Tools & Resources WG

with Christian Busse & Chaim Schramm

- Biological Standards WG: was Sarah Taylor; Christian Busse (acting)
- File Formats WG: Uri Laserson
- Germline WG: Corey Watson and Andrew Collins
- Software WG: Frederick “Erick” Matsen (me)

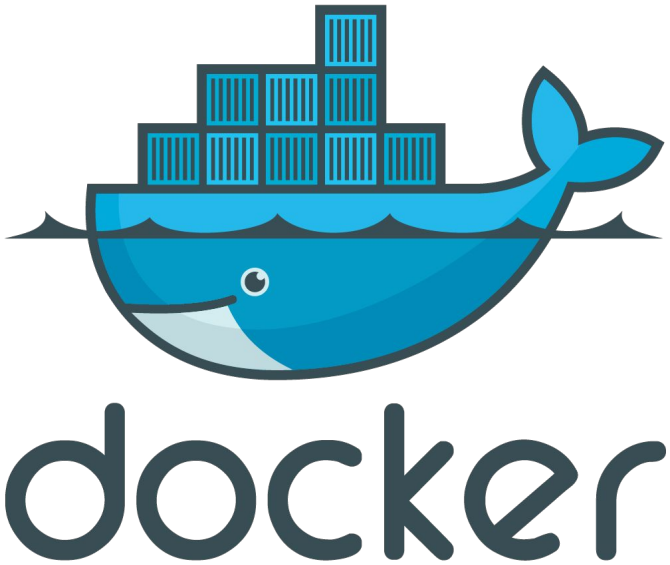
Software WG

Christian Busse, Victor Greiff, Uri Laserson, William Lees, Enkelejda Miho, Branden Olson, Chaim Schramm, Adrian Shepherd, Mikhail Shugay, Inimary Toby, Jason Vander Heiden, Corey Watson, Jian Ye

Frederick “Erick” Matsen (Fred Hutch)

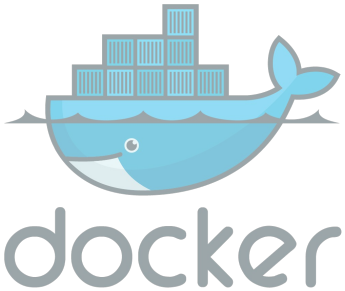
Goal:
make it *easy* to do
rigorous analysis
of AIRR-seq data.

We started thinking about how to make things *easy*



COMMON
WORKFLOW
LANGUAGE

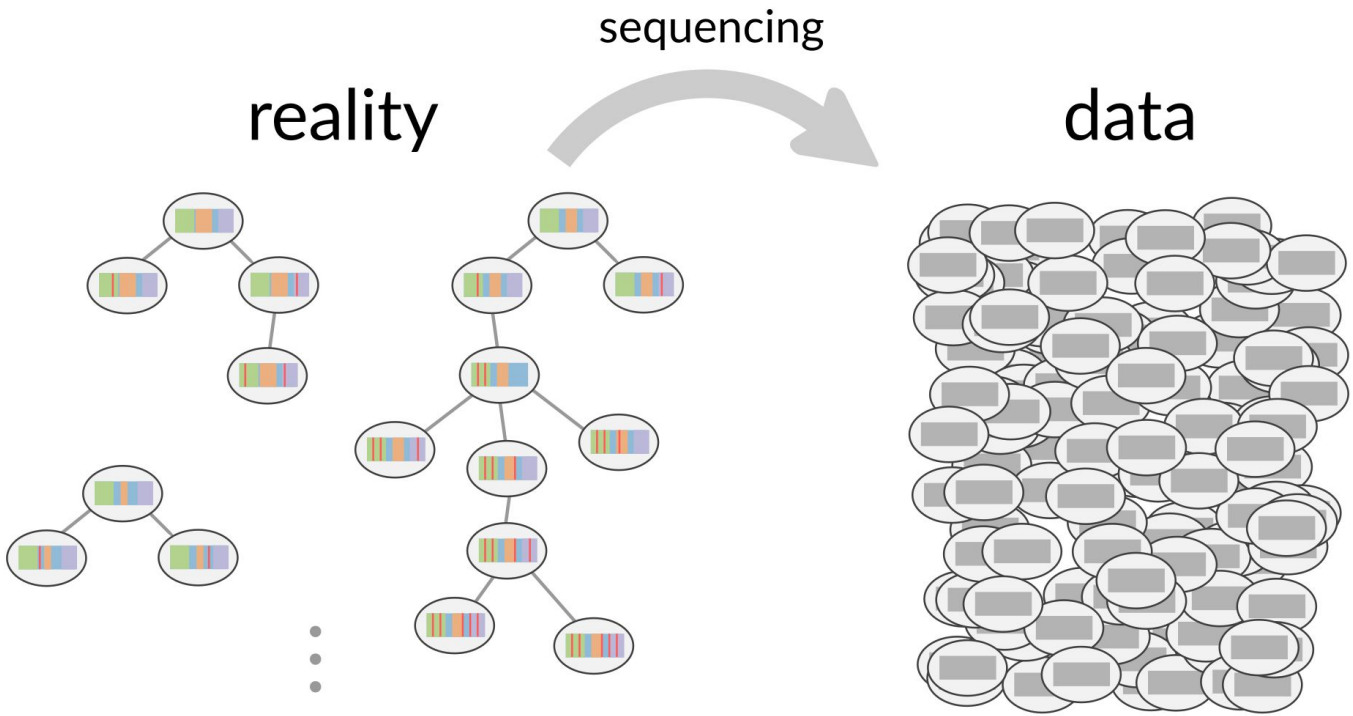
by containerization and standardized ways for tools to interact.

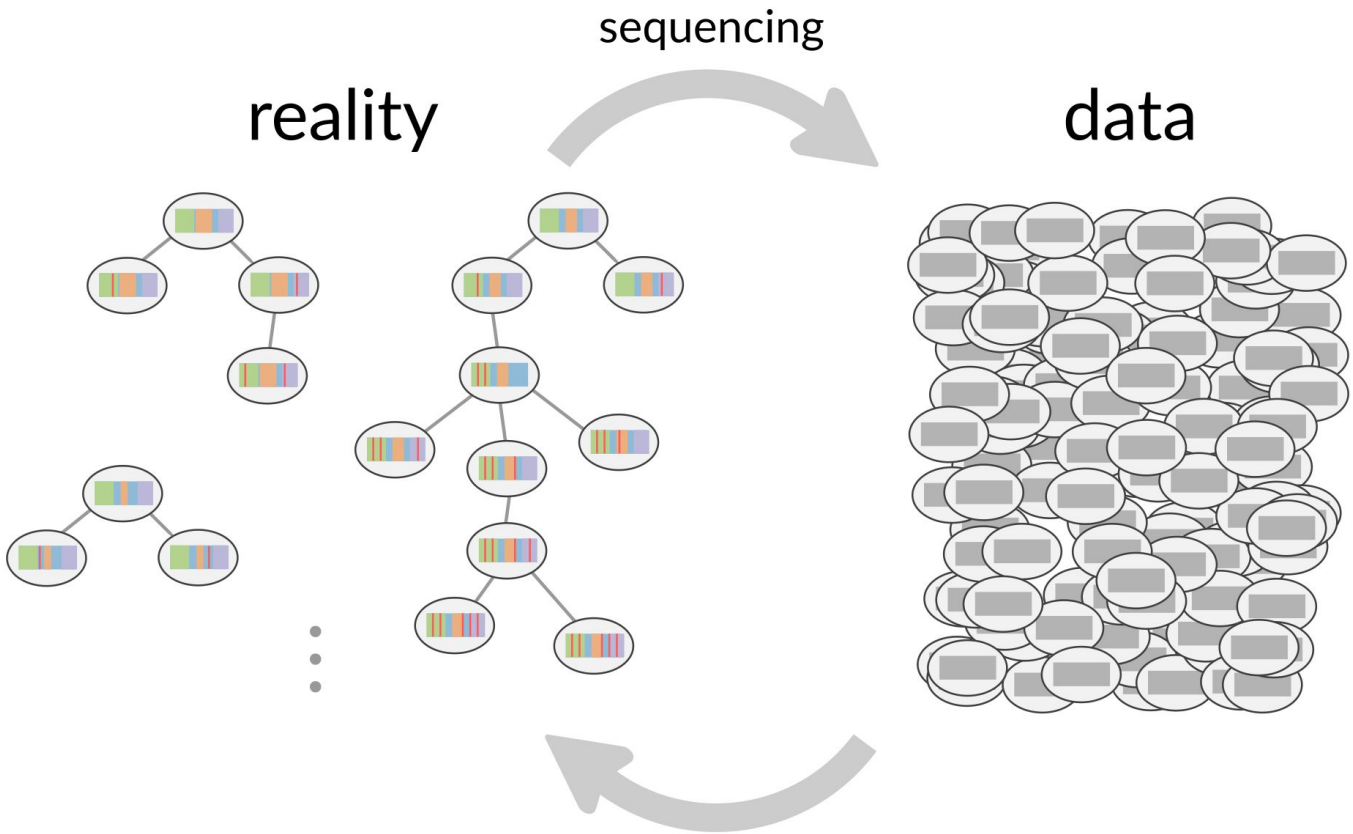


COMMON
WORKFLOW
LANGUAGE

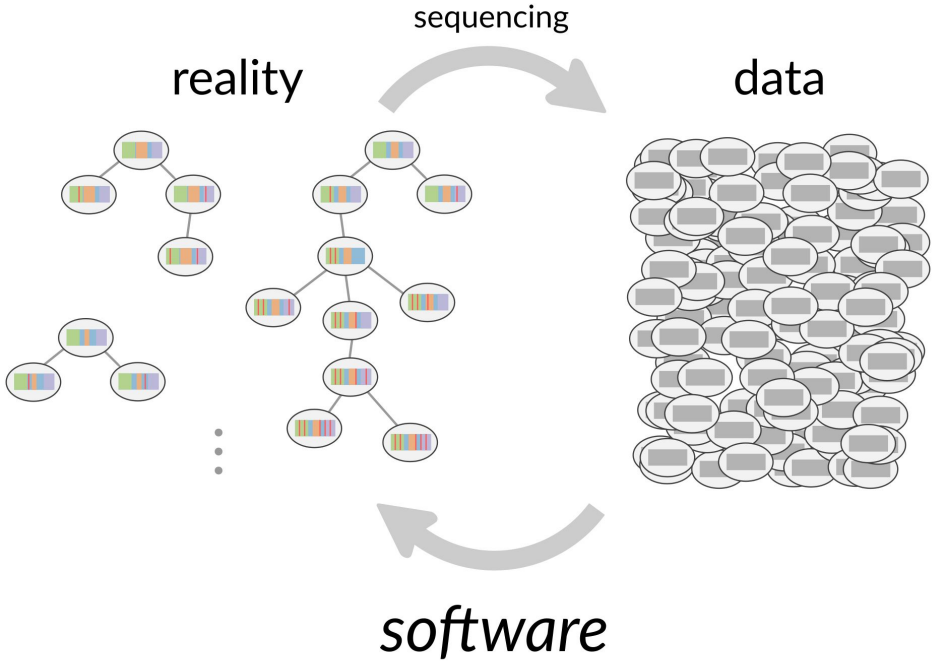
But after a while we decided our most important task was to help make things more *rigorous*.

What does that mean in this context?

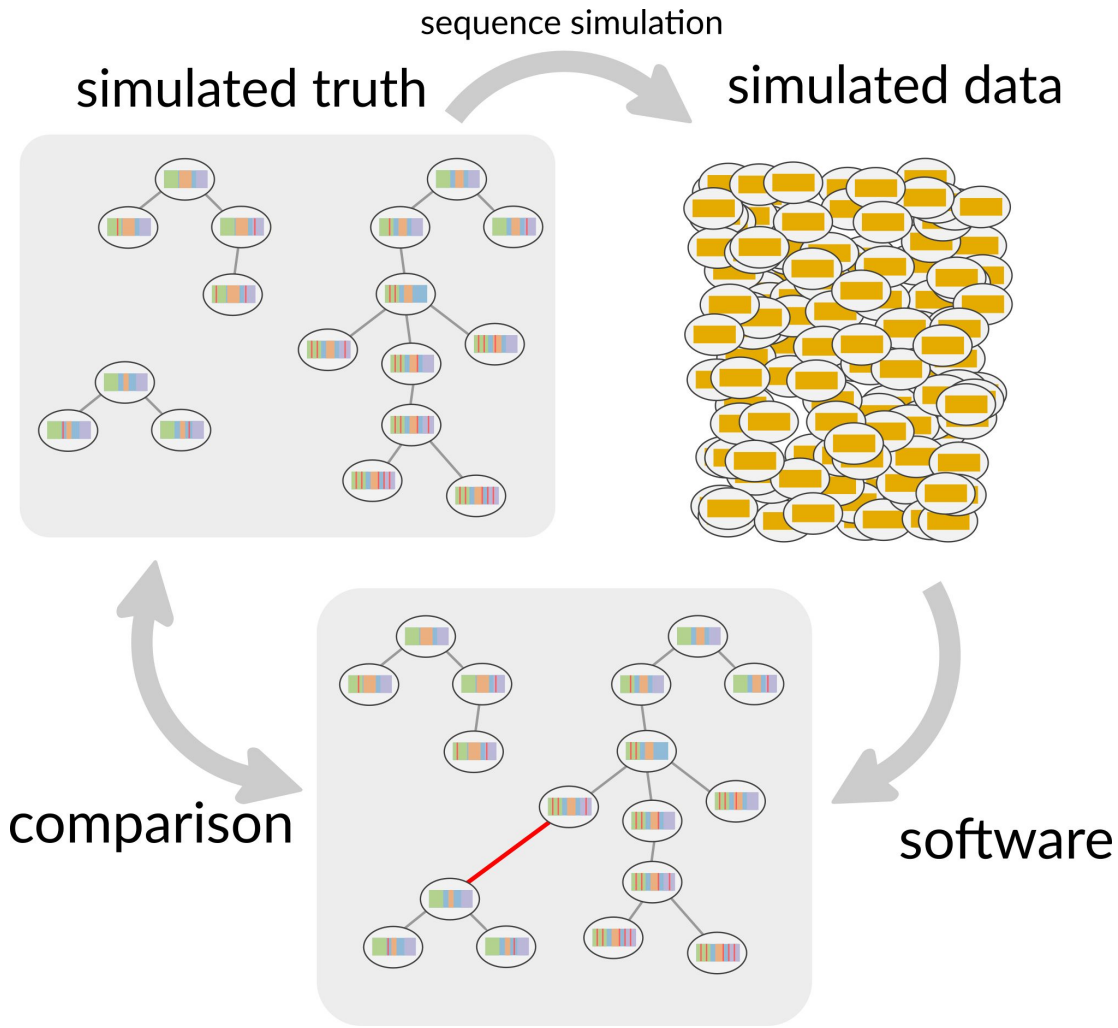

sequencing
reality
data

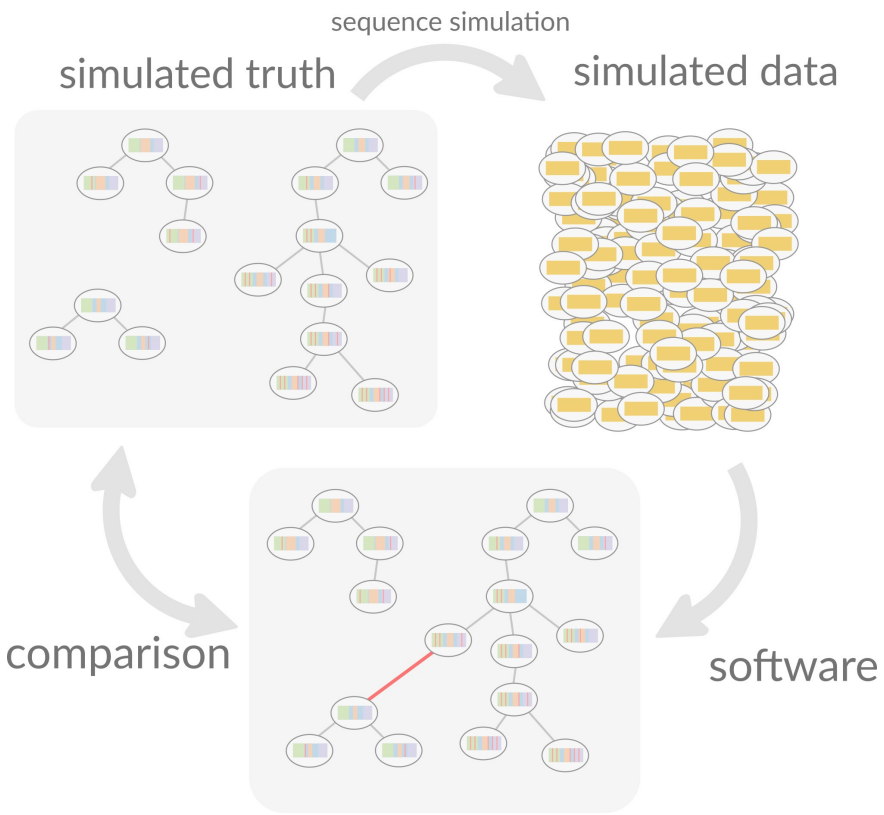


annotation, germline inference, phylogenetics, clonal diversity, networks, machine learning, etc....



Which software tools work well under what conditions?


sequence simulation
simulated truth
simulated data
comparison
software



This only works if simulated data accurately mimics properties of experimental data.

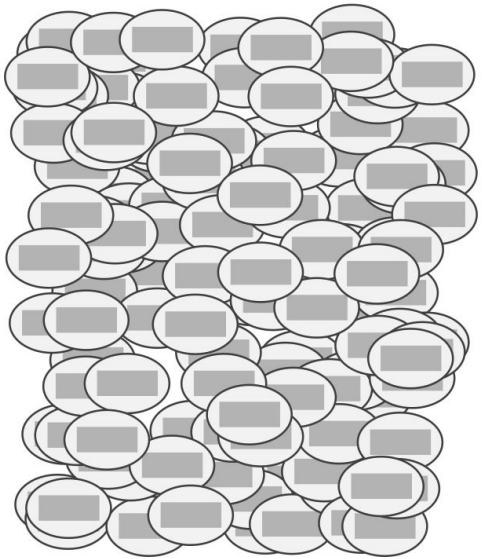
The current goal of the Software WG:

Develop criteria for accurate repertoire sequence simulation, in order to enable rigorous benchmarking studies.

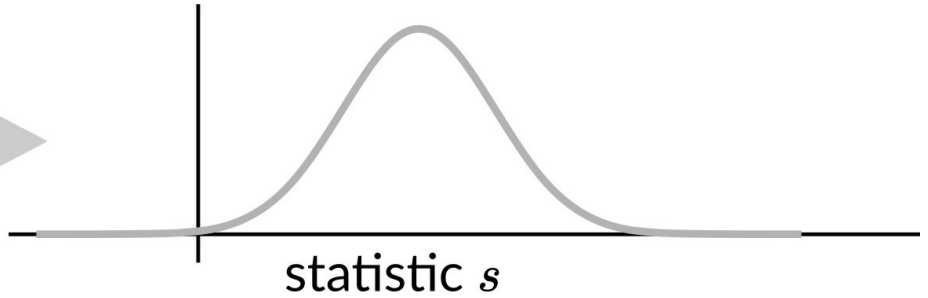
We will do this via “summary statistics.”

Summary statistics quantify some aspect of repertoire data

experimental data

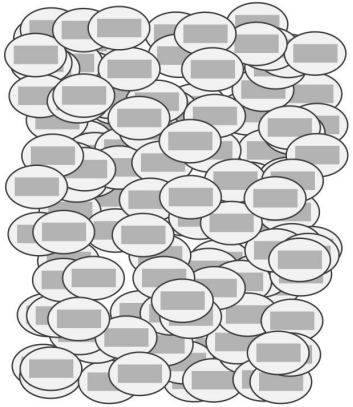


summary statistic

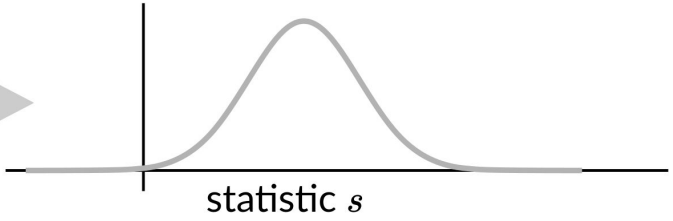


(for example, GC content)

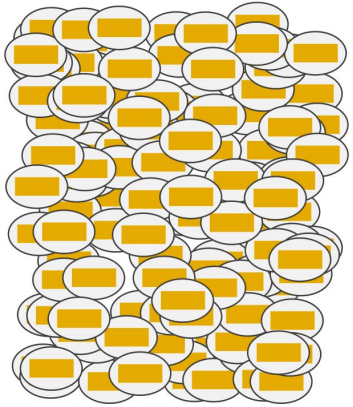
experimental data



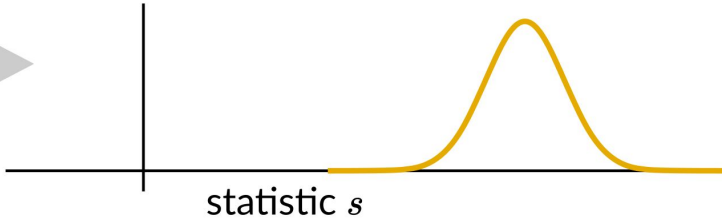
summary statistic



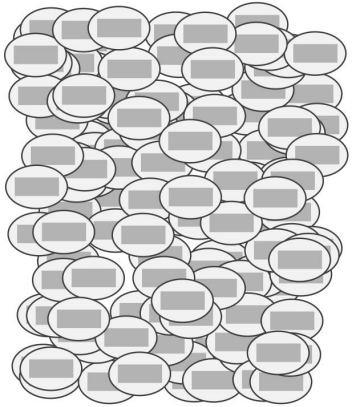
simulated data



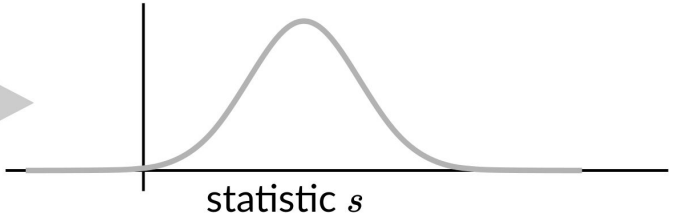
summary statistic



experimental data

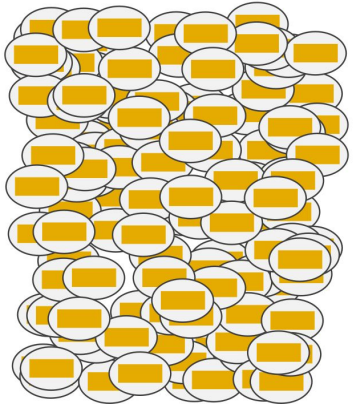


summary statistic

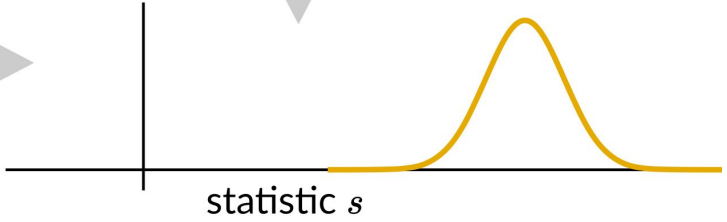


numerical summaries are easy to compare!

simulated data



summary statistic



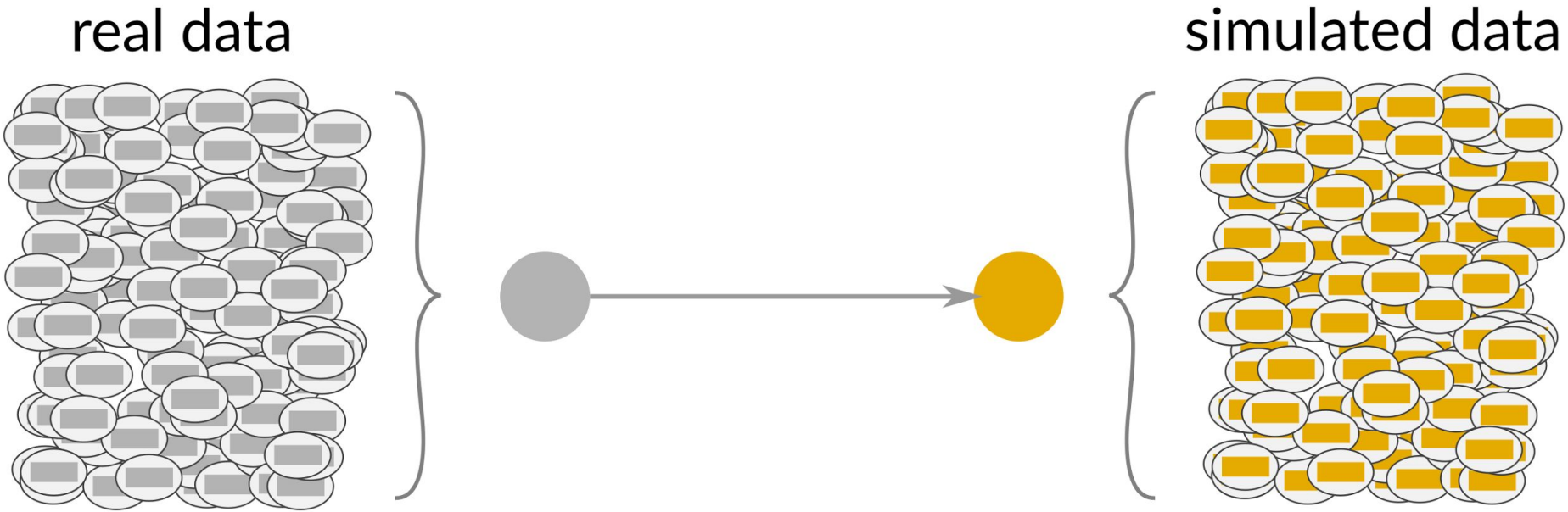
The Software WG selected 31 summary statistics

- Some act on sequences directly, like GC content
- Some require alignment, such as germline gene use
- Some require clone clustering, such as clonal family size distribution
- Some require phylogenetics, such as tree balance

https://goo.gl/oKGxLu ← statistics

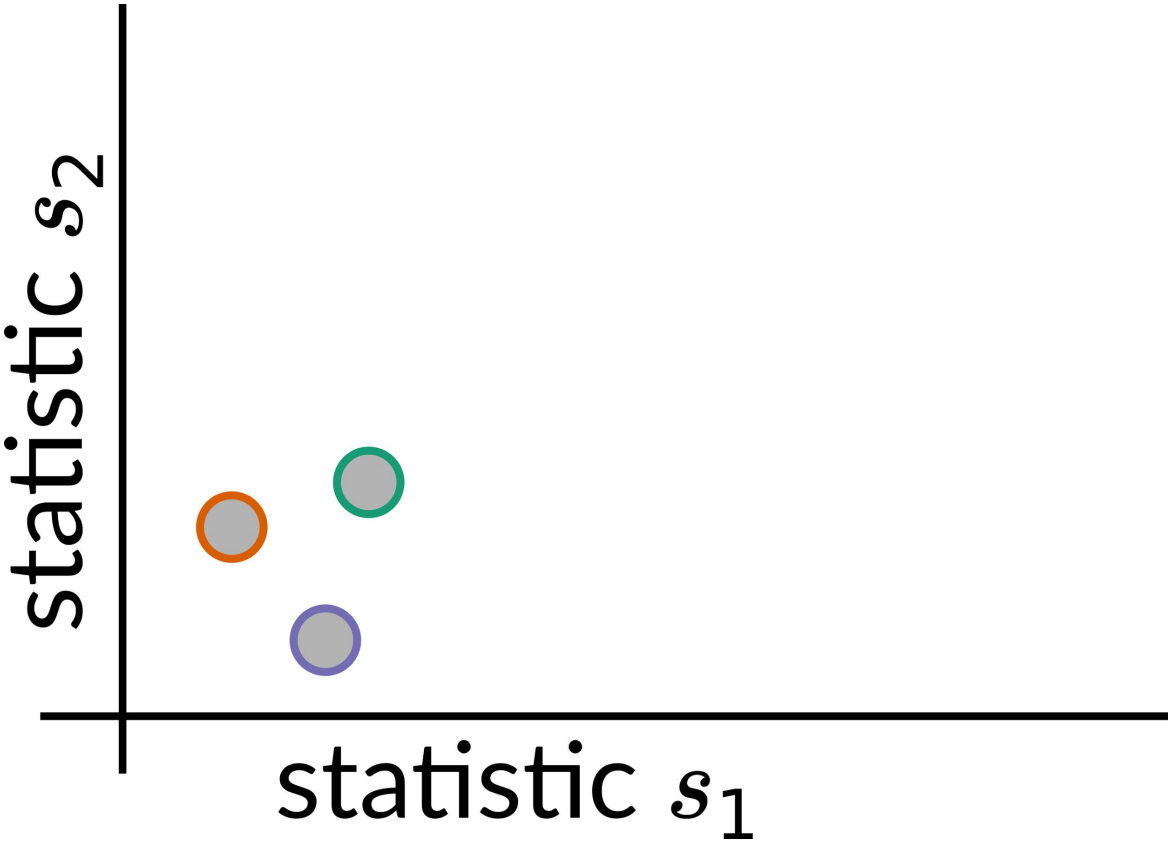
https://github.com/matsengrp/sumrep ← R package

Good simulators fit their simulation to an observed repertoire and then simulate based on that fit.

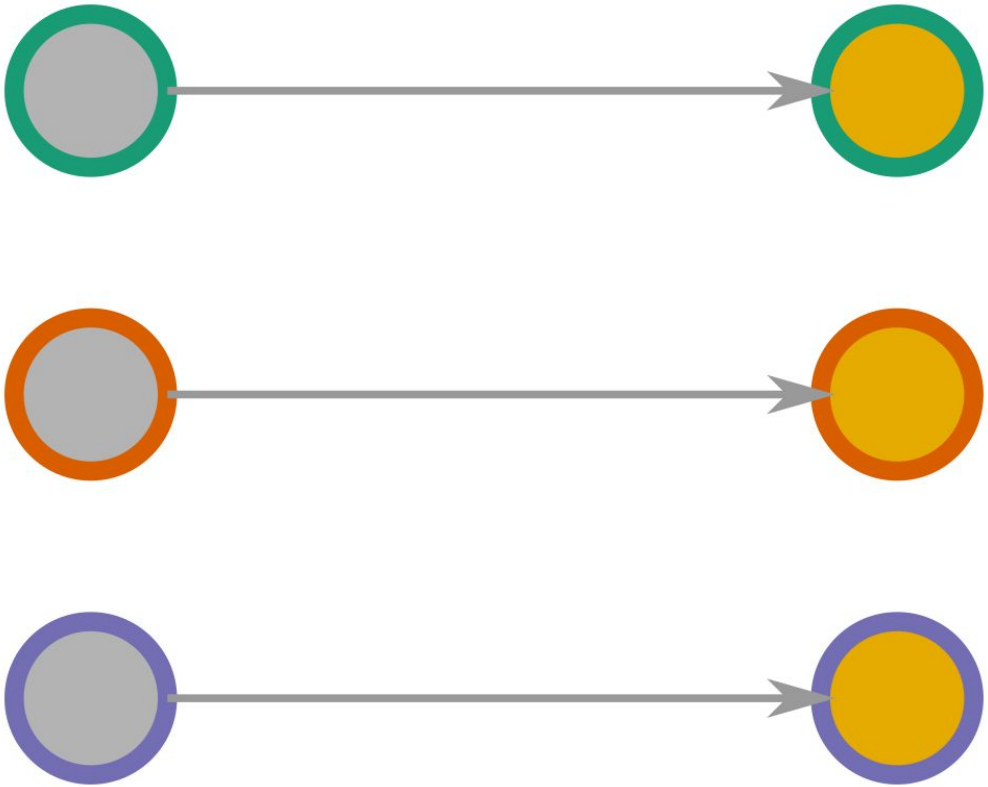


Say we have three data sets

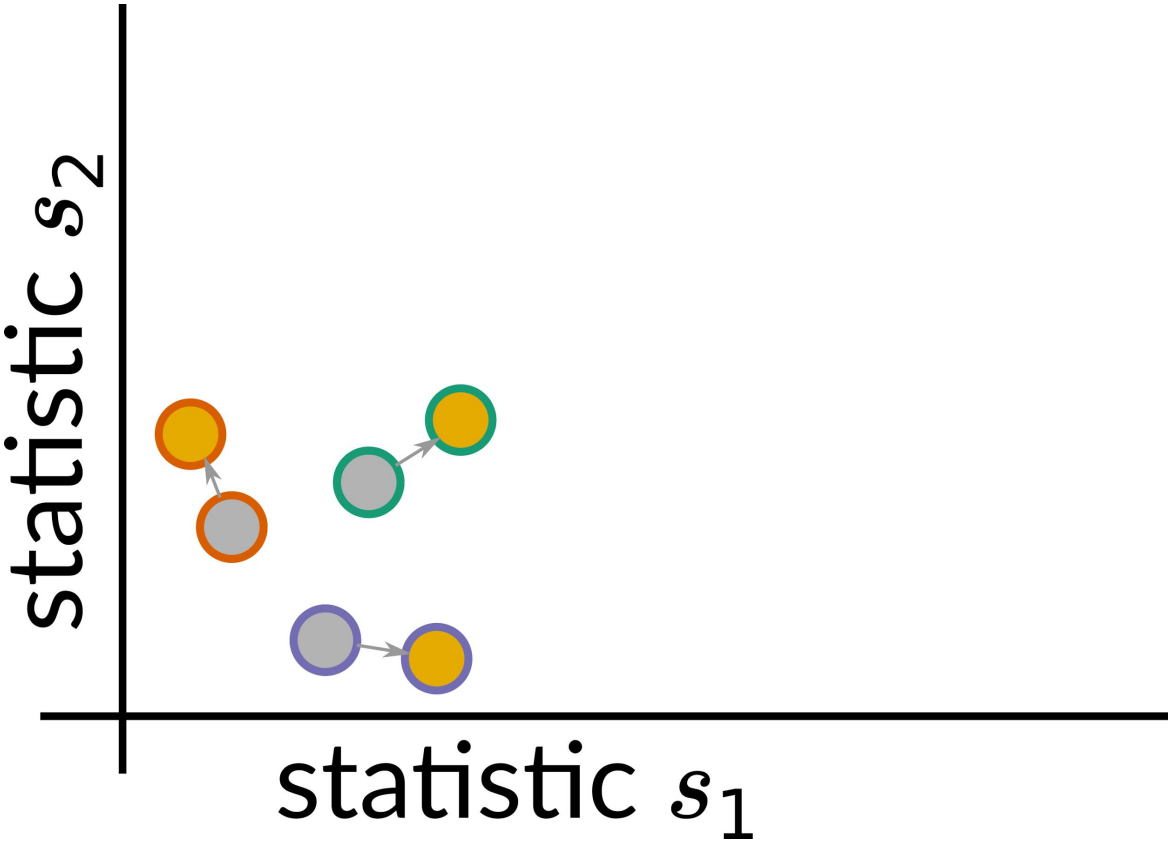
Apply summary statistics to real data



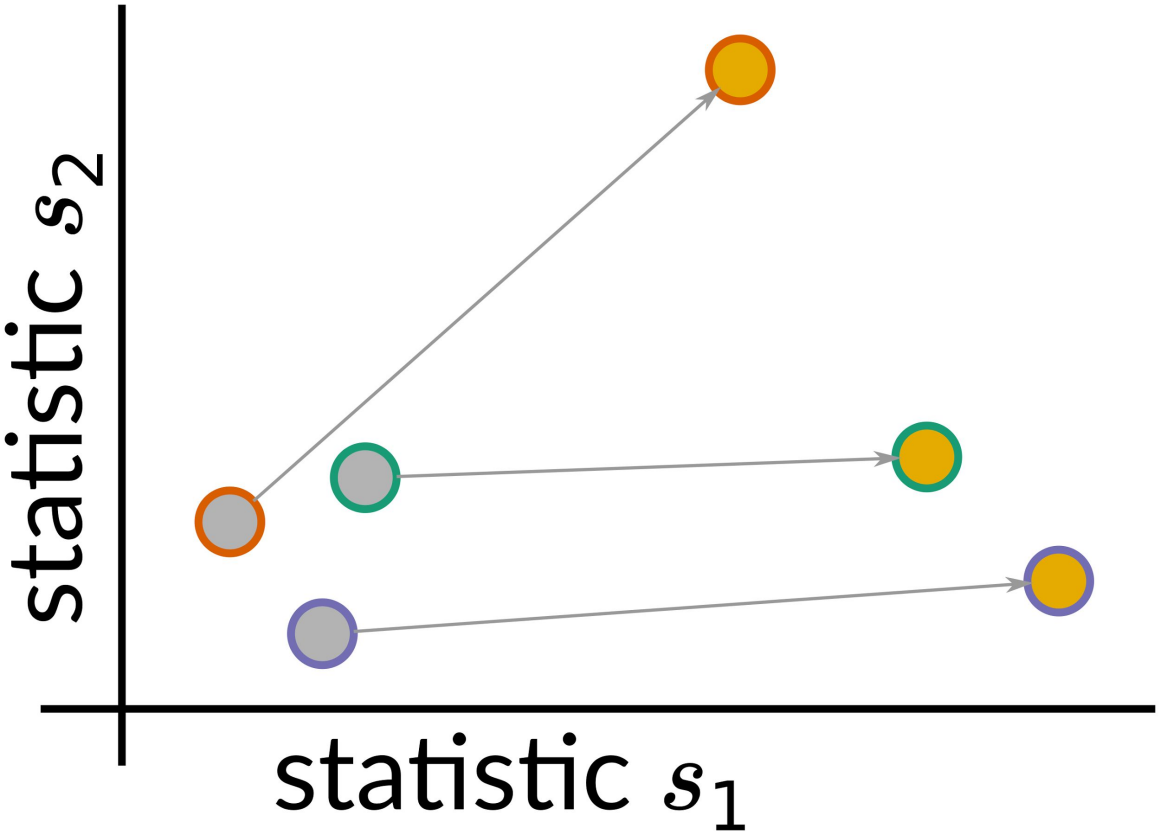
Simulate one data set from each of those three



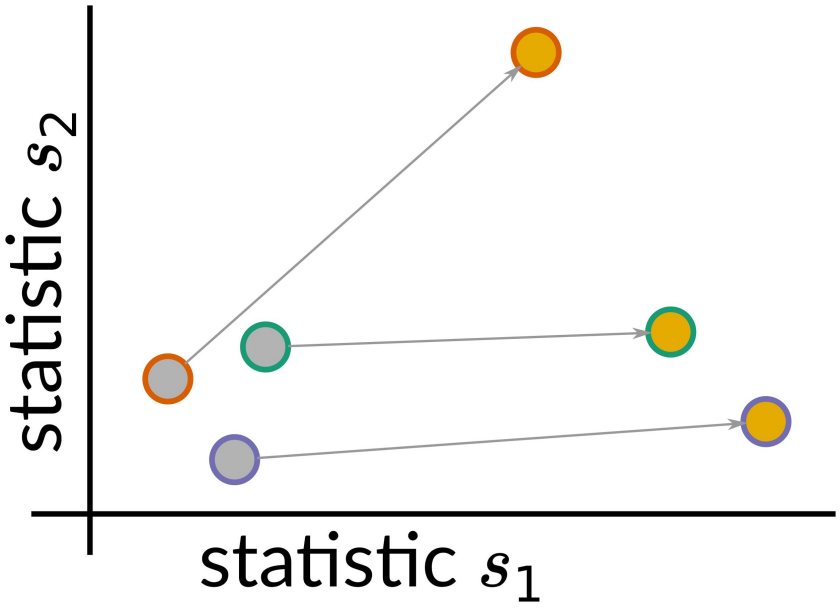
Simulation looking pretty good!



Simulation *not* looking so good.



Branden Olson is building an R package, `sumrep`



`https://github.com/matsengrp/sumrep`

16 summary stats so far.
Uses Immcantation a lot!

Recap:

- Everyone wants software that performs well
- We can use simulation to validate software
- Simulation methods are often insufficiently described and not publicly available, simulated sequences not available
- Summary statistics quantify repertoire characteristics; we can use them to compare to experimental data
- Use these statistics to benchmark simulation tools
- ... and eventually benchmark software confidently!

Simulation needs to become a first-class enterprise

Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees [PDF] oup.com

A Rambaut, NC Grass - Bioinformatics, 1997 - academic.oup.com Paperpile

Abstract Motivation: **Seq-Gen** is a program that will simulate the evolution of nucleotide sequences along a phylogeny, using common models of the substitution process. A range of models of molecular evolution are implemented, including the general reversible model.

☆ Cited by 1349 Related articles All 10 versions

look, citations!

Accurate simulation is a type of *understanding*.

How you can help

- Make beautiful data, use the MiAIRR standard, and make it public! We need sorted T/B cell populations with high-quality PCR/sequencing workflow, high technological/biological sampling depth, probing of different immune states, antigen immunizations, etc.
- Post simulated data to https://zenodo.org/communities/airr
- Use the AIRR format for your software (see next talk)
- Join the group and contribute code!

Reconstructing Antibody Repertoires from Error-Prone Immunosequencing Reads

Alexander Shlemov,*,[1] Sergey Bankevich,*,[1] Andrey Bzikadze,* Maria A. Turchaninova,[†] Yana Safonova,*,[‡] and Pavel A. Pevzner*,[§]

YAY!

Availability

All datasets analyzed in this paper are publicly available under the following accession numbers or DOIs. Simulated: doi.org/10.5281/zenodo.823351 (https://zenodo.org/record/823351#.WYN34dOGPBI); simulated TCR: doi.org/10.5281/zenodo.823347 (https://zenodo.org/record/823347#.WYN3-NOGPBI); simulated barcoded: doi.org/10.5281/zenodo.826956 (https://zenodo.org/record/826956#.WYN4B9OGPBI); synthetic: SRR4431793 (https://www.ncbi.nlm.nih.gov/sra/SRX2251687); real: SRR5851422 (https://

Goals for 2018

- Evaluate simulators: which reproduce features of real data sets?
- Evaluate summary statistics: which are robust to noise? Which are “orthogonal” to each other?
- Write paper with whole Software WG (!)

`Describe the point at which your WG will have achieved its goals and can be dissolved`

Software WG work will be done when

- we have standards for software evaluation
- we have done such evaluation
- tools can talk to each other and fit easily into pipelines
- we have continuously running evaluations

(... I’m not necessarily going to lead all of this.)

THANK YOU Software WG

Christian Busse, Victor Greiff, Uri Laserson, William Lees, Enkelejda Miho, Branden Olson, Chaim Schramm, Adrian Shepherd, Mikhail Shugay, Inimary Toby, Jason Vander Heiden, Corey Watson, Jian Ye

The following slides are not part of the regular presentation, but are proposed arguments in response to questions.

Objection #1:

Your summary statistics will never be able to capture the complexity of repertoire data.

1. Unless you stare at your sequences one by one, you use summary statistics to analyze your data already.
2. If there is some aspect of complexity missing, we can simply quantify and add it. (This is scientific development.)

Objection #2:

Your simulations will never be able to recapitulate the complexity of repertoire data.

1. Simulation is strictly easier than inference, because we don't have to search over models or parameters. If we can do the latter, we can do the former.
2. Have we actually tried? Are the correct motivations in place? Right now there are zero benchmarks. Is that better?
3. Better simulators mean more robust validations, even if we can't get everything right.

Objection #3:

Simulators will overfit the summary statistics.

1. If we require that simulators are able to generate an arbitrarily large amount of data that fits observed summary statistics, this will ensure that there is an underlying probabilistic model.
2. We can always add more summary statistics and then re-evaluate!

Objection #4:

Inference tools will overfit your simulations.

1. If the simulations are very realistic, that means the tools are working very well!
2. There are many types of repertoires, and so tools will have to be good at many types of simulations.

Objection #5:

There are many different types of repertoires.
So your notion of good/bad is an oversimplification.

1. Yes, yes, yes, yes! That's why we need simulations that can can be fit to repertoires and then simulate from them.
2. And yes, some tools may work better in some regimes than others. We need to simulate in a variety of parameter regimes, which we may classify into "types" if that's helpful.

Objection #6:

Why not use real data sets rather than simulated ones?

1. This is an excellent idea for certain types of analyses (e.g. H/L data for phylogenetics), but is different than that which we are going after here.
2. No real data set exists for which all of the hidden aspects of receptor sequences are revealed.

Objection #7:

Why not use simplified data sets for specific tests even if they are unrealistic?

1. That's a great approach for certain settings, and we aren't excluding that approach. However, we are going after something broadly applicable here.
2. Newer methods are using entire-repertoire properties (e.g. germline allele set & their usage probabilities) to do even per-sequence tasks such as annotations. Therefore, the whole repertoire properties need to be realistic.

Objection #8:

You should be focusing more on raw data processing.

1. Definitely. As a first step we are starting from “preprocessed” data as a way to simplify the task.
2. Sequencing technology moves very quickly!