BIOINFORMATICS

Vol. 00 no. 00 2015
Pages 1–3

IgRepertoireConstructor: a novel algorithm for antibody repertoire construction and immunoproteogenomics analysis. Appendix figures

Yana Safonova [1,4,6], Stefano Bonissone [2,6] Eugene Kurpilyansky [4], Ekaterina Starostina [1,4], Alla Lapidus [1,4], Wendy Sandoval [5], Jennie Lill [5] and Pavel A. Pevzner [1,3,4,*]

[1]Center for Algorithmic Biotechnology, St. Petersburg State University, Russia
[2]Bioinformatics Program, University of California at San Diego, USA
[3]Dept. of Computer Science and Engineering, University of California at San Diego, USA
[4]Algorithmic Biology Laboratory, St. Petersburg Academic University, Russia
[5]Genentech, South San Francisco, California, USA
[6]These authors have contributed equally.

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Associate Editor: XXXXXXX

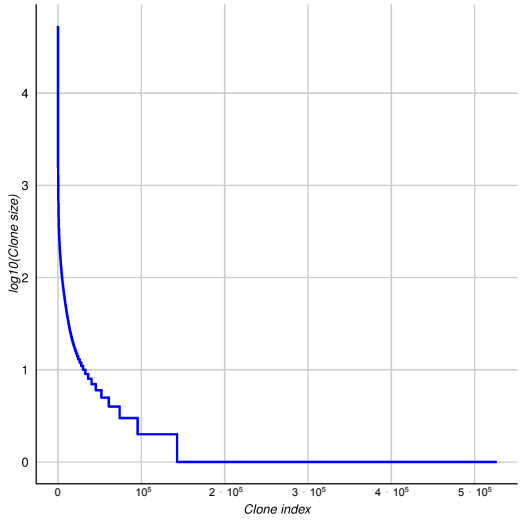

Fig. A1: Clone size distribution for the Ig-seq heavy chain dataset.

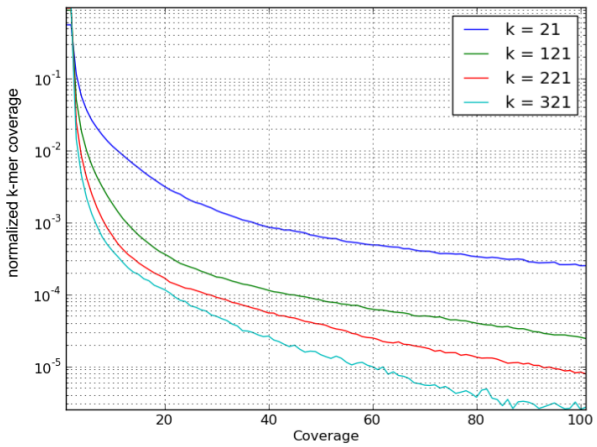

Fig. A2: Histograms of normalized k-mer coverage distribution for $k = 21, 121, 221$, and 321 illustrate that the coverage of the antibody repertoire by short k-mers is orders of magnitude higher compared to the coverage by long k-mers. Note that the y-axis is given in logarithmic scale.

*to whom correspondence should be addressed

```
s1 CAGGTCTGATGCAGTCTGGGACTTAGCTGGGGCGCT
s2 CAGGTCTGGTGCAATCTGGGACTGAGCTGGGTCGCT
```

$d(s1, s2) = 3$

(a)

```
s1 CTCAGGTCTGATGCAGTCTGGGACTTAGCTGGGGCGCT
s2    CAGGTCTGGTGCAATCAGGGACTGAGCTGGGTCGCTA
```

$\tilde{d}(s1, s2) = 4$

(b)

Fig. A3: Hamming distance $d(s_1, s_2)$ (a) and generalized Hamming distance $\tilde{d}(s_1, s_2)$ (b). Note that $d(s_1, s_2)$ and $\tilde{d}(s_1, s_2)$ for sequences s_1 and s_2 of equal length are not necessarily the same.

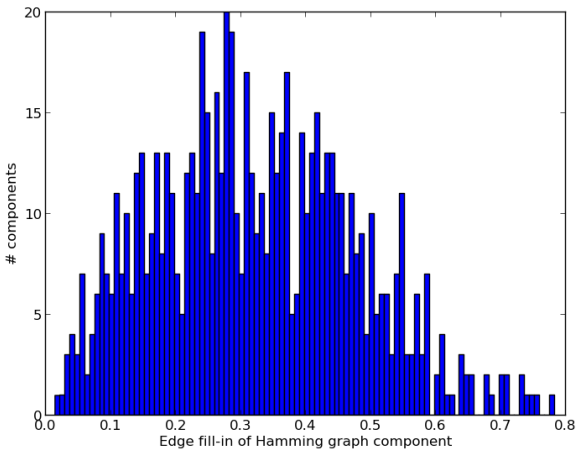

Fig. A4: Histogram of the distribution of edge fill-ins computed for 721 large ($\geq$ 100 vertices) connected components of the Bounded Hamming graph with $\tau = 3$. The average size of components with edge fill-ins > 0.7 is 151.

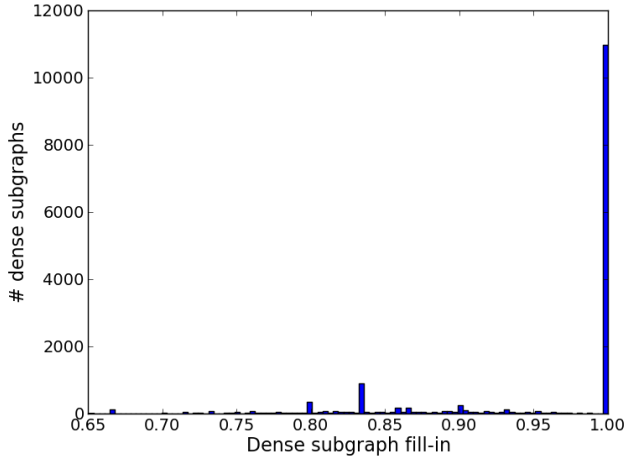

Fig. A5: Histogram of edge fill-in distribution for non-trivial dense subgraphs constructed from 721 large ($\geq$ 100 vertices) connected components. The total number of the constructed non-trivial dense subgraphs is $15,996$.

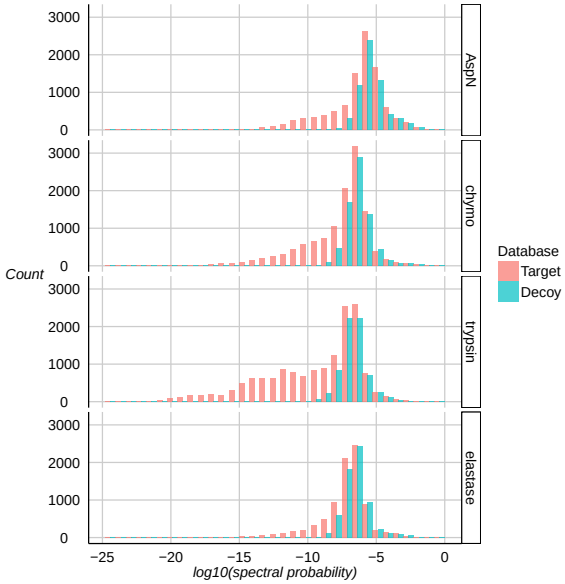

Fig. A6: Spectral probability distributions for Peptide-Spectrum Matches found in target/decoy identifications for four spectral datasets.

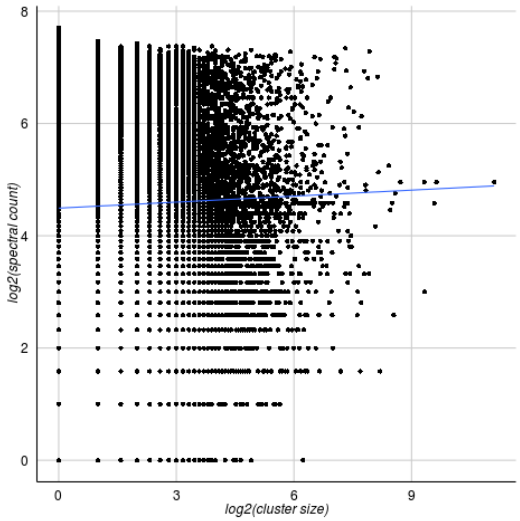

Fig. A9: Scatterplot of NGS-based and MS-based abundances of antibodies (cluster size vs. spectral count). Pearson correlation $\rho = 0.007544031$.

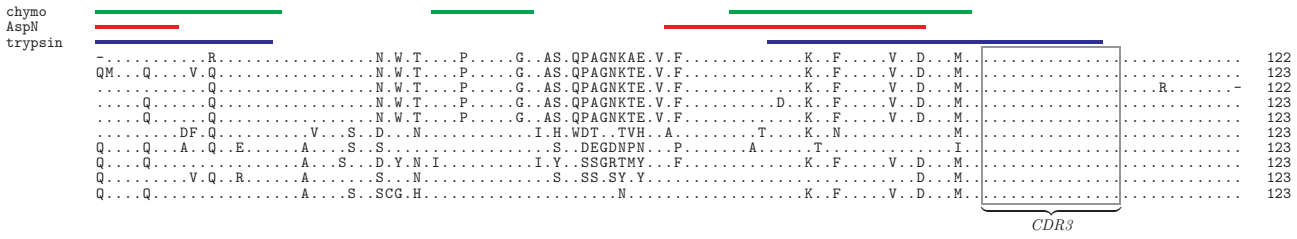

Fig. A7: Alignment of sequences of a single clone with peptide evidence. The 11 antibodies with most peptide evidence within this clone are aligned to one another to show the sequence diversity, while the antibody with most peptide evidence is omitted. Identified peptides for the omitted sequence are shown above the alignment. The CDR3 region is noted, and shown in gray box. Positions differing from the sequence with peptide evidence shown are noted, positions agreeing with the omitted sequence are shown as a dot.

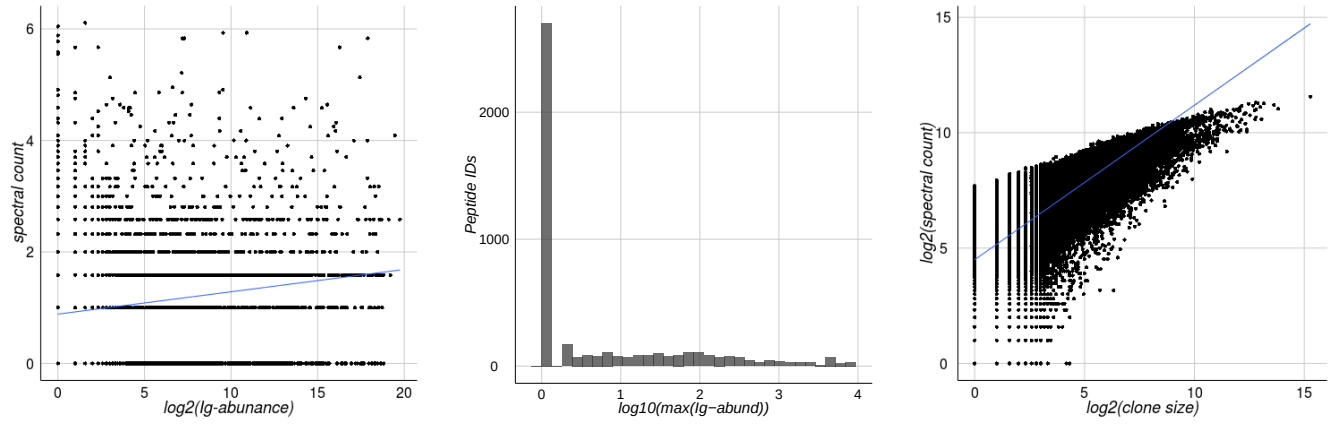

Fig. A8: Scatterplot of genomics-based and proteomics-based abundances. Cluster size compared to spectral counts of each cluster; Pearson correlation $\rho = 0.007544031$. (a) The total Ig-abundance compared against the spectral count for all peptide. Pearson correlation $\rho = 0.1724002$. (b) Histogram of peptide IDs by the maximal Ig-abundance of each peptide. (c) Spectral count of each clone, related to clone size; Pearson correlation $\rho = 0.5687614$. The spectral count of a clone is the total number of PSMs originating from all antibodies within that clone. Normalization of spectra for shared peptides is not performed in these plots.