., Routier, E., Lambotte, O., Ladurie, F. L., Carbonnel, F., Izzeddine, H., Marabelle, A., Champiat, S., Berdelou, A., Lanoy, E., Texier, M., Libenciuc, C., Eggermont, A. M., Soria, J. C., Mateus, C., and Robert, C. (2016) Safety profiles of anti-CTLA-4 and anti-PD-1 antibodies alone and in combination. *Nat. Rev. Clin. Oncol.* **13,** 473–486
11. Camacho, L. H., Antonia, S., Sosman, J., Kirkwood, J. M., Gajewski, T. F., Redman, B., Pavlov, D., Bulanhagui, C., Bozon, V. A., Gomez-Navarro, J., and Ribas, A. (2009) Phase I/II trial of tremelimumab in patients with metastatic melanoma. *J. Clin. Oncol.* **27,** 1075–1081
12. Hodi, F. S., O'Day, S. J., McDermott, D. F., Weber, R. W., Sosman, J. A., Haanen, J. B., Gonzalez, R., Robert, C., Schadendorf, D., Hassel, J. C., Akerley, W., van den Eertwegh, A. J., Lutzky, J., Lorigan, P., Vaubel, J. M., Linette, G. P., Hogg, D., Ottensmeier, C. H., Lebbe, C., Peschel, C., Quirt, I., Clark, J. I., Wolchok, J. D., Weber, J. S., Tian, J., Yellin, M. J., Nichol, G. M., Hoos, A., and Urba, W. J. (2010) Improved survival with ipilimumab in patients with metastatic melanoma. N. *Engl. J. Med.* **363,** 711–723
13. Koff, W. C., Burton, D. R., Johnson, P. R., Walker, B. D., King, C. R., Nabel, G. J., Ahmed, R., Bhan, M. K., and Plotkin, S. A. (2013) Accelerating next-generation vaccine development for global disease prevention. Science **340,** 1232910
14. Hueber, W., and Robinson, W. H. (2006) Proteomic biomarkers for autoimmune disease. *Proteomics* **6,** 4100–4105
15. Robinson, W. H., DiGennaro, C., Hueber, W., Haab, B. B., Kamachi, M., Dean, E. J., Fournel, S., Fong, D., Genovese, M. C., de Vegvar, H. E., Skriner, K., Hirschberg, D. L., Morris, R. I., Muller, S., Pruijn, G. J., van

Venrooij, W. J., Smolen, J. S., Brown, P. O., Steinman, L., and Utz, P. J. (2002) Autoantigen microarrays for multiplex characterization of autoantibody responses. *Nat. Med.* **8,** 295–301

16. Legutki, J. B., Zhao, Z. G., Greving, M., Woodbury, N., Johnston, S. A., and Stafford, P. (2014) Scalable high-density peptide arrays for comprehensive health monitoring. *Nat. Commun.* **5,** 4785
17. Price, J. V., Tangsombatvisit, S., Xu, G., Yu, J., Levy, D., Baechler, E. C., Gozani, O., Varma, M., Utz, P. J., and Liu, C. L. (2012) On silico peptide microarrays for high-resolution mapping of antibody epitopes and diverse protein-protein interactions. *Nat. Med.* **18,** 1434–1440
18. Phizicky, E., Bastiaens, P. I., Zhu, H., Snyder, M., and Fields, S. (2003) Protein analysis on a proteomic scale. *Nature* **422,** 208–215
19. Larman, H. B., Zhao, Z., Laserson, U., Li, M. Z., Ciccia, A., Gakidis, M. A., Church, G. M., Kesari, S., Leproust, E. M., Solimini, N. L., and Elledge, S. J. (2011) Autoantigen discovery with a synthetic human peptidome. *Nat. Biotechnol.* **29,** 535–541
20. Tan, Y. C., Kongpachith, S., Blum, L. K., Ju, C. H., Lahey, L. J., Lu, D. R., Cai, X., Wagner, C. A., Lindstrom, T. M., Sokolove, J., and Robinson, W. H. (2014) Barcode-enabled sequencing of plasmablast antibody repertoires in rheumatoid arthritis. *Arthritis Rheumatol.* **66,** 2706–2715
21. Kerkman, P. F., Rombouts, Y., van der Voort, E. I., Trouw, L. A., Huizinga, T. W., Toes, R. E., and Scherer, H. U. (2013) Circulating plasmablasts/plasmacells as a source of anticitrullinated protein antibodies in patients with rheumatoid arthritis. *Ann. Rheum. Dis.* **72,** 1259–1263
22. Utz, P. J., and Anderson, P. (1998) Posttranslational protein modifications, apoptosis, and the bypass of tolerance to autoantigens. *Arthritis Rheum.* **41,** 1152–1160
23. Wrammert, J., Smith, K., Miller, J., Langley, W. A., Kokko, K., Larsen, C., Zheng, N. Y., Mays, I., Garman, L., Helms, C., James, J., Air, G. M., Capra, J. D., Ahmed, R., and Wilson, P. C. (2008) Rapid cloning of high-affinity human monoclonal antibodies against influenza virus. *Nature* **453,** 667–671
24. McHeyzer-Williams, M., Okitsu, S., Wang, N., and McHeyzer-Williams, L. (2011) Molecular programming of B cell memory. *Nat. Rev. Immunol.* **12,** 24–34
25. Franz, B., May, K. F., Dranoff, G., and Wucherpfennig, K. (2011) Ex vivo characterization and isolation of rare memory B cells with antigen tetramers. *Blood* **118,** 348–357
26. Doria-Rose, N. A., Klein, R. M., Manion, M. M., O'Dell, S., Phogat, A., Chakrabarti, B., Hallahan, C. W., Migueles, S. A., Wrammert, J., Ahmed, R., Nason, M., Wyatt, R. T., Mascola, J. R., and Connors, M. (2009) Frequency and phenotype of human immunodeficiency virus envelope-specific B cells from patients with broadly cross-neutralizing antibodies. *J. Virol.* **83,** 188–199
27. Amara, K., Steen, J., Murray, F., Morbach, H., Fernandez-Rodriguez, B. M., Joshua, V., Engstrom, M., Snir, O., Israelsson, L., Catrina, A. I., Wardemann, H., Corti, D., Meffre, E., Klareskog, L., and Malmstrom, V. (2013) Monoclonal IgG antibodies generated from joint-derived B cells of RA patients have a strong bias toward citrullinated autoantigen recognition. *J. Exp. Med.* **210,** 445–455
28. Stern, J. N., Yaari, G., Vander Heiden, J. A., Church, G., Donahue, W. F., Hintzen, R. Q., Huttner, A. J., Laman, J. D., Nagra, R. M., Nylander, A., Pitt, D., Ramanan, S., Siddiqui, B. A., Vigneault, F., Kleinstein, S. H., Hafler, D. A., and O'Connor, K. C. (2014) B cells populating the multiple sclerosis brain mature in the draining cervical lymph nodes. *Sci. Transl. Med.* **6,** 248ra107
29. Shigematsu, Y., Hanagiri, T., Kuroda, K., Baba, T., Mizukami, M., Ichiki, Y., Yasuda, M., Takenoyama, M., Sugio, K., and Yasumoto, K. (2009) Malignant mesothelioma-associated antigens recognized by tumor-infiltrating B cells and the clinical significance of the antibody titers. *Cancer Sci.* **100,** 1326–1334
30. Robinson, W. H. (2015) Sequencing the functional antibody repertoire–diagnostic and therapeutic discovery. *Nat. Rev. Rheumatol.* **11,** 171–182
31. Georgiou, G., Ippolito, G. C., Beausang, J., Busse, C. E., Wardemann, H., and Quake, S. R. (2014) The promise and challenge of high-throughput sequencing of the antibody repertoire. *Nat. Biotechnol.* **32,** 158–168
32. Woo, S., Cha, S. W., Merrihew, G., He, Y., Castellana, N., Guest, C., MacCoss, M., and Bafna, V. (2014) Proteogenomic database construction driven from large scale RNA-seq data. *J. Proteome Res.* **13,** 21–28
33. Woo, S., Cha, S. W., Na, S., Guest, C., Liu, T., Smith, R. D., Rodland, K. D., Payne, S., and Bafna, V. (2014) Proteogenomic strategies for identification of aberrant cancer peptides using large-scale next- generation sequencing data. *Proteomics* **14,** 2719–2730
34. Woo, S., Cha, S. W., Bonissone, S., Na, S., Tabb, D. L., Pevzner, P. A., and Bafna, V. (2015) Advanced Proteogenomic Analysis Reveals Multiple Peptide Mutations and Complex Immunoglobulin Peptides in Colon Cancer. *J. Proteome Res.* **14,** 3555–3567
35. Zhang, B., Wang, J., Wang, X., Zhu, J., Liu, Q., Shi, Z., Chambers, M. C., Zimmerman, L. J., Shaddox, K. F., Kim, S., Davies, S. R., Wang, S., Wang, P., Kinsinger, C. R., Rivers, R. C., Rodriguez, H., Townsend, R. R., Ellis, M. J., Carr, S. A., Tabb, D. L., Coffey, R. J., Slebos, R. J., Liebler, D. C., Carr, S. A., Gillette, M. A., Klauser, K. R., Kuhn, E., Mani, D. R., Mertins, P., Ketchum, K. A., Paulovich, A. G., Whiteaker, J. R., Edwards, N. J., McGarvey, P. B., Madhavan, S., Wang, P., Chan, D., Pandey, A., Shih Ie M., Zhang, H., Zhang, Z., Zhu, H., Whiteley, G. A., Skates, S. J., White, F. M., Levine, D. A., Boja, E. S., Kinsinger, C. R., Hiltke, T., Mesri, M., Rivers, R. C., Rodriguez, H., Shaw, K. M., Stein, S. E., Fenyo, D., Liu, T., McDermott, J. E., Payne, S. H., Rodland, K. D., Smith, R. D., Rudnick, P., Snyder, M., Zhao, Y., Chen, X., Ransohoff, D. F., Hoofnagle, A. N., Liebler, D. C., Sanders, M. E., Shi, Z., Slebos, R. J., Tabb, D. L., Zhang, B., Zimmerman, L. J., Wang, Y., Davies, S. R., Ding, L., Ellis, M. J., and Townsend, R. R. (2014) Proteogenomic characterization of human colon and rectal cancer. *Nature* **513,** 382–387
36. Fanayan, S., Smith, J. T., Lee, L. Y., Yan, F., Snyder, M., Hancock, W. S., and Nice, E. (2013) Proteogenomic analysis of human colon carcinoma cell lines LIM1215, *LIM1899, and LIM2405. J. Proteome Res.* **12,** 1732–1742
37. Safonova, Y., Bonissone, S., Kurpilyansky, E., Starostina, E., Lapidus, A., Stinson, J., DePalatis, L., Sandoval, W., Lill, J., and Pevzner, P. A. (2015) IgRepertoireConstructor: a novel algorithm for antibody repertoire construction and immunoproteogenomics analysis. Bioinformatics, 31(12):53–61.
38. Cunningham, F., Amode, M. R., Barrell, D., Beal, K., Billis, K., Brent, S., Carvalho-Silva, D., Clapham, P., Coates, G., Fitzgerald, S., Gil, L., Giron, C. G., Gordon, L., Hourlier, T., Hunt, S. E., Janacek, S. H., Johnson, N., Juettemann, T., Kahari, A. K., Keenan, S., Martin, F. J., Maurel, T., McLaren, W., Murphy, D. N., Nag, R., Overduin, B., Parker, A., Patricio, M., Perry, E., Pignatelli, M., Riat, H. S., Sheppard, D., Taylor, K., Thormann, A., Vullo, A., Wilder, S. P., Zadissa, A., Aken, B. L., Birney, E., Harrow, J., Kinsella, R., Muffato, M., Ruffier, M., Searle, S. M., Spudich, G., Trevanion, S. J., Yates, A., Zerbino, D. R., and Flicek, P. (2015)Ensembl 2015. *Nucleic Acids Res*. **43,** D662–D669
39. Kim, S., and Pevzner, P. A. (2014) MS-GF+ makes progress towards a universal database search tool for proteomics. *Nat. Commun.* **5,** 5277
40. Kessner, D., Chambers, M., Burke, R., Agus, D., and Mallick, P. (2008) ProteoWizard: open source software for rapid proteomics tools development. *Bioinformatics* **24,** 2534–2536
41. Pedrioli, P. G., Eng, J. K., Hubley, R., Vogelzang, M., Deutsch, E. W., Raught, B., Pratt, B., Nilsson, E., Angeletti, R. H., Apweiler, R., Cheung, K., Costello, C. E., Hermjakob, H., Huang, S., Julian, R. K., Kapp, E., McComb, M. E., Oliver, S. G., Omenn, G., Paton, N. W., Simpson, R., Smith, R., Taylor, C. F., Zhu, W., and Aebersold, R. (2004) A common open representation of mass spectrometry data and its application to proteomics research. *Nat. Biotechnol.* **22,** 1459–1466
42. Ruiz, M., and Lefranc, M. P. (2002) IMGT gene identification and Colliers de Perles of human immunoglobulins with known 3D structures. *Immunogenetics* **53,** 857–883
43. Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., Funke, R., Gage, D., Harris, K., Heaford, A., Howland, J., Kann, L., Lehoczky, J., LeVine, R., McEwan, P., McKernan, K., Meldrim, J., Mesirov, J. P., Miranda, C., Morris, W., Naylor, J., Raymond, C., Rosetti, M., Santos, R., Sheridan, A., Sougnez, C., Stange-Thomann, Y., Stojanovic, N., Subramanian, A., Wyman, D., Rogers, J., Sulston, J., Ainscough, R., Beck, S., Bentley, D., Burton, J., Clee, C., Carter, N., Coulson, A., Deadman, R., Deloukas, P., Dunham, A., Dunham, I., Durbin, R., French, L., Grafham, D., Gregory, S., Hubbard, T., Humphray, S., Hunt, A., Jones, M., Lloyd, C., McMurray, A., Matthews, L., Mercer, S., Milne, S., Mullikin, J. C., Mungall, A., Plumb, R., Ross, M., Shownkeen, R., Sims, S., Waterston, R. H., Wilson, R. K., Hillier, L. W., McPherson, J. D., Marra, M. A., Mardis, E. R., Fulton, L. A., Chinwalla, A. T., Pepin, K. H., Gish,

W. R., Chissoe, S. L., Wendl, M. C., Delehaunty, K. D., Miner, T. L., Delehaunty, A., Kramer, J. B., Cook, L. L., Fulton, R. S., Johnson, D. L., Minx, P. J., Clifton, S. W., Hawkins, T., Branscomb, E., Predki, P., Richardson, P., Wenning, S., Slezak, T., Doggett, N., Cheng, J. F., Olsen, A., Lucas, S., Elkin, C., Uberbacher, E., Frazier, M., Gibbs, R. A., Muzny, D. M., Scherer, S. E., Bouck, J. B., Sodergren, E. J., Worley, K. C., Rives, C. M., Gorrell, J. H., Metzker, M. L., Naylor, S. L., Kucherlapati, R. S., Nelson, D. L., Weinstock, G. M., Sakaki, Y., Fujiyama, A., Hattori, M., Yada, T., Toyoda, A., Itoh, T., Kawagoe, C., Watanabe, H., Totoki, Y., Taylor, T., Weissenbach, J., Heilig, R., Saurin, W., Artiguenave, F., Brottier, P., Bruls, T., Pelletier, E., Robert, C., Wincker, P., Smith, D. R., Doucette-Stamm, L., Rubenfield, M., Weinstock, K., Lee, H. M., Dubois, J., Rosenthal, A., Platzer, M., Nyakatura, G., Taudien, S., Rump, A., Yang, H., Yu, J., Wang, J., Huang, G., Gu, J., Hood, L., Rowen, L., Madan, A., Qin, S., Davis, R. W., Federspiel, N. A., Abola, A. P., Proctor, M. J., Myers, R. M., Schmutz, J., Dickson, M., Grimwood, J., Cox, D. R., Olson, M. V., Kaul, R., Raymond, C., Shimizu, N., Kawasaki, K., Minoshima, S., Evans, G. A., Athanasiou, M., Schultz, R., Roe, B. A., Chen, F., Pan, H., Ramser, J., Lehrach, H., Reinhardt, R., McCombie, W. R., de la Bastide, M., Dedhia, N., Blocker, H., Hornischer, K., Nordsiek, G., Agarwala, R., Aravind, L., Bailey, J. A., Bateman, A., Batzoglou, S., Birney, E., Bork, P., Brown, D. G., Burge, C. B., Cerutti, L., Chen, H. C., Church, D., Clamp, M., Copley, R. R., Doerks, T., Eddy, S. R., Eichler, E. E., Furey, T. S., Galagan, J., Gilbert, J. G., Harmon, C., Hayashizaki, Y., Haussler, D., Hermjakob, H., Hokamp, K., Jang, W., Johnson, L. S., Jones, T. A., Kasif, S., Kaspryzk, A., Kennedy, S., Kent, W. J., Kitts, P., Koonin, E. V., Korf, I., Kulp, D., Lancet, D., Lowe, T. M., McLysaght, A., Mikkelsen, T., Moran, J. V., Mulder, N., Pollara, V. J., Ponting, C. P., Schuler, G., Schultz, J., Slater, G., Smit, A. F., Stupka, E., Szustakowki, J., Thierry-Mieg, D., Thierry-Mieg, J., Wagner, L., Wallis, J., Wheeler, R., Williams, A., Wolf, Y. I., Wolfe, K. H., Yang, S. P., Yeh, R. F., Collins, F., Guyer, M. S., Peterson, J., Felsenfeld, A., Wetterstrand, K. A., Patrinos, A., Morgan, M. J., de Jong, P., Catanese, J. J., Osoegawa, K., Shizuya, H., Choi, S., Chen, Y. J., and Szustakowki, J. (2001) Initial sequencing and analysis of the human genome. *Nature* **409,** 860–921

44. Pevzner, P. A., Tang, H., and Waterman, M. S. (2001) An Eulerian path approach to DNA fragment assembly. *Proc. Natl. Acad. Sci. U.S.A.* **98,** 9748–9753
45. Butler, J., MacCallum, I., Kleber, M., Shlyakhter, I. A., Belmonte, M. K., Lander, E. S., Nusbaum, C., and Jaffe, D. B. (2008) ALLPATHS: de novo assembly of whole-genome shotgun microreads. *Genome Res.* **18,** 810–820
46. Ronen, R., Boucher, C., Chitsaz, H., and Pevzner, P. (2012) SEQuel: improving the accuracy of genome assemblies. *Bioinformatics* **28,** i188–i196
47. Nikolenko, S. I., Korobeynikov, A. I., and Alekseyev, M. A. (2013) BayesHammer: Bayesian clustering for error correction in single-cell sequencing. *BMC Genomics* **14,** S7
48. Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q., Chen, Z., Mauceli, E., Hacohen, N., Gnirke, A., Rhind, N., di Palma, F., Birren, B. W., Nusbaum, C., Lindblad-Toh, K., Friedman, N., and Regev, A. (2011) Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.* **29,** 644–652
49. Peng, Y., Leung, H. C. M., Yiu, S.-M., Lv, M.-J., Zhu, X.-G., and Chin, F. Y. L. (2013) IDBA-tran: a more robust de novo de Bruijn graph assembler for transcriptomes with uneven expression levels. *Bioinformatics* **29,** i326–i334
50. Bankevich, A., Nurk, S., Antipov, D., Gurevich, A. A., Dvorkin, M., Kulikov, A. S., Lesin, V. M., Nikolenko, S. I., Pham, S., Prjibelski, A. D., Pyshkin, A. V., Sirotkin, A. V., Vyahhi, N., Tesler, G., Alekseyev, M. A., and Pevzner, P. A. (2012) SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J. Comput. Biol.* **19,** 455–477
51. Scaviner, D., Barbie, V., Ruiz, M., and Lefranc, M. P. (1999) Protein displays of the human immunoglobulin heavy, kappa and lambda variable and joining regions. *Exp. Clin. Immunogenet.* **16,** 234–240
52. Muzny, D. M., Bainbridge, M. N., Chang, K., Dinh, H. H., Drummond, J. A., Fowler, G., Kovar, C. L., Lewis, L. R., Morgan, M. B., Newsham, I. F., Reid, J. G., Santibanez, J., Shinbrot, E., Trevino, L. R., Wu, Y. Q., Wang, M., Gunaratne, P., Donehower, L. A., Creighton, C. J., Wheeler, D. A., Gibbs, R. A., Lawrence, M. S., Voet, D., Jing, R., Cibulskis, K., Sivachenko, A., Stojanov, P., McKenna, A., Lander, E. S., Gabriel, S., Getz, G., Ding, L., Fulton, R. S., Koboldt, D. C., Wylie, T., Walker, J., Dooling, D. J., Fulton, L., Delehaunty, K. D., Fronick, C. C., Demeter, R., Mardis, E. R., Wilson, R. K., Chu, A., Chun, H. J., Mungall, A. J., Pleasance, E., Robertson, A., Stoll, D., Balasundaram, M., Birol, I., Butterfield, Y. S., Chuah, E., Coope, R. J., Dhalla, N., Guin, R., Hirst, C., Hirst, M., Holt, R. A., Lee, D., Li, H. I., Mayo, M., Moore, R. A., Schein, J. E., Slobodan, J. R., Tam, A., Thiessen, N., Varhol, R., Zeng, T., Zhao, Y., Jones, S. J., Marra, M. A., Bass, A. J., Ramos, A. H., Saksena, G., Cherniack, A. D., Schumacher, S. E., Tabak, B., Carter, S. L., Pho, N. H., Nguyen, H., Onofrio, R. C., Crenshaw, A., Ardlie, K., Beroukhim, R., Winckler, W., Getz, G., Meyerson, M., Protopopov, A., Zhang, J., Hadjipanayis, A., Lee, E., Xi, R., Yang, L., Ren, X., Zhang, H., Sathiamoorthy, N., Shukla, S., Chen, P. C., Haseley, P., Xiao, Y., Lee, S., Seidman, J., Chin, L., Park, P. J., Kucherlapati, R., Auman, J. T., Hoadley, K. A., Du, Y., Wilkerson, M. D., Shi, Y., Liquori, C., Meng, S., Li, L., Turman, Y. J., Topal, M. D., Tan, D., Waring, S., Buda, E., Walsh, J., Jones, C. D., Mieczkowski, P. A., Singh, D., Wu, J., Gulabani, A., Dolina, P., Bodenheimer, T., Hoyle, A. P., Simons, J. V., Soloway, M., Mose, L. E., Jefferys, S. R., Balu, S., O'Connor, B. D., Prins, J. F., Chiang, D. Y., Hayes, D., Perou, C. M., Hinoue, T., Weisenberger, D. J., Maglinte, D. T., Pan, F., Berman, B. P., Van Den Berg, D. J., Shen, H., Triche, T., Baylin, S. B., Laird, P. W., Getz, G., Noble, M., Voet, D., Saksena, G., Gehlenborg, N., DiCara, D., Zhang, J., Zhang, H., Wu, C. J., Liu, S. Y., Shukla, S., Lawrence, M. S., Zhou, L., Sivachenko, A., Lin, P., Stojanov, P., Jing, R., Park, R. W., Nazaire, M. D., Robinson, J., Thorvaldsdottir, H., Mesirov, J., Park, P. J., Chin, L., Thorsson, V., Reynolds, S. M., Bernard, B., Kreisberg, R., Lin, J., Iype, L., Bressler, R., Erkkila, T., Gundapuneni, M., Liu, Y., Norberg, A., Robinson, T., Yang, D., Zhang, W., Shmulevich, I., de Ronde, J. J., Schultz, N., Cerami, E., Ciriello, G., Goldberg, A. P., Gross, B., Jacobsen, A., Gao, J., Kaczkowski, B., Sinha, R., Aksoy, B., Antipin, Y., Reva, B., Shen, R., Taylor, B. S., Chan, T. A., Ladanyi, M., Sander, C., Akbani, R., Zhang, N., Broom, B. M., Casasent, T., Unruh, A., Wakefield, C., Hamilton, S. R., Cason, R., Baggerly, K. A., Weinstein, J. N., Haussler, D., Benz, C. C., Stuart, J. M., Benz, S. C., Sanborn, J., Vaske, C. J., Zhu, J., Szeto, C., Scott, G. K., Yau, C., Ng, S., Goldstein, T., Ellrott, K., Collisson, E., Cozen, A. E., Zerbino, D., Wilks, C., Craft, B., Spellman, P., Penny, R., Shelton, T., Hatfield, M., Morris, S., Yena, P., Shelton, C., Sherman, M., Paulauskis, J., Gastier-Foster, J. M., Bowen, J., Ramirez, N. C., Black, A., Pyatt, R., Wise, L., White, P., Bertagnolli, M., Brown, J., Chan, T. A., Chu, G. C., Czerwinski, C., Denstman, F., Dhir, R., Dorner, A., Fuchs, C. S., Guillem, J. G., Iacocca, M., Juhl, H., Kaufman, A., Kohl, B., Van Le, X., Mariano, M. C., Medina, E. N., Meyers, M., Nash, G. M., Paty, P. B., Petrelli, N., Rabeno, B., Richards, W. G., Solit, D., Swanson, P., Temple, L., Tepper, J. E., Thorp, R., Vakiani, E., Weiser, M. R., Willis, J. E., Witkin, G., Zeng, Z., Zinner, M. J., Zornig, C., Jensen, M. A., Sfeir, R., Kahn, A. B., Chu, A. L., Kothiyal, P., Wang, Z., Snyder, E. E., Pontius, J., Pihl, T. D., Ayala, B., Backus, M., Walton, J., Whitmore, J., Baboud, J., Berton, D. L., Nicholls, M. C., Srinivasan, D., Raman, R., Girshik, S., Kigonya, P. A., Alonso, S., Sanbhadti, R. N., Barletta, S. P., Greene, J. M., Pot, D. A., Shaw, K. R., Dillon, L. A., Buetow, K., Davidsen, T., Demchok, J. A., Eley, G., Ferguson, M., Fielding, P., Schaefer, C., Sheth, M., Yang, L., Guyer, M. S., Ozenberger, B. A., Palchik, J. D., Peterson, J., Sofia, H. J., and Thomson, E. (2012) Comprehensive molecular characterization of human colon and rectal cancer. *Nature* **487,** 330–337
53. Orciani, M., Trubiani, O., Guarnieri, S., Ferrero, E., and Di Primio, R. (2008) CD38 is constitutively expressed in the nucleus of human hematopoietic cells. *J. Cell. Biochem.* **105,** 905–912
54. Partida-Sanchez, S., Rivero-Nava, L., Shi, G., and Lund, F. E. (2007) CD38: an ecto-enzyme at the crossroads of innate and adaptive immune responses. *Adv. Exp. Med. Biol.* **590,** 171–183
55. Wang, Z. Q., Milne, K., Webb, J. R., and Watson, P. H. (2017) CD74 and intratumoral immune response in breast cancer. *Oncotarget* **8,** 12664–12674
56. Xu, Z., Chen, H., Liu, D., and Huo, J. (2015) Fibulin-1 is downregulated through promoter hypermethylation in colorectal cancer: a CONSORT study. *Medicine* **94,** e663
57. Sherry, S. T., Ward, M. H., Kholodov, M., Baker, J., Phan, L., Smigielski, E. M., and Sirotkin, K. (2001) dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.* **29,** 308–311
58. Houghton, A. N., and Guevara-Patino, J. A. (2004) Immune recognition of self in immunity against cancer. *J. Clin. Invest.* **114,** 468–471
59. Vanky, F., Klein, E., Willems, J., Book, K., Ivert, T., Peterffy, A., Nilsonne, U., Kreicbergs, A., and Aparisi, T. (1986) Lysis of autologous tumor

cells by blood lymphocytes tested at the time of surgery. Correlation with the postsurgical clinical course. *Cancer Immunol. Immunother.* **21,** 69–76

60. Zhang, L., Conejo-Garcia, J. R., Katsaros, D., Gimotty, P. A., Massobrio, M., Regnani, G., Makrigiannakis, A., Gray, H., Schlienger, K., Liebman, M. N., Rubin, S. C., and Coukos, G. (2003) Intratumoral T cells, recurrence, and survival in epithelial ovarian cancer. N. *Engl. J. Med.* **348,** 203–213

61. Salgado, R., Denkert, C., Demaria, S., Sirtaine, N., Klauschen, F., Pruneri, G., Wienert, S., Van den Eynden, G., Baehner, F. L., Penault-Llorca, F., Perez, E. A., Thompson, E. A., Symmans, W. F., Richardson, A. L., Brock, J., Criscitiello, C., Bailey, H., Ignatiadis, M., Floris, G., Sparano, J., Kos, Z., Nielsen, T., Rimm, D. L., Allison, K. H., Reis-Filho, J. S., Loibl, S., Sotiriou, C., Viale, G., Badve, S., Adams, S., Willard-Gallo, K., and Loi, S. (2015) The evaluation of tumor-infiltrating lymphocytes (TILs) in breast cancer: recommendations by an International TILs Working Group 2014. *Ann. Oncol.* **26,** 259–271