CAP 5510: Introduction to Bioinformatics

Giri Narasimhan

ECS 254; Phone: x3748

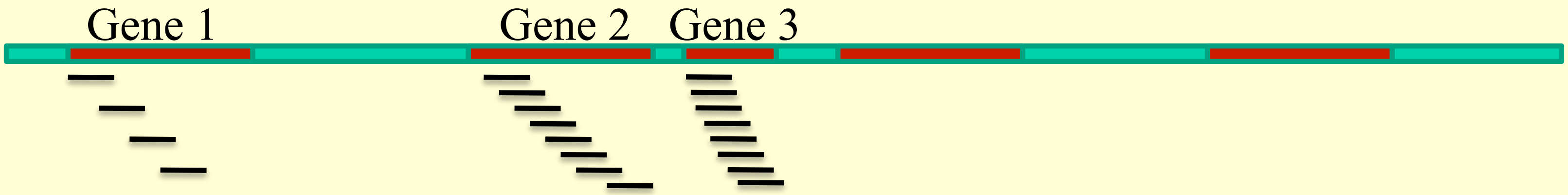
giri@cis.fiu.edu

www.cis.fiu.edu/~giri/teach/BioinfS11.html

Applications of NGS

- RNA-Seq
- ChIP-Seq
- SNP-Seq
- Metagenomics
- Alternative Splicing
- Copy Number Variations (CNV)
- …

RNA-Seq



- Align reads to genes and count
- Assume uniform sampling
  - Count of number of reads mapped per gene is a measure of its expression level
  - Expression of Gene 2 is twice that of Gene 1
  - Expression of Gene 3 is twice that of Gene 2

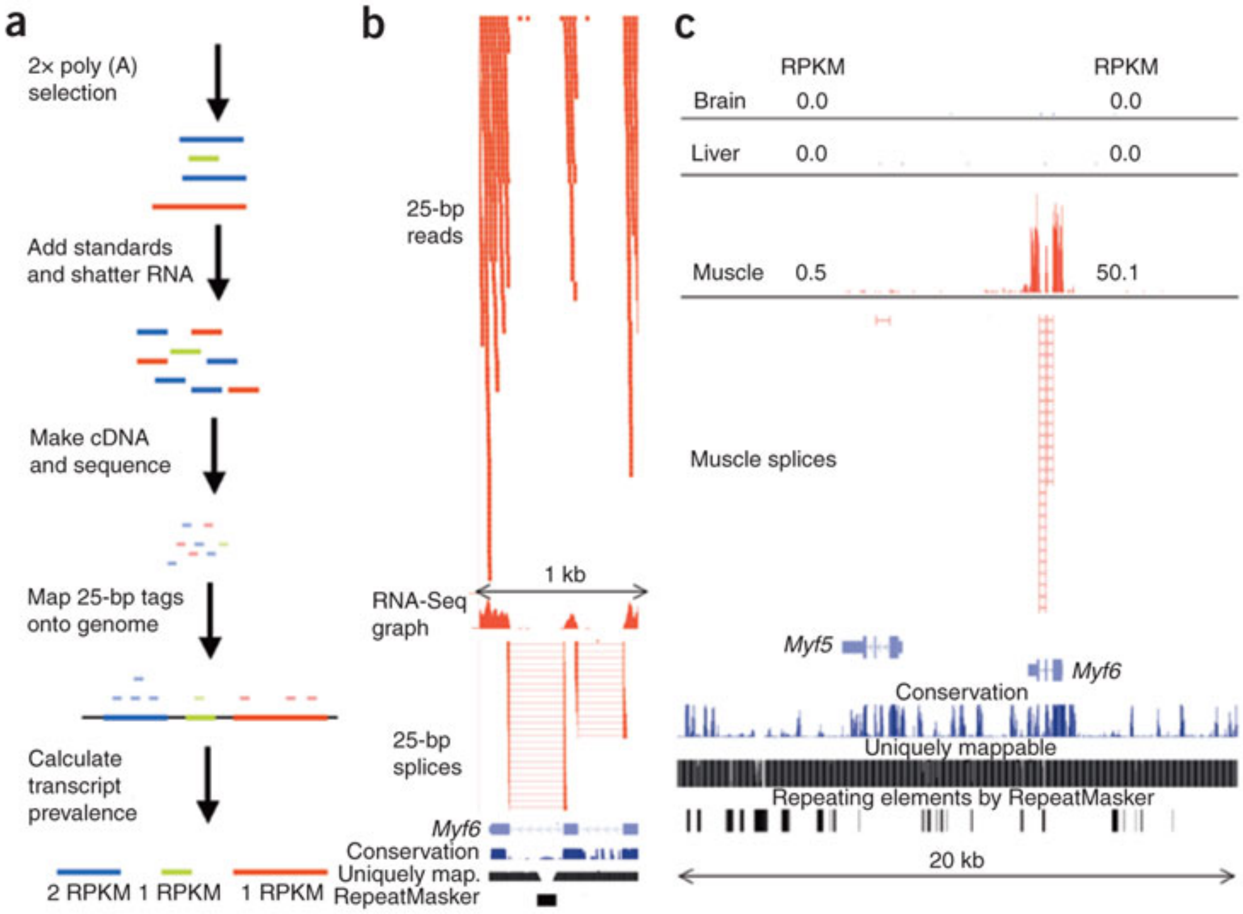
Expression Level of Gene

- RPKM = Ng / (N X L)
  - Ng = Number of reads mapped to gene
  - N = Total number of mapped reads (in millions)
  - L = Length of gene in KB
  - [Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B., Nat Methods. 2008 Jul;5(7):621-8. **Mapping and quantifying mammalian transcriptomes by RNA-Seq.**]

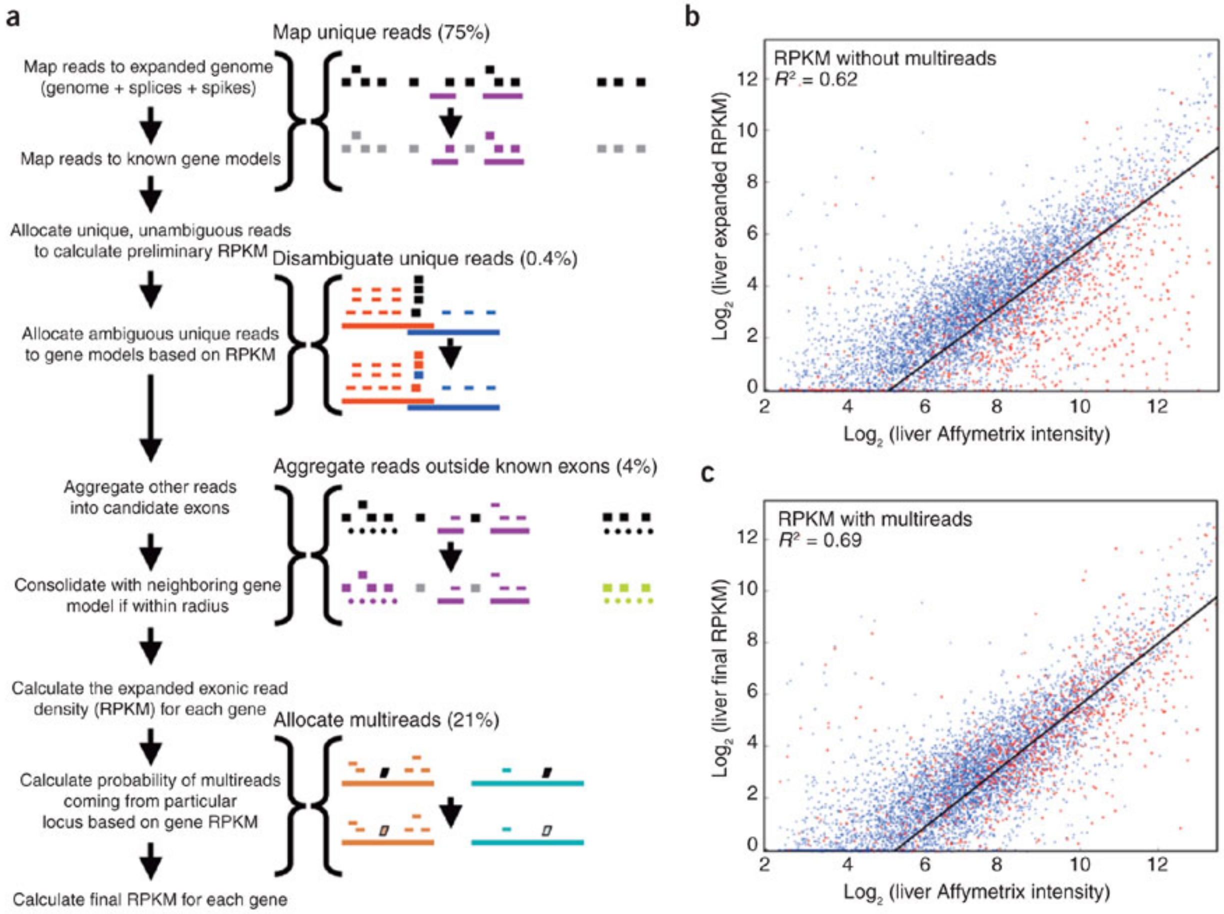
Complications

- Repeat regions
  - Paralogs and other homologous regions in genes
  - Ambiguities in maps
- Introns and Exons
  - Aligning reads to genome is more complex
- Alternative Splicing
- Transcription start site is upstream of ORFs
- Unknown ORFs and Small RNAs
- Other transcripts

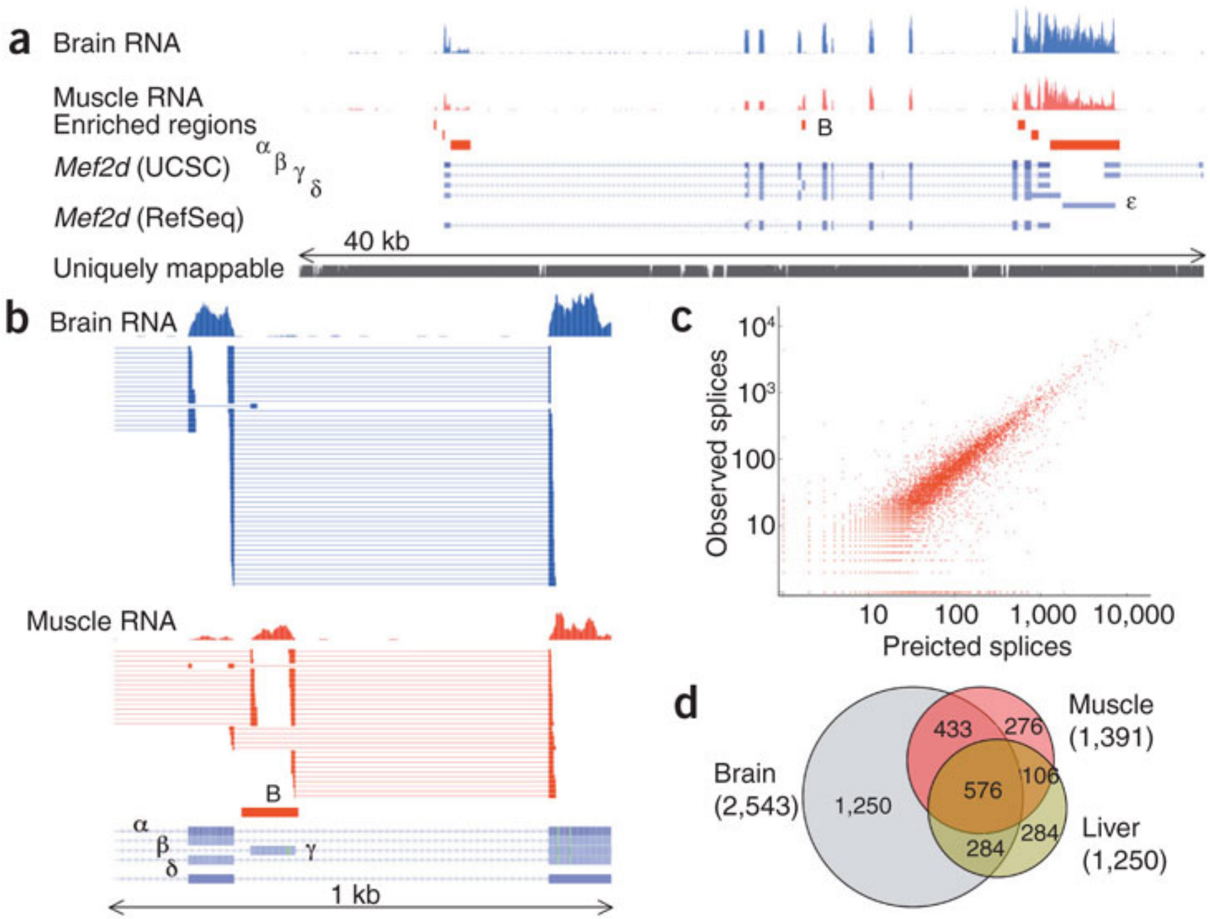
RNA-Seq Procedure



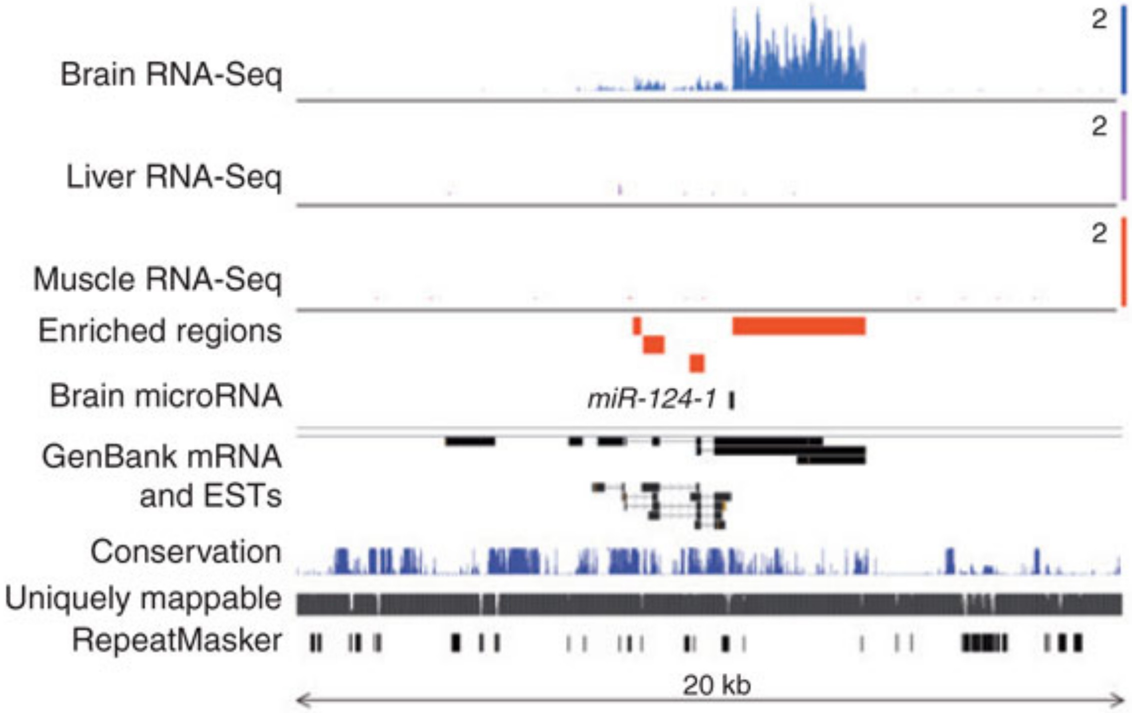
Mapping Reads to Reference


a
Map unique reads (75%)
Map reads to expanded genome (genome + splices + spikes)
Map reads to known gene models
Allocate unique, unambiguous reads to calculate preliminary RPKM
Disambiguate unique reads (0.4%)
Allocate ambiguous unique reads to gene models based on RPKM
Aggregate reads outside known exons (4%)
Aggregate other reads into candidate exons
Consolidate with neighboring gene model if within radius
Calculate the expanded exonic read density (RPKM) for each gene
Allocate multireads (21%)
Calculate probability of multireads coming from particular locus based on gene RPKM
Calculate final RPKM for each gene
b
RPKM without multireads
$R^2 = 0.62$
Log$_2$ (liver expanded RPKM)
Log$_2$ (liver Affymetrix intensity)
0
2
4
6
8
10
12
c
RPKM with multireads
$R^2 = 0.69$
Log$_2$ (liver final RPKM)
Log$_2$ (liver Affymetrix intensity)

Alternative Splicing



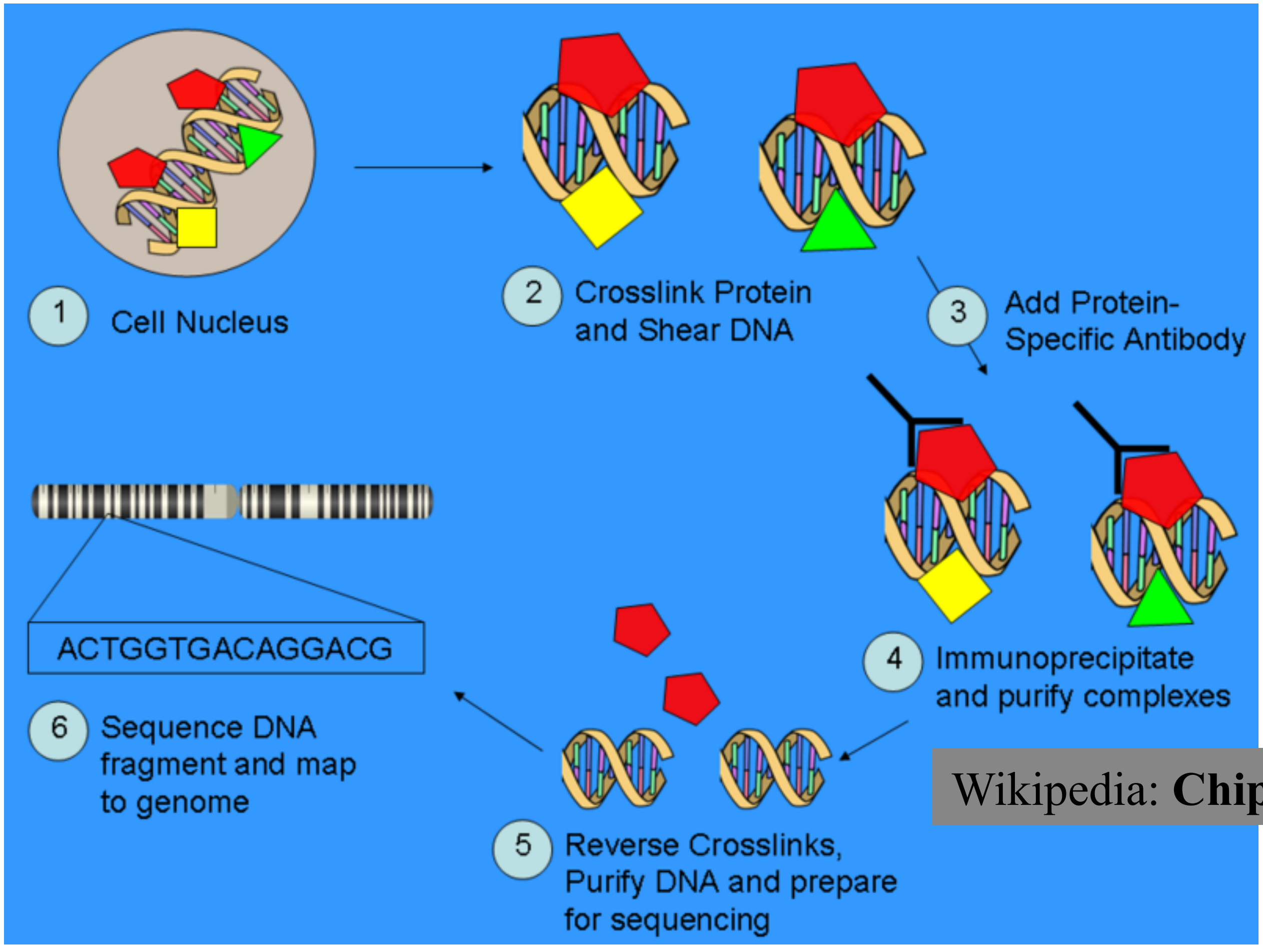
microRNA


Brain RNA-Seq
Liver RNA-Seq
Muscle RNA-Seq
Enriched regions
Brain microRNA
miR-124-1
GenBank mRNA and ESTs
Conservation
Uniquely mappable
RepeatMasker
2
2
2
20 kb

Chromatin Immunoprecipitation

- Useful for pinpointing location of TFBS for TF
- High-throughput method to find all binding sites for a specific TF under specific conditions
- Identify sites using
  - ChIP-on-chip (Microarray technique)
  - ChIP-Seq (Sequencing technique)
- Problems: TFs bind to specific TFBS only under specific conditions – hard to predict

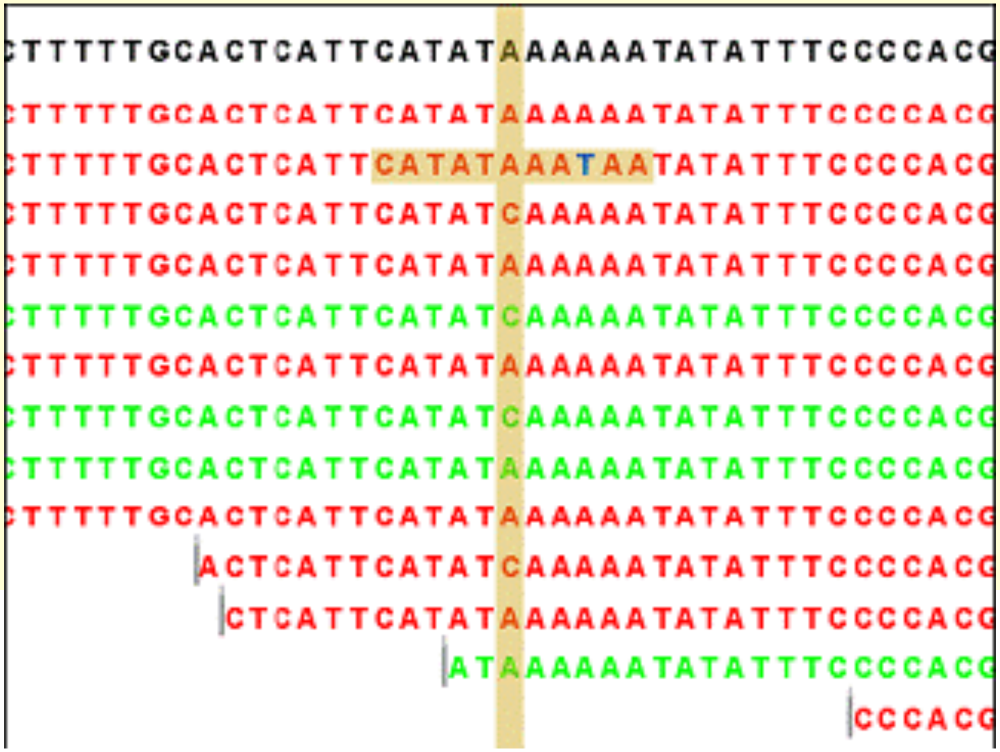
ChIP-Seq


1 Cell Nucleus
2 Crosslink Protein and Shear DNA
3 Add Protein-Specific Antibody
4 Immunoprecipitate and purify complexes
5 Reverse Crosslinks, Purify DNA and prepare for sequencing
6 Sequence DNA fragment and map to genome
ACTGGTGACAGGACG

Wikipedia: **Chip-Sequencing**

SNP-Seq

- ❑ Align reads and look for differences
  - Differences to reference
    - Align reads to reference sequence first
  - Differences within reads
  - Differences between samples or sets of reads



```
CTTTTTGCACTCATTCATATAAAAAATATATTTCCCCACG

CTTTTTGCACTCATTCATATAAAAAATATATTTCCCCACG
CTTTTTGCACTCATTCATATAAATAATATATTTCCCCACG
CTTTTTGCACTCATTCATATCAAAAATATATTTCCCCACG
CTTTTTGCACTCATTCATATAAAAAATATATTTCCCCACG
CTTTTTGCACTCATTCATATCAAAAATATATTTCCCCACG
CTTTTTGCACTCATTCATATAAAAAATATATTTCCCCACG
CTTTTTGCACTCATTCATATCAAAAATATATTTCCCCACG
CTTTTTGCACTCATTCATATAAAAAATATATTTCCCCACG
CTTTTTGCACTCATTCATATAAAAAATATATTTCCCCACG
        ACTCATTCATATCAAAAATATATTTCCCCACG
         CTCATTCATATAAAAAATATATTTCCCCACG
                  ATAAAAAATATATTTCCCCACG
                                  CCCACG
```

Environmental Microbiology

- Conventional methods
  - Culture, then identify
    - Slow, expensive, labor intensive, unculturable microbes
  - PCR-based length heterogeneity studies
- Microarray-based methods
  - Unique probes for organisms (e.g., Virochip)
    - Only works for sequenced regions of known organisms
- NGS-based methods

Metagenomics

- Detect known pathogens
- Diversity
  - Identity of individual species not needed
- Functional profile of community

NGS-based method

- ❑ Map reads against appropriate database
- ❑ Identify closest hits for each read
- ❑ Generate contigs
- ❑ Generate abundance information
- ❑ Clustering of reads can be beneficial to estimate abundance