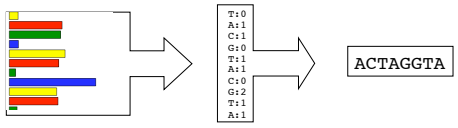
# Elements of amplicon data analysis pipelines: Brief introduction to QIIME

## Initial processing of sequences

- Basecalling
- Filtering: short sequences (<~200bp) tend to be artifacts
- Trimming: the quality of reads is typically worse at the ends
- Masking: low quality areas of sequences are masked with N's
- Chimera detection: Certain artifacts of PCR need to be removed
- Demultiplexing: separate sequences by barcode

## Base Calling

- Alternatives:
  - Native 454 base-caller
  - Denoiser: http://www.microbio.me/denoiser/
  - Pyronoise: http://userweb.eng.gla.ac.uk/christopher.quince//PyroNoise.html
- Issues:
  - Quality of sequence drops towards the end due to error accumulation
  - Long homo-polymers are not always called correctly

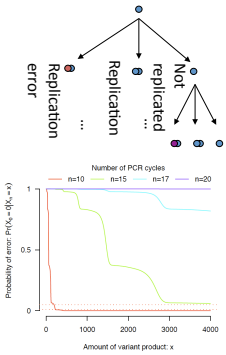



## Problems caused by homo-polymer

- Insertions or deletions
- Non-trivial substitutions
  - True sequence: … AAAAGT …
  - May be called as: … AAAGAT …
- Solutions:
  - Trimming
  - Filtering
  - Masking

## PCR artifacts in sequencing

- Unfaithful replication: errors introduced by the polymerase enzyme (Taq)
- Chimera: "In vitro recombination products"
  - Templates: XXXXXXXXXX and YYYYYYYYYY
  - May result in amplified product: XXXXXYYYYY




Acinas, SG., Sarma-Rupavtarm, R., Klepac-Ceraj, V., Polz., MF. PCR-Induced Sequence Artifacts and Bias: Insights from Comparison of Two 16S rRNA Clone Libraries Constructed from the Same Sample. Appl Environ Microbiol. 2005 December; 71(12): 8966–8969.

## Chimera removal

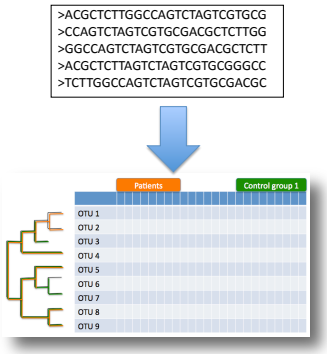
- Tools:
  - Uchime: http://www.drive5.com/uchime/
  - ChimeraSlayer: http://microbiomeutil.sourceforge.net/
  - Perseus: Quince, C., Lanzen, A., Davenport, R., Turnbaugh, P. Removing Noise From Pyrosequenced Amplicons. BMC Bioinformatics 2011, **12**:38.
- Basic idea:
  - Match each sequence to all putative generating sequences: ChimeraSlayer uses a database; Perseus finds 'parents' in the output of sequencing run; Uchime does both.
  - Evaluate the probability that the sequence is a product of two or more generating sequences.

# 16S reference databases

- NCBI Microbial Genomes http://www.ncbi.nlm.nih.gov/genomes/lproks.cgi?view=1
- Ribosomal Database Project: http://rdp.cme.msu.edu/
- Greengenes: http://greengenes.lbl.gov/
- Human Oral Microbiome Database: http://www.homd.org/
- HMP Data Analysis and Coordination Center (DACC): http://hmpdacc.org/

# Defining units of microbiomic analysis Operational Taxonomic Units

- Reference based OTU:
  - Construct OTUs based on matching to a reference taxonomic database
- Phylogeny based OTU:
  - Construct a phylogeny and define OTUs based on monophyly




# Reference based OTU

- Cluster sequences based on similarity
- Assign a taxonomy to each cluster based on matching a representative sequence to a reference database
- Both clustering and assignment rely on sequence matching methods:
  - Sequence alignment: e.g. Blast, Smith-Waterman, Muscle, PyNAST, etc.
  - Fast similarity searches (k-mer): e.g. Desantis TZ, Keller K, Karaoz U, Alekseyenko AV, Singh NN, Brodie EL, Pei Z, Simrank: Rapid and sensitive general-purpose k-mer search tool. Andersen GL, Larsen N. BMC Ecol. 2011 Apr 27;11:11.

# Phylogeny based OTUs

Phylogenetic inference:

- PHYLIP: http://evolution.genetics.washington.edu/phylip.html
- FASTTREE: http://www.microbesonline.org/fasttree/
- BEAST: http://beast.bio.ed.ac.uk/
- Mr.Bayes: http://mrbayes.csit.fsu.edu/

- Likelihood based phylogenetics: Based on models of evolution
  - Maximum likelihood: find the tree that maximizes the likelihood of the data
  - Bayesian methods: explore the distribution of phylogenetic trees, given the observed data and model of evolution
- Distance based: Use a distance between leaves matrix to cluster them into phylogeny
  - Neighbor-joining
  - Unweighted Pair Group Method with Arithmetic mean (UPGMA)

# QIIME: Quantitative Insights Into Microbial Ecology

- Collection of scripts and programs to automate the processing of next generation sequencing data
- Uses most up-to-date tools for specific sub-tasks
- Options for steps of upstream analysis
- Some downstream analysis: not as much control as one has when making analysis themselves in R

http://qiime.sourceforge.net/

# QIIME programs

- PyNAST alignment, tree-building, taxonomy assignment, OTU picking, and other data generation steps
- Alignment, tree-building, taxonomy assignment, OTU picking, and other data generation steps
- Denoising 454 data
- Visualization and plotting
- Supervised learning (very basic)