Supplementary Text 1: A set of 530,104 IGHV sequences were submitted to IMGT HighV-Quest and also analyzed with ImmuneDB without local alignment. Sequences annotated with pseduo-genes and those missing over 100 nucleotides or with more than 20 insertions/deletions were excluded.

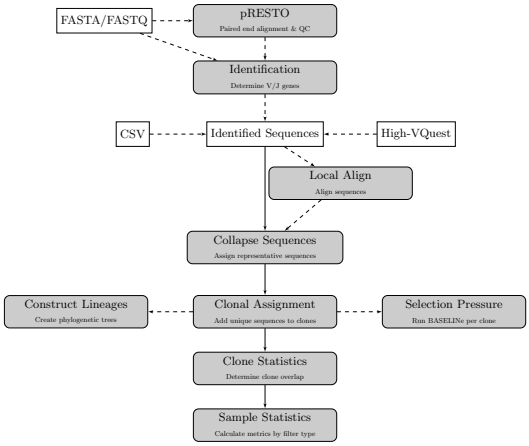
A total of 128,135 were identified: 110,656 by both methods, 15,275 only by HighV-Quest, and 2,204 only by ImmuneDB as shown in Supplementary Table 1.

Identical sequences were then collapsed and those with at least two copies were compared as detailed in Supplementary Table 2. HighV-Quest identified 1,379 that ImmuneDB did not, and ImmuneDB identified 1,388 that HighV-Quest did not. Of those identified by both, 9,228 agreed upon which sequences collapsed and 7,430 also agreed on both V- and J-gene assignments.

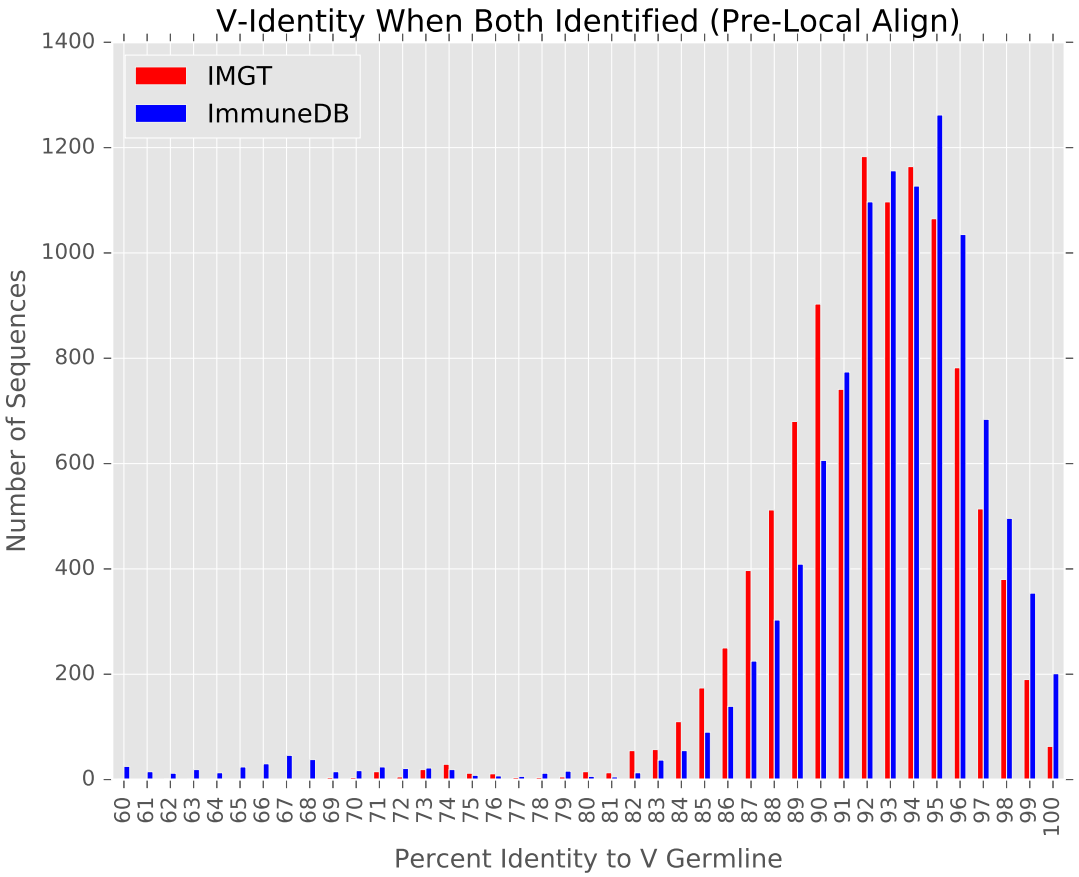
Of the 1,798 differing sequences, 1,544 were due to HighV-Quest identifying a pseudo-gene for the V-region where ImmuneDB identified a functional gene. While in this example ImmuneDB did not check for pseudo-genes, this is a parameter that users can specify. Additionally, among the sequences that differed in gene identification, ImmuneDB generally achieved higher V-identities, although this may be due to different calculation methods.

ImmuneDB's local-alignment step was then run on the sequences that were either not identified or flagged as having an insertion or deletion (Supplementary Table 3). After this, 12,727 unique sequences with at least two copies were identified by both methods, HighV-Quest identified 58 sequences that ImmuneDB did not and ImmuneDB identifying 3,684 that HighV-Quest did not. Of the sequences that both identified, 10,432 collapsed to the same set of sequences of which 8,504 had the same assigned V- and J-gene. From the 1,928 sequences with different genes, 1,577 differed because HighV-Quest assigned pseudo-genes. Although ImmuneDB did not search for pseudo-genes, it had a higher V-identity in 891 of the cases and 518 had the same V-identity.

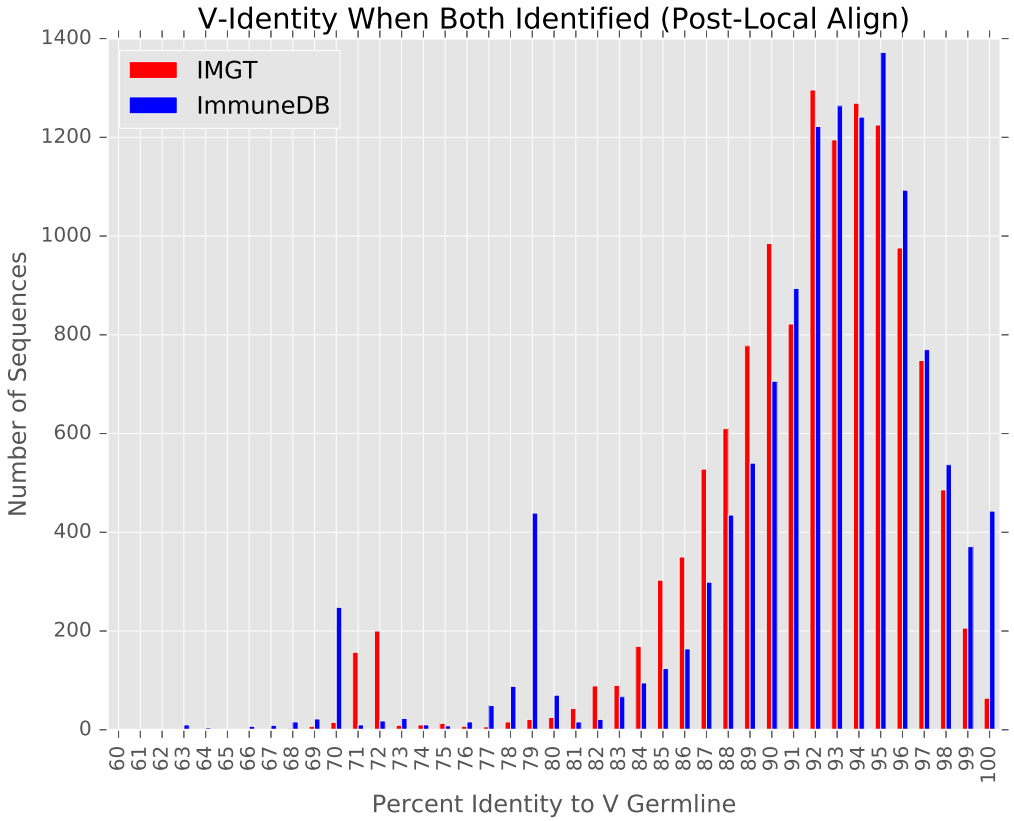
ImmuneDB processed the sequences by anchoring in approximately 20 minutes with an additional 4 hours required for local alignment. It is unknown how long IMGT HighV-Quest took to process all the sequences due to it queueing jobs from multiple users.



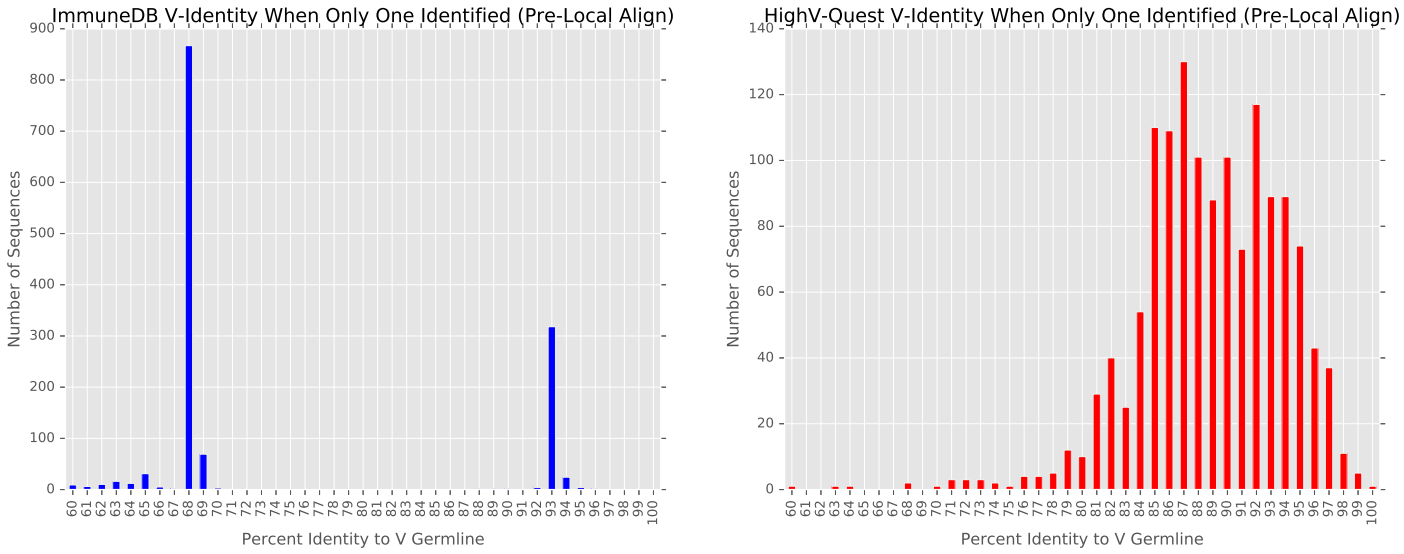
Supplementary Figure 1: The ImmuneDB pipeline. Solid lines represent required steps and dashed lines are optional steps. Three sets of files can be used as input to ImmuneDB: raw FASTA/FASTQ files (which may be pre-processed with pRESTO (Vander Heiden *et al.*, 2014)), Comma-Separated Value (CSV) files with appropriate fields, or HighV-Quest output. During identification V and J genes are identified for each sequence using the method in (Zhang*et al.*, 2015). Optionally, local-alignment can then be used to correct any sequences with insertions/deletions or mutations in their anchors. Identical sequences within each subject are then collapsed taking into account ambiguous "N" bases. At this point sequences are aggregated into clones based on their subject, V/J genes, and CDR3 amino-acid similarity. Each clone can optionally have selection pressure and a lineage calculated. Finally, clone and sample statistics calculate distributions of various sequence features (e.g. CDR3 length) for later analysis.



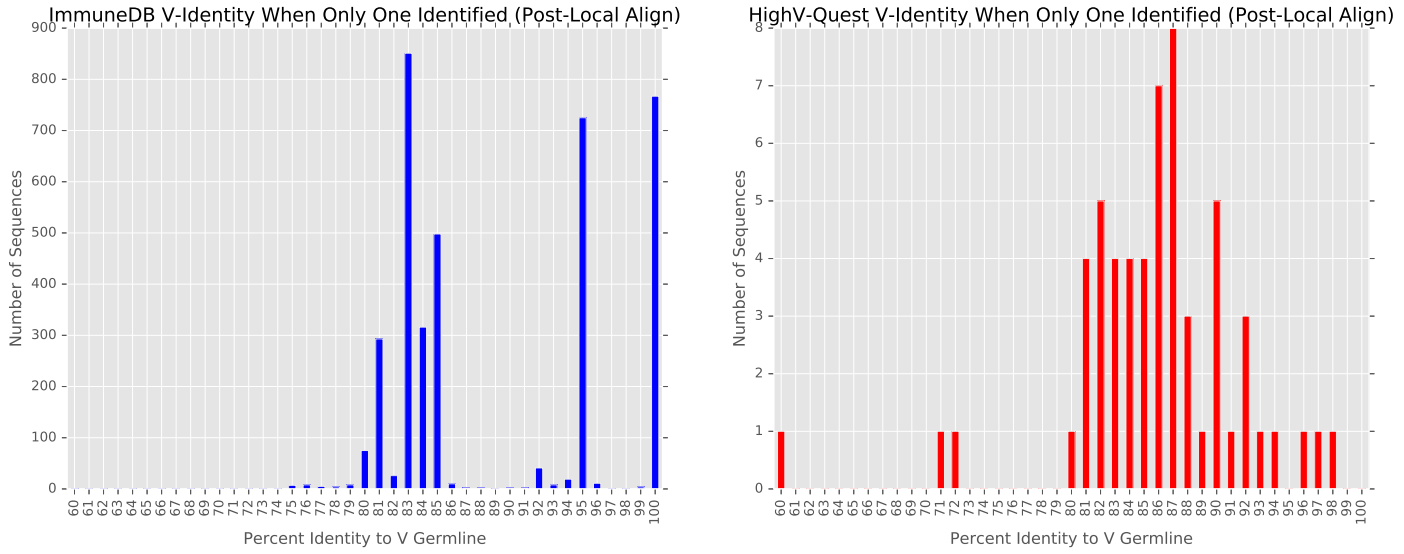
Supplementary Figure 2: The identity to the assigned V-gene for all unique sequence identified by both ImmuneDB using the anchoring method and HighV-Quest.



Supplementary Figure 3: The identity to the assigned V-gene for all unique sequence identified by both ImmuneDB after local-alignment and HighV-Quest.



Supplementary Figure 4: The identity to the assigned V-gene for all unique sequence identified by only ImmuneDB using the anchoring method or HighV-Quest but not both.



Supplementary Figure 5: The identity to the assigned V-gene for all unique sequence identified by only ImmuneDB after local-alignment or HighV-Quest but not both.

```
Total Sequences: 530104
Neither Identified: 392213
Either Identified: 128135
Both Identified: 110656
Only HighV-Quest Identified: 24820
    Not pseudogene or large insertion/deletion: 15275
Only ImmuneDB Identified: 2415
    With < 100 padding: 2204
ImmuneDB and HighV-Quest Both Identified & Agree on Collapsed Sequence: 108504
    Same V & J: 89403 (Best-identity: Equal=28232, HighV-Quest=2213, ImmuneDB=58958)
    Differing V only: 16568 (Best-identity: Equal=5184, HighV-Quest=2545, ImmuneDB=8839)
        Because of pseduogene: 15223 (Best-identity: Equal=5131, HighV-Quest=1342, ImmuneDB=8750)
    Differing J only: 2058 (Best-identity: Equal=631, HighV-Quest=55, ImmuneDB=1372)
    Differing V & J: 475 (Best-identity: Equal=166, HighV-Quest=63, ImmuneDB=246)
        Because of pseduogene: 439 (Best-identity: Equal=166, HighV-Quest=32, ImmuneDB=241)
    Both found indel: 2584 (Best-identity: Equal=1, HighV-Quest=2582, ImmuneDB=1)
    Only HighV-Quest found indel: 681 (Best-identity: Equal=25, HighV-Quest=610, ImmuneDB=46)
    Only ImmuneDB found indel: 1677 (Best-identity: Equal=102, HighV-Quest=1368, ImmuneDB=207)
```

Supplementary Table 1: A breakdown of how many total sequences were identified by ImmuneDB with the anchoring method or HighV-Quest. For sequences both identified, the number of times HighV-Quest and ImmuneDB had a higher V-identity is listed (or if there was a tie).

```
Either Identified 13230
Both Identified: 10463
Only HighV-Quest Identified: 3162
        Not pseudogene or large insertion/deletion 1379
Only ImmuneDB Identified: 1490
        With < 100 padding 1388
Removed because CN=1 in both: 99276
Removed because CN=1 in HighV-Quest: 626
Removed because CN=1 in ImmuneDB: 291
ImmuneDB and HighV-Quest Both Identified & Agree on Collapsed Sequence: 9228
        Same V & J: 7430 (Best-identity: Equal=2154, HighV-Quest=118, ImmuneDB=5158)
        Differing V only: 1544 (Best-identity: Equal=496, HighV-Quest=203, ImmuneDB=845)
                Because of pseduogene: 1439 (Best-identity: Equal=491, HighV-Quest=108, ImmuneDB=840)
        Differing J only: 203 (Best-identity: Equal=52, HighV-Quest=6, ImmuneDB=145)
        Differing V & J: 51 (Best-identity: Equal=15, HighV-Quest=9, ImmuneDB=27)
                Because of pseduogene: 44 (Best-identity: Equal=15, HighV-Quest=2, ImmuneDB=27)
        Both found indel: 149 (Best-identity: Equal=0, HighV-Quest=149, ImmuneDB=0)
        Only HighV-Quest found indel: 53 (Best-identity: Equal=2, HighV-Quest=48, ImmuneDB=3)
        Only ImmuneDB found indel: 131 (Best-identity: Equal=9, HighV-Quest=111, ImmuneDB=11)
```

Supplementary Table 2: A breakdown of how many total sequences were identified by ImmuneDB with the anchoring method or HighV-Quest after collapsing identical sequences. Sequences with a copy number of one in either method were removed for further analysis.

```
Either Identified 16469
Both Identified: 12727
Only HighV-Quest Identified: 337
        Not pseudogene or large insertion/deletion 58
Only ImmuneDB Identified: 90961
        With < 100 padding 3684
Removed because CN=1 in both: 116513
Removed because CN=1 in HighV-Quest: 2698
Removed because CN=1 in ImmuneDB: 852
ImmuneDB and HighV-Quest Both Identified & Agree on Collapsed Sequence: 10432
        Same V & J: 8504 (Best-identity: Equal=2403, HighV-Quest=294, ImmuneDB=5807)
        Differing V only: 1673 (Best-identity: Equal=523, HighV-Quest=230, ImmuneDB=920)
                Because of pseduogene: 1577 (Best-identity: Equal=518, HighV-Quest=168, ImmuneDB=891)
        Differing J only: 205 (Best-identity: Equal=55, HighV-Quest=3, ImmuneDB=147)
        Differing V & J: 50 (Best-identity: Equal=16, HighV-Quest=5, ImmuneDB=29)
                Because of pseduogene: 45 (Best-identity: Equal=16, HighV-Quest=1, ImmuneDB=28)
        When indel found by either:
                Equal number of indels: 195 (Best-identity: Equal=33, HighV-Quest=6, ImmuneDB=156)
                HighV-Quest has more indels: 508 (Best-identity: Equal=11, HighV-Quest=478, ImmuneDB=19)
                ImmuneDB has more indels: 196 (Best-identity: Equal=22, HighV-Quest=14, ImmuneDB=160)
ImmuneDB is Subset of HighV-Quest: 743
        Same V & J: 613 (Best-identity: Equal=131, HighV-Quest=43, ImmuneDB=439)
        Differing V only: 110 (Best-identity: Equal=12, HighV-Quest=55, ImmuneDB=43)
                Because of pseduogene: 89 (Best-identity: Equal=11, HighV-Quest=47, ImmuneDB=31)
        Differing J only: 13 (Best-identity: Equal=2, HighV-Quest=0, ImmuneDB=11)
        Differing V & J: 7 (Best-identity: Equal=1, HighV-Quest=1, ImmuneDB=5)
                Because of pseduogene: 4 (Best-identity: Equal=1, HighV-Quest=0, ImmuneDB=3)
        When indel found by either:
                Equal number of indels: 22 (Best-identity: Equal=0, HighV-Quest=0, ImmuneDB=22)
                HighV-Quest has more indels: 109 (Best-identity: Equal=6, HighV-Quest=90, ImmuneDB=13)
                ImmuneDB has more indels: 31 (Best-identity: Equal=0, HighV-Quest=4, ImmuneDB=27)
```

Supplementary Table 3: A breakdown of how many total sequences were identified by ImmuneDB after local-alignment or HighV-Quest after collapsing identical sequences. Sequences with a copy number of one in either method were removed for further analysis.