NG-Tax

OTU/ASV picking
Most abundant sequence picking
feature picking

Open & reproducible microbiome data analysis spring school
Wageningen, The Netherlands, May 28-30, 2018

Gerben DA Hermes, PhD

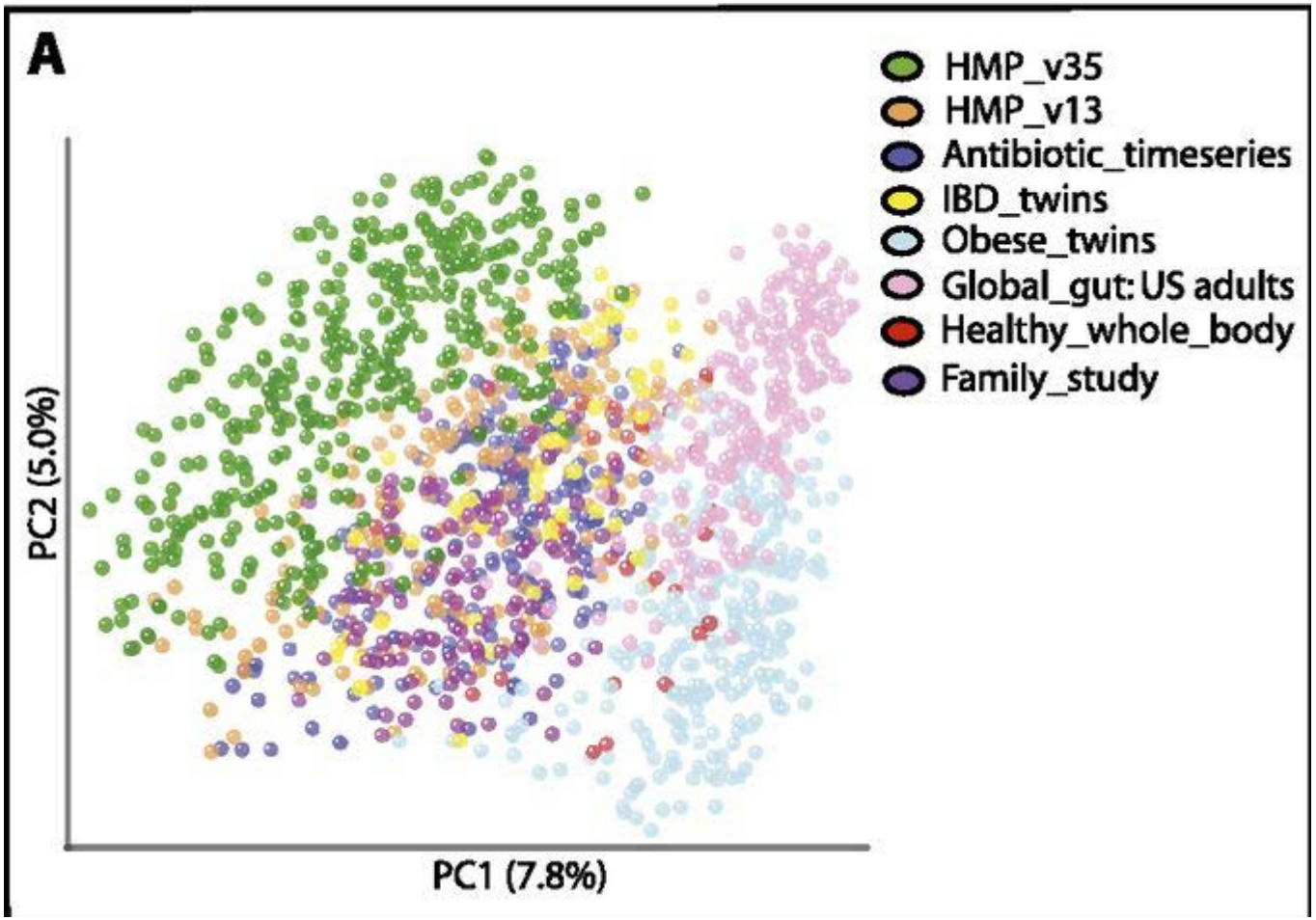
Laboratory of Microbiology,

Wageningen University & Research

Why NG-tax?

An increasing number of studies have shown that the chosen methodology rather than the natural variance is responsible for the greatest variance in microbiome studies

Meta analysis

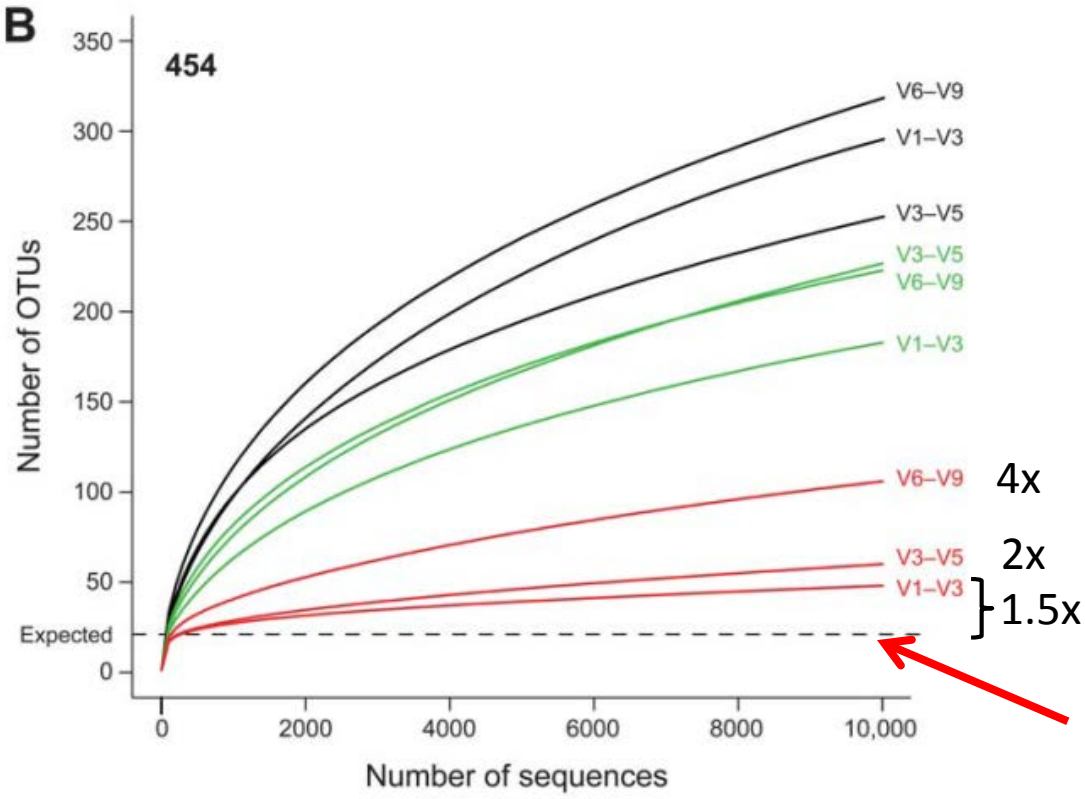


Catherine A. Lozupone et al. Genome Res. 2013;23:1704-1714

Objective

- Accuracy of taxonomic classifications
- Quantification potential (abundance)
- Estimate true richness/diversity
- Compare performance of 2 regions (V5-V6, V4)

Evaluation of 16S rDNA-based community profiling for human microbiome research (jumpstart consortium human microbiome project data generation working group, Plos 2012)



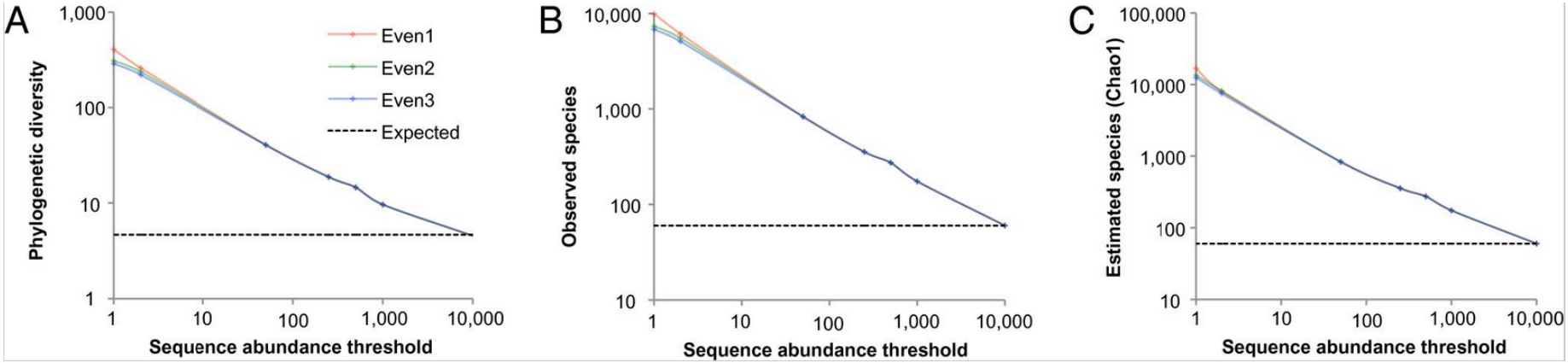
unfiltered

Quality filtering

Chimera + quality filtering

"improved estimation of community diversity after quality filtering and chimera checking"

Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample (Caporaso, Pnas 2010)

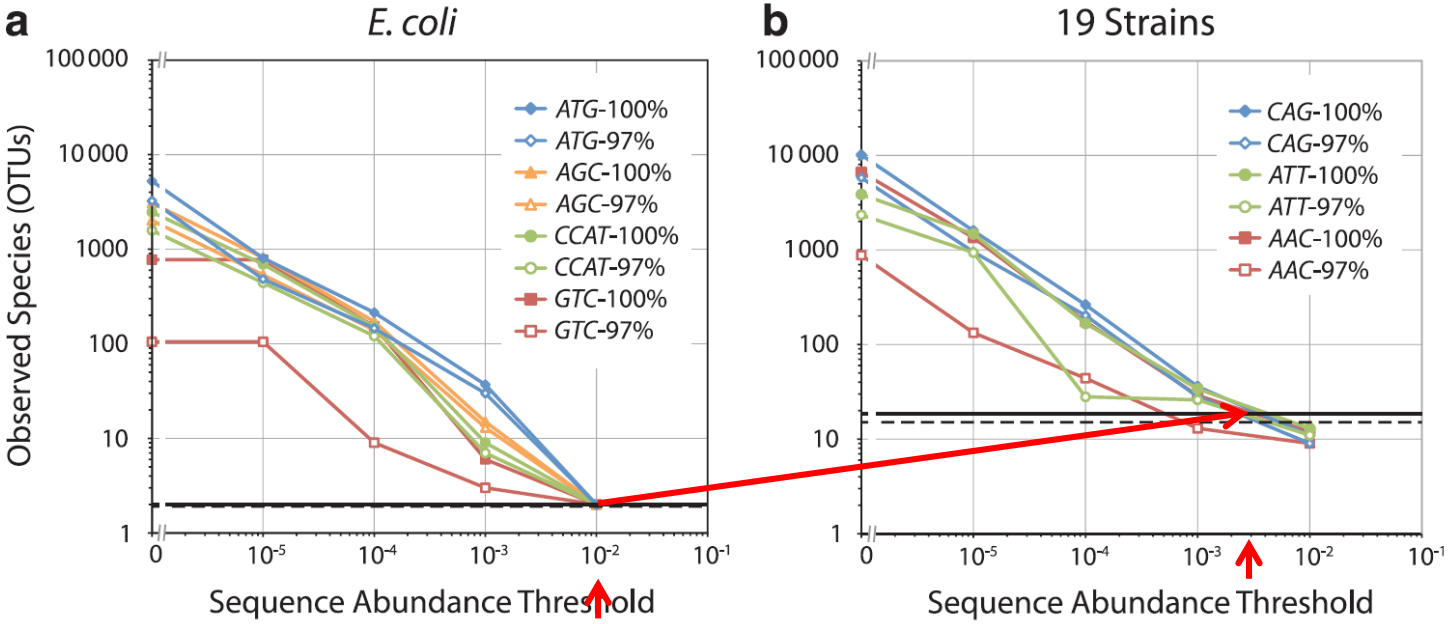


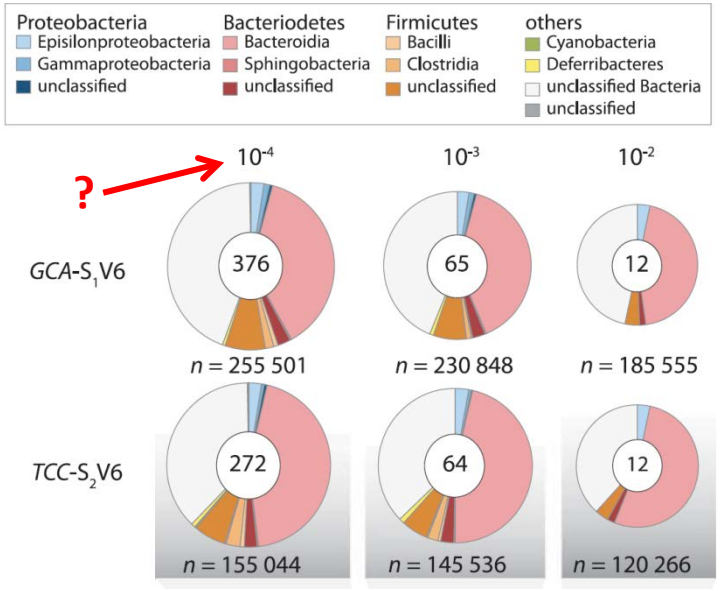
‘Correct’ flawed data

Sequence abundance threshold

Illumina-based analysis of microbial community diversity

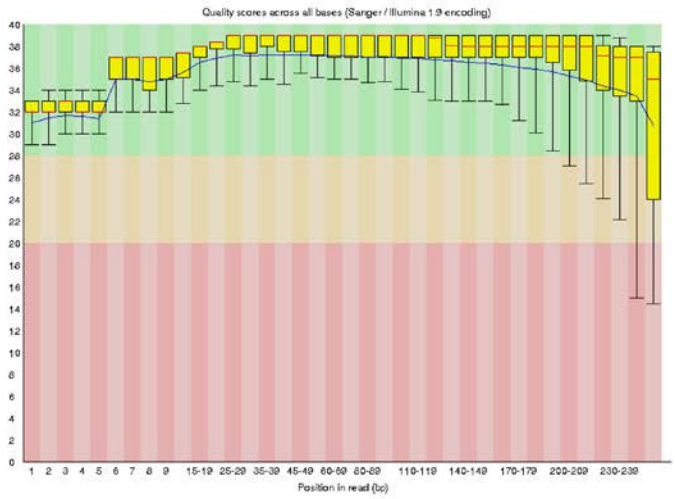
Degnan & Ochman, 2011, Isme j)



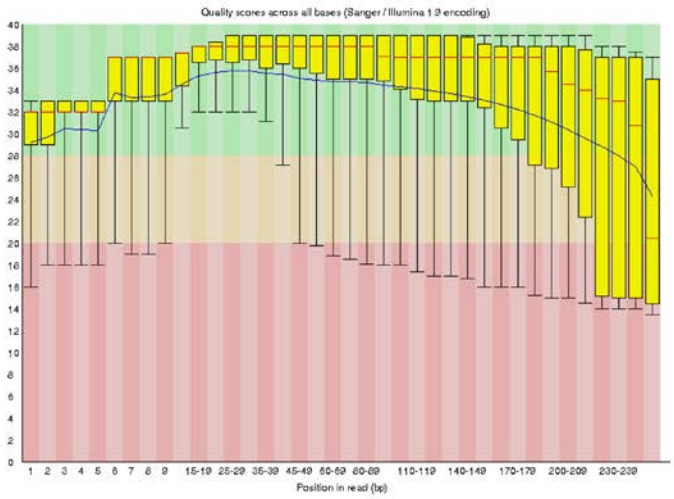


Identification vs Composition

- Default = optimal
- Optimum between length (resolution) & quality (repeated sequences)
- Phred 10 = 90% accuracy, 30 99.9%



Forward reads



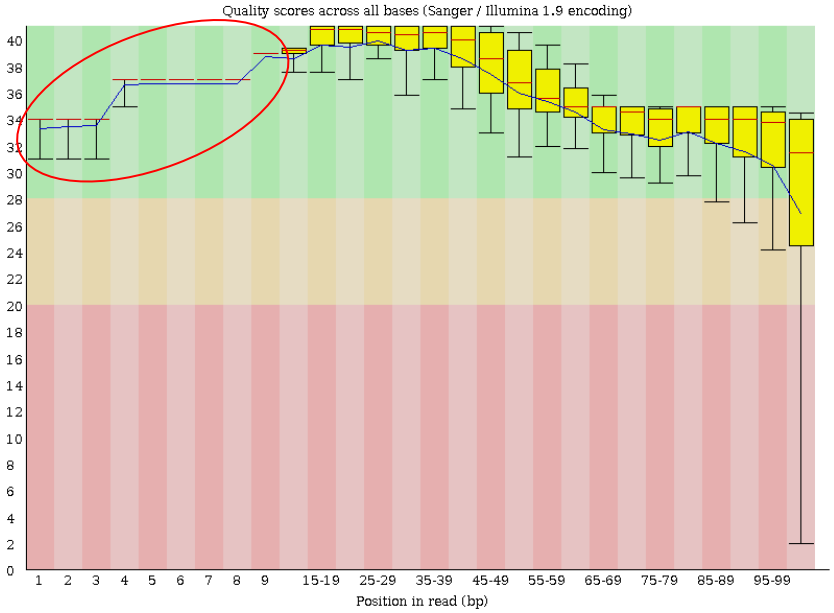
Reverse reads

Sequencing quality: the truth

	Old library 4	%	New Library 4	%
Both invalid barcodes*	80483	4.2	7137707	12.9
Only one valid barcode	560049	29.4	15224394	27.6
Both valid barcodes but not matching	884080	46.4	5326330	9.7
Both valid barcodes and matching	379231	19.9	27476627	49.8
Total	1903843	100	55165058	100

>99.9% accuracy

- 40-70% error in primer and/or barcode (1/3th read)
- 26-40% error in the barcode region (1/10th)
- 10-50% matching barcodes



Sequencing quality: biology

Sequencing error is not random(ly distributed)

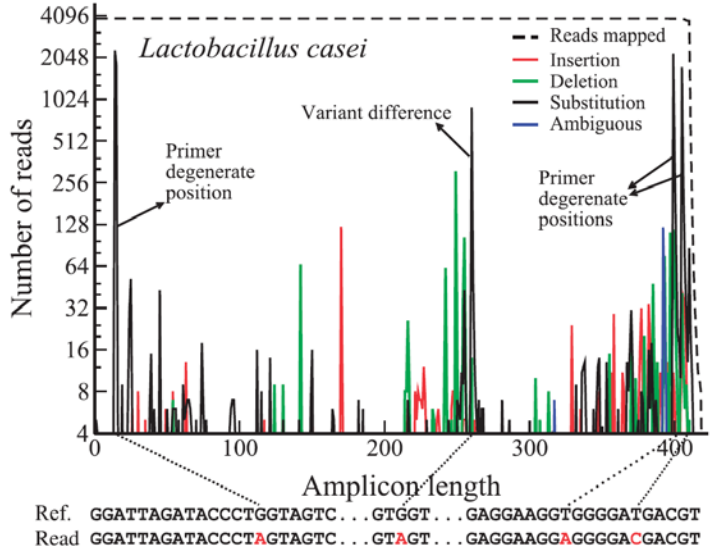
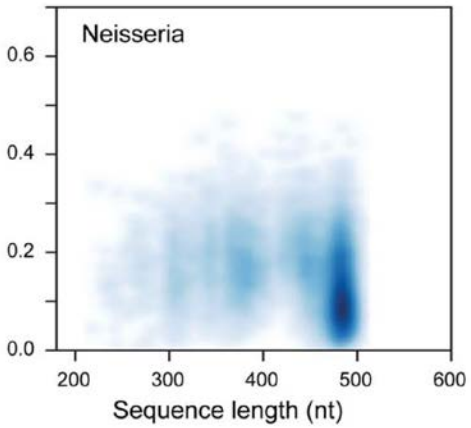
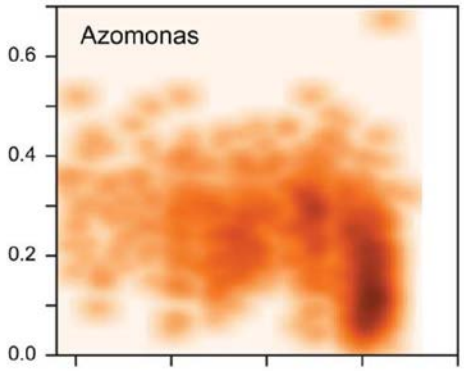


Fig. 2. Errors occurring along the *Lactobacillus casei* reference sequence.





Evaluation of 16S rDNA-based community profiling for human microbiome research (jumpstart consortium human microbiome project data generation working group, Plos 2012)

Unraveling the outcome of 16S rDNA –based taxonomy analysis through mock data and simulations Bioinformatics 2014, May et al.

Illumina base-caller specific error

- Schirmer, Insight into biases and sequencing errors for amplicon sequencing with the Illumina MiSeq platform, 2015

Table 1. A selection of substitutions that occurred at a very high rate in data set *DS 35*

R1			R2		
A –> G	pos 226	Rate 0.25	A –> G	pos 57	Rate 0.03
T –> G	pos 162	Rate 0.02	T –> C	pos 136	Rate 0.02
T –> G	pos 179	Rate 0.01	G –> A	pos 57	Rate 0.03
C –> G	pos 118	Rate 0.18	G –> C	pos 174	Rate 0.14

Columns 1–3 specify the type of substitution, its position and the substitution rate for the R1 reads. Columns 4–6 detail the respective information for the R2 reads.

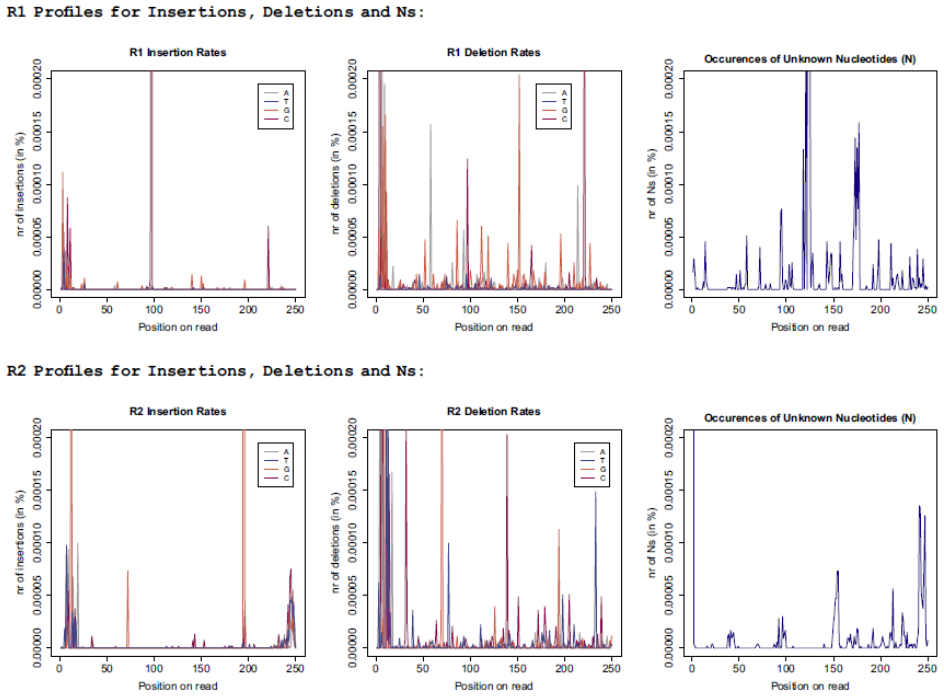
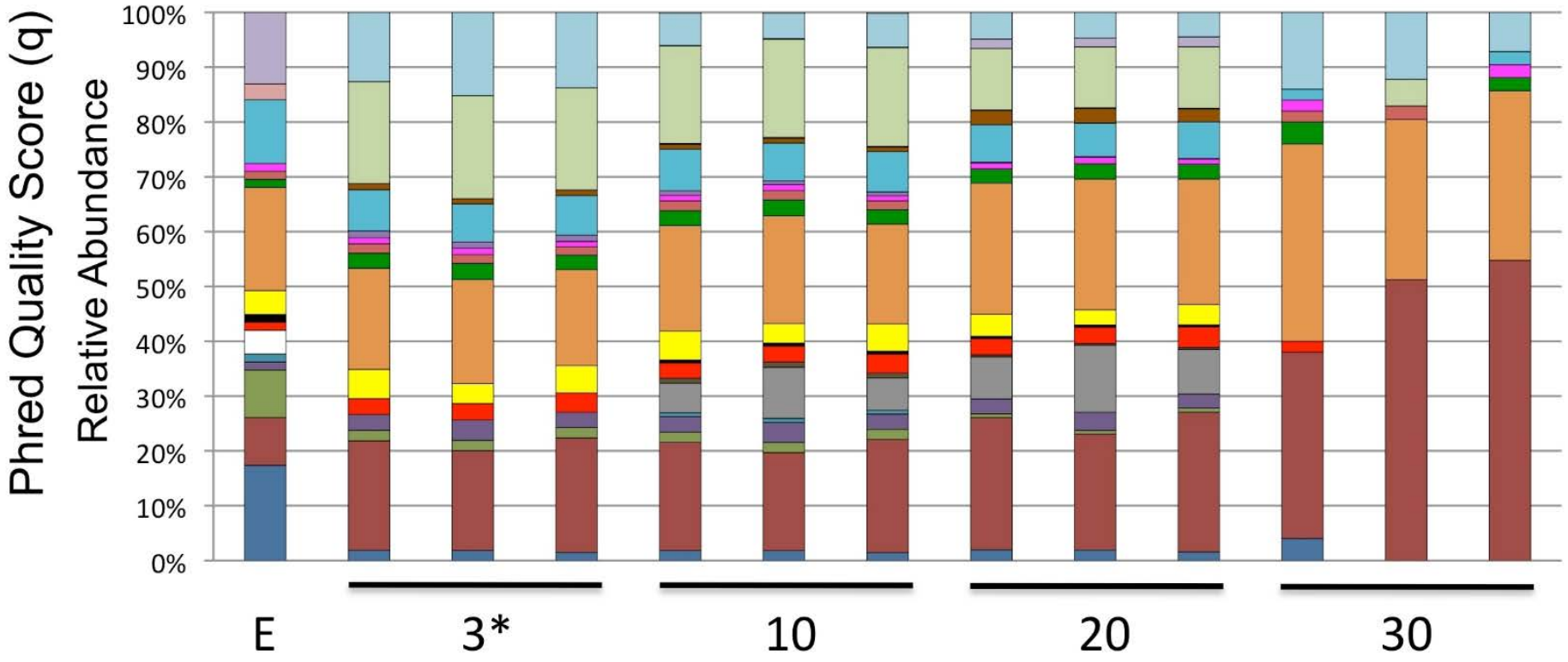


Figure 2. Error profiles for insertions, deletions and unknown nucleotides (Ns): the first three graphs show the R1 error profiles. For insertions the colour identifies the inserted nucleotide and for deletions the colour refers to the type of nucleotide that was deleted. The lower three graphs display the error

Effect of quality filtering

- Bockulich 2013, Quality-filtering vastly improves diversity estimates from Illumina amplicon sequencing, supplementals



Filtering based on quality good idea?

Known/'improve'

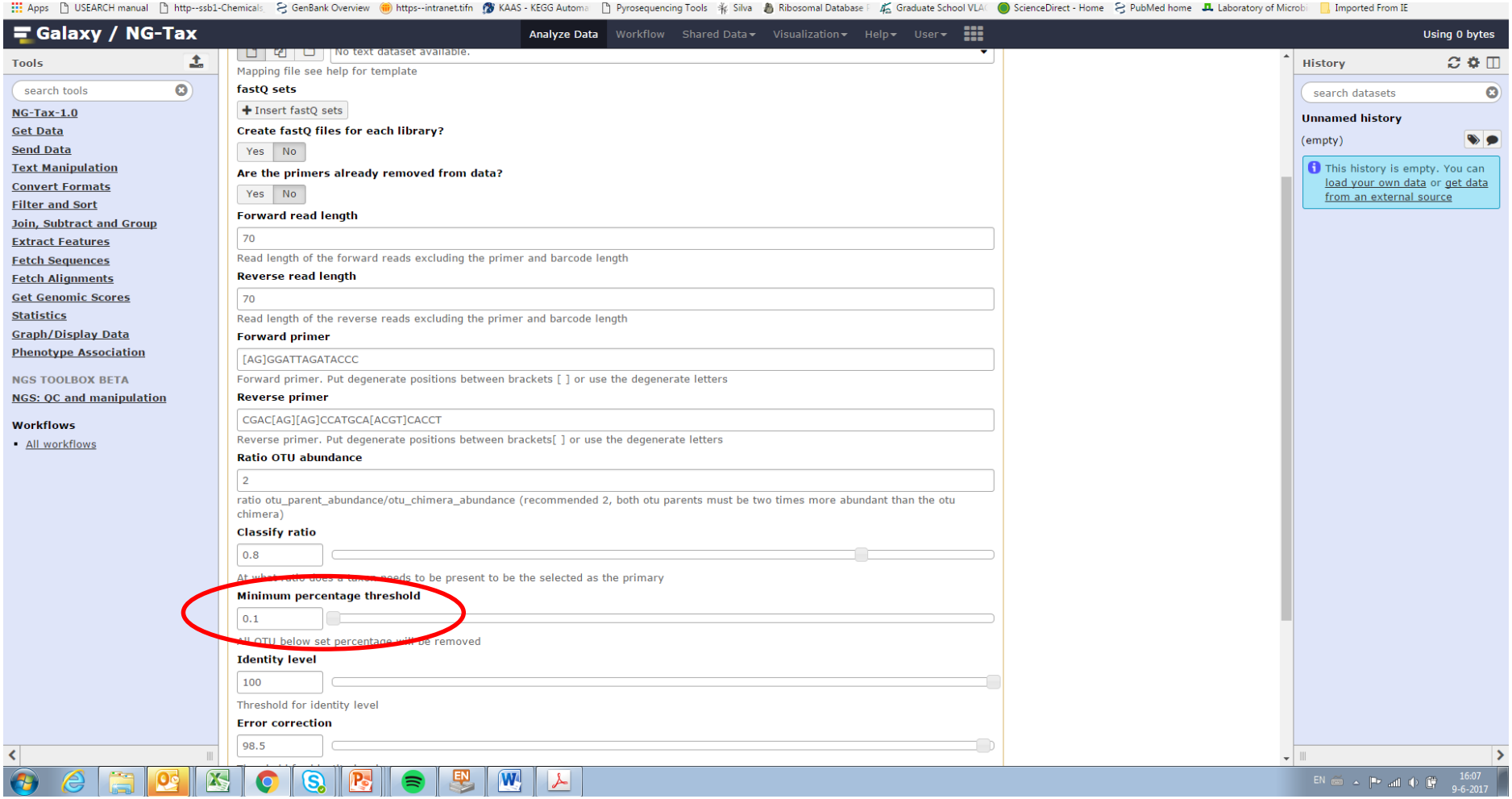
- PCR (high fidelity polymerase)
- Little cycles=less chance of error
- 1 PCR (2^{nd} step PCR=~50% don't contain barcode)

Unknown/no control whatsoever

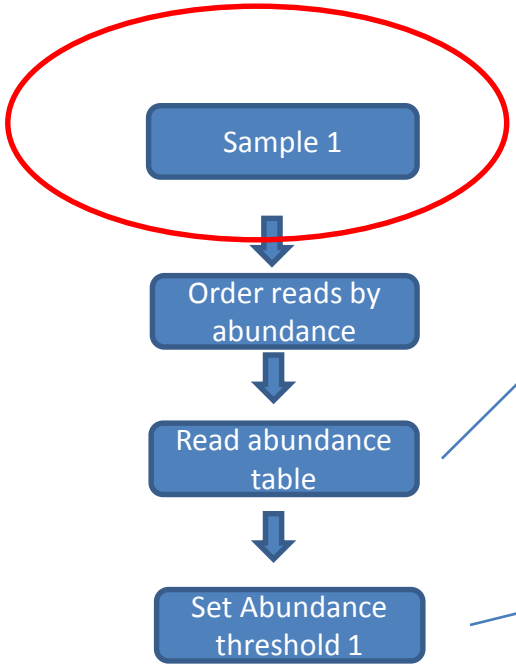
- Sequences in your sample (primer dependent)
- Illumina base-caller (improve with PhyX and equal base distribution of barcodes?)

Optimal settings

- No quality filtering
- Filter by abundance -> high quality
- ‘Short’ reads (enough for identification) = less chance for error


Galaxy / NG-Tax
Analyze Data
Workflow
Shared Data
Visualization
Help
User
Using 0 bytes
Tools
search tools
NG-Tax-1.0
Get Data
Send Data
Text Manipulation
Convert Formats
Filter and Sort
Join, Subtract and Group
Extract Features
Fetch Sequences
Fetch Alignments
Get Genomic Scores
Statistics
Graph/Display Data
Phenotype Association
NGS TOOLBOX BETA
NGS: QC and manipulation
Workflows
All workflows
No text dataset available.
Mapping file see help for template
fastQ sets
Insert fastQ sets
Create fastQ files for each library?
Yes
No
Are the primers already removed from data?
Yes
No
Forward read length
70
Read length of the forward reads excluding the primer and barcode length
Reverse read length
70
Read length of the reverse reads excluding the primer and barcode length
Forward primer
[AG]GGATTAGATACCC
Forward primer. Put degenerate positions between brackets [ ] or use the degenerate letters
Reverse primer
CGAC[AG][AG]CCATGCA[ACGT]CACCT
Reverse primer. Put degenerate positions between brackets[ ] or use the degenerate letters
Ratio OTU abundance
2
ratio otu_parent_abundance/otu_chimera_abundance (recommended 2, both otu parents must be two times more abundant than the otu chimera)
Classify ratio
0.8
Minimum percentage threshold
0.1
Identity level
100
Threshold for identity level
Error correction
98.5
History
search datasets
Unnamed history
(empty)
This history is empty. You can load your own data or get data from an external source

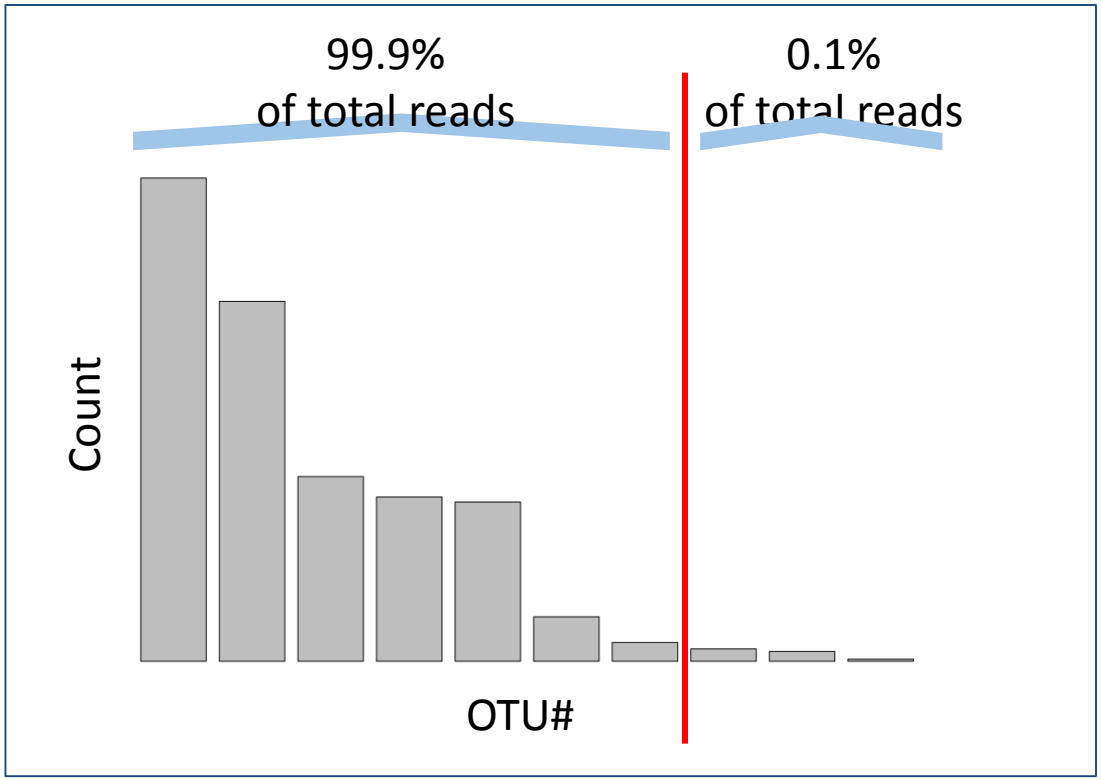
Threshold determination

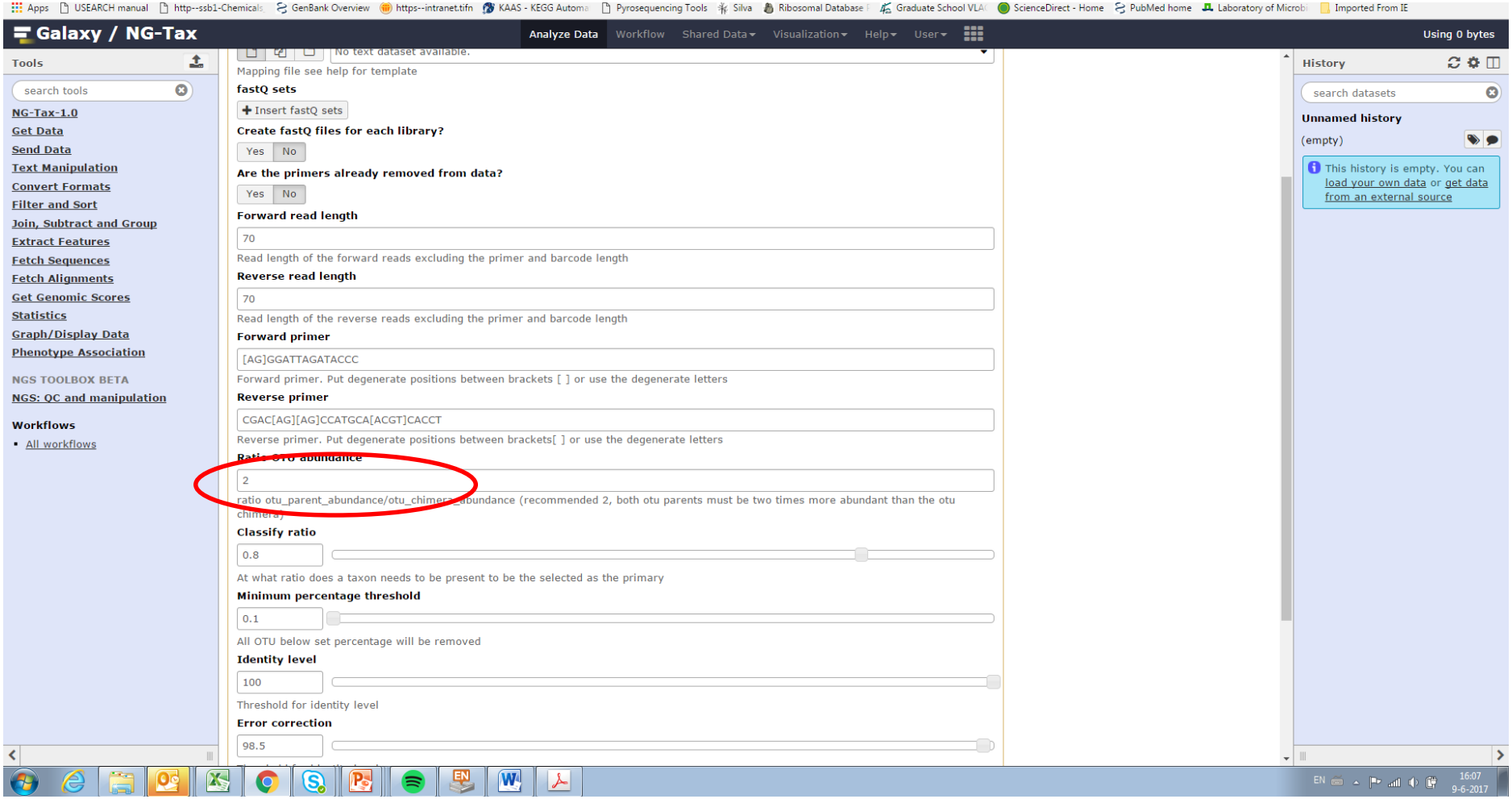


Sequence	Counts	Accumulated counts	Counts /Accumulated counts > 0.001
TACTATGCCA	658	658	658/658>0.001 Yes
GTCTAGTACA	523	1171	523/1171>0.001 Yes
…	…	…	…
GTATTAGCCA	75	74560	75/74560>0.001 Yes
GTCTATGCCA	72	74632	72/74632>0.001 No
ATCTATGCCT	70	74701	No
…	…	…	…
CTTTACGCCT	1	159240	No

OTU selection (per sample)

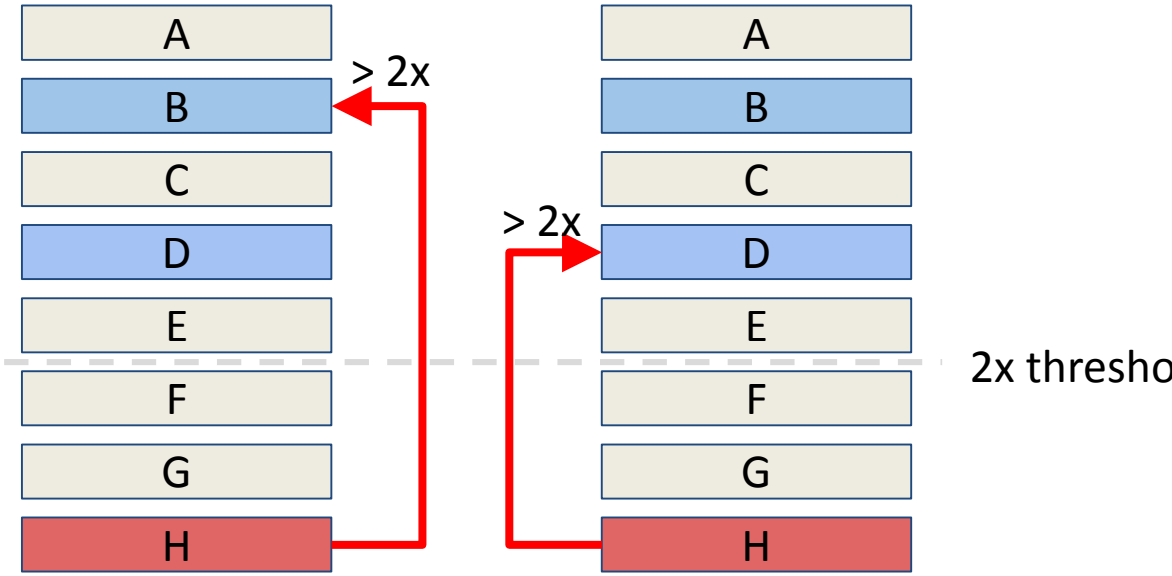
- Create unique OTU list
  - Rank by abundance
- Abundance threshold
  - Above threshold included
  - Below threshold rejected

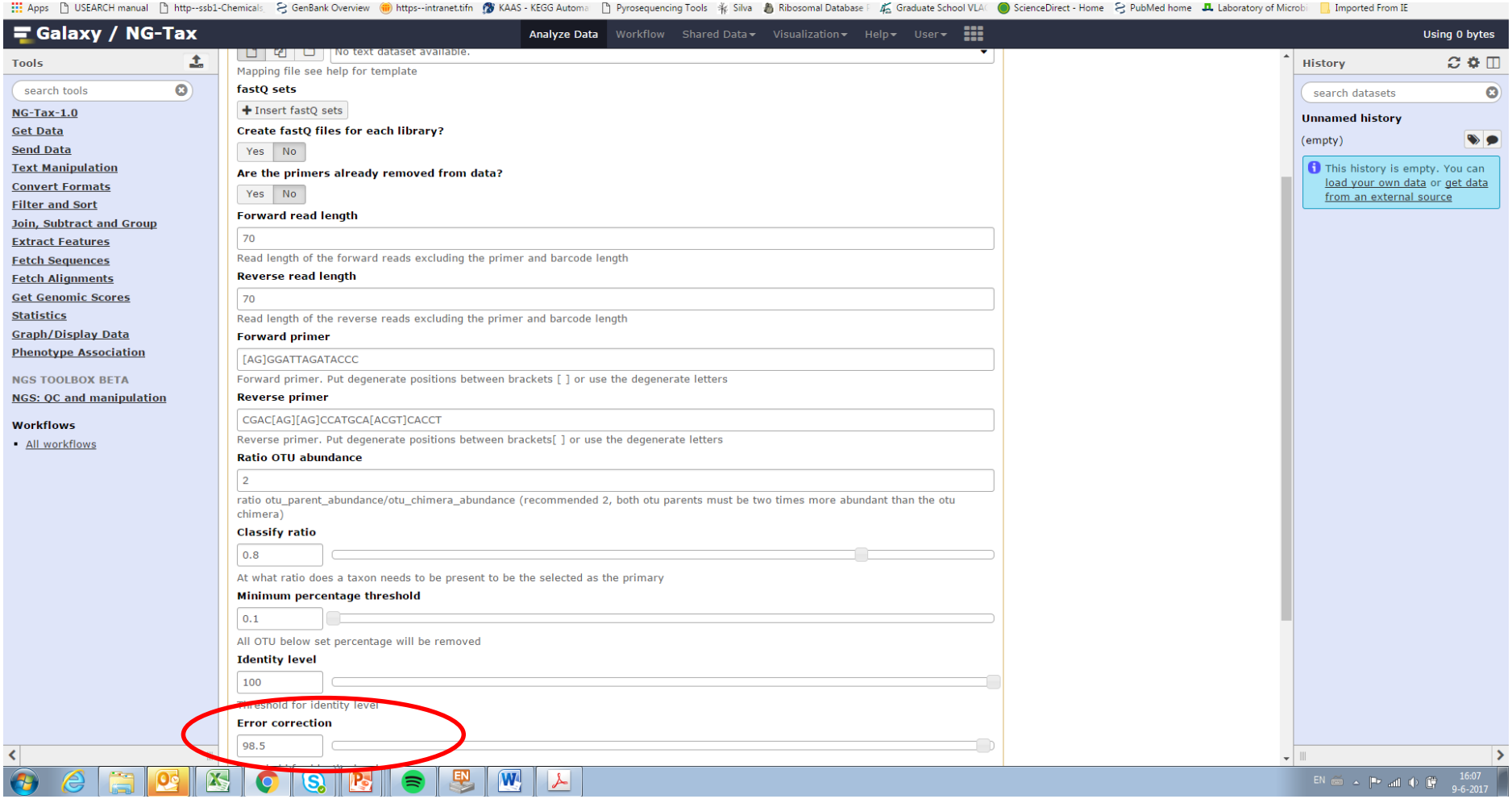



Galaxy / NG-Tax
Analyze Data
Workflow
Shared Data
Visualization
Help
User
Using 0 bytes
Tools
search tools
NG-Tax-1.0
Get Data
Send Data
Text Manipulation
Convert Formats
Filter and Sort
Join, Subtract and Group
Extract Features
Fetch Sequences
Fetch Alignments
Get Genomic Scores
Statistics
Graph/Display Data
Phenotype Association
NGS TOOLBOX BETA
NGS: QC and manipulation
Workflows
All workflows
No text dataset available.
Mapping file see help for template
fastQ sets
Insert fastQ sets
Create fastQ files for each library?
Yes
No
Are the primers already removed from data?
Yes
No
Forward read length
70
Read length of the forward reads excluding the primer and barcode length
Reverse read length
70
Read length of the reverse reads excluding the primer and barcode length
Forward primer
[AG]GGATTAGATACCC
Forward primer. Put degenerate positions between brackets [ ] or use the degenerate letters
Reverse primer
CGAC[AG][AG]CCATGCA[ACGT]CACCT
Reverse primer. Put degenerate positions between brackets[ ] or use the degenerate letters
Ratio OTU abundance
2
ratio otu_parent_abundance/otu_chimera_abundance (recommended 2, both otu parents must be two times more abundant than the otu chimera)
Classify ratio
0.8
At what ratio does a taxon needs to be present to be the selected as the primary
Minimum percentage threshold
0.1
All OTU below set percentage will be removed
Identity level
100
Threshold for identity level
Error correction
98.5
History
search datasets
Unnamed history
(empty)
This history is empty. You can load your own data or get data from an external source

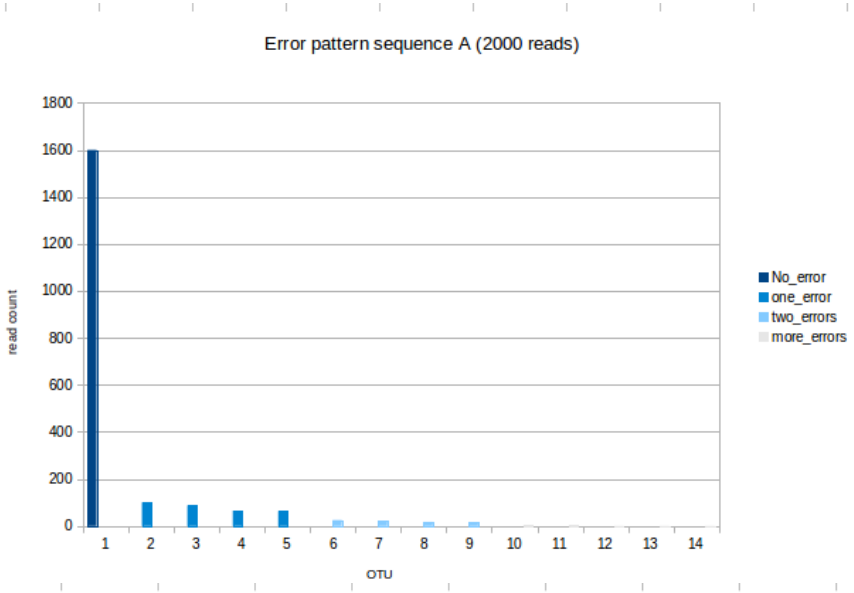
Chimera filtering

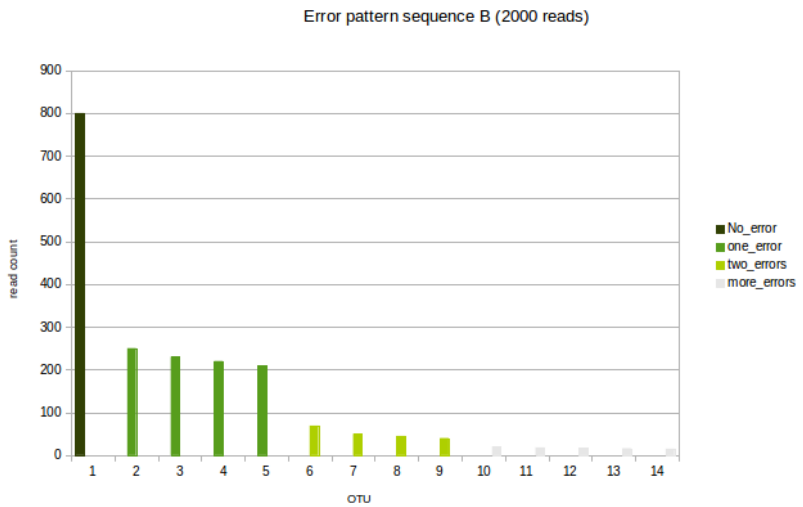
- Only accepted OTU's
- The filtering
  - 2x threshold
- H is chimera of B & D

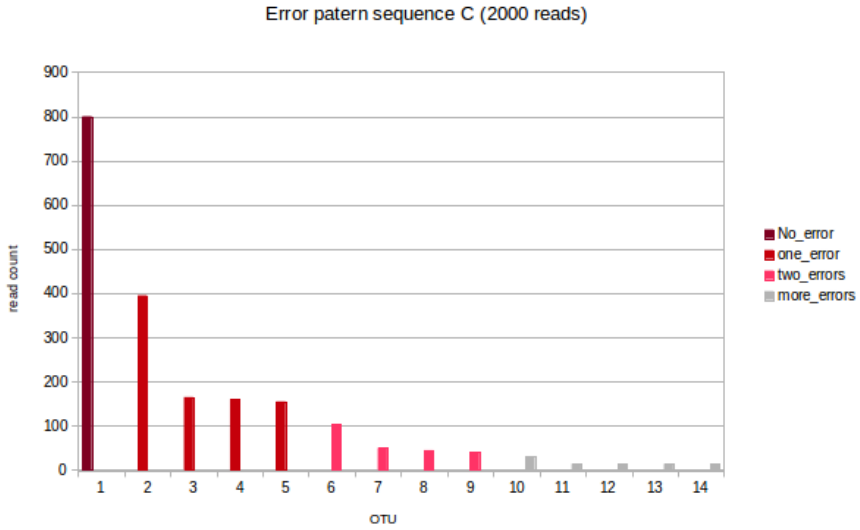


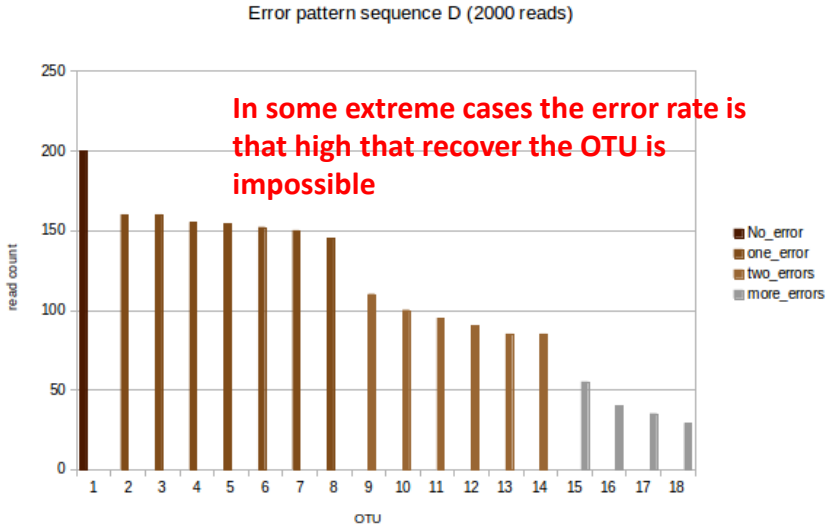

Galaxy / NG-Tax
Analyze Data
Workflow
Shared Data
Visualization
Help
User
Using 0 bytes
Tools
search tools
NG-Tax-1.0
Get Data
Send Data
Text Manipulation
Convert Formats
Filter and Sort
Join, Subtract and Group
Extract Features
Fetch Sequences
Fetch Alignments
Get Genomic Scores
Statistics
Graph/Display Data
Phenotype Association
NGS TOOLBOX BETA
NGS: QC and manipulation
Workflows
All workflows
No text dataset available.
Mapping file see help for template
fastQ sets
Insert fastQ sets
Create fastQ files for each library?
Yes
No
Are the primers already removed from data?
Yes
No
Forward read length
70
Read length of the forward reads excluding the primer and barcode length
Reverse read length
70
Read length of the reverse reads excluding the primer and barcode length
Forward primer
[AG]GGATTAGATACCC
Forward primer. Put degenerate positions between brackets [ ] or use the degenerate letters
Reverse primer
CGAC[AG][AG]CCATGCA[ACGT]CACCT
Reverse primer. Put degenerate positions between brackets[ ] or use the degenerate letters
Ratio OTU abundance
2
ratio otu_parent_abundance/otu_chimera_abundance (recommended 2, both otu parents must be two times more abundant than the otu chimera)
Classify ratio
0.8
At what ratio does a taxon needs to be present to be the selected as the primary
Minimum percentage threshold
0.1
All OTU below set percentage will be removed
Identity level
100
Threshold for identity level
Error correction
98.5
History
search datasets
Unnamed history
(empty)
This history is empty. You can load your own data or get data from an external source

Error pattern is sequence-specific

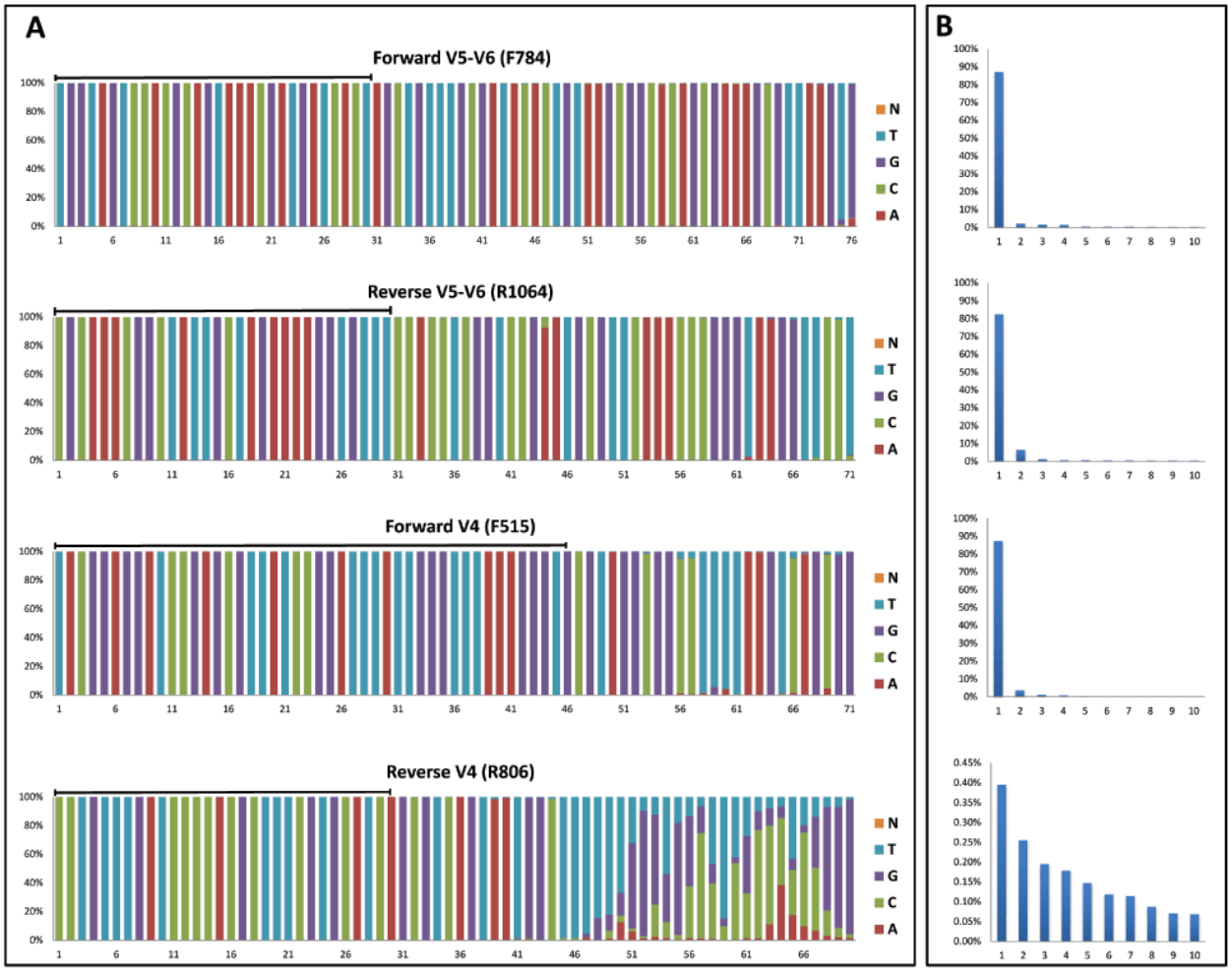






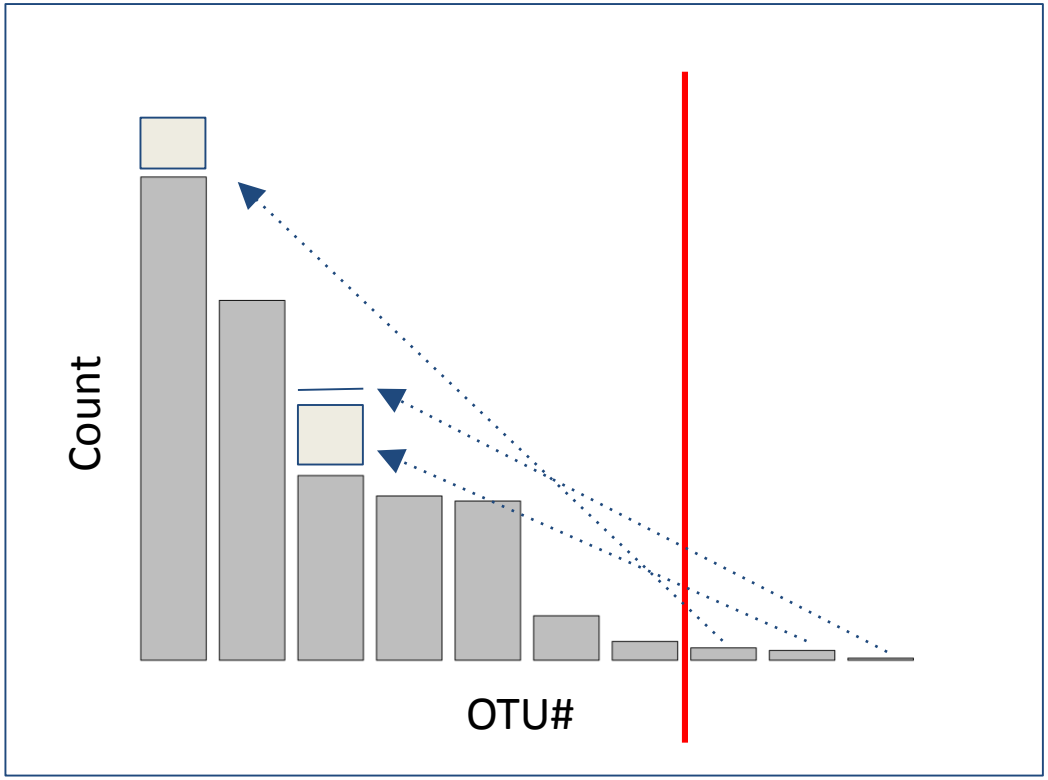


Example of unknown ‘things’: in silico ≠ in vitro Parabacteroides



Error correction

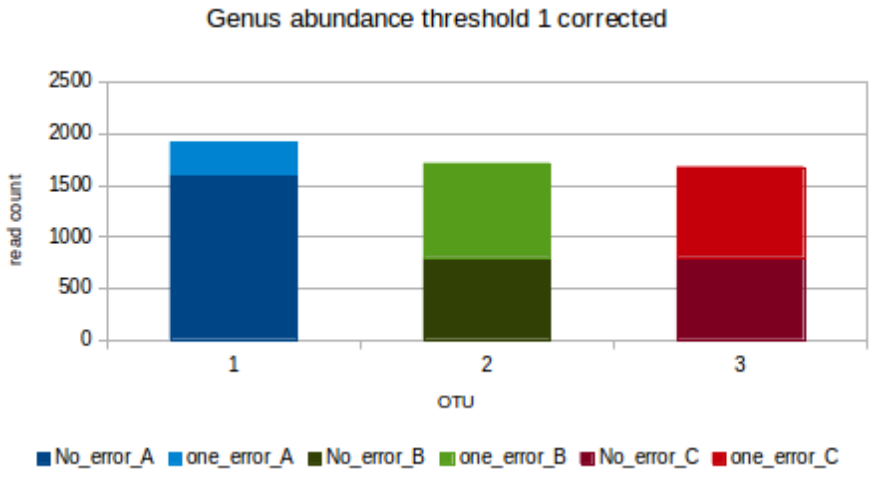
- Correction of abundance profiles
- For each rejected Otu
  - Find most abundant accepted OTU with 1 mismatch
  - Add count of rejected OTU to the count of accepted OTU



Abundance determination step allowing one mismatch -> correct for differential error pattern and reduce the impact of the abundance threshold

Sample with equimolar
Blue, Green, Red and Brown genera

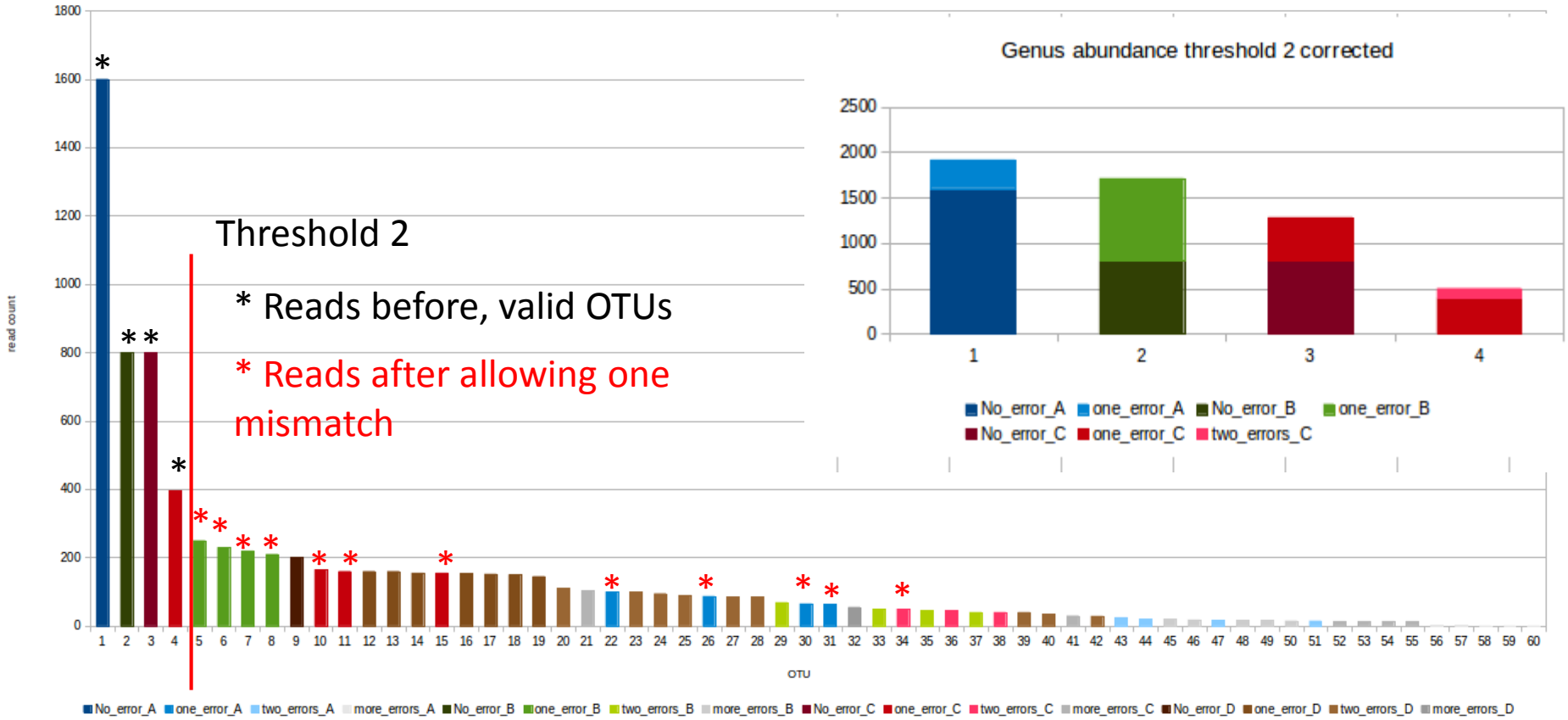
Before correction -> dark color
“Blue” genus >> “Green” genus = “red” genus
After correction -> added light color
“Blue” genus ≈ “Green” genus ≈ “red” genus



Abundance determination step allowing one mismatch -> correct for differential error pattern and reduce the impact of the abundance threshold

Sample with equimolar Blue, Green, Red and Brown genera

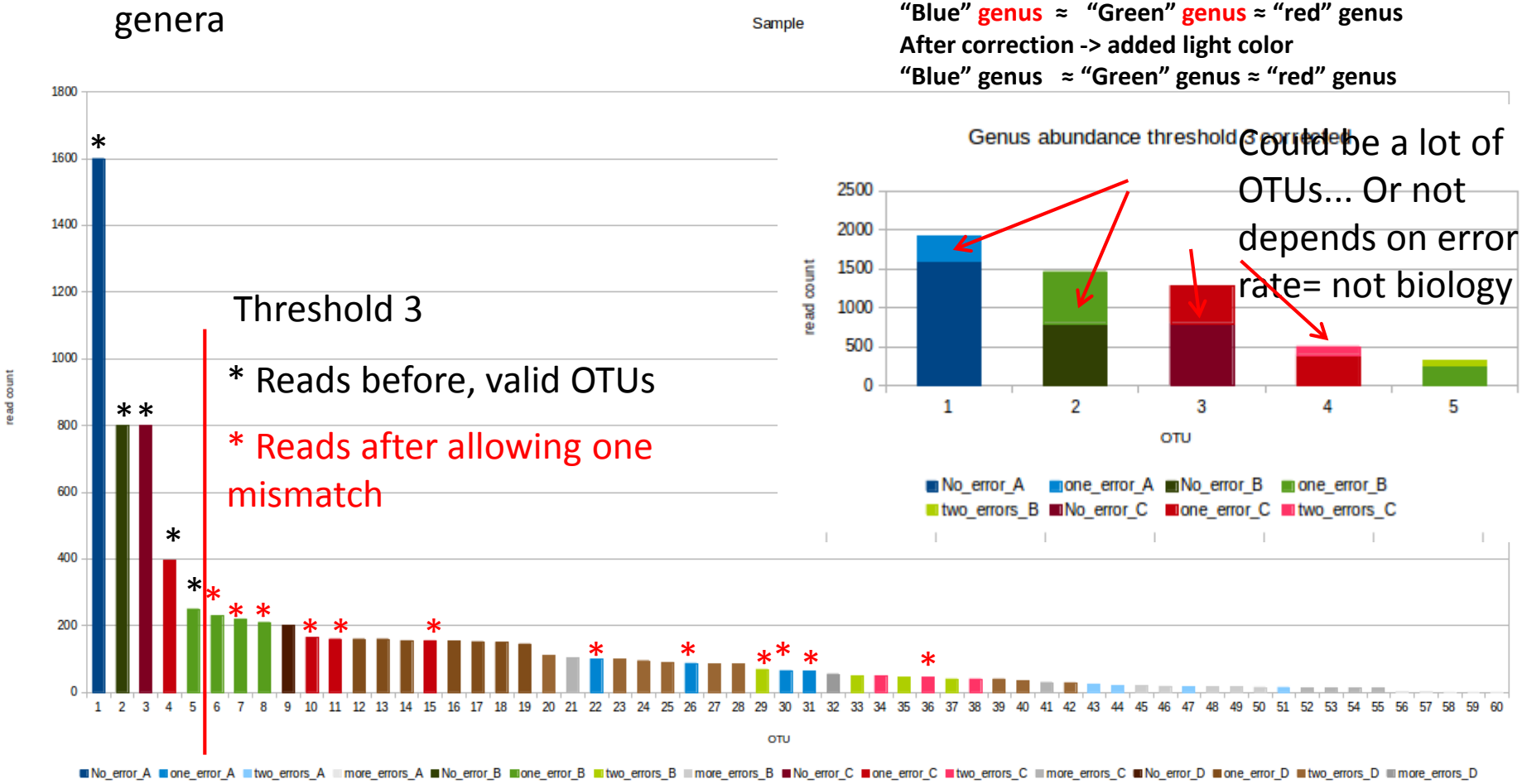
Before correction -> dark color
"Blue" genus >> "Green" genus < "red" genus
After correction -> added light color
"Blue" genus ≈ "Green" genus ≈ "red" genus



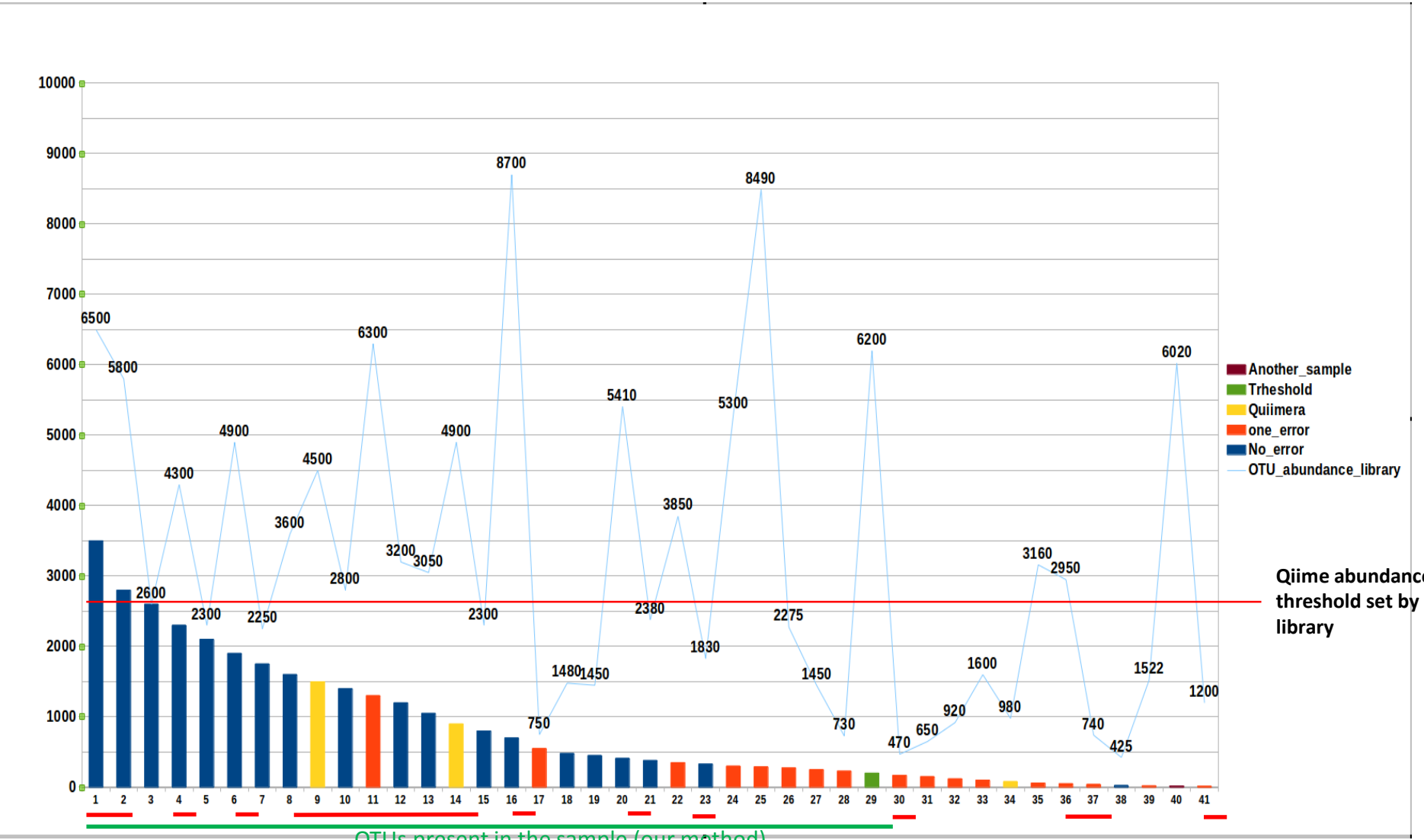
Abundance determination step allowing one mismatch -> correct for differential error pattern and reduce the impact of the abundance threshold

Sample with equimolar Blue, Green, Red and Brown genera

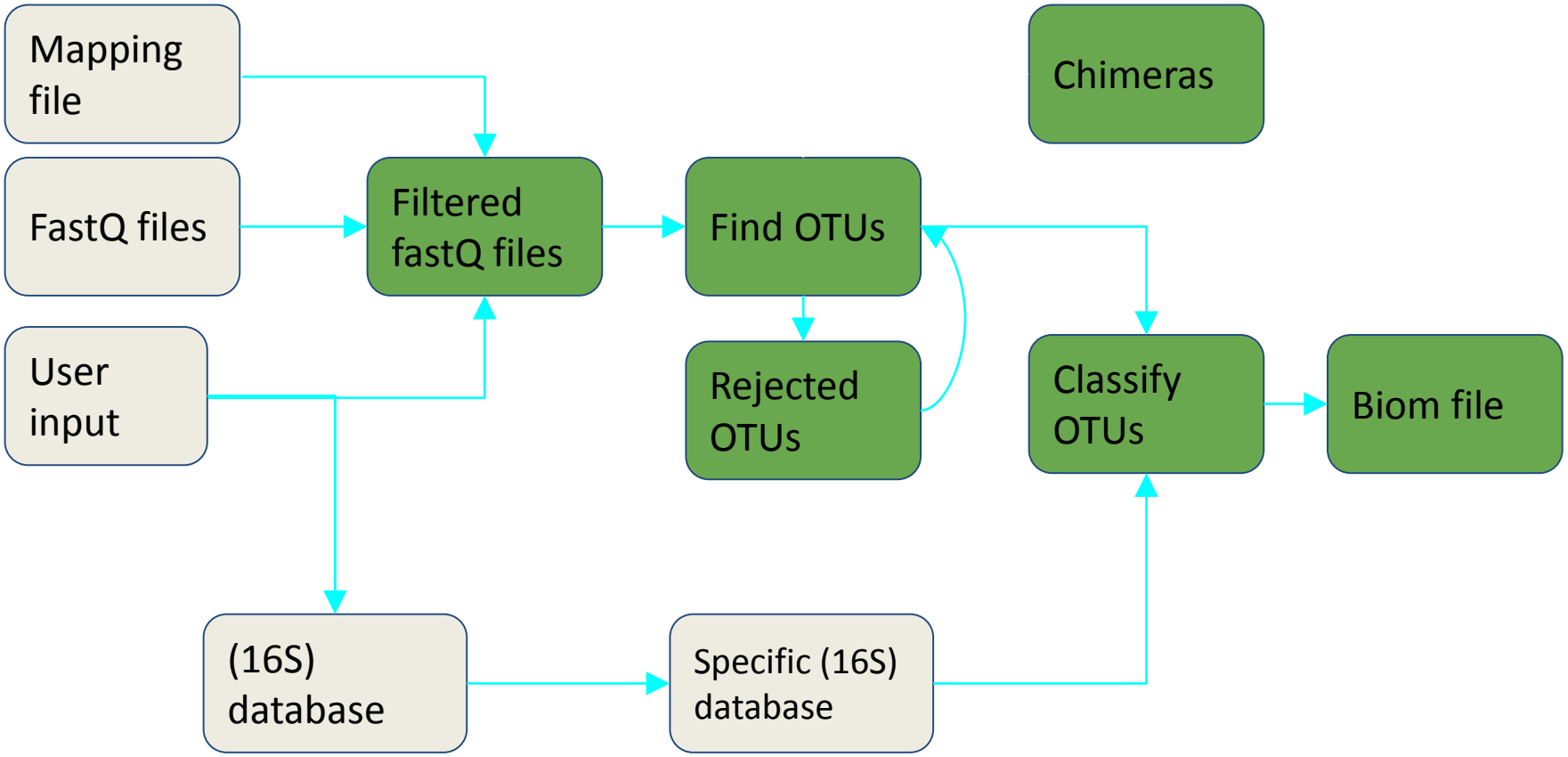
Before correction -> dark color
"Blue" genus ≈ "Green" genus ≈ "red" genus
After correction -> added light color
"Blue" genus ≈ "Green" genus ≈ "red" genus



Manual OTU manipulation (for instance QIIME)
OTU distribution per dataset (add samples/different results)



Overview pipeline



NG-Tax

OTU picking
Most abundant sequence picking
feature picking
ASV picking

from amplicon NGS data (repeated features from very noisy data)
(You can use it without a database)

- Per sample (independent of the dataset or other samples). Replicability
- Independent of sequencing depth
- sample composition/distribution (diversity)
  (correct for sequencing specific error)
- Independent of region (removing noise)
- No arbitrary ‘filtering’ parameters.
  reproducibility