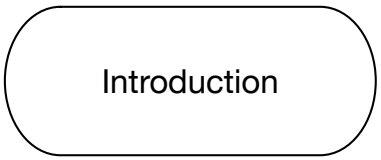
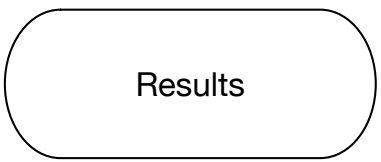



Biological goals

MinHash is awesome, but…

New requirements, introduce…

subsampling
DBG?




Some theory + error/missingness analysis

Gather description and demonstration

metrics?

Basic evaluation (podar: reads x known, reads x gen bank)

Comparison with kaiju/etc performance.

metrics? include time/space of DB, updates...


It basically works (we hope)

This is practical and convenient on a laptop level

Tradeoff with hashset size is not so bad

↳ unknown hashes?
containment analysis more useful for metagenomes?

Issues:

- we don't show that mash/MinHash approach is insensitive here. (e.g. for metagenomes)
- practicality has proven to be very important. Should we emphasize? (DB construction/updating, signature format, scriptability, etc.)
- more SIST? (containment only)