

## REAL TIME METAGENOMICS

Rob Edwards

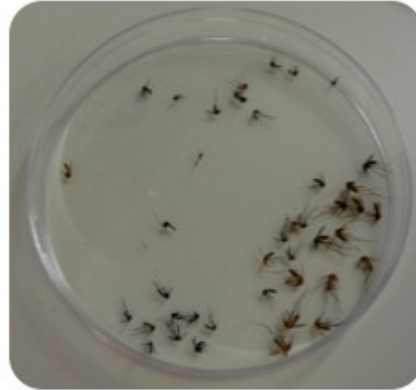
San Diego State University, San Diego, CA  
Argonne National Laboratory, Argonne, IL

Terry Disz, Bob Olsen, Ross Overbeek – Argonne  
Daniel Cuevas, Josh Hoffman – SDSU

# How to do metagenomics



Collect mosquitoes



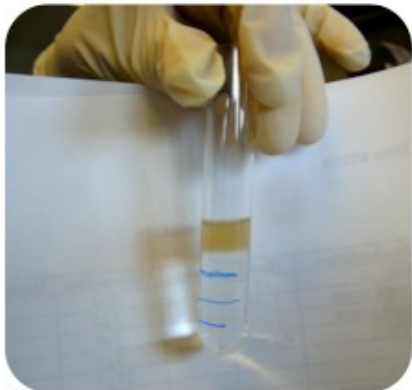
Separate or pool mosquitoes



Homogenize mosquitoes



Filter out eukaryotic cells



Density-gradient centrifugation



DNA extraction & amplification



Sequencing on GS 20/GS FLX



Bioinformatics

# THE SEED FAMILY

## THE SEED Environmental, Viral, Bacterial, Archaeal, and Eukaryal Genome Interpretation

The SEED Viewer interface displays a table of genes with columns for Name, Location, Status, Feature, and Type. The table lists various genes such as *SEED00000001*, *SEED00000002*, etc., with their corresponding locations and features.

The NMPDR website displays a list of pathogens under the heading "National Microbial Pathogen Data Resource". The list includes various organisms such as *Escherichia coli*, *Salmonella enterica*, etc., with brief descriptions of their characteristics and associated diseases.

**NMPDR**  
Display of complete genomes  
Focus on pathogenesis

The RAST website displays a list of genomes under the heading "RAST (Rapid Annotation using Subsystem Technology)". The list includes various organisms such as *Escherichia coli*, *Salmonella enterica*, etc., with their corresponding genome sizes and annotations.

**RAST**  
Annotation and analysis of  
complete genomes

The MG-RAST website displays a list of metagenomes under the heading "MG-RAST (Metagenome Rapid Annotation using Subsystem Technology)". The list includes various metagenomes such as *Human Gut*, *Soil*, etc., with their corresponding genome sizes and annotations.

**MG-RAST**  
Annotation and analysis of  
metagenomes

# METAGENOME SEQUENCE ANALYSIS 2006 - 2009

## Total:

3,565 metagenomes

334,168,924 sequences

88,311,139,391 bp (88 Gbp)

Largest metagenome: 729 Mbp, 11,719,618 reads

## Public:

394 Metagenomes

54,414,564 sequences

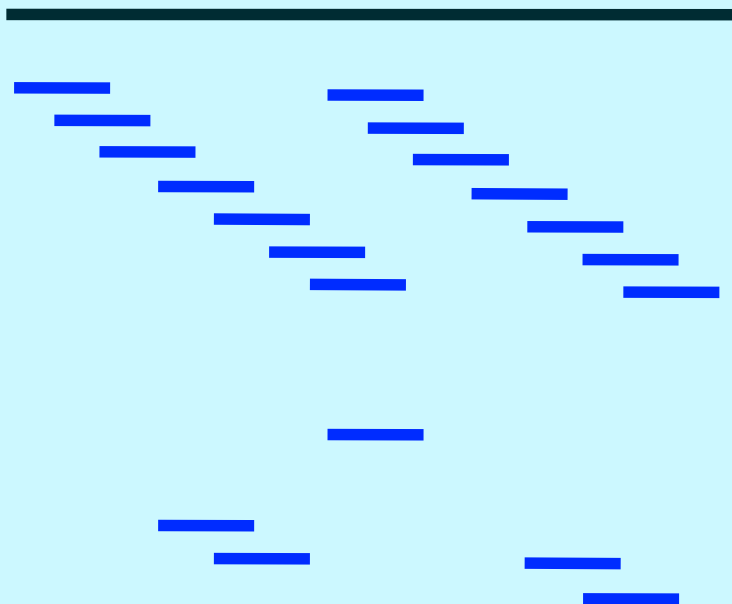
22,234,298,797 bp (22 Gbp)

Compute time (on a single CPU):

1,677,911 hours = 69,912 days = 191 years

# How BLASTX Works

**Map**

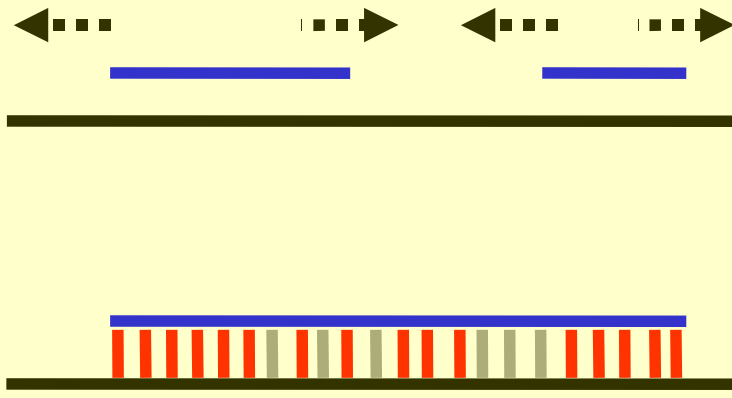


Protein sequence

Find all words in the protein sequence (>3 letters by default)

Filter for words above a threshold

**Reduce**



Extend while score is above another threshold

Calculate & report final score for alignment high scoring pairs

# Identify unique oligos

## FIGFam 57

... LQRTVPAPPAERQAALWPCV ...  
... LQRTVPAPPAERQAALWPCV ...  
... LQKSVPAFPAERQAALWPCV ...  
... L-RSVPAPPAERQAALWPCV ...  
... L-RPVPAPPAERQALLWHCV ...  
... VQKTVPAPPAERQAILWHCV ...  
... VQRSVPVF-AERQAVLWHCV ...

Oligomer is unique to this family

Doesn't have to contain all members

All members should be represented by oligos

Suffix tree to rapidly search

# SEED k-mer Annotation Server

## Parameters:

length of k-mer

Number of k-mer hits per sequence

## Returns:

Annotation (function)

Most likely organism // taxonomic level

Likelihood

YAML // SOAP Servers: <http://servers.nmpdr.org>