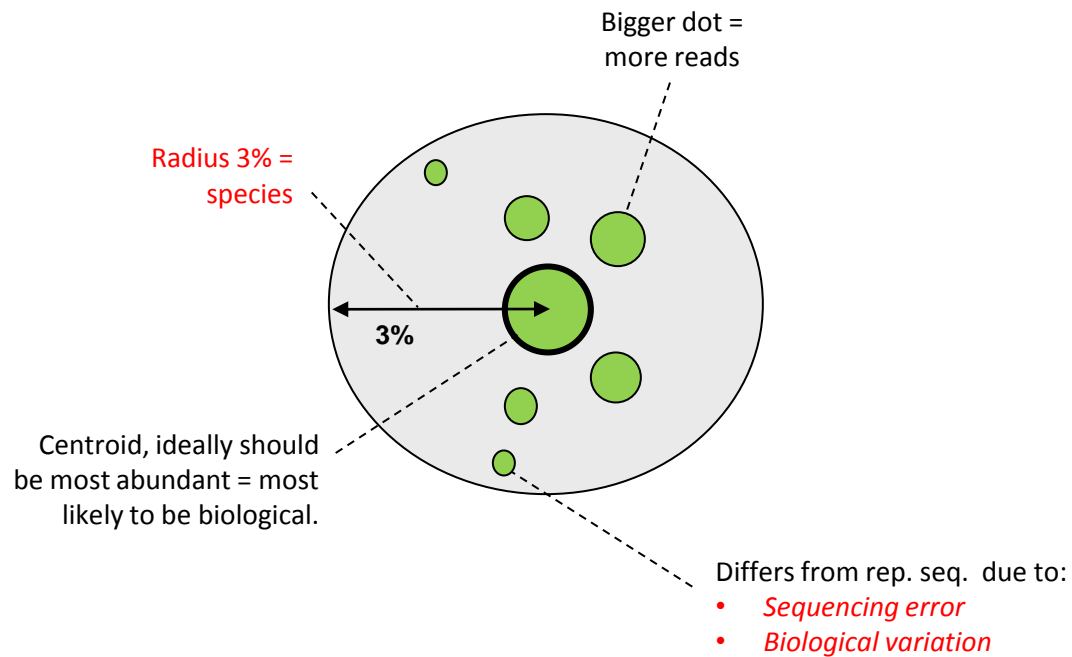


CHALLENGES IN OTU CLUSTERING

Robert Edgar

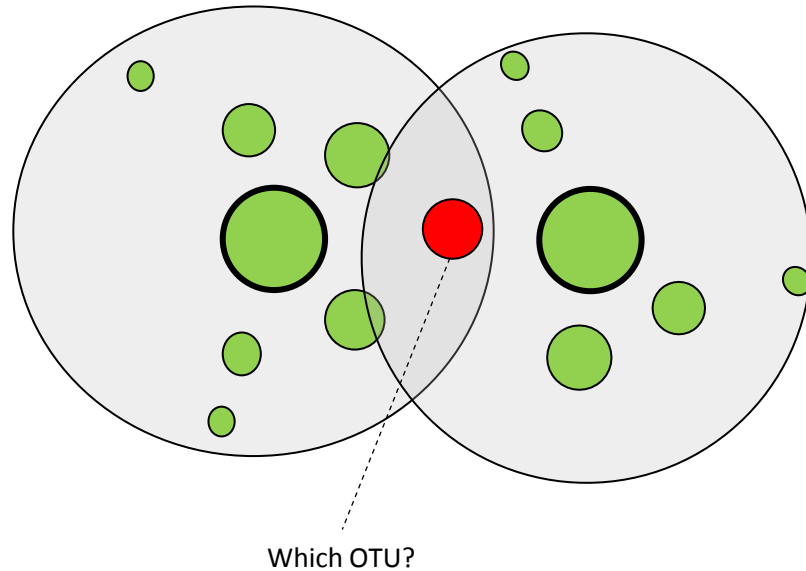
Independent scientist
robert@drive5.com
www.drive5.com

OTUs: Huse "dot plots"



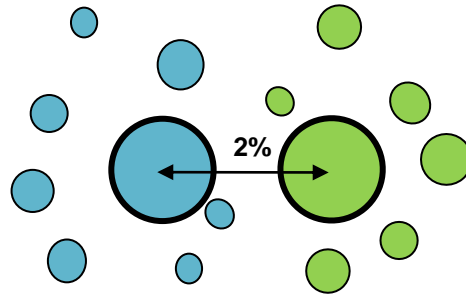
Challenges in OTU clustering

Ambiguous assignments



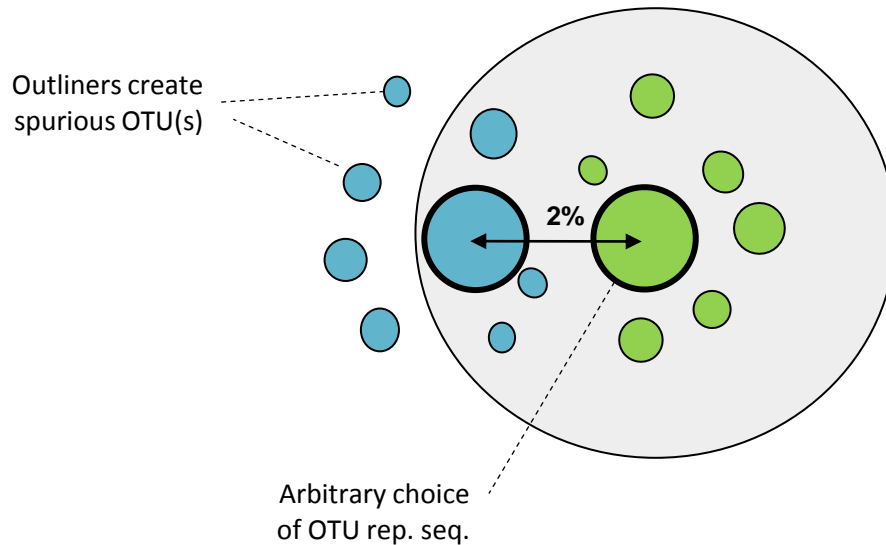
Challenges in OTU clustering

Abundant sequences <3% different



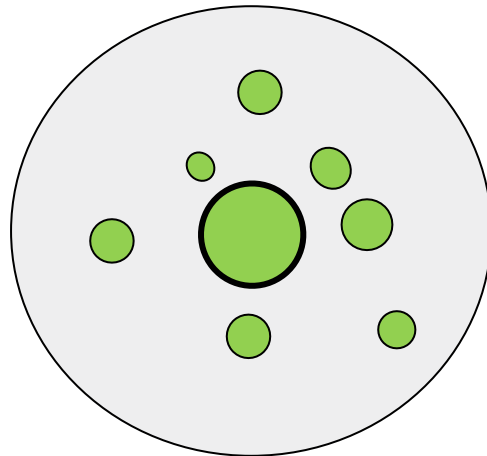
Challenges in OTU clustering

Abundant sequences <3% different



Challenges in OTU clustering

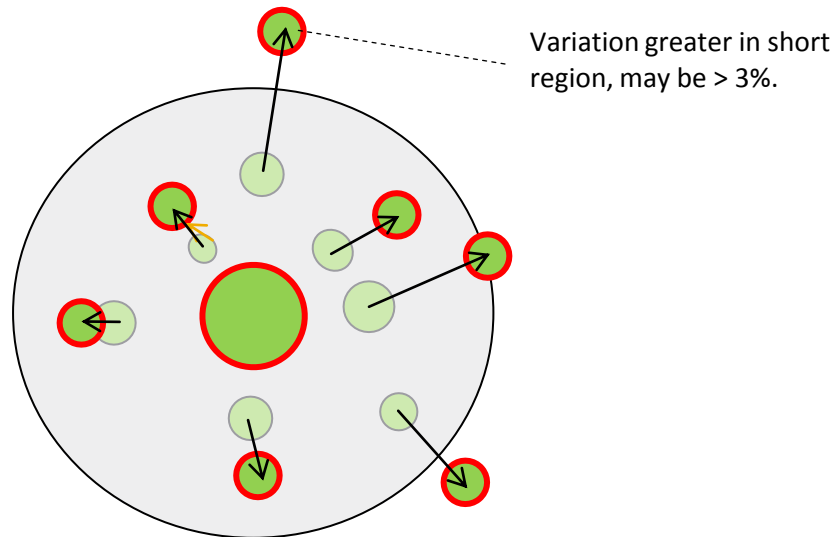
Full-length 16S gene (~1500nt)



Challenges in OTU clustering

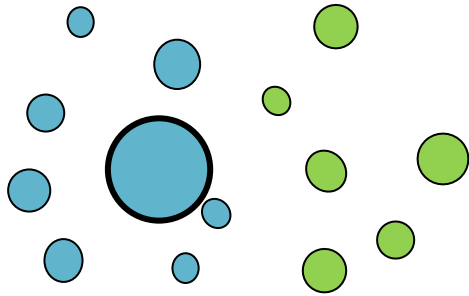
Full-length 16S gene (~1500nt)

Next-gen reads of hypervariable region (~300nt)

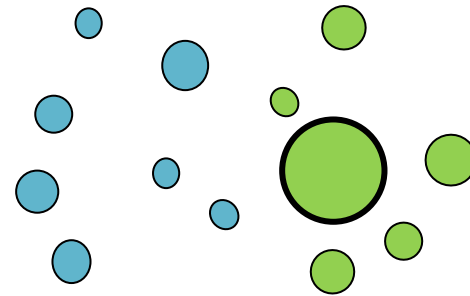


Challenges in OTU clustering

Variation between populations



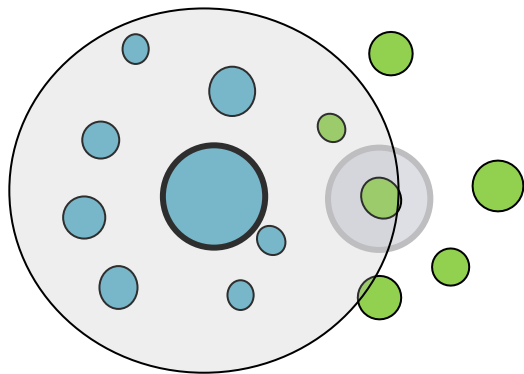
Healthy



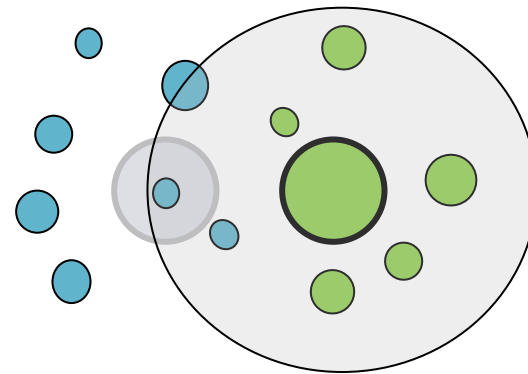
Diseased

Challenges in OTU clustering

Variation between populations



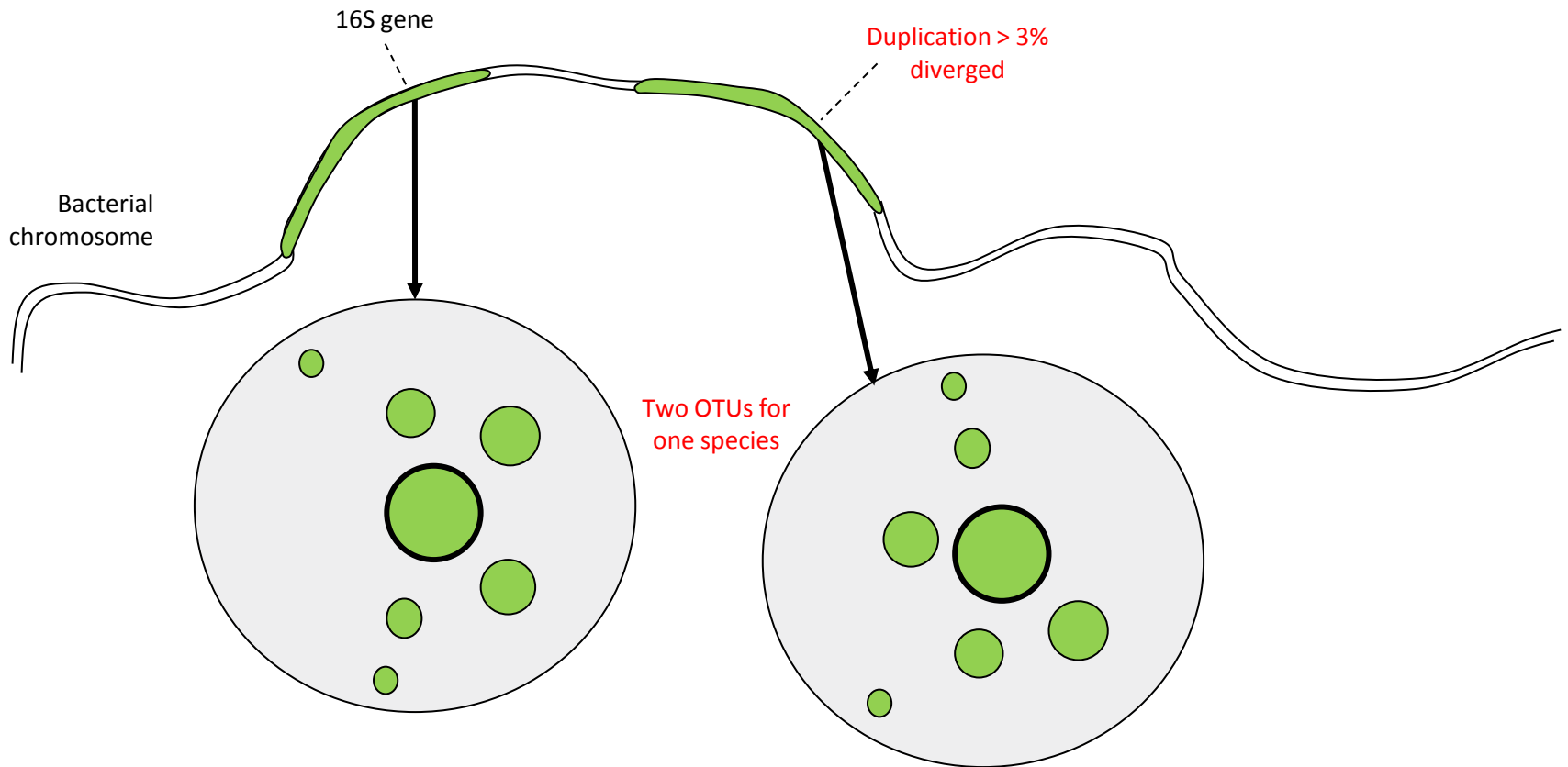
Healthy



Diseased

Challenges in OTU clustering

Paralogs and segmental duplications



Challenges in OTU clustering

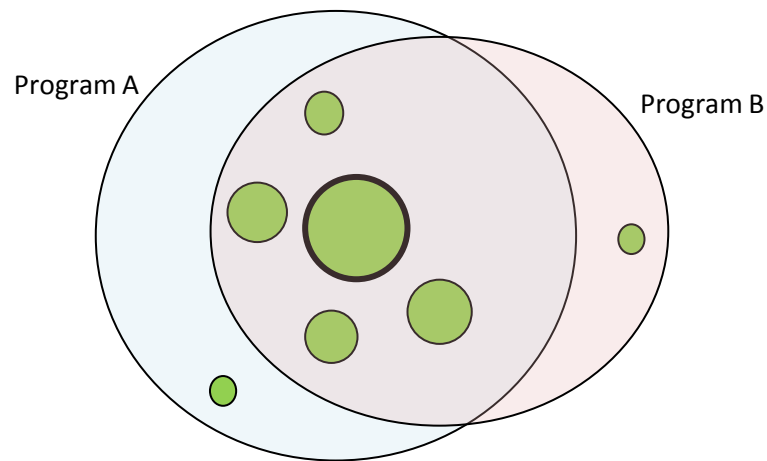
Alignment variation and defining % identity

G A T T A C A - -
G A A T T A A C A

3 diffs or 5 diffs?

G A - T T A - C A
G A A T T A A C A

No diffs or 2 diffs?

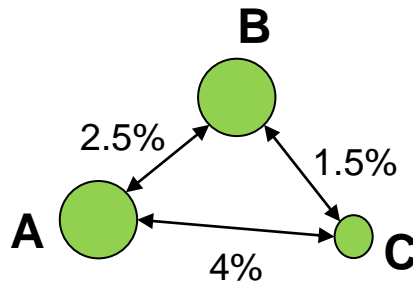
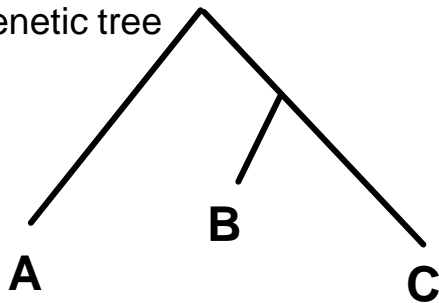


Different programs produce different results from the same algorithm & same input data because alignments and %id definition vary. This can bias validation, e.g. *Schloss & Westcott (2011) AEM*.

Challenges in OTU clustering

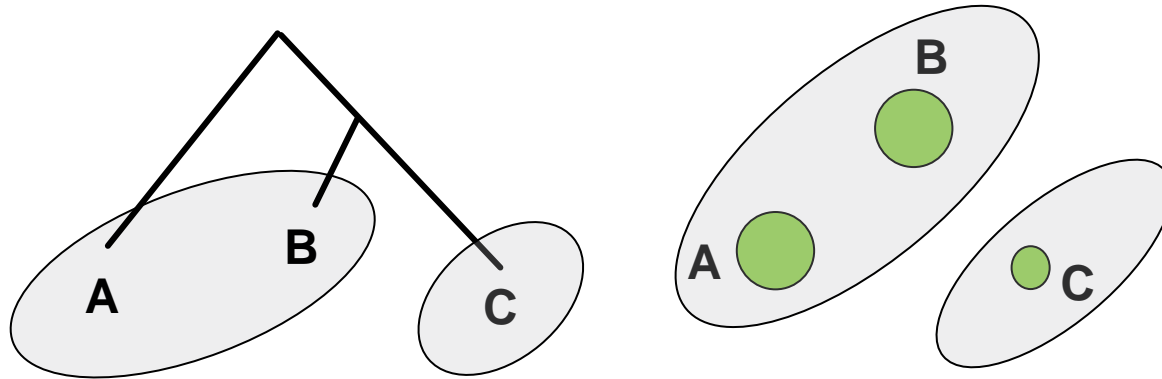
Hard to define an OTU or an optimal set of OTUs

Phylogenetic tree



Challenges in OTU clustering

Hard to define an OTU or an optimal set of OTUs



Optimal OTUs per Schloss & Westcott's
MCC measure can be non-monophyletic.

Challenges in OTU clustering

- OTUs are hacks
- Do not exist in nature
- Cannot be defined and validated robustly
- But can still be useful!