

hillrep — validation & use cases

Coverage-based Hill-number diversity for immune repertoires.

Comparing the clonal diversity of TCR/BCR repertoires sequenced to different depths is a routine but error-prone step: richness, Shannon and Simpson indices all rise with sequencing depth simply because more rare clones are observed, so comparing them across samples of unequal depth is biased.

The statistically correct fix is to report Hill numbers (effective numbers of clones) standardized to a common sample coverage, via rarefaction and extrapolation. That framework (the iNEXT method of Chao et al. 2014; Hsieh, Ma & Chao 2016) has lived almost entirely in R. hillrep brings it to Python with first-class AIRR-seq support.

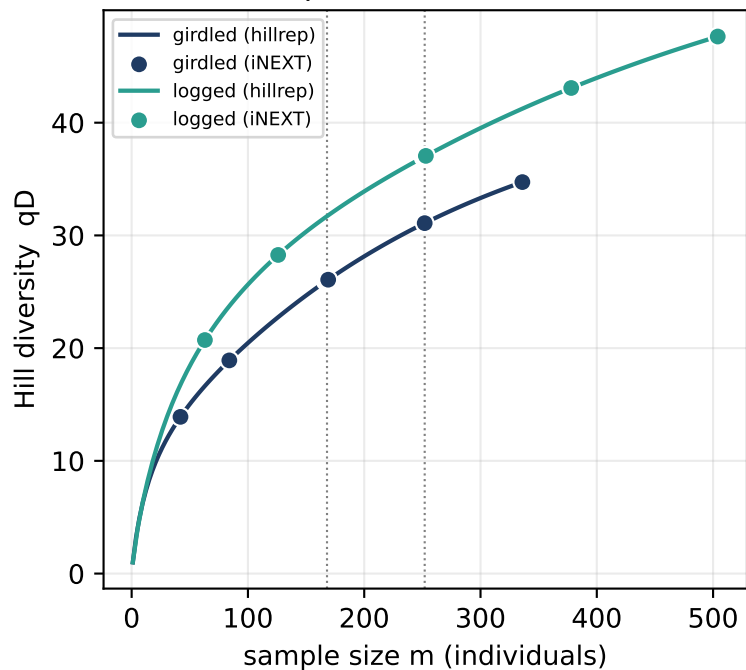
This report (1) validates hillrep against R iNEXT 3.0.2 on canonical data, then demonstrates six use cases: (2) depth-bias correction, (3) fair multi-sample comparison, (4) an AIRR Rearrangement pipeline, (5) real public TCR-beta repertoires from the AIRR Data Commons, (6) a comparison-reliability assessment (assess), and (7) robustness on extreme inputs.

Every figure is produced by scripts/make_report.py and is fully reproducible.

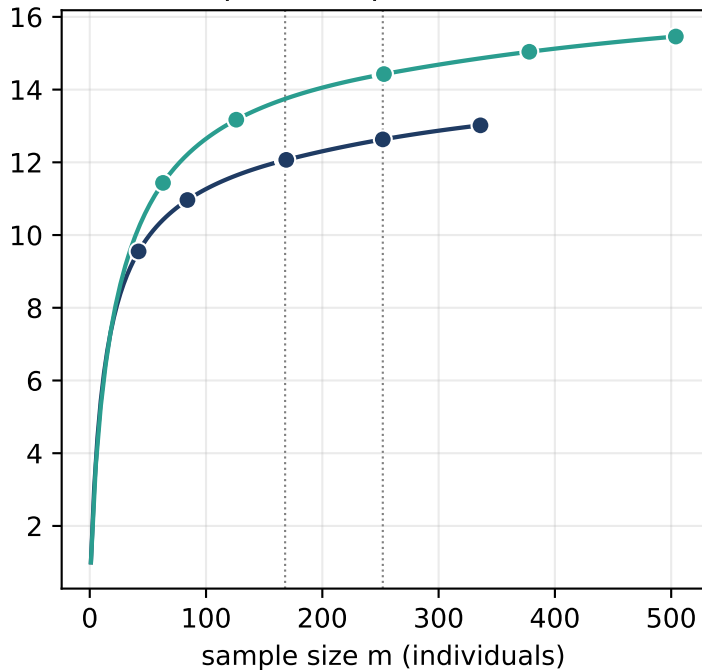
Data are public (the ecology 'spider' dataset and real repertoires from the AIRR Data Commons) or simulated from a realistic log-normal clone-size model with known ground truth; no unpublished patient data is used.

Validation: hillrep vs R iNEXT 3.0.2 (canonical 'spider' ecology data)

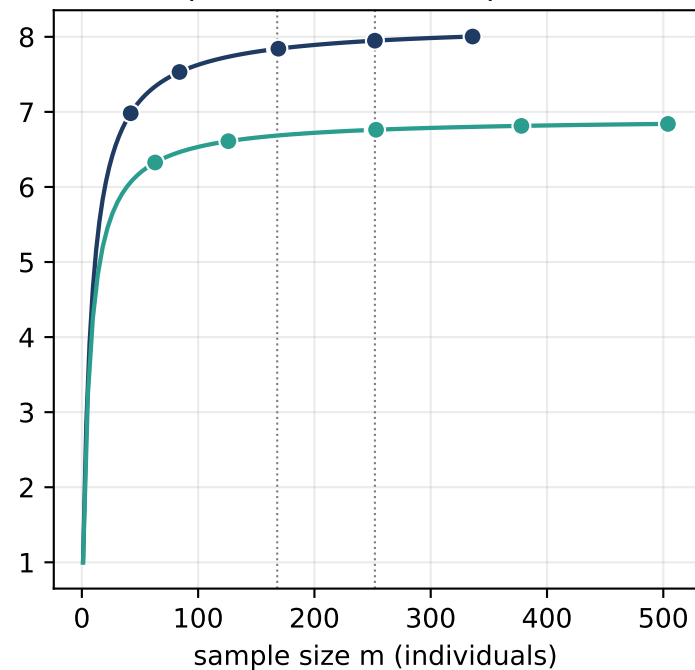
q = 0 (richness)



q = 1 (exp-Shannon)

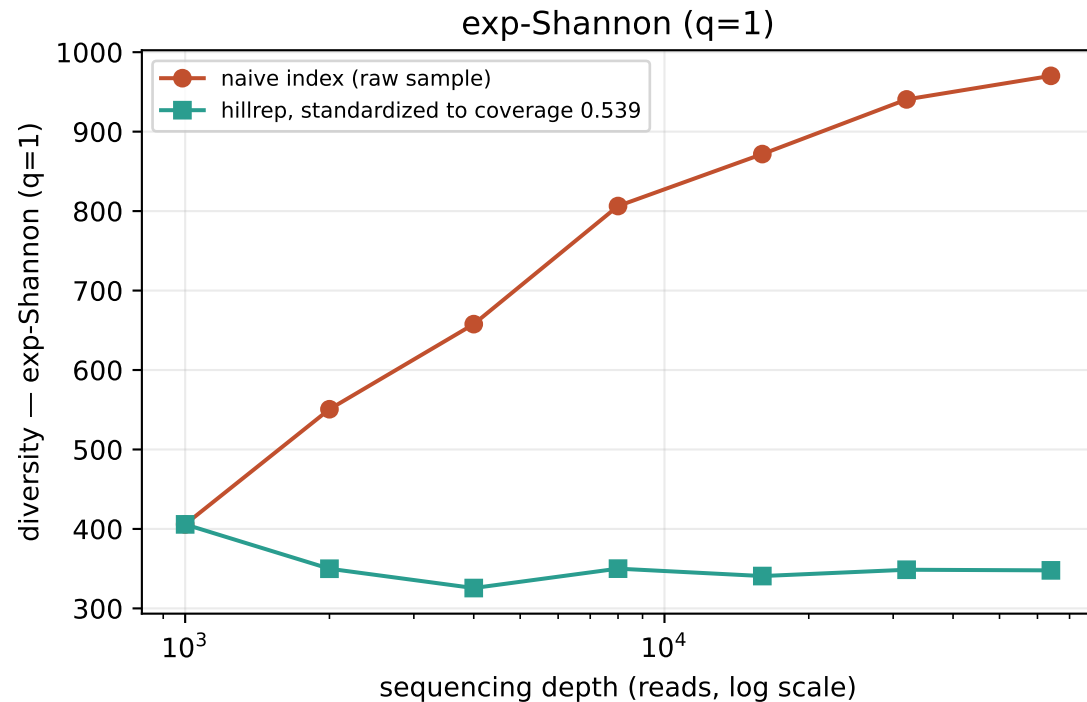
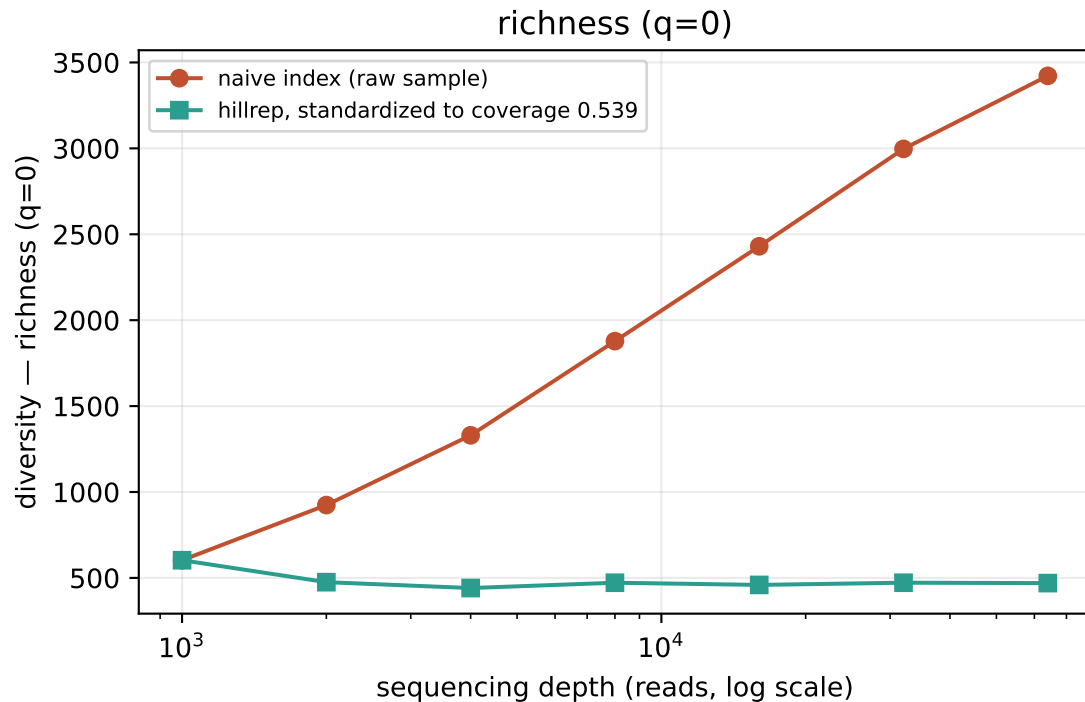


q = 2 (inverse-Simpson)



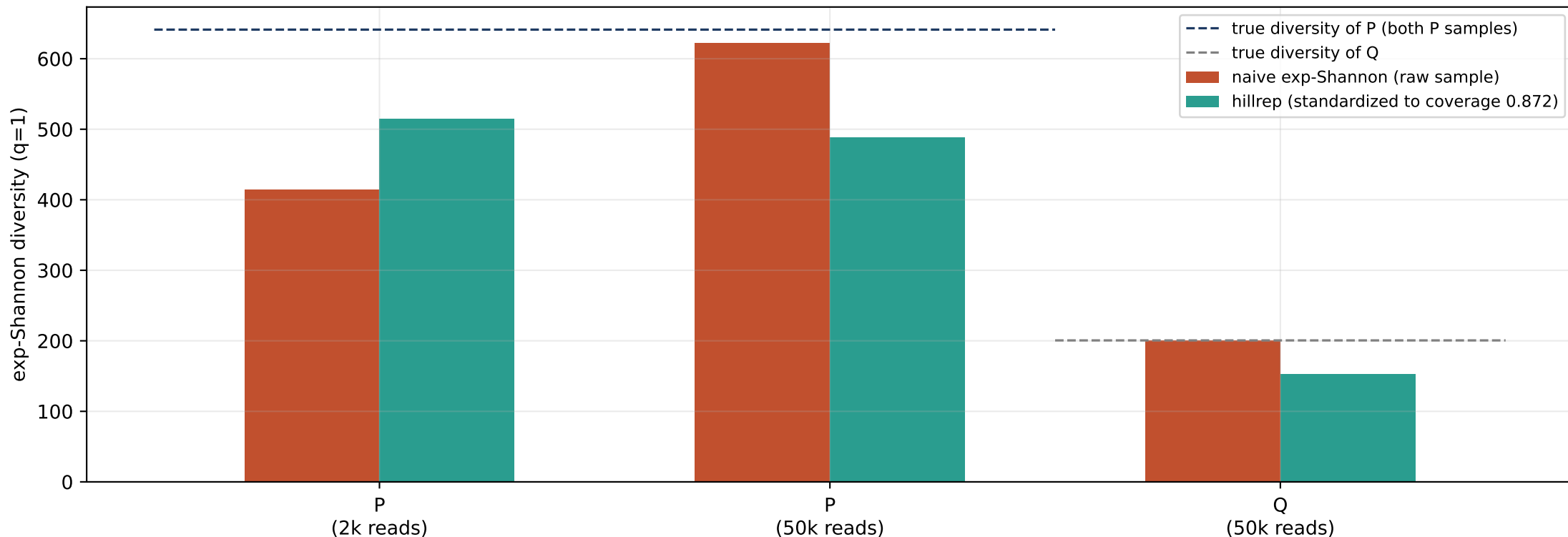
Lines = hillrep continuous curve; dots = iNEXT reference points; dotted vertical line = observed sample size n .
Maximum relative error vs iNEXT across all reference points: $2.78e-11$.

Use case 1 — sequencing-depth bias and its correction



The naive index climbs with depth (more rare clones are seen); the coverage-standardized estimate is stable and comparable across depths. Richness coefficient of variation across depths: naive 51% vs standardized 10%.

Use case 2 — fair comparison across unequal sequencing depths



P sampled at two depths is the SAME repertoire, so its true diversity is identical (dashed line). The naive index makes the deeper P sample look 150% as diverse as the shallow one (a pure depth artifact). hillrep standardizes both to a common coverage, correctly reports them as equal, and keeps both well above the genuinely less-diverse repertoire Q.

Use case 3 — AIRR Rearrangement TSV, end to end

Input: an AIRR Rearrangement table
(1300 rows, 2 repertoires,
clonotype = junction_aa + V-gene + J-gene)

```
from hillrep.airr import read_rearrangement, clonotype_counts
from hillrep import compare

df = read_rearrangement('rearrangements.tsv')
reps = clonotype_counts(df, by='repertoire_id')
out = compare(reps, level='coverage')
```

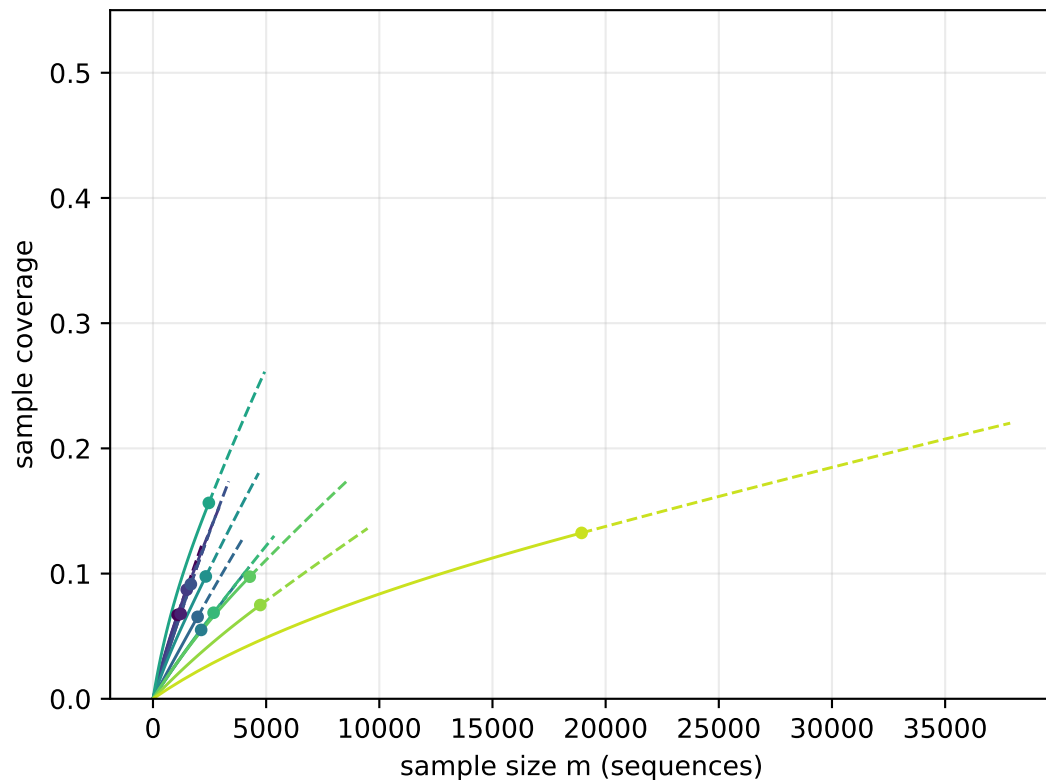
compare() output (standardized to common coverage)

assemblage	order_q	qD	qD_lcl	qD_ucl	coverage
donor1	0.0	710.35	689.1	731.6	0.9749
donor1	1.0	291.26	284.4	298.1	0.9749
donor1	2.0	162.48	157.4	167.6	0.9749
donor2	0.0	643.32	583.3	703.4	0.9749
donor2	1.0	165.72	157.7	173.7	0.9749
donor2	2.0	56.69	52.0	61.4	0.9749

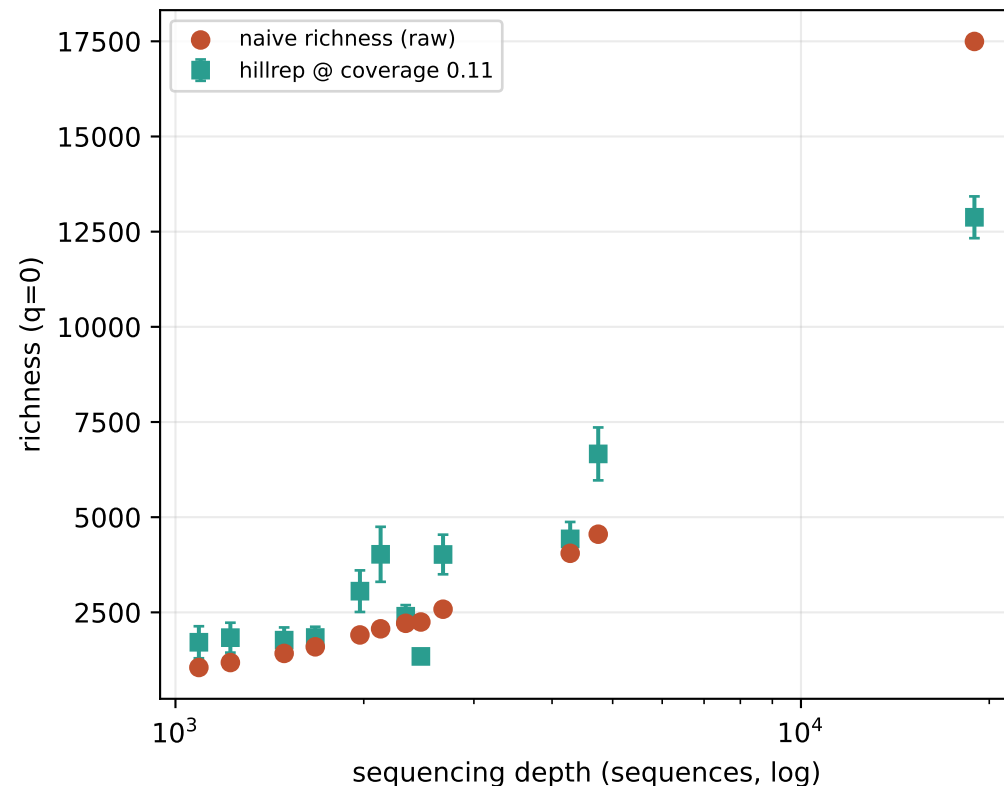
Both donors are standardized to the same coverage before comparison, so the diversity numbers are directly comparable despite the 3x depth difference.

Use case 5 — real public repertoires (AIRR Data Commons)

sample-completeness curves (all 12 are far from saturated)



naive vs coverage-standardized



12 real TCR-beta repertoires from the AIRR Data Commons (study: T cell Receptor Repertoires Acquired via Routine Pap Testing May Help ...), depths 1,090-18,937 sequences. Naive richness tracks depth almost perfectly (you are measuring sequencing effort); standardizing to a common coverage compresses the spread (coefficient of variation 1.2 -> 0.8). hillrep also reports that these repertoires are coverage-limited (6%-16%), flagging that deeper sequencing is needed.

Use case 6 — is the comparison reliable? (assess)

per-sample reliability table (sorted by coverage)

verdict: CAUTION

target coverage (Cmax): 0.105

why:

- 12 of 12 samples: sample coverage < 0.5 (badly undersampled)

```
from hillrep import assess
report = assess(reps)          # reps: {label: counts}
report.verdict                 # 'caution'
report.table                   # per-sample flags

# straight from a scanpy/scirpy AnnData:
from hillrep import from_anndata
reps = from_anndata(adata, groupby='sample')
assess(reps)
```

assemblage	n	coverage	coverage_2n	extrap_factor	flag
59811545	2128	0.055	0.105	2.0	low_coverage
59392786	1972	0.065	0.127	1.64	low_coverage
60186066	1090	0.067	0.124	1.67	low_coverage
59866950	1224	0.068	0.131	1.58	low_coverage
60454502	2678	0.069	0.13	1.59	low_coverage
59863944	4741	0.075	0.136	1.49	low_coverage
59936958	1492	0.087	0.156	1.26	low_coverage
60460515	1673	0.092	0.174	1.16	low_coverage
59585630	4276	0.098	0.174	1.1	low_coverage
59856213	2334	0.098	0.18	1.09	low_coverage
59591213	18937	0.132	0.22	0.72	low_coverage
60189073	2468	0.156	0.261	0.58	low_coverage

assess turns hillrep's coverage knowledge into a verdict. On these real, coverage-limited repertoires it does not just return numbers: it flags that they are badly undersampled and that a fair comparison must be read with caution, the honest message a naive richness index never gives. Same call works on AIRR counts or a single-cell AnnData.

Use case 4 — robustness on extreme and degenerate inputs

input	result	wall time
megaclone n=1.0M, rarefaction q=1 @ n/2	1.011	107.8 ms
f1=100000, f2=1, asymptotic Shannon (q=1)	5e+09	6.3 ms
all singletons (n=2000), asymptotic Simpson (q=2)	inf	0.4 ms
single clonotype (n=7), observed richness	1	0.0 ms

No crash, no overflow, bounded time: inputs that broke a naive implementation (large n, dominant clone, all-singletons) return finite or correctly-infinite values.

Summary & reproduction

Findings

- hillrep reproduces R iNEXT 3.0.2 to a maximum relative error of $2.8e-11$ across all reference points.
- Depth-bias use case: naive richness varies by 51% across sequencing depths, while the coverage-standardized estimate varies by only 10%.
- Fair-comparison use case: the naive index makes the same repertoire sequenced deeper look 150% as diverse as the shallow sample (a depth artifact); hillrep standardizes both to a common coverage and reports them as equal (yes), both above the less-diverse repertoire (yes).
- Real public data: on 12 TCR-beta repertoires from the AIRR Data Commons (depths 1,090-18,937), naive richness varies by coefficient of variation 1.2; standardizing to a common coverage compresses it to 0.8. hillrep also flags that the repertoires are coverage-limited (6%-16%).
- Reliability assessment: `assess()` returns the verdict 'caution' on those real repertoires, flagging 12 of 12 samples as coverage-limited rather than reporting them as if comparable. The same call works on AIRR counts or a single-cell AnnData (`from_anndata`).
- Robustness: large-n, dominant-clone, and all-singleton inputs return finite or correctly-infinite values in bounded time (no crash, no overflow).

Reproduce

```
pip install -e '[dev,plot]'  
python scripts/make_report.py
```

Correctness is additionally checked by the test suite (pytest), which asserts agreement with committed iNEXT golden values to $1e-6$ on every estimator.