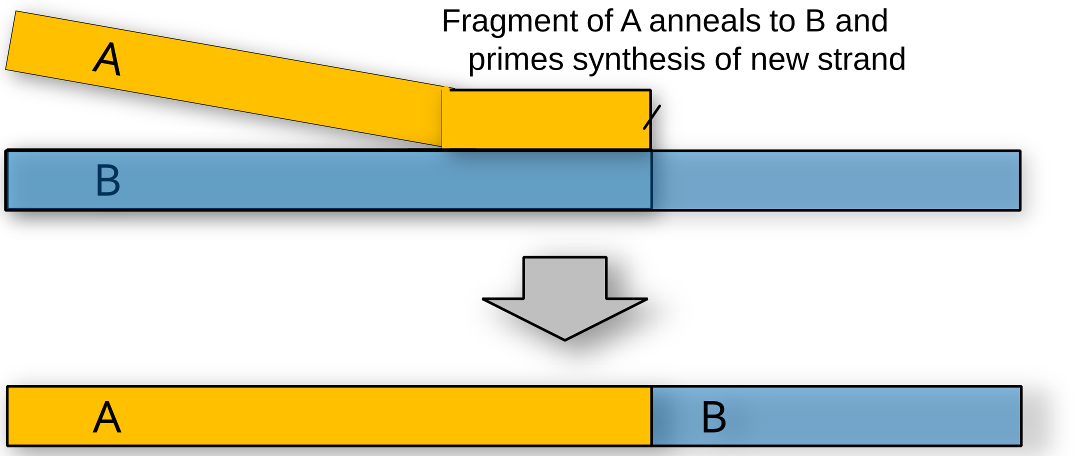
CHIMERAS
STAMPS 2016

Robert Edgar

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

Chimeras

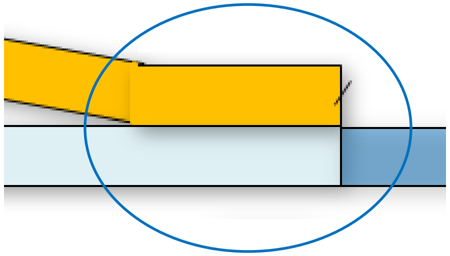
- Created during PCR
- Fragment primes different extension


Fragment of A anneals to B and primes synthesis of new strand

Chimeric A+B template, amplified in following rounds of PCR

Chimeras

- Annealing requires complementary bases
- Cross-over at conserved, homologous locus
- Chimeras align well to known sequences
- Hard to distinguish from biological variants

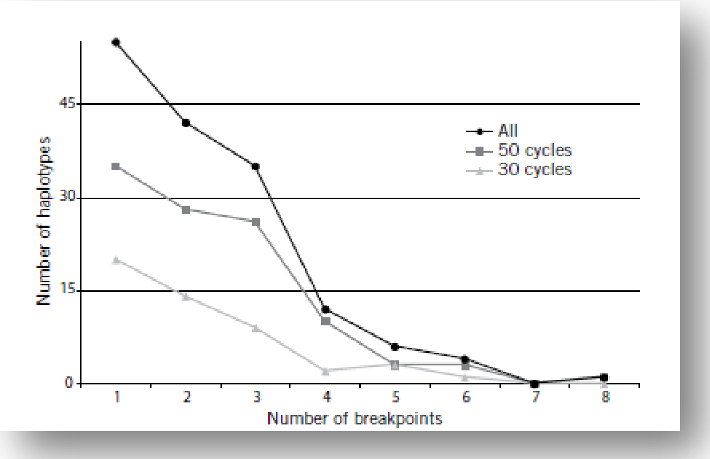


Chimeras in practice

- Frequency depends on PCR conditions
  - choice of polymerase, template concentration
  - also on community structure (less so)
- Typical frequencies
  - 5% of reads
  - 50% of OTUs -- even if high diversity (e.g. soil)
- Lower freq. possible but unusual
  - "Extreme" mock community (DADA2 paper)

Most chimeras are bi

- Bimera=2 segs, trimera=3…
  - >2 form when parent is chimeric
- Lahr & Katz (2009) found many 3+ in 700bp amplicons
- Very rare in V4 (250bp)
  - >2 almost always singleton reads
  - which should be discarded before clustering anyway



Lahr & Katz (2009) doi 10.2144/000113219

Detection algorithms

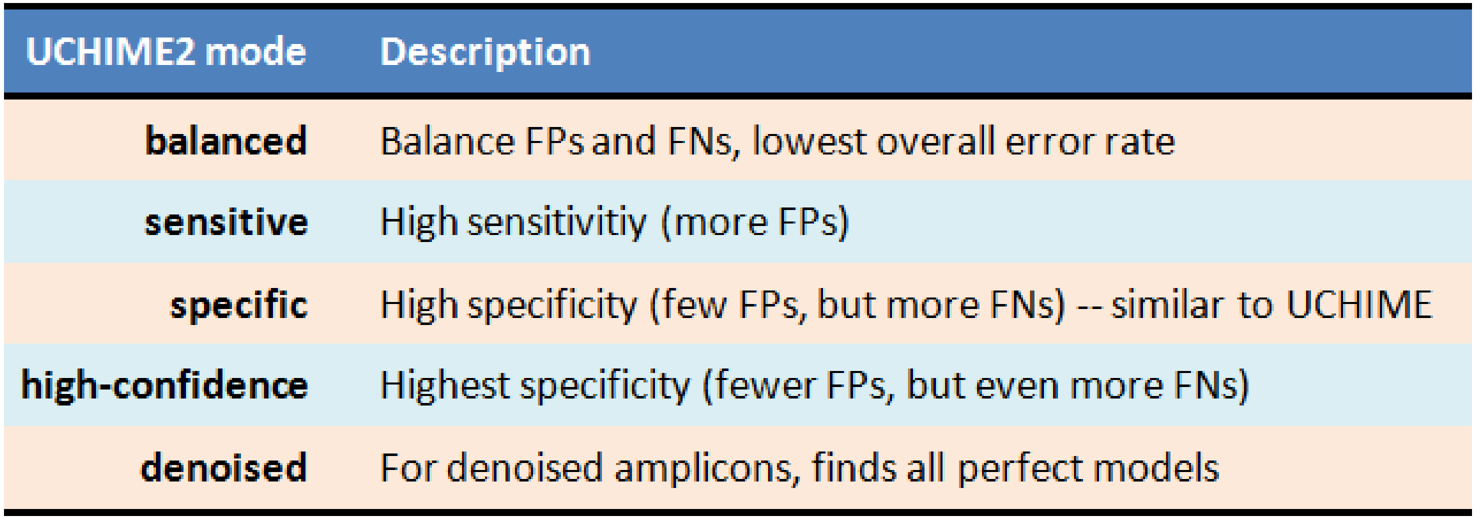
- "Reference"
  - Reference database provided by the user
  - Ideally should be free of chimeras
    - can be a circular problem…
- "*De-novo*"
  - Database constructed from sequences in the reads

Chimera detection algorithms

Algorithm	Paper	Ref/dn	Method	Comments
Bellerophon	Huber *et al.* 2004	Ref	"Partial treeing"	Low sensitivity, obsolete
Pintail	Ashelford *et al.* 2005	Ref	Divergence from ref seq over sliding window	Low sensitivity, obsolete
ChimeraSlayer	Haas *et al.* 2011	Ref	Make 2-seg "model"	Re-implemented in mothur (much faster)
AmpliconNoise	Quince *et al.* 2011	De-novo	Make 2-seg "model"	454 only
UCHIME	Edgar *et al.* 2011	Ref & de-novo	Make 2-seg "model"	Better accuracy than ChimeraSlayer
DECIPHER	Wright *et al.* 2011	Ref	*k*-mer freq. in subtrees	Very low sensitivity
UPARSE	Edgar 2013	De-novo	Max parsimony	Better than UCHIME for OTU clustering
UCHIME2	Edgar 2016 (preprint)	Ref & de-novo	Make 2-seg "model"	Improved accuracy over UCHIME

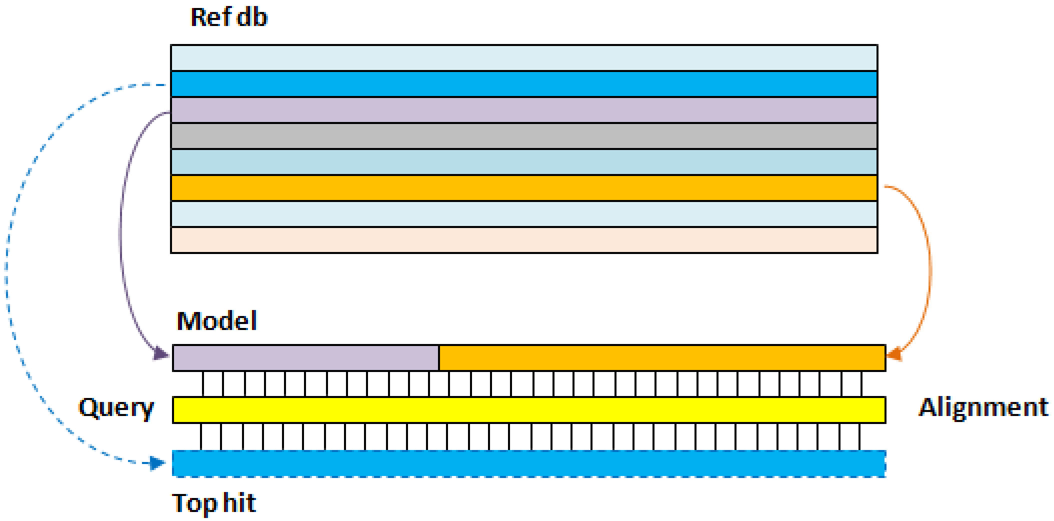
UCHIME2

- Update of UCHIME
  - uses top hit as a control
  - new modes = heuristics + parameter settings

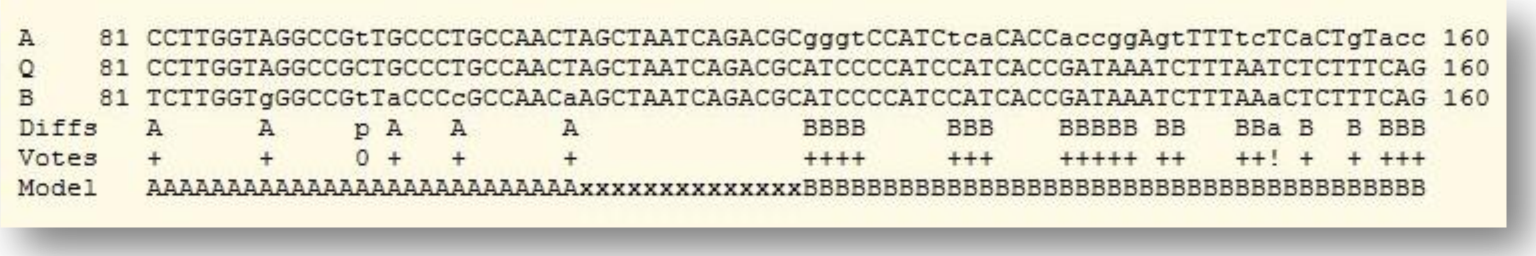


UCHIME2 mode	Description
balanced	Balance FPs and FNs, lowest overall error rate
sensitive	High sensitivitiy (more FPs)
specific	High specificity (few FPs, but more FNs) -- similar to UCHIME
high-confidence	Highest specificity (fewer FPs, but even more FNs)
denoised	For denoised amplicons, finds all perfect models

UCHIME2 algorithm



Query predicted to be chimera
if alignment score > threshold

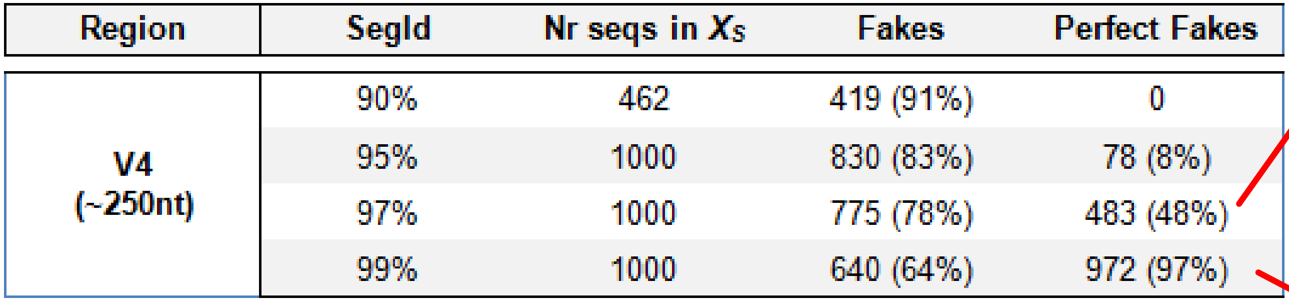


```
A    81 CCTTGGTAGGCCGtTGCCCTGCCAACTAGCTAATCAGACGCgggtCCATCtcaCACCaccggAgtTTTtcTCaCTgTacc 160
Q    81 CCTTGGTAGGCCGCTGCCCTGCCAACTAGCTAATCAGACGCATCCCCATCCATCACCGATAAATCTTTAATCTCTTTCAG 160
B    81 TCTTGGTgGGCCGtTaCCCcGCCAACaAGCTAATCAGACGCATCCCCATCCATCACCGATAAATCTTTAAaCTCTTTCAG 160
Diffs   A      A     p A   A      A              BBBB     BBB    BBBBB BB   BBa B  B BBB
Votes   +      +     0 +   +      +              ++++     +++    +++++ ++   ++! +  + +++
Model   AAAAAAAAAAAAAAAAAAAAAAAAAAAxxxxxxxxxxxxxxBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBB
```

Perfect and fake models

- **Perfect** model if identical to query
  - query may or may not be chimeric
- **Fake** model if query not chimeric & score > 0
  - model is better match than top hit
- **Perfect fake** if not chimeric & exact match
- Fake and perfect fake models very common
- Error-free prediction impossible in principle!

Fake models



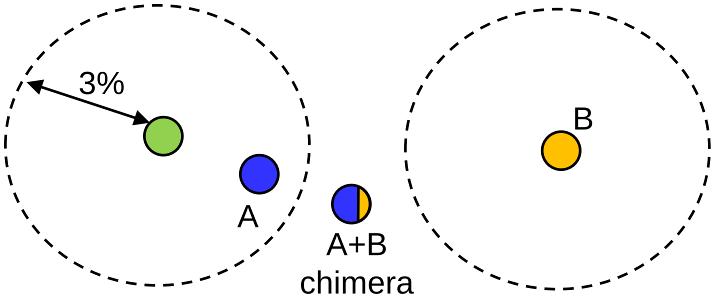
Region	SegId	Nr seqs in X_S	Fakes	Perfect Fakes
V4 (~250nt)	90%	462	419 (91%)	0
	95%	1000	830 (83%)	78 (8%)
	97%	1000	775 (78%)	483 (48%)
	99%	1000	640 (64%)	972 (97%)

If query is *not* chimeric and is 97% identical to ref. db., 48% probability of a *perfect* fake.

At 99% id, almost always a perfect fake, so better coverage makes problem *worse*!

Goals for chimera filtering

- How to compromise FPs and FNs?
- OTU pipeline, 97% clusters
- Chimera >3% diverged harmful
  - always causes spurious OTU
- Chimeras <3% diverged can also be harmful
  - sometimes cause spurious OTU



Goals for chimera filtering

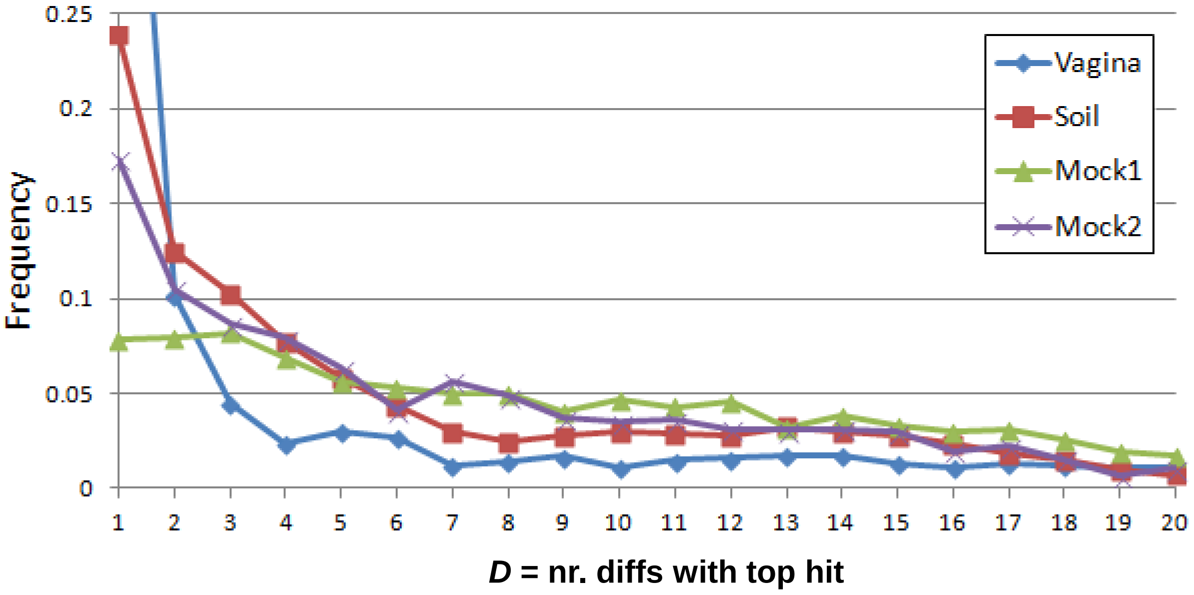
- False positives: discard good OTUs
- False negatives: cause spurious OTUs
- FPs and FNs equally harmful
  - Not typical for bioinformatics!
- Sensitivity of 90% sounds good, but…
  - 90% sensitivity = 10% FNs
  - hundreds or thousands of spurious chimeric OTUs

Chimera divergence

- Parent divergence
  - *PD* = 100% – (parent identity)
  - If similar parents, harder to detect
  - Chimera can very similar to one parent even if large *PD*
- Top-hit divergence
  - *D* = nr. diffs between chimera & top hit
  - Better indicates hard to detect (small *D*)
  - *De novo*: top hit usually a parent

Chimera divergence

- Low-divergence chimeras most common
  - hardest to detect
- Majority have D < 10, most common is D=1



Measuring accuracy

- ChimeraSlayer & UCHIME benchmark
- Sensitivity to simulated bimeras
    - parents <u>always</u> in reference database
    - not realistic! coverage is sparse in practice
- Error rate
    - false positives on leave-one-out test
    - not realistic!

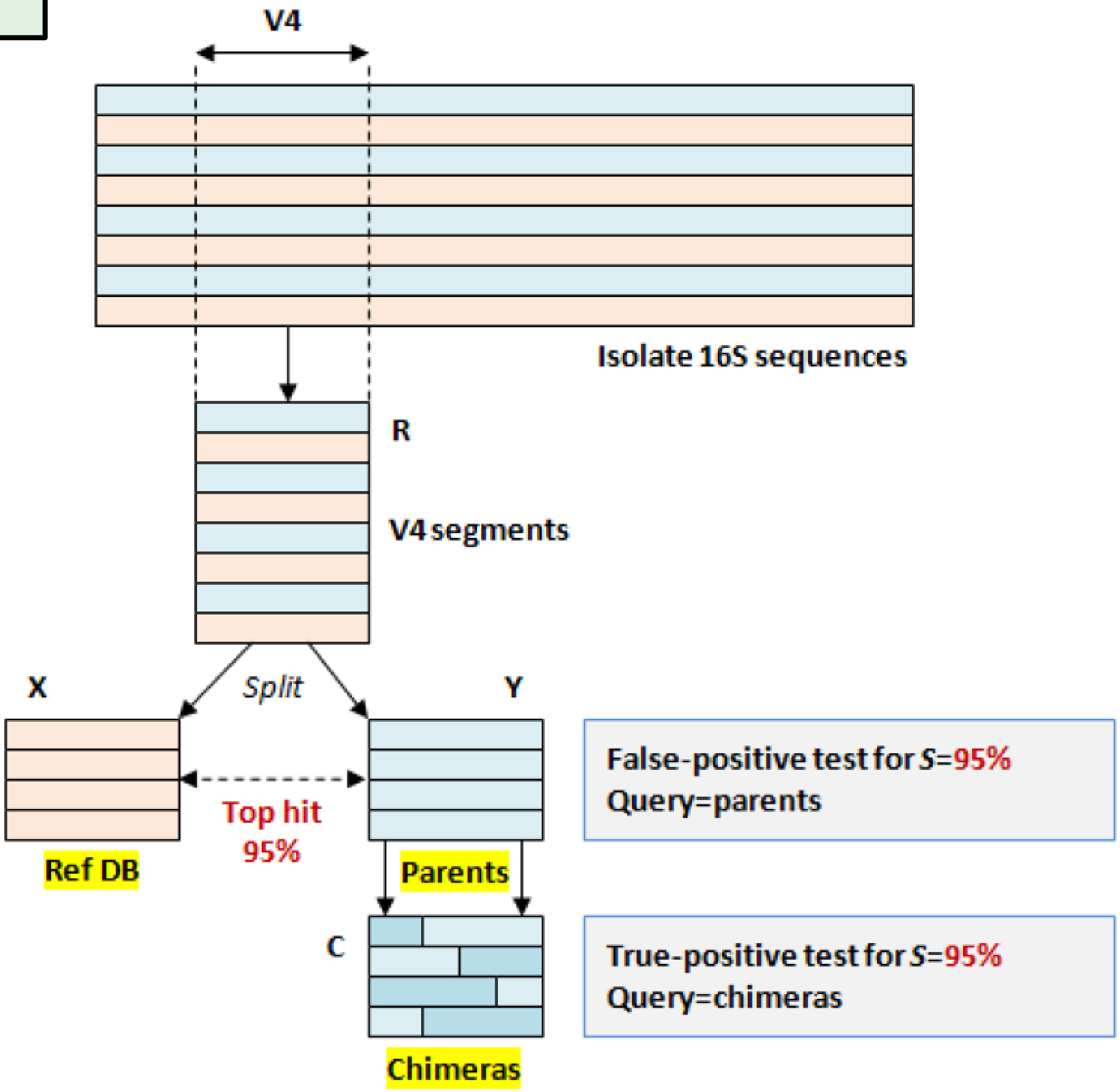
New benchmark design

- Measure dependence of accuracy on:
  - *D* = divergence, especially small *D*
  - *S* = similarity to closest reference sequence
- Sensitivity when:
  - "Step-parent" for segment is 100%, 99% ... 90% id (*S*)
- False-positives when:
  - Closest reference sequence is 100%, 99% ... 90% id (*S*)

New benchmark design

- Split reference db. into subsets X and Y
  - so that top hit similarity $X \leftrightarrow Y = S$, e.g. S=95%
- Make simulated bimeras C from parents in Y
  - with divergences D = 1%, 2% … 10%
- Measure TPs with query=C, db=X
- Measure FPs with query=Y, db=X

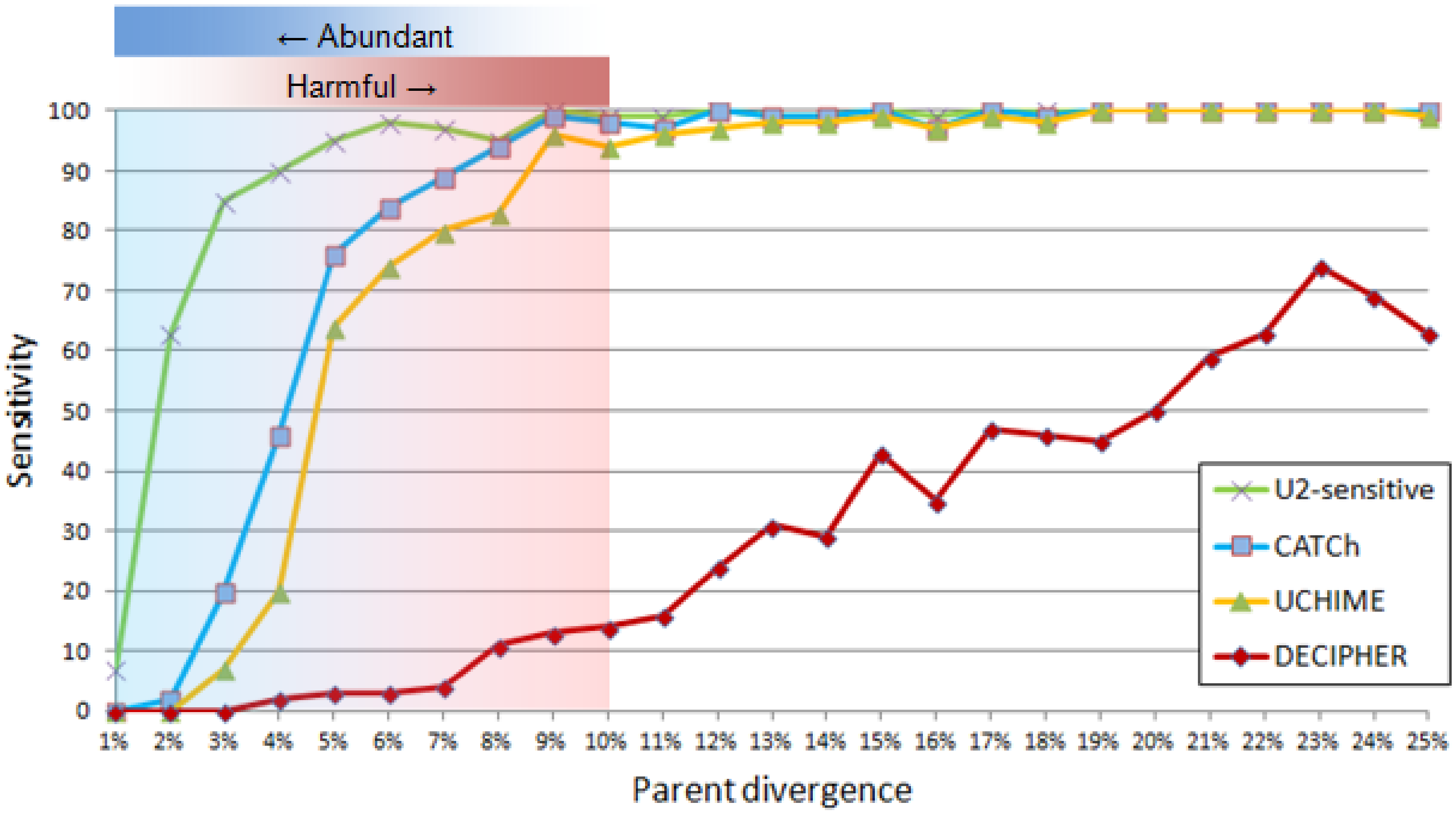
V4,*S* = 95%



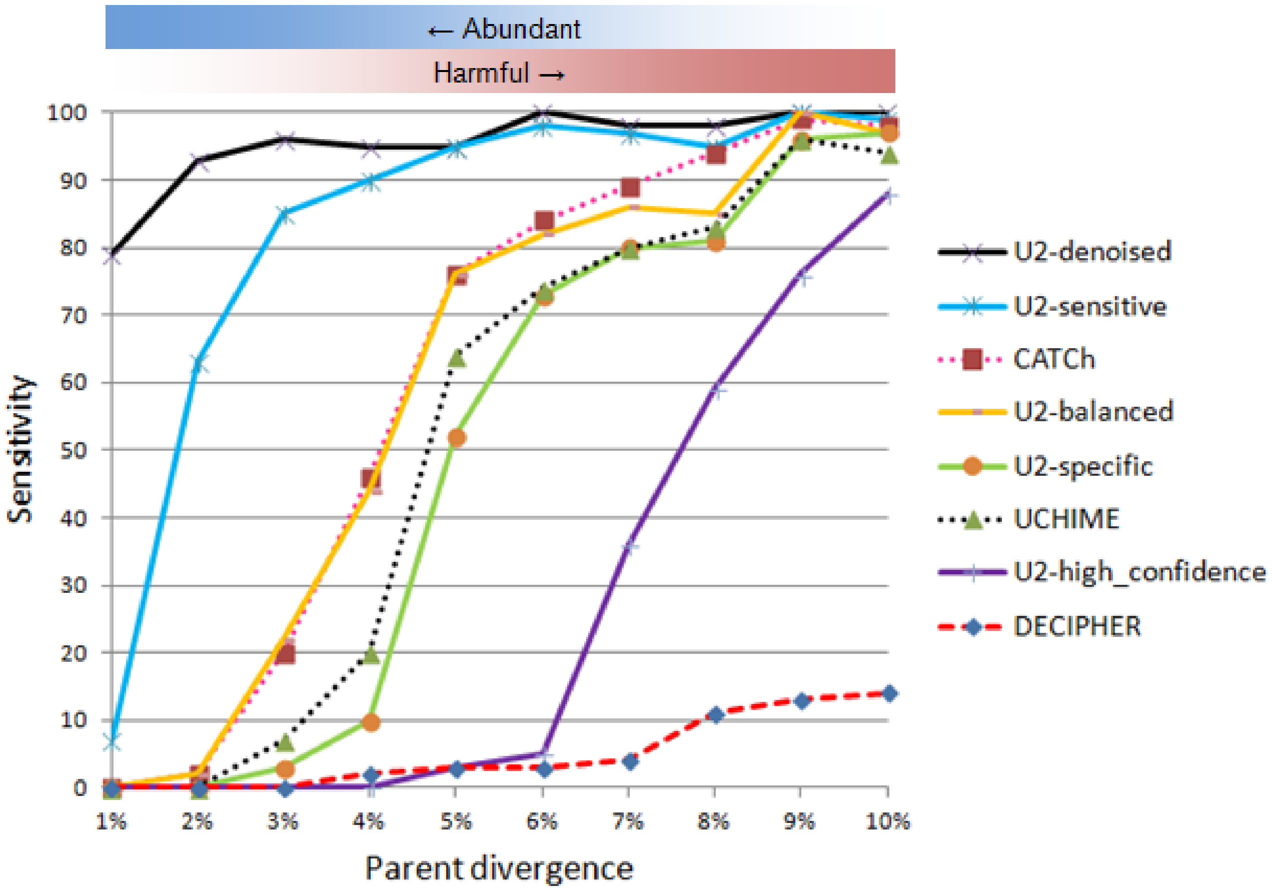
False-positive test for *S*=95%
Query=parents

True-positive test for *S*=95%
Query=chimeras

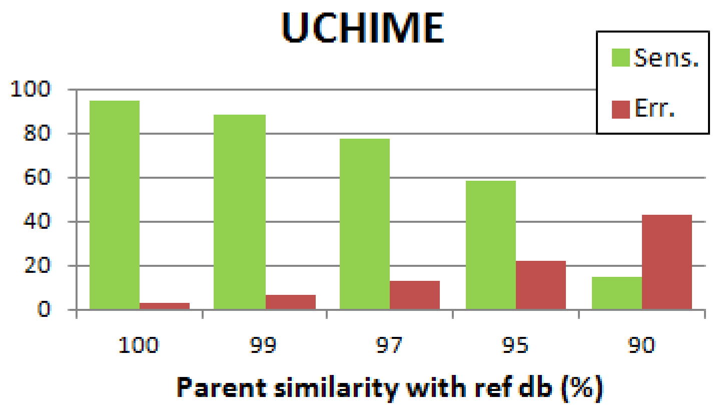
Benchmark results

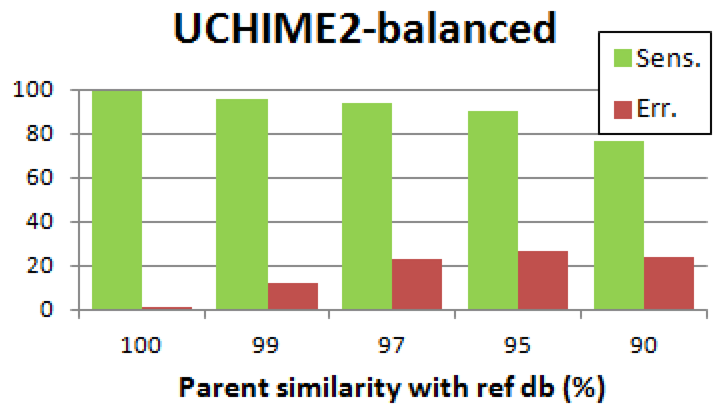


Benchmark results



Benchmark results


UCHIME


UCHIME2-balanced

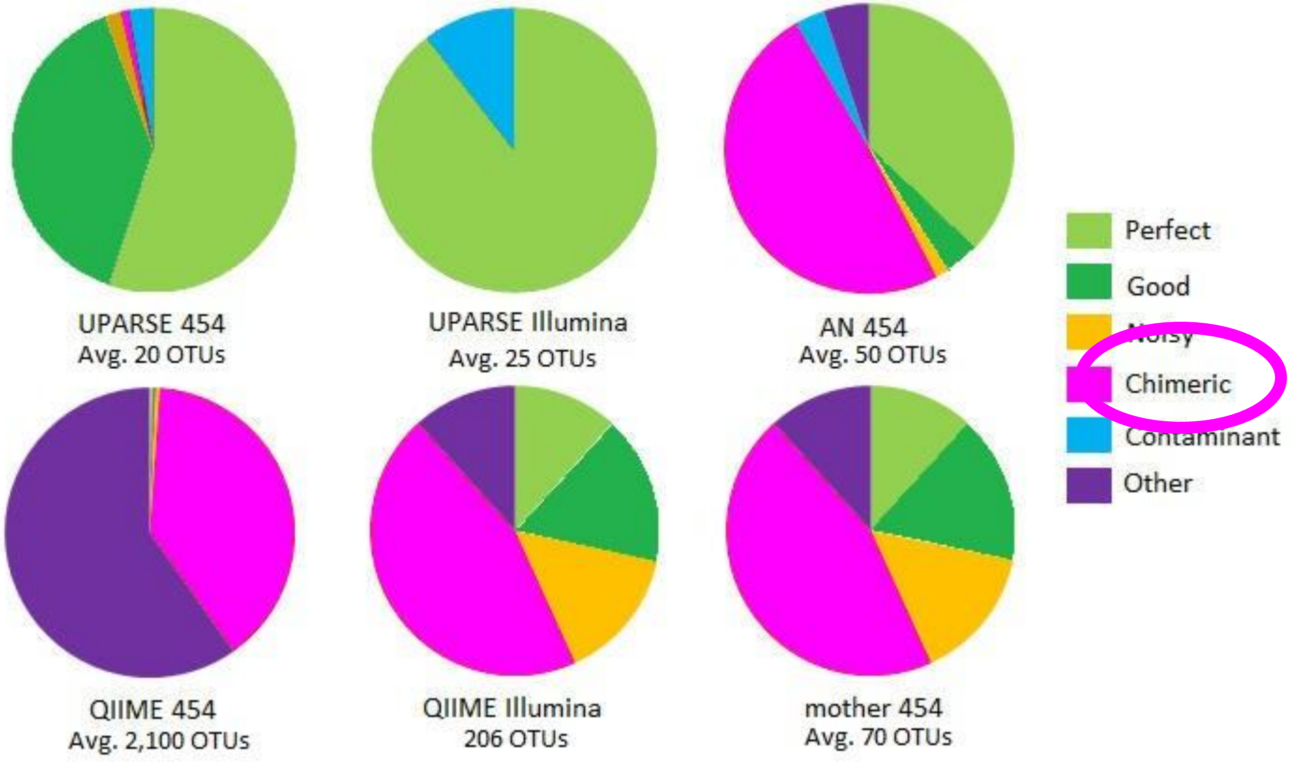
- High error rates if parents not in db
- Should use largest possible db (SILVA 1.8M)
- Gold (5k) misguided default for CS &UCHIME

Reference or *de-novo*?

- At <100% identity, fake models common
- All databases have sparse coverage
  - Even SILVA
- Reference mode has high error rates
- *De-novo* on filtered reads also high error rates
  - Because diffs. due to errors rapidly degrade accuracy
- *De-novo* on denoised reads very effective

OTU clustering: use UPARSE

- Better than UCHIME & UCHIME2 for OTUs
  - No need to distinguish read errors from low-div chimeras


UPARSE 454
Avg. 20 OTUs
UPARSE Illumina
Avg. 25 OTUs
AN 454
Avg. 50 OTUs
QIIME 454
Avg. 2,100 OTUs
QIIME Illumina
206 OTUs
mother 454
Avg. 70 OTUs
Perfect
Good
Noisy
Chimeric
Contaminant
Other