Repertoire analysis

Clonal abundance, diversity and V-family gene usage

February 26, 2022

Contents

Bcellmagic analysis pipeline

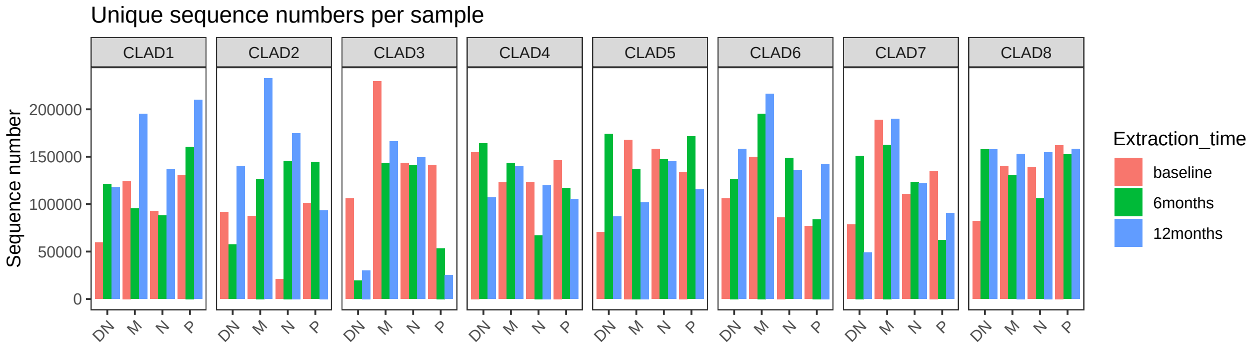
Pipeline overview

Number of sequences

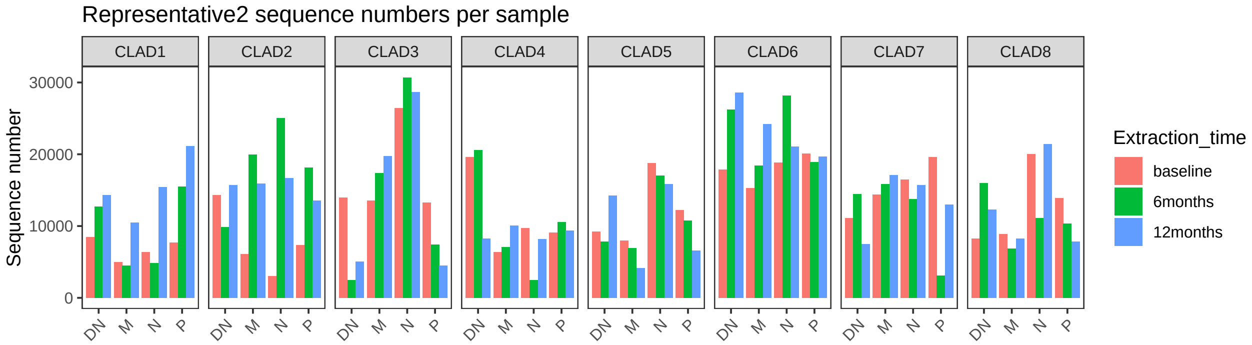
Number of reads for each of the samples and number of sequences left after representative analysis steps.

ID	Source	Treatment	Extraction_time	Population	Sequences_R1	Sequences_R2	Filtered_qu
QMKMK229AC	CLAD1	Cladribin	baseline	DN	1001901	1001901	
QMKMK230AF	CLAD1	Cladribin	baseline	N	601885	601885	
QMKMK231AN	CLAD1	Cladribin	baseline	M	1036290	1036290	
QMKMK232AV	CLAD1	Cladribin	baseline	P	1352414	1352414	
QMKMK233A5	CLAD1	Cladribin	6months	DN	1223690	1223690	
QMKMK234AD	CLAD1	Cladribin	6months	N	577697	577697	
QMKMK235AL	CLAD1	Cladribin	6months	M	801072	801072	
QMKMK236AT	CLAD1	Cladribin	6months	P	2043123	2043123	
QMKMK241AU	CLAD1	Cladribin	12months	DN	1857453	1857453	
QMKMK242A4	CLAD1	Cladribin	12months	M	1850771	1850771	
QMKMK243AC	CLAD1	Cladribin	12months	N	1244736	1244736	
QMKMK244AK	CLAD1	Cladribin	12months	P	1811754	1811754	
QMKMK533AN	CLAD2	Cladribin	baseline	N	299299	299299	
QMKMK534AV	CLAD2	Cladribin	baseline	M	911354	911354	
QMKMK535A5	CLAD2	Cladribin	baseline	DN	1548122	1548122	
QMKMK536AD	CLAD2	Cladribin	baseline	P	828075	828075	
QMKMK537AL	CLAD2	Cladribin	6months	N	1695948	1695948	
QMKMK538AT	CLAD2	Cladribin	6months	M	2467900	2467900	
QMKMK539A3	CLAD2	Cladribin	6months	DN	1670407	1670407	
QMKMK540A6	CLAD2	Cladribin	6months	P	1418658	1418658	
QMKMK541AE	CLAD2	Cladribin	12months	N	1254799	1254799	
QMKMK542AM	CLAD2	Cladribin	12months	M	2480119	2480119	
QMKMK543AU	CLAD2	Cladribin	12months	DN	1492340	1492340	
QMKMK544A4	CLAD2	Cladribin	12months	P	1104002	1104002	
QMKMK545AC	CLAD3	Cladribin	baseline	N	1571620	1571620	
QMKMK546AK	CLAD3	Cladribin	baseline	M	1798031	1798031	
QMKMK547AS	CLAD3	Cladribin	baseline	DN	1325826	1325826	
QMKMK548A2	CLAD3	Cladribin	baseline	P	1375460	1375460	
QMKMK549AA	CLAD3	Cladribin	6months	N	1385899	1385899	
QMKMK550AD	CLAD3	Cladribin	6months	M	1240876	1240876	
QMKMK551AL	CLAD3	Cladribin	6months	DN	261349	261349	
QMKMK552AT	CLAD3	Cladribin	6months	P	235034	235034	
QMKMK553A3	CLAD3	Cladribin	12months	N	1591756	1591756	
QMKMK554AB	CLAD3	Cladribin	12months	M	1564019	1564019	
QMKMK555AJ	CLAD3	Cladribin	12months	DN	111483	111483	
QMKMK556AR	CLAD3	Cladribin	12months	P	155013	155013	
QMKMK557A1	CLAD4	Cladribin	baseline	N	1398965	1398965	
QMKMK558A9	CLAD4	Cladribin	baseline	M	1374928	1374928	
QMKMK559AH	CLAD4	Cladribin	baseline	DN	2409179	2409179	
QMKMK560AK	CLAD4	Cladribin	baseline	P	1068173	1068173	
QMKMK561AS	CLAD4	Cladribin	6months	N	1641107	1641107	
QMKMK562A2	CLAD4	Cladribin	6months	M	1385399	1385399	
QMKMK563AA	CLAD4	Cladribin	6months	DN	2589345	2589345	
QMKMK564AI	CLAD4	Cladribin	6months	P	1801823	1801823	
QMKMK565AQ	CLAD4	Cladribin	12months	N	1233912	1233912	
QMKMK566A0	CLAD4	Cladribin	12months	M	1519136	1519136	
QMKMK567A8	CLAD4	Cladribin	12months	DN	1084835	1084835	
QMKMK568AG	CLAD4	Cladribin	12months	P	1322433	1322433	
QMKMK569AO	CLAD5	Cladribin	baseline	N	1447197	1447197	
QMKMK570AR	CLAD5	Cladribin	baseline	M	1468661	1468661	
QMKMK571A1	CLAD5	Cladribin	baseline	DN	1126741	1126741	
QMKMK572A9	CLAD5	Cladribin	baseline	P	1397460	1397460	
QMKMK573AH	CLAD5	Cladribin	6months	N	1329042	1329042	
QMKMK574AP	CLAD5	Cladribin	6months	M	1237833	1237833	
QMKMK575AX	CLAD5	Cladribin	6months	DN	1383178	1383178	
QMKMK576A7	CLAD5	Cladribin	6months	P	1262705	1262705	
QMKMK577AF	CLAD5	Cladribin	12months	N	1351967	1351967	
QMKMK578AN	CLAD5	Cladribin	12months	M	[illegible]	[illegible]	

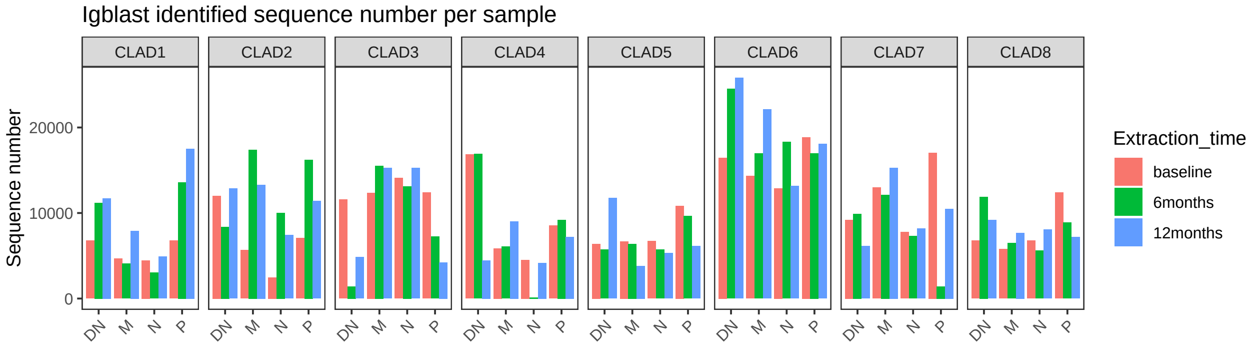
Plotting number of unique sequences



Plotting number of representative 2 sequences



Plotting number of Igblast identified sequences



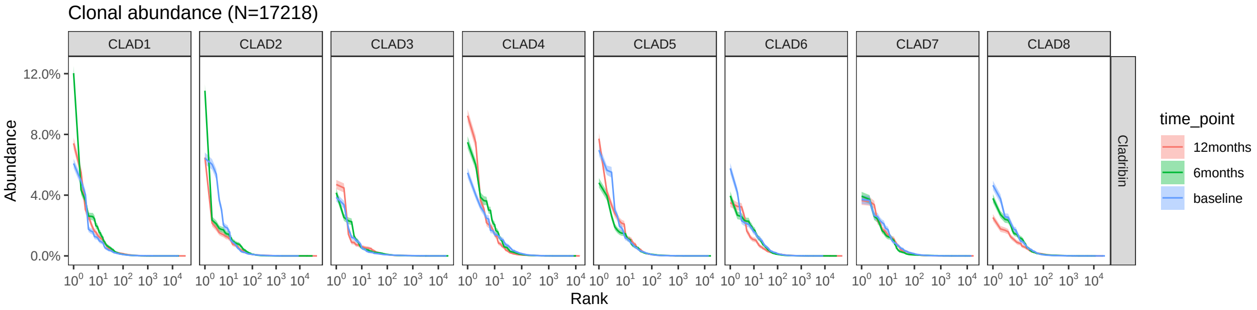
Clonal abundance

For plotting the clonal abundance, the clones were ordered by size from bigger clones to smaller clones (x-axis, Rank). The Abundance of each clone was represented as the percentage of unique sequences in the clone, with respect to the total number of unique sequences in that subject (By Patient) or in the B-cell or T-cell sample (By Cell Population).

To correct for the different number of sequences in each of the samples, the Bootstrapping technique was employed, in which 200 random bootstrap samples were taken, with size the number of sequences in the sample with less sequences (N). The solid line shows the mean Abundance of the bootstrap samples, whereas the transparent area shows the full Abundance range of the bootstrap samples.

All clonal abundance plots and tables with abundance values can be found under `repertoire_analysis/Abundance`.

Clonal abundance per subject

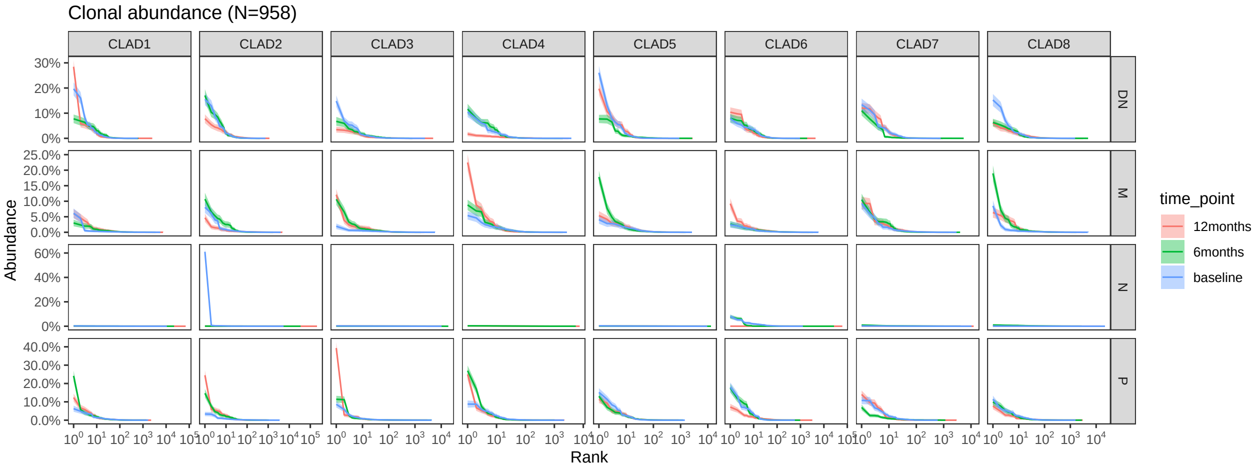


Calculate area under the curve for abundance

Count clones per subject

Clonal abundance per cell population

If different types of B-cell or T-cell populations are provided, here the clonal abundance is plotted for each patient and B / T-cell population.



Count clones per population

Clonal diversity

The clonal diversity D of the repertoire was calculated according to the general formula of Hill Diversity numbers:

$$^{q}D = \left(\sum_{i=1}^{R} p_i^q\right)^{1/(1-q)}$$

where:

- p_i is the proportion of unique sequences belonging to clone i.
- q are the values of the different diversity numbers.
- R is the Richness, the number of different clones in the sample.

At $q = 1$ the function is undefined and the limit to zero equals the exponential of the Shannon Entropy:

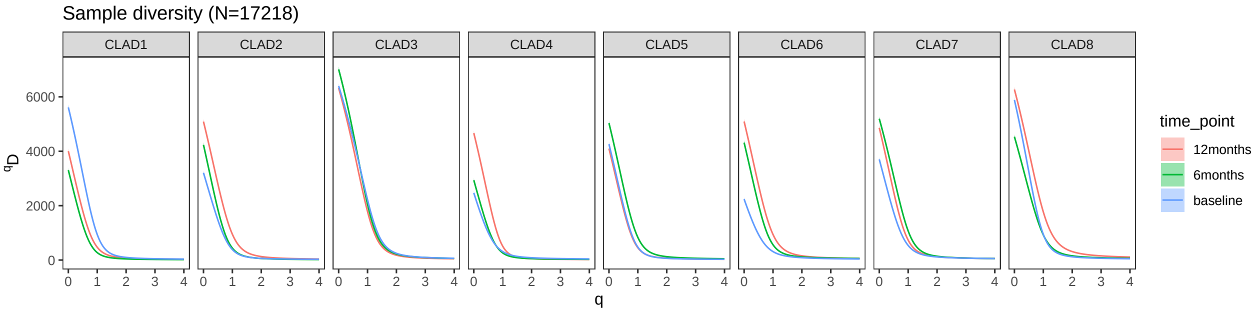
$$^{1}D = exp\left(\sum_{i=1}^{R} p_i ln(p_i)\right)$$

The intuition about the different Hill Diversity values is the following:

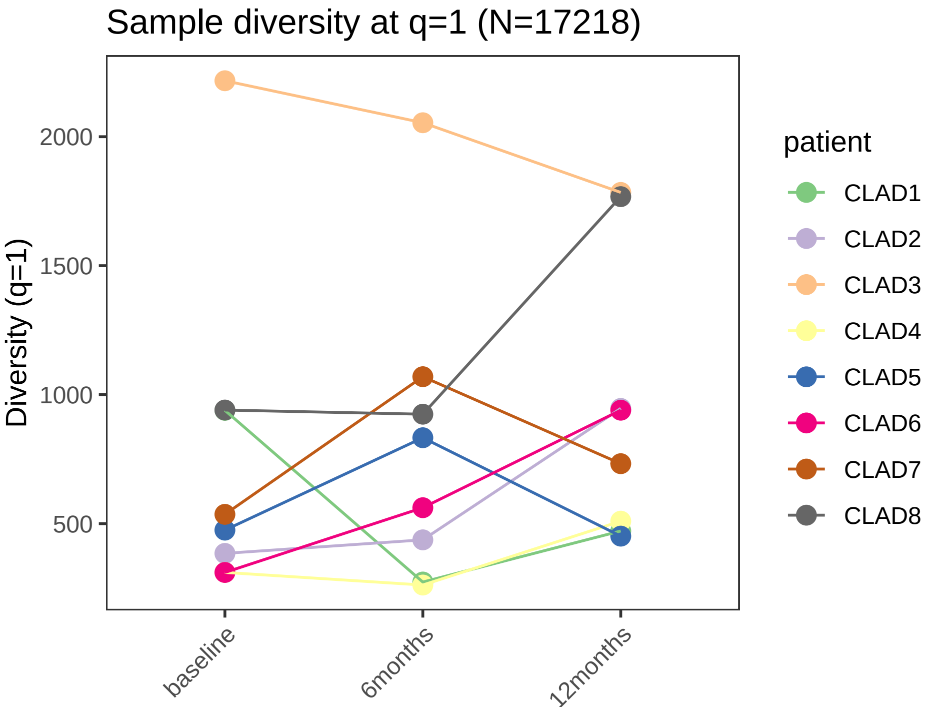
- At $q = 0$ the diversity index equals the number of clones in the sample.
- At $q = 1$ the diversity index is the geometric mean of the clones in the sample, weighted by their proportion in the sample.
- At $q > 1$ more weight is given to the clones with higher proportions in the sample.

All clonal diversity plots and tables with diversity values can be found under `repertoire_analysis/Diversity`. To correct for the different number of sequences in each of the samples, the Bootstrapping technique was employed, in which 200 random bootstrap samples were taken, with size the number of sequences in the sample with less sequences (N). The solid line shows the mean Diversity of the bootstrap samples, whereas the transparent area shows the full Diversity range of the bootstrap samples.

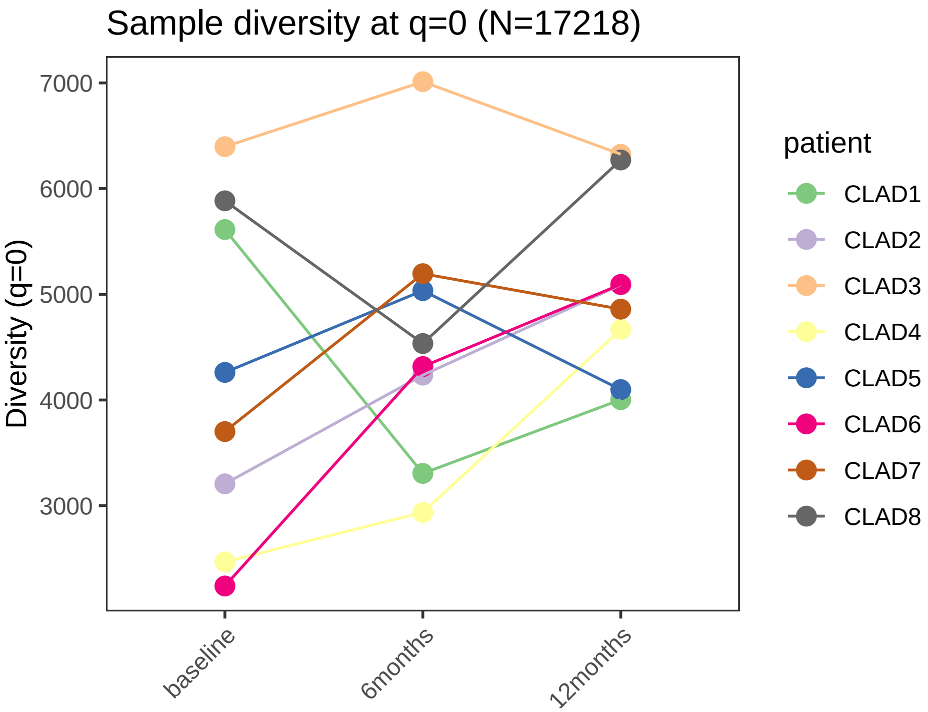
Clonal diversity per subject



Clonal diversity at specific q values

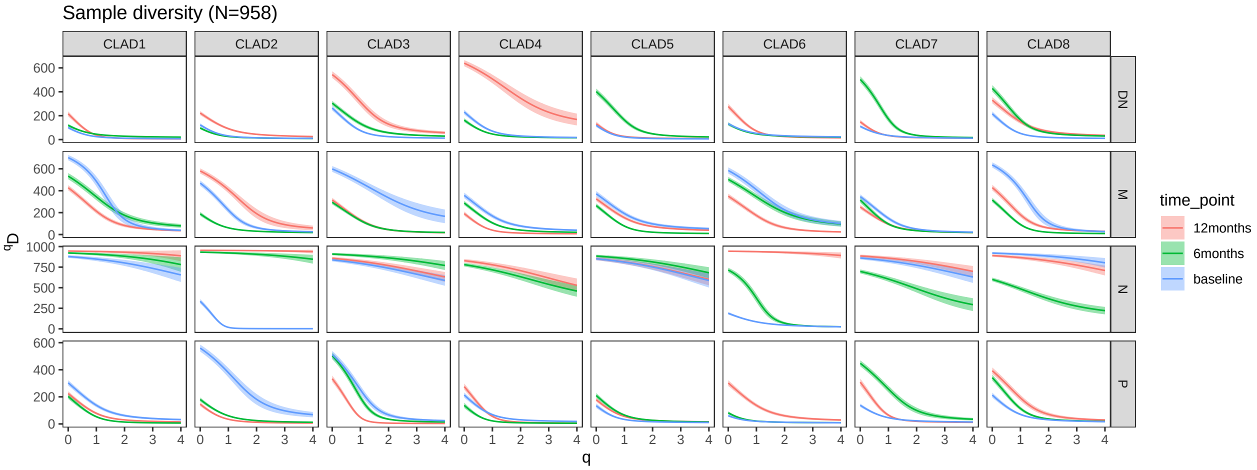


q=1

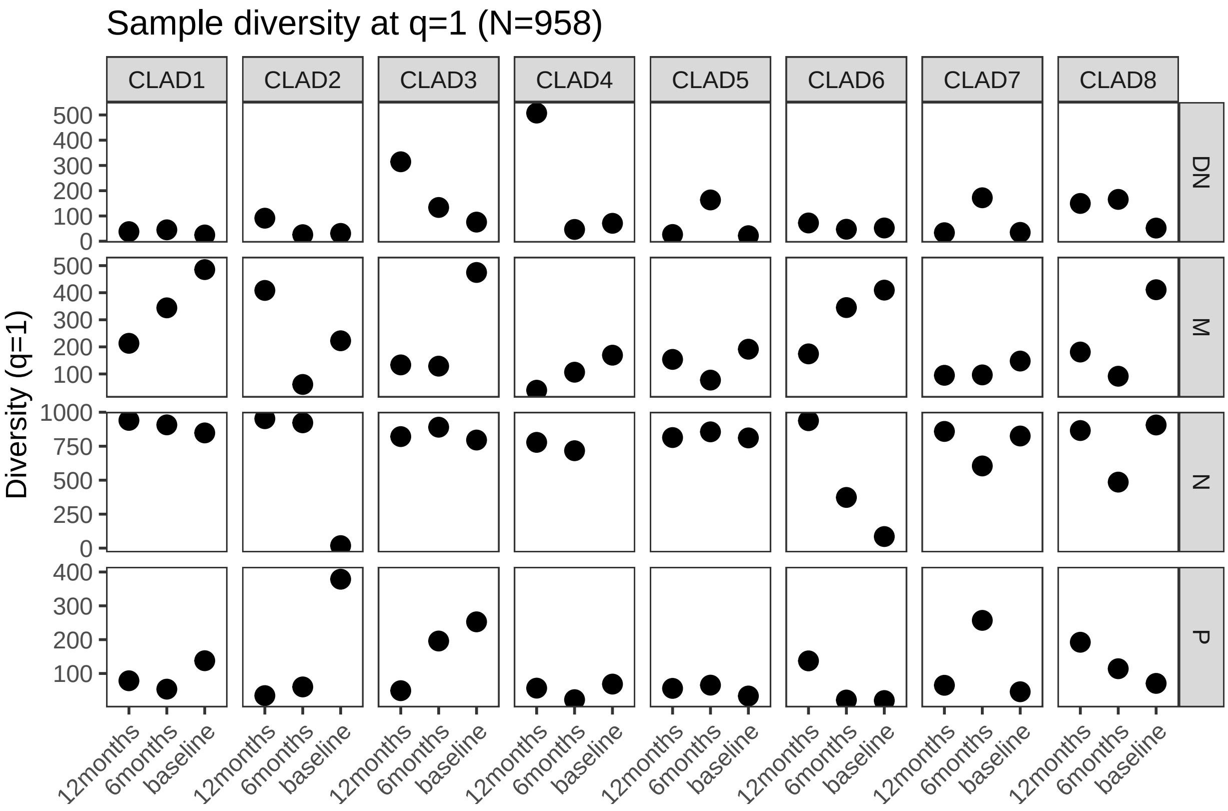


q=0

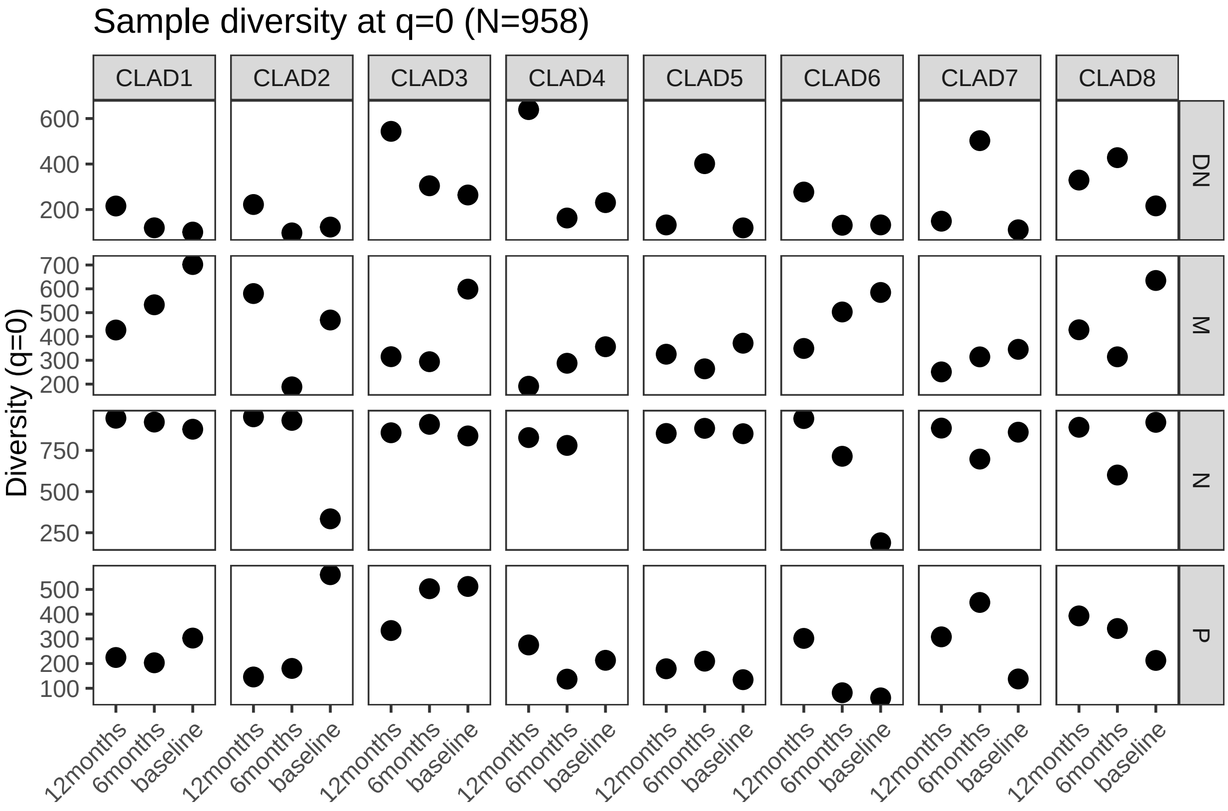
Clonal diversity per cell population



Clonal diversity per population at specific q values


Sample diversity at q=1 (N=958)
CLAD1
CLAD2
CLAD3
CLAD4
CLAD5
CLAD6
CLAD7
CLAD8
DN
M
N
P
Diversity (q=1)
0
100
200
300
400
500
250
750
1000
12months
6months
baseline

q=1


Sample diversity at q=0 (N=958)
CLAD1
CLAD2
CLAD3
CLAD4
CLAD5
CLAD6
CLAD7
CLAD8
DN
M
N
P
Diversity (q=0)
100
200
300
400
500
600
700
250
750
12months
6months
baseline

q=0

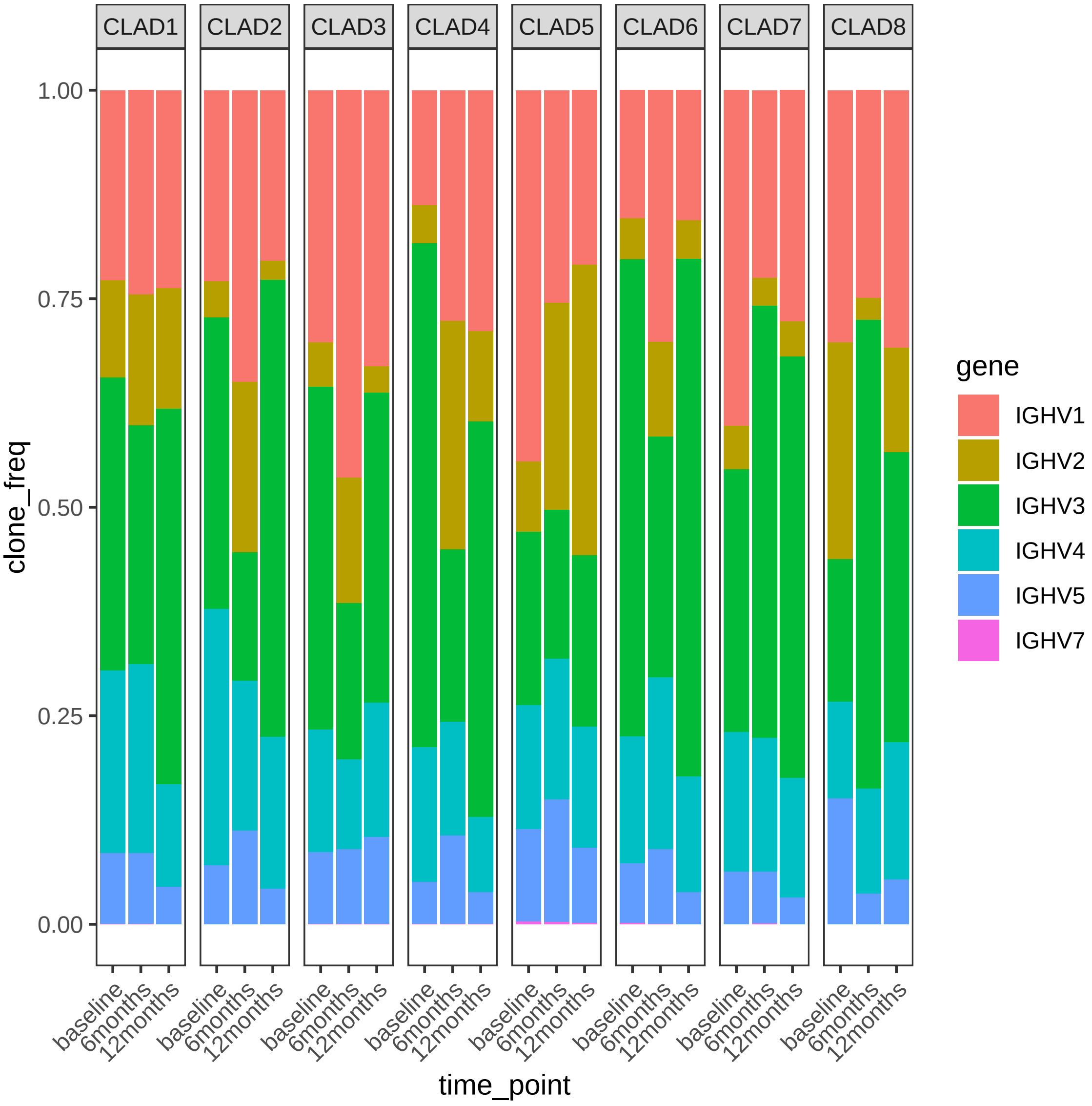
V gene usage

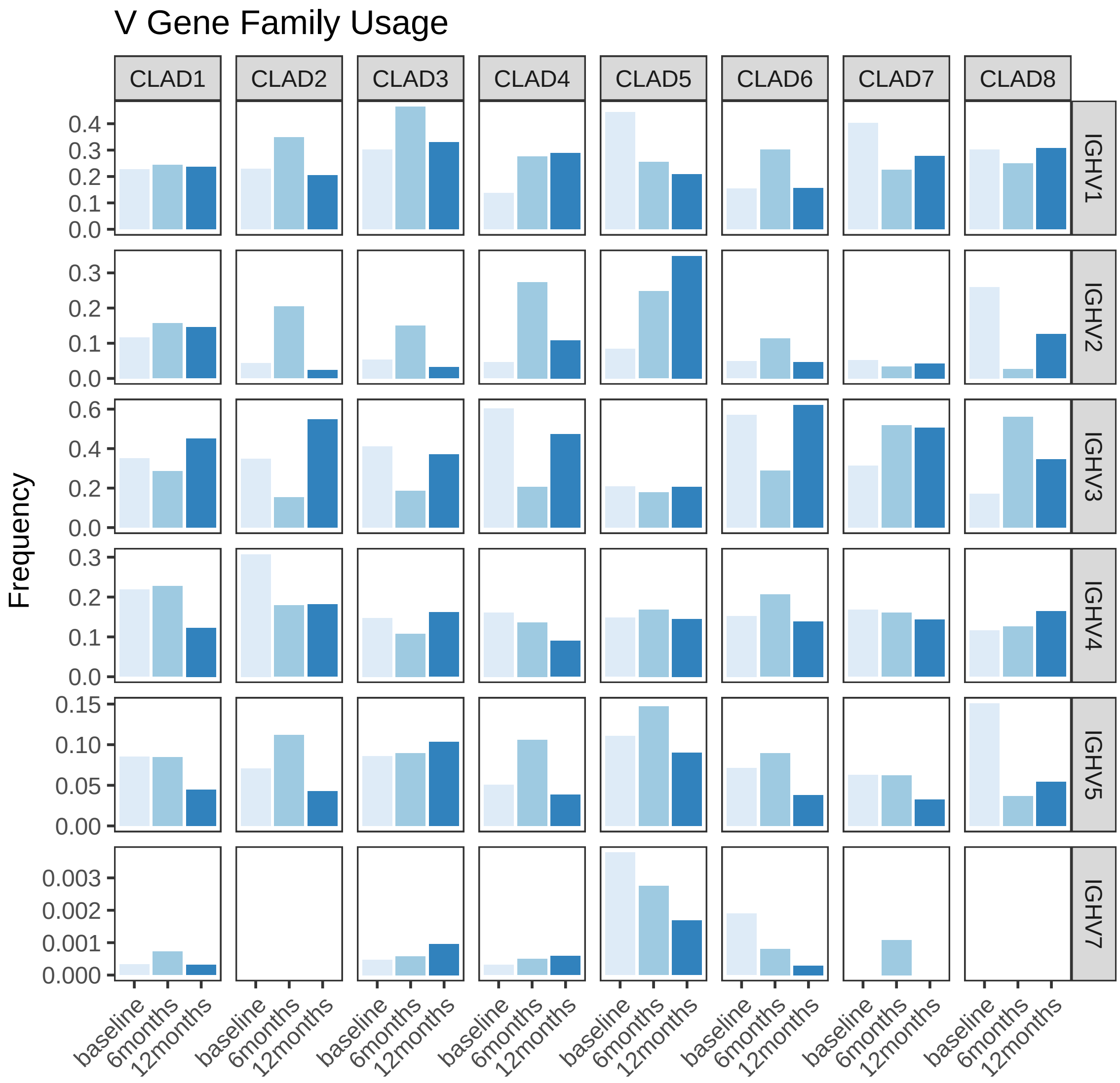
V gene family usage

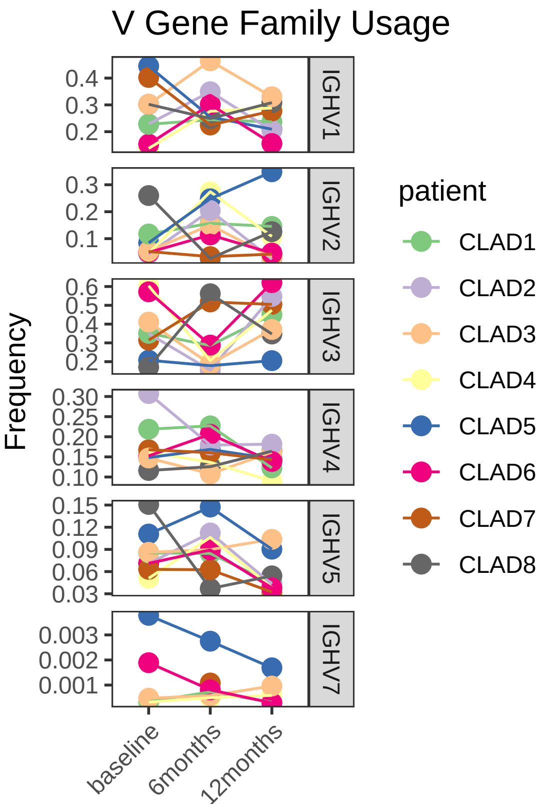
The V gene usage (in percentage) in each of the samples is represented below. All plots and tables can be found here.

Gene family usage is normalized by the number of clones.

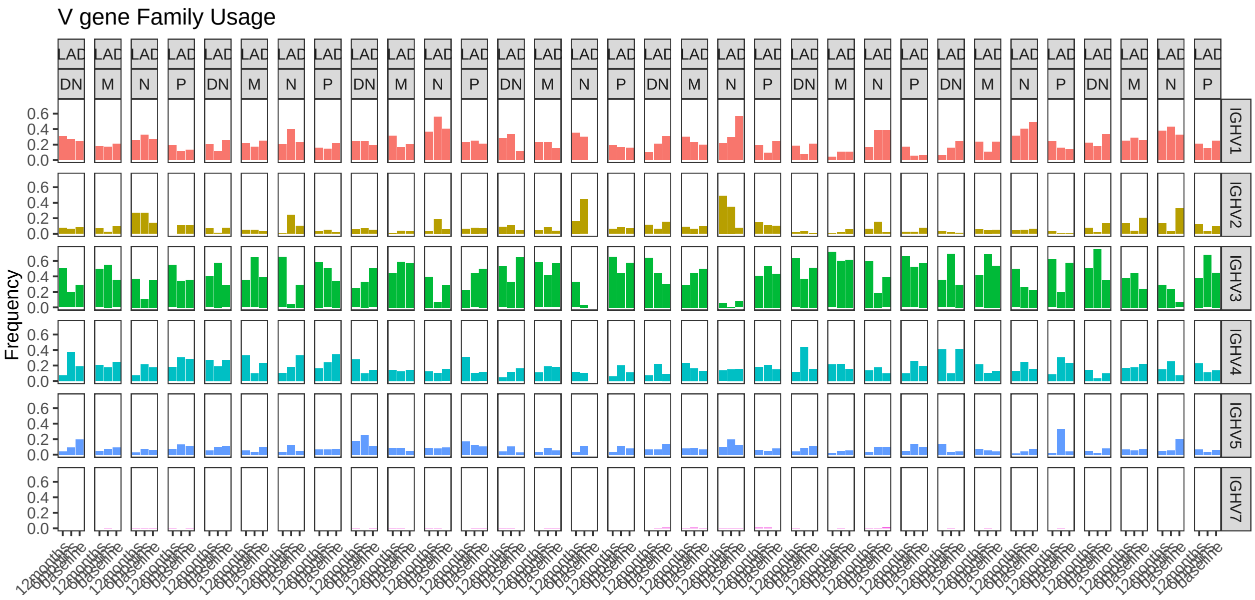
By patient




V Gene Family Usage
CLAD1
CLAD2
CLAD3
CLAD4
CLAD5
CLAD6
CLAD7
CLAD8
IGHV1
IGHV2
IGHV3
IGHV4
IGHV5
IGHV7
Frequency
0.0
0.1
0.2
0.3
0.4
0.0
0.1
0.2
0.3
0.0
0.2
0.4
0.6
0.0
0.1
0.2
0.3
0.00
0.05
0.10
0.15
0.000
0.001
0.002
0.003
baseline
6months
12months



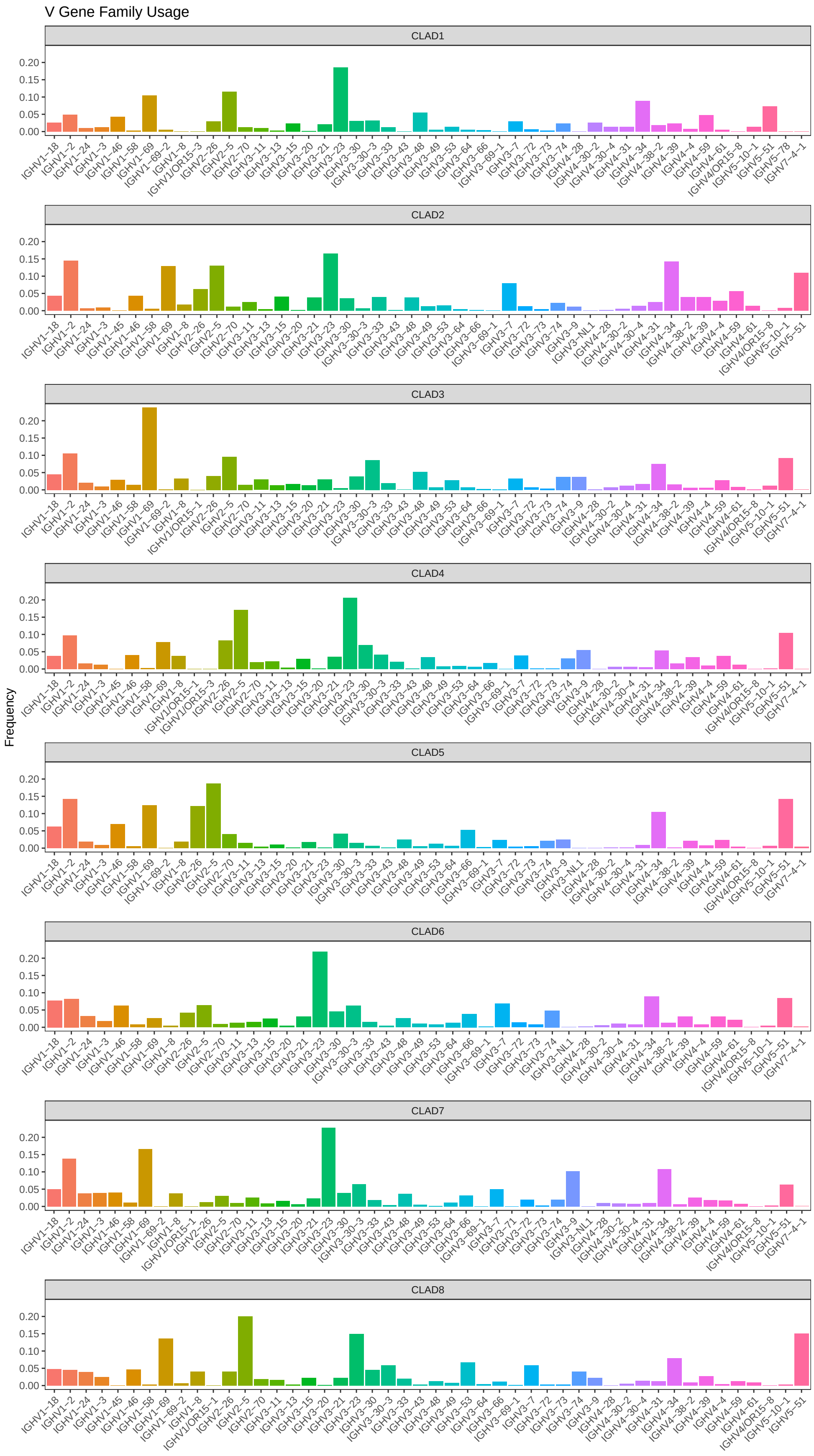
By Population



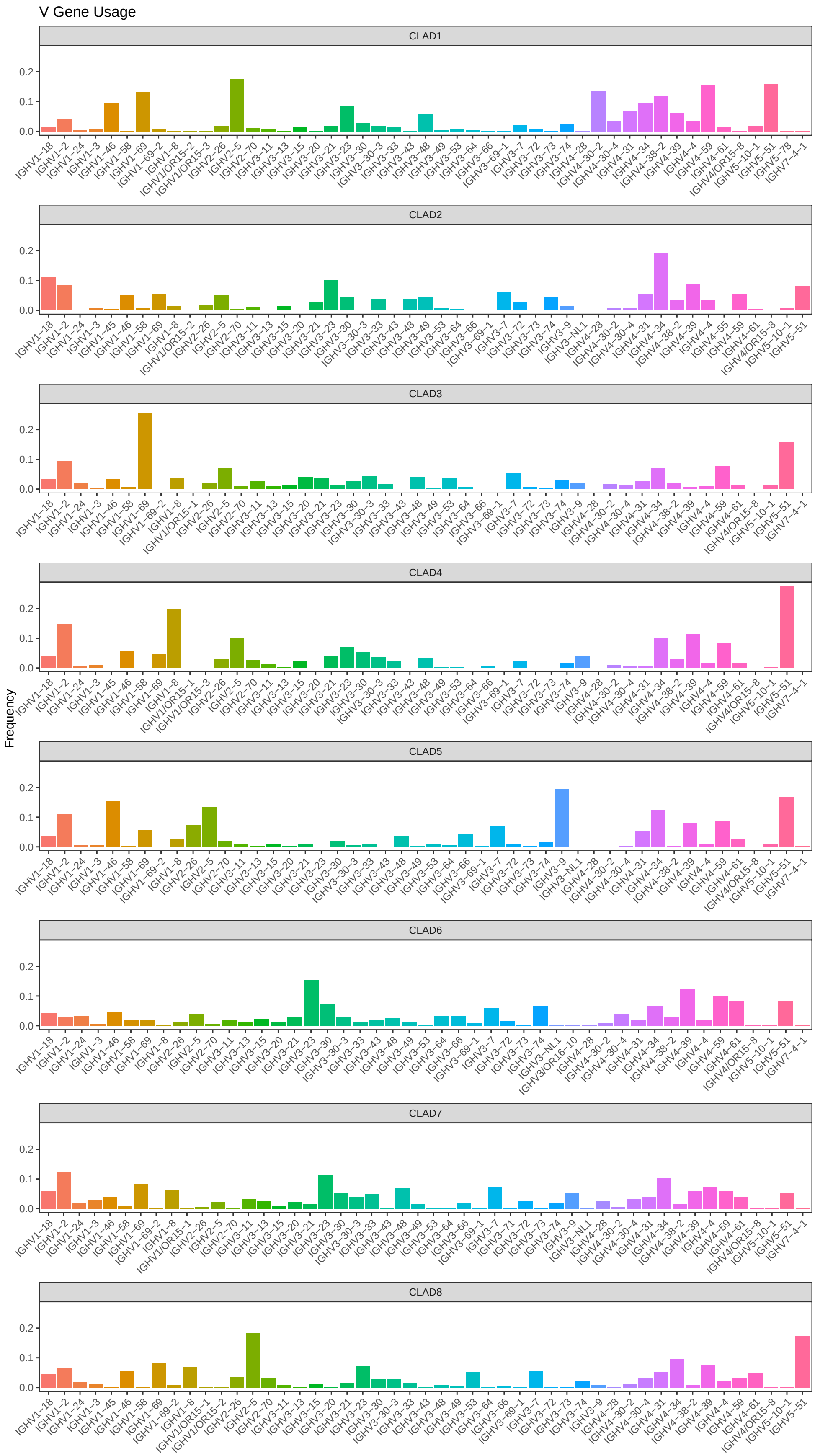
V gene usage

The V gene usage (in percentage) in each of the samples is represented below. All plots and tables can be found here.

By clones


V Gene Family Usage
Frequency
0.00 0.05 0.10 0.15 0.20
CLAD1
IGHV1−18 IGHV1−2 IGHV1−24 IGHV1−3 IGHV1−46 IGHV1−58 IGHV1−69 IGHV1−69−2 IGHV1−8 IGHV1/OR15−3 IGHV2−26 IGHV2−5 IGHV2−70 IGHV3−11 IGHV3−13 IGHV3−15 IGHV3−20 IGHV3−21 IGHV3−23 IGHV3−30 IGHV3−30−3 IGHV3−33 IGHV3−43 IGHV3−48 IGHV3−49 IGHV3−53 IGHV3−64 IGHV3−66 IGHV3−69−1 IGHV3−7 IGHV3−72 IGHV3−73 IGHV3−74 IGHV4−28 IGHV4−30−2 IGHV4−30−4 IGHV4−31 IGHV4−34 IGHV4−38−2 IGHV4−39 IGHV4−4 IGHV4−59 IGHV4−61 IGHV4/OR15−8 IGHV5−10−1 IGHV5−51 IGHV5−78 IGHV7−4−1
CLAD2
IGHV1−18 IGHV1−2 IGHV1−24 IGHV1−3 IGHV1−45 IGHV1−46 IGHV1−58 IGHV1−69 IGHV1−8 IGHV2−26 IGHV2−5 IGHV2−70 IGHV3−11 IGHV3−13 IGHV3−15 IGHV3−20 IGHV3−21 IGHV3−23 IGHV3−30 IGHV3−30−3 IGHV3−33 IGHV3−43 IGHV3−48 IGHV3−49 IGHV3−53 IGHV3−64 IGHV3−66 IGHV3−69−1 IGHV3−7 IGHV3−72 IGHV3−73 IGHV3−74 IGHV3−9 IGHV3−NL1 IGHV4−28 IGHV4−30−2 IGHV4−30−4 IGHV4−31 IGHV4−34 IGHV4−38−2 IGHV4−39 IGHV4−4 IGHV4−59 IGHV4−61 IGHV4/OR15−8 IGHV5−10−1 IGHV5−51
CLAD3
IGHV1−18 IGHV1−2 IGHV1−24 IGHV1−3 IGHV1−46 IGHV1−58 IGHV1−69 IGHV1−69−2 IGHV1−8 IGHV1/OR15−1 IGHV2−26 IGHV2−5 IGHV2−70 IGHV3−11 IGHV3−13 IGHV3−15 IGHV3−20 IGHV3−21 IGHV3−23 IGHV3−30 IGHV3−30−3 IGHV3−33 IGHV3−43 IGHV3−48 IGHV3−49 IGHV3−53 IGHV3−64 IGHV3−66 IGHV3−69−1 IGHV3−7 IGHV3−72 IGHV3−73 IGHV3−74 IGHV3−9 IGHV4−28 IGHV4−30−2 IGHV4−30−4 IGHV4−31 IGHV4−34 IGHV4−38−2 IGHV4−39 IGHV4−4 IGHV4−59 IGHV4−61 IGHV4/OR15−8 IGHV5−10−1 IGHV5−51 IGHV7−4−1
CLAD4
IGHV1−18 IGHV1−2 IGHV1−24 IGHV1−3 IGHV1−45 IGHV1−46 IGHV1−58 IGHV1−69 IGHV1−8 IGHV1/OR15−1 IGHV1/OR15−3 IGHV2−26 IGHV2−5 IGHV2−70 IGHV3−11 IGHV3−13 IGHV3−15 IGHV3−20 IGHV3−21 IGHV3−23 IGHV3−30 IGHV3−30−3 IGHV3−33 IGHV3−43 IGHV3−48 IGHV3−49 IGHV3−53 IGHV3−64 IGHV3−66 IGHV3−69−1 IGHV3−7 IGHV3−72 IGHV3−73 IGHV3−74 IGHV3−9 IGHV4−28 IGHV4−30−2 IGHV4−30−4 IGHV4−31 IGHV4−34 IGHV4−38−2 IGHV4−39 IGHV4−4 IGHV4−59 IGHV4−61 IGHV4/OR15−8 IGHV5−10−1 IGHV5−51 IGHV7−4−1
CLAD5
IGHV1−18 IGHV1−2 IGHV1−24 IGHV1−3 IGHV1−46 IGHV1−58 IGHV1−69 IGHV1−69−2 IGHV1−8 IGHV2−26 IGHV2−5 IGHV2−70 IGHV3−11 IGHV3−13 IGHV3−15 IGHV3−20 IGHV3−21 IGHV3−23 IGHV3−30 IGHV3−30−3 IGHV3−33 IGHV3−43 IGHV3−48 IGHV3−49 IGHV3−53 IGHV3−64 IGHV3−66 IGHV3−69−1 IGHV3−7 IGHV3−72 IGHV3−73 IGHV3−74 IGHV3−9 IGHV3−NL1 IGHV4−28 IGHV4−30−2 IGHV4−30−4 IGHV4−31 IGHV4−34 IGHV4−38−2 IGHV4−39 IGHV4−4 IGHV4−59 IGHV4−61 IGHV4/OR15−8 IGHV5−10−1 IGHV5−51 IGHV7−4−1
CLAD6
IGHV1−18 IGHV1−2 IGHV1−24 IGHV1−3 IGHV1−46 IGHV1−58 IGHV1−69 IGHV1−8 IGHV2−26 IGHV2−5 IGHV2−70 IGHV3−11 IGHV3−13 IGHV3−15 IGHV3−20 IGHV3−21 IGHV3−23 IGHV3−30 IGHV3−30−3 IGHV3−33 IGHV3−43 IGHV3−48 IGHV3−49 IGHV3−53 IGHV3−64 IGHV3−66 IGHV3−69−1 IGHV3−7 IGHV3−72 IGHV3−73 IGHV3−74 IGHV3−NL1 IGHV4−28 IGHV4−30−2 IGHV4−30−4 IGHV4−31 IGHV4−34 IGHV4−38−2 IGHV4−39 IGHV4−4 IGHV4−59 IGHV4−61 IGHV4/OR15−8 IGHV5−10−1 IGHV5−51 IGHV7−4−1
CLAD7
IGHV1−18 IGHV1−2 IGHV1−24 IGHV1−3 IGHV1−46 IGHV1−58 IGHV1−69 IGHV1−69−2 IGHV1−8 IGHV1/OR15−1 IGHV2−26 IGHV2−5 IGHV2−70 IGHV3−11 IGHV3−13 IGHV3−15 IGHV3−20 IGHV3−21 IGHV3−23 IGHV3−30 IGHV3−30−3 IGHV3−33 IGHV3−43 IGHV3−48 IGHV3−49 IGHV3−53 IGHV3−64 IGHV3−66 IGHV3−69−1 IGHV3−7 IGHV3−71 IGHV3−72 IGHV3−73 IGHV3−74 IGHV3−9 IGHV3−NL1 IGHV4−28 IGHV4−30−2 IGHV4−30−4 IGHV4−31 IGHV4−34 IGHV4−38−2 IGHV4−39 IGHV4−4 IGHV4−59 IGHV4−61 IGHV4/OR15−8 IGHV5−10−1 IGHV5−51 IGHV7−4−1
CLAD8
IGHV1−18 IGHV1−2 IGHV1−24 IGHV1−3 IGHV1−45 IGHV1−46 IGHV1−58 IGHV1−69 IGHV1−69−2 IGHV1−8 IGHV1/OR15−1 IGHV2−26 IGHV2−5 IGHV2−70 IGHV3−11 IGHV3−13 IGHV3−15 IGHV3−20 IGHV3−21 IGHV3−23 IGHV3−30 IGHV3−30−3 IGHV3−33 IGHV3−43 IGHV3−48 IGHV3−49 IGHV3−53 IGHV3−64 IGHV3−66 IGHV3−69−1 IGHV3−7 IGHV3−72 IGHV3−73 IGHV3−74 IGHV3−9 IGHV4−28 IGHV4−30−2 IGHV4−30−4 IGHV4−31 IGHV4−34 IGHV4−38−2 IGHV4−39 IGHV4−4 IGHV4−59 IGHV4−61 IGHV4/OR15−8 IGHV5−10−1 IGHV5−51

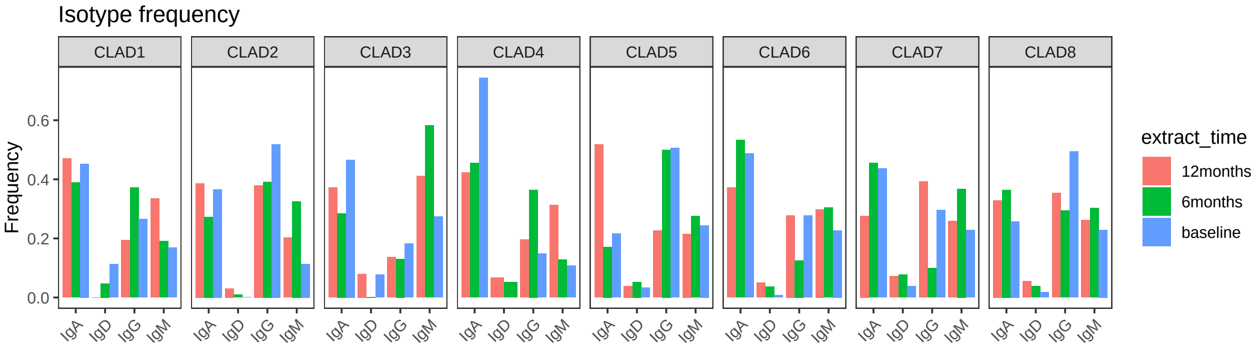
By sequences



Isotype usage

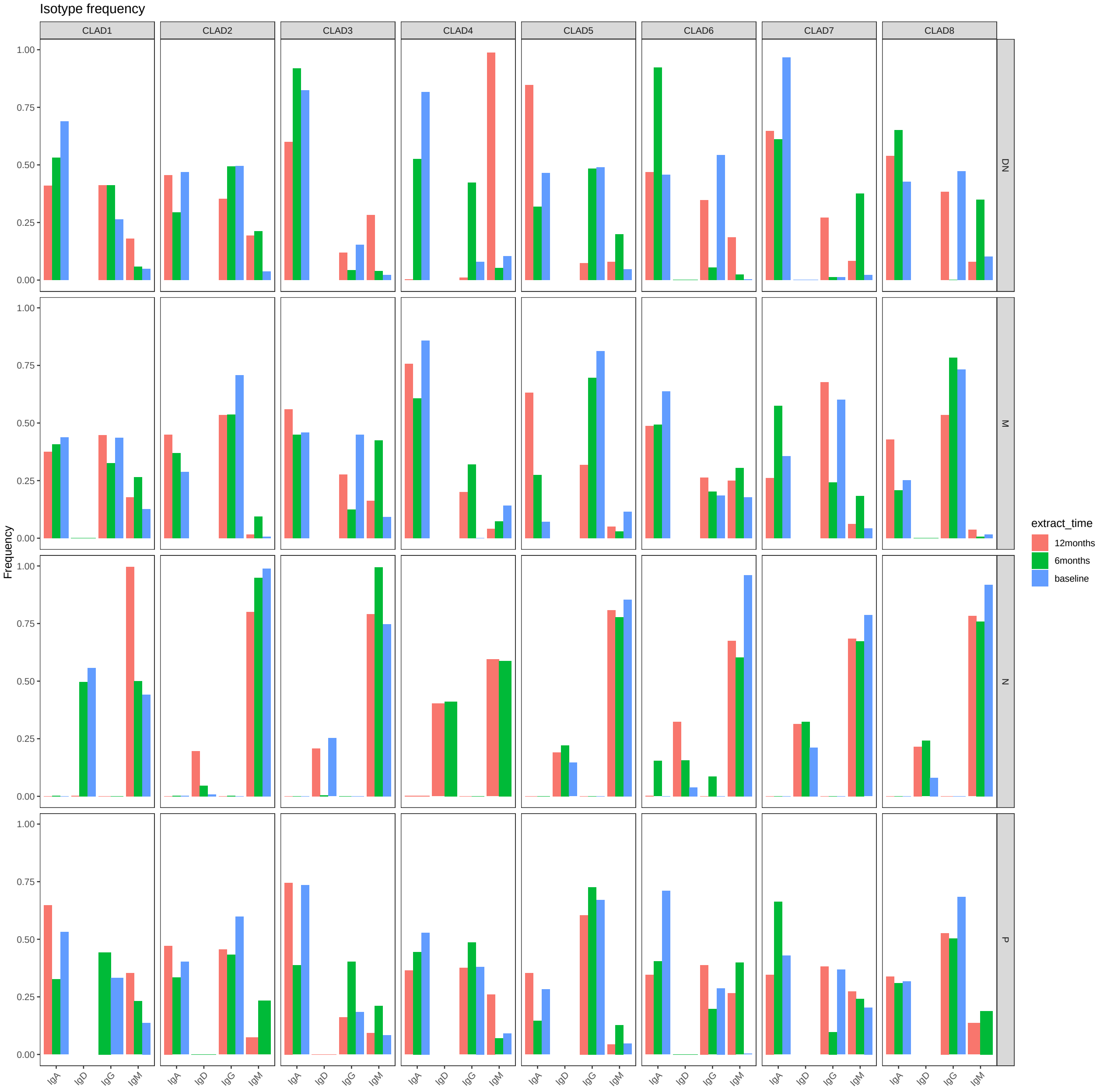
Isotype usage per subject

```
## `summarise()` has grouped output by 'isotype', 'sample', 'source', 'treatment'. You can override usi
```



Isotype usage per cell population

```
## `summarise()` has grouped output by 'isotype', 'sample_pop', 'source', 'treatment', 'extract_time'.
```



Clonal overlap analysis

Citations

If you use nf-core/bcellmagic for your analysis, please cite it using the following DOI: 10.5281/zenodo.3607408

Please also cite the `nf-core` publication (P. A. Ewels et al. 2020).

In addition, citations for the tools and data used in this pipeline are as follows:

- **pRESTO** (Vander Heiden et al. 2014)
- **SHazaM, Change-O** (Gupta et al. 2015)
- **Alakazam** (Stern et al. 2014)
- **TIgGER** (Gadala-Maria et al. 2015)
- **FastQC** (Andrews et al. 2010)

- **MultiQC** (P. Ewels et al. 2016)

Andrews, Simon et al. 2010. "FastQC: A Quality Control Tool for High Throughput Sequence Data."

Ewels, Philip A., Alexander Peltzer, Sven Fillinger, Harshil Patel, Johannes Alneberg, Andreas Wilm, Maxime Ulysse Garcia, Paolo Di Tommaso, and Sven Nahnsen. 2020. "The Nf-Core Framework for Community-Curated Bioinformatics Pipelines." *Nature Biotechnology* 38 (3): 276–78. https://doi.org/10.1038/s41587-020-0439-x.

Ewels, Philip, Måns Magnusson, Sverker Lundin, and Max Käller. 2016. "MultiQC: Summarize Analysis Results for Multiple Tools and Samples in a Single Report." *Bioinformatics* 32 (19): 3047–48.

Gadala-Maria, Daniel, Gur Yaari, Mohamed Uduman, and Steven H. Kleinstein. 2015. "Automated Analysis of High-Throughput b-Cell Sequencing Data Reveals a High Frequency of Novel Immunoglobulin v Gene Segment Alleles." *Proceedings of the National Academy of Sciences of the United States of America* 112 (8): E862–870. https://doi.org/10.1073/pnas.1417683112.

Gupta, Namita T., Jason A. Vander Heiden, Mohamed Uduman, Daniel Gadala-Maria, Gur Yaari, and Steven H. Kleinstein. 2015. "Change-o: A Toolkit for Analyzing Large-Scale b Cell Immunoglobulin Repertoire Sequencing Data." *Bioinformatics* 31 (20): 3356–58. https://doi.org/10.1093/bioinformatics/btv359.

Stern, Joel N. H., Gur Yaari, Jason A. Vander Heiden, George Church, William F. Donahue, Rogier Q. Hintzen, Anita J. Huttner, et al. 2014. "B Cells Populating the Multiple Sclerosis Brain Mature in the Draining Cervical Lymph Nodes." *Science Translational Medicine* 6 (248). https://doi.org/10.1126/scitranslmed.3008879.

Vander Heiden, Jason A., Gur Yaari, Mohamed Uduman, Joel N. H. Stern, Kevin C. O'Connor, David A. Hafler, Francois Vigneault, and Steven H. Kleinstein. 2014. "pRESTO: A Toolkit for Processing High-Throughput Sequencing Raw Reads of Lymphocyte Receptor Repertoires." *Bioinformatics* 30 (13): 1930–32. https://doi.org/10.1093/bioinformatics/btu138.