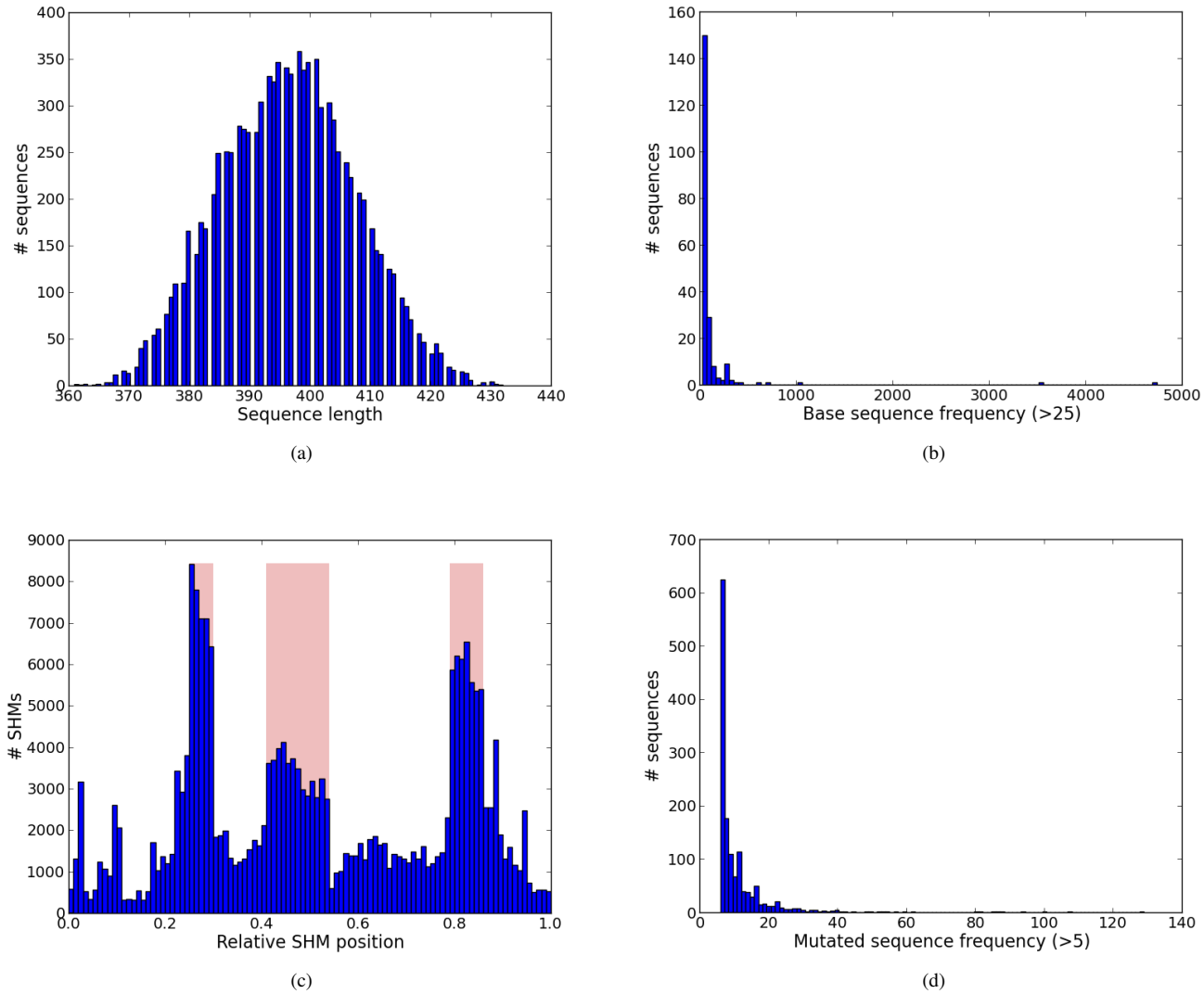
Metrics	Highly diverged repertoire	Highly abundant repertoire
# base sequences	50 000	10 000
# mutated sequences	100 000	50 000
Repertoire size	500 000	500 000
# clusters	82 147	50 040
Avg cluster size	5.28	8.03
Max cluster size	22 333	11 230
# trivial clusters	45 489	25 826
# clusters (> 10*)*	5011	4201
# clusters (> 100*)*	367	437

Table A1. Examples of highly diverged and abundant repertoires. The first repertoire (highly diverged repertoire) was simulated with the following parameters: *# base sequences* $= 50\,000$, *# mutated sequences* $= 100\,000$, and *repertoire size* $= 500\,000$. Big number of *base sequences* and slight difference between *mutated sequences* and *base sequences* lead to simulation of repertoire with big number of different lowly abundant clusters. The second repertoire (highly abundant repertoire) was simulated with the following parameters: *# base sequences* $= 10\,000$, *# mutated sequences* $= 50\,000$, and *repertoire size* $= 500\,000$. Small number of *base sequences* and big difference between *mutated sequences* and *base sequences* lead to simulation of repertoire with small number of different highly abundant clusters.

```
>cluster___101___size___3                                    MISEQ:14345:28882    101
CAGGTGCAGCTGGTGCAATCTGGGGCTGAGGTGAAGAAGCCTGGGTCCTCGGTGA     MISEQ:14374:28884    101
>cluster___102___size___2                                    MISEQ:14393:28886    101
TGAGGAGACGGTGACCAGGGTTCCCTGGCCCCAGTAGTCAAAGAAGATCCCGA       MISEQ:16454:28882    102
>cluster___103___size___1                                    MISEQ:16426:28886    102
GAGGTGCAGCTGGTGGAGTCTGGGGGAGGCGTGGTCCAGCCTGGGAGGTCCCTG      MISEQ:15812:28886    103
```

Fig. A1: Made-up example illustrating the representation of an antibody repertoire by two files: a FASTA file with a set of antibody sequences (left column) and a Read-Cluster Map (RCM) file (right column). The first file describes three antibody clusters: cluster $\#101$ of size 3, cluster $\#102$ of size 2, and cluster $\#103$ of size 1. Each cluster is given a unique id (encoded in header line) that is used to identify the cluster in the RCM file. For example, cluster $\#101$ was constructed by reads MISEQ:14345:28882, MISEQ:14374:28884, and MISEQ:14393:28886.



(a) (b) (c) (d)

Fig. A2: Statistics for the heavy chain repertoire with the following parameters: *# base sequences* = 10 000, *# mutated sequences* = 50 000, and *repertoire size* = 500 000. (Upper left) The histogram of distribution of the lengths of simulated base antibody sequences with mean value 377.4. (Upper right) The histogram of distribution of large frequencies (> 25) of base antibody sequences with mean frequency 5 and maximal frequency 4743. (Bottom left) The histogram of distribution of positions of somatic mutations with three peaks corresponding to positions of CDRs. The positions within an antibody (varying from 0 to 1) have been normalized by dividing them by the total antibody length. (Bottom right) The histogram of distribution of large frequencies (> 5) of mutated antibody sequences with mean frequency 1.64 and maximal frequency 129.