**Class Genome Evolution 17.01.2019**

Evolutionary genome biology to understand evolution of genome and genome function

3 courses in 1:

**Genomics course -> in last 2-3 years single lab can sequence their own genome**

1. How to sequence, assemble and annotate genomic data PART 1. Work with public available data or own data.
2. Comand line, wrangle data file, HPC, version control PART 2.
3. Comparative part, phylogenetics PART 3

Syllabus is living document throughout the course -> he wants other profs to grab his syllabus -> each class has a reading -> students present paper for 15 mins -> challenges of working with de novo sequence data AND comparative methods

Get raw reads from genome -> assemble -> do annotations -> analyze

Use git:

Final project will be in git and github to share docs -> projects are open to everyone in the course throughout the course

Fork a github repo -> make a copy of this and work on that -> can browse other ppls project AND do project submission.

Some work goes into HPC

Ben can help us set up accounts.

Systematics ->

Evolution of morphology AND physiology!

Evo biologist interest og history of genomes

When was there diversification of genes ->

How often have characteristic arisen during evolution

1. History of genome evolution (patterns and processes)
2. Phenotype evolution -> evolution of morphology OR physiology -> understand the evolution of the sequences of the genes and the evolution of the trait of interest
3. Understand genome function via comparative approaches
4. Try to connect genotypes and phrnotypes -> morphology OR a trait -> what genes involved in this -> QTL is classical genetics to understand functional links phenotypes genotypes -> phylogenetics will be new genetics -> do the same within a clade of organisms -> monophylogenetics group zB some are bioluminescent -> query species and find evolutionary shifts. Phylogenetics is the new genetics -> look at broader scale character data -> ELTON TRAITS ->
5. Ecological question -> genome by environmental -> how much variation is heritable or just response to environment. Environmental interactions. GBE gene by environmental interactions. Whats the environmental impact of a set of interests
6. Understand by comparing genomes of extant organisms that live across dif environmental gradients -> how will individuals respond to climate change -> very applied
7. Applied stuff zB understand wild variation in traits -> pharmaceutical or genes pool into wild relatives of cops. -> industrial applications.

Paper 1: PNAS 2018 pairwise species comparisons: Casey Dunn PNAS 2018

Felsenstein 1985 -> set and questions of problems.

Functional genomics data collected at broad scale \_> promoter location, protein abundance -> collected at broad scale -> comparisons cross species -> what is the diversity of these traits.

Paper 2018: Compared analized studies of functional genomics with phylogenetic methods -> does it matter ? ->

Most published functional genomic studies do pairwise comparisons: -> 7 species but there are multiple genes.

Orthologs: genes related from shared common ancestor.

Single species comparison -> not independent -> each branch of phylogeny Is included in multiple pairwise comparisons.

One small shift might appear to be a replicated result.

You diluted the signal of that change with pairwise comparisos -> problem when NOT accounting for phylogenies and instead doing multiple non independent pairwise comparisons ->

Need to know branch length. -> use fossils to calibrate speciation nodes ->

Sensitivity analysis.

Climate vhange things -> look at a monophyletic clade that has generalist and specialists -> we assume generalists more likely to be buffering climate change -> BUT lets see if there is a phylogenetic signal to this?

To genes are ortholog -> I know the function of one, the other gene likely is similar -> hypothesis -> orthologs are more similar to each other than paralog -> gene duplication leads to more rapid change in the attribute I look at than speciation.

NCBI -> get gene tree.

The method they tought was looking at trait evolution is just reflecting the phylogeny and analyzing the phylogeny. Independent contrast often a limiting case for methods.

Einstein -> greatest thing you can hope for is that your grand theory becomes the limiting thing of another grant theory.

2 ideas -> Eidolon OR 1 monophyletic clade with generalists and specialists -> climate change resistance.

Do what authors said they did and come to the same conclusion than they did -> THEN I can do posthoc analysis.

Plos computational biology paper had a git repository ->

Post it Yellow -> when im working

Post it Yellow -> when im done move it down.

Post it Red -> when I have a problem.

Atom browser.

Evolutionary question of genome function -> need to apply phylogeny data tools

Compare genomes -> stats differenly

Look for conserved elements

docker run -it rocker/rstudio /bin/bash

# Run the R studio container and in there run the command BASH

DOCKER: genomic would be virtually unreproducible -> run analysis in a container with analysis code and stuff and share it with people ->

24.01.2019

Add on issue tracker:

TLR genes

Send papers not just link also how they are formatted ->

Nucleotide has sugar and polymerase -> energy driving process is cleavage of triphosphate -> cleave away sugar and phosphase.

**Illumina Sequencer :** attaches nucleotide and then takes bunch of images -> raw data, looks at color of dots -> where is each dots which correspond to cluster of identical molecules

Error: Phasing error: -> sometimes polymerase doesn’t add something, template might be bad nucleotides used in PCR are bad. ->

Data Analysis: align back to reference genome

Illumina: raw reads are of high quality > 99% accuracy.

If we don’t do sample preparation, we get bad quality data.

If I want 150bp on each end -> do 150 bp of sequencing on both sides.

Ilumina -> Need adapters to prepare the sequence data

PacBio: No synchronized sequencing, every single well collects at diferent rate.

Flowcell -> each well has a single polymerase, that has a strand threated through -> illuminates and measures residence time of single molecules at high resolution -> PacBIO engineers them.

What destrozs polzmerase -> getting it anchored and stil be the right configuration -> hard

Polymerase at the bottom with epoxysilene chemistry to link glass with nucleotides.

Is fluorescent -> excited wit laser.

Pac bio -> ligate hairpin loops to the end -> end up with circular DNA molecule.

Can get really high coverage of a single molecule.

Error rate is 10% each read is not very accurate but each read gives you the structure (adapter, reverse read, adapter, forward read) -> stack up lower quality read and stacked up get high quality reads through this circular concensus.

**This is easier to assemble than Illumina (illumine parts of the genome come out at different error.)**

For cDNA for transcripts is great, get lots of copy. ISOSEC thecjnology for cDNA.

Oxford Nanopore: DNA mixed with enzyme -> make a complex -> 1 base a time goes through nanopore -> disrupts electrical current ->

Sequencing cell is plastic surface with holes -> engineered protein nucleopore complexes sitting in holes -> is ibonic (engineered proteins and plastic) -> apply voltage -> have electrode on each pore complex -> can apply a voltage across and keep it constant and measure current -> current determine by ion flux moving through pores -> bases have different physical confirmations and charges -> disrupts ion curents -> can measure this disruption. -> current in Amps gets translated into basepairs -> digital signal processing and turne into squiggle.   
Highly error prone

Ilumina -> Average over population of molecules -> Take majority signal, doesn’t matter if very few error prone.

Economy move from batch processing to stream processes.

Discrepancy instrument sizes -> Illumina and pac bio are glorified microscopes with many liquids.

data out of oxford nanopores is electric signal. Is a single signal -> leverages advance of microchip and semiconductor industry.

Nanopore has no nucleotide, illumina and pacbio depend on highly engineered and pure nucleotides.

Ebola outbreak: Not exporting samples and take porable sequencers in the field. GOOD vs international propotcal transport -> NOW I can just sequence and leave with a harddrive ->

Some species is really hard to get molecular DNA and I need a lot of molecular material often.

Tomato talk oxford nanopore ->

Pacbio: Huge and very expensive. Is the best at the moment

But everyone roots for Nanopore.

Nanopore -> is like a faucet -> you keep it open till your done.

Ebola:

<https://medicine.yale.edu/intmed/people/onyema_ogbuagu.profile>

<https://news.yale.edu/2018/02/23/after-ebola-crisis-yale-works-strengthen-liberian-health-system>

<https://news.yale.edu/2014/12/16/yale-researchers-reveal-ebola-virus-spreads-social-clusters>

<https://news.yale.edu/2015/11/20/emerging-epidemics-experts-discuss-ebola-and-other-infectious-disease-threats>

Climate change misinformation -> XXX

Join yale climate change information AND yale data driven