

Ecological Genomics Notes

P. Alexander Burnham

January 23, 2017

January 23, 2017:

Info Update: Mellisa

- Advantages in Seq tech.
- Range of applications:
 - WGS (whole genome sequencing)
 - RNAseq (sequencing RNA converted to cDNA)
 - Targeted capture seq. (string of probes mixed with sample, pulls immune related genes from organism and seq just those)
 - ChIPseq (chromatin immunoprecipitation seq, recognizes and antibody and pulls out all DNA bound to that protein)

Why one or the other?

Genetic variation

- phenotypes

number of samples

- population
- individual
- comparative studies
- model or note

Demographic history

Adaptive genetic variation

gene expression var.

length of reads

number of reads

distribution

Reads:

- short = 50bp
- long 100 bp, 150 bp, 300 bp (miseq)
- 10,000-60,000bp = SMRT

Single vs. paired end

- General library Prep. Workflow
 - extraction (DNA, RNA -> to cDNA)
 - fragment sample
 - ligate adaptors (individual barcodes)
 - add seq. adaptors

Reduced Rep

- RNA -> coding
- GBS/RAD-seq
- near restriction sites
- Sequencing-by-synthesis (SBS)
 - bridge amp
 - cluster gen.
 - labeled dNTP (ATCG)
- Other Technologies
- Learning Activity

Human Genome Project (2001-2003)

- ABI = Sanger
- 15 years
- 1 genome (one person)
- \$3 billion

Uses PCR and sequences broken by faulty base pairs to work backwards

2014 X-Ten releases

- HiSeq by Illumina (look up video of how it works)
- 1 day
- 45 whole genomes
- \$1000 bucks each

Sheet of glass with 8 lanes with flow cells... (look up)

Take home messages

- Likely using Illumina seq. (just library tech. changes) usually SBS
- adaptors are markers (barcodes) used to identify samples during sequencing
 - first thing that's sample and gives ID barcode = seq adaptor
 - alago is a sequence of DNA that is attached to plate binds to sample
- Model vs. non Model:
 - short reads (assembly to create long sequence based on short reads that shift)
 - denovo assembly -> computer program (added variability with mixed sample)
 - 15% error for SMRT can be reduced to less than 1 with repeated passes
 - illumina is much smaller but more accurate (0.05% error)
 - combine the two to have a higher degree of confidence

Paper Discussion:

Three advances in biology:

- Modern Synthesis (evolution (Darwin) and population genetics (Mendel))
- Watson and Crick (molecular biology DNA)
- Omics Era (genomics, proteomics etc.)

What do we think about this?

- Phylogenies -> reducing error bars or reshaping question?
- Do you throw out experimental design and scientific method for large scale shotgun blast sequencing?
- > can still be used to do hypothesis driven science, but also something to be said for sending a telescope into deep space.
- Most journals and funding sources require all seq data be made public - leads to a storage space issue
-> genbank ran out of space and needed emergency funding from congress.
- data are not reviewed very well some gen bank sequences are kinda rough
- reference genomes are a sample size of 1 so it might not be representative

On Wednesday two papers to talk about:

- 1) where do genomics data have limitations and why we use them
- 2) discussion leaders (sign up via blackboard)
- 3) next week talk about some of library preps (four of them) four different update people

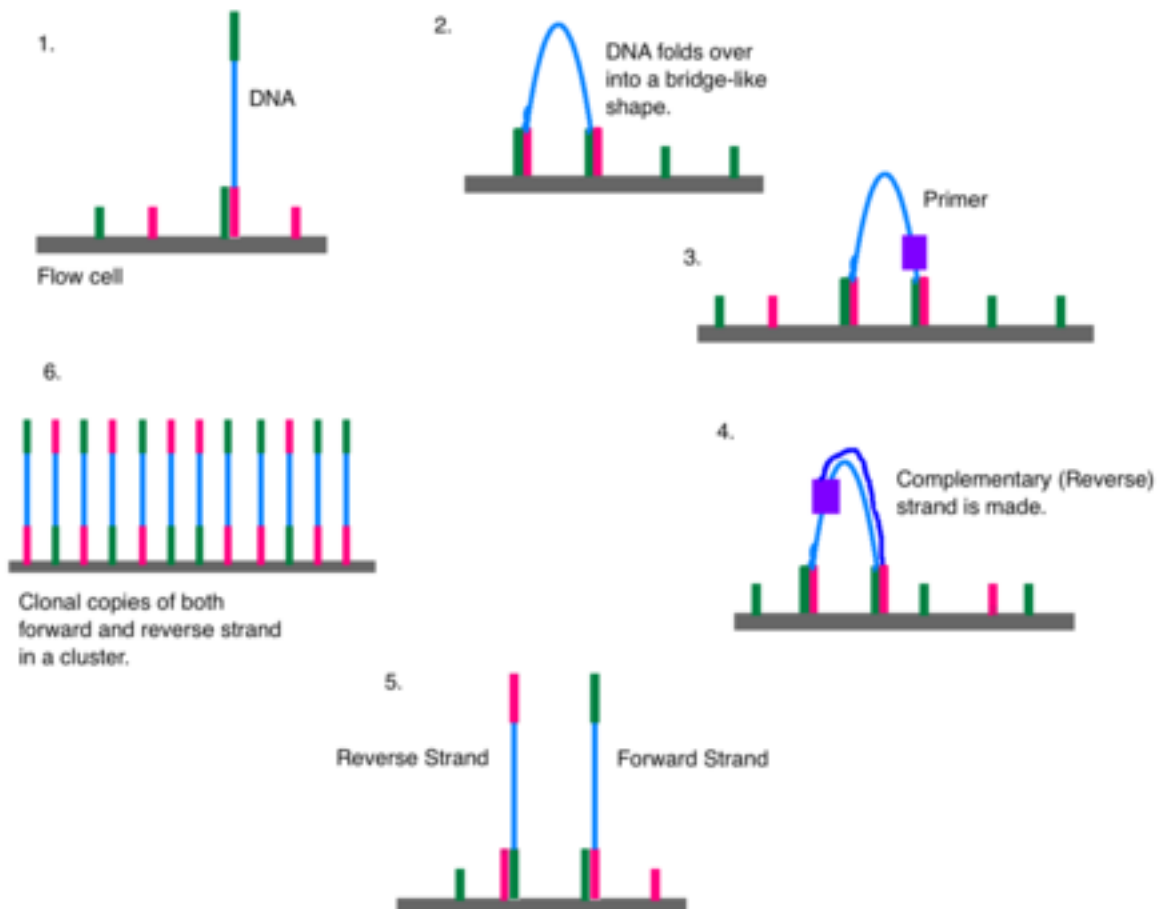


Figure 1: Illumina Sequencing

January 25, 2017:

Info Update: Steve

Outline:

- What are QTN
- Quantitative Genetics
 - theory of adaptive traits
 - * V_a
 - * h^2
- Methods
 - linkage mapping
 - GWAS
 - selection scans

Notes on topics in outline:

- QTN = “quantitative trait nucleotides”
 - Flowering times (quantitative traits)
 - Flower color (more discrete - mendelian major effects traits)
 - Thermal tolerance
 - venom potency
 - defense compounds
 - toxin tolerance
 - drought tolerance
 - altitude tolerance

Looking at haldane and fishers work on Quant genetics:

- additive effect at each locus and additive effect average out Additive genetic variance or V_a
- the additive portion of the additive variance or phenotypic variance equals heritability

$$(V_a/V_p = h^2)$$

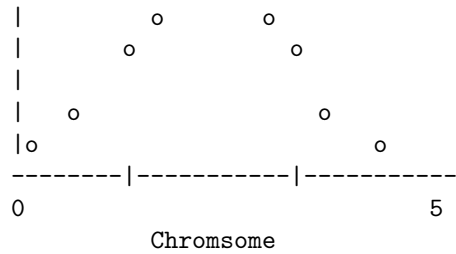
- the difference between phenotypic variance between loci is additive effect and average out to average effect which is usually a linear function
- most populations are at the value of trait where fitness is max (local adaptation) - mutation with small effect on trait (fisher thought most important mutation 50% of being beneficial)
- large effect mutation -> tend not to be selected -> to extreme and usually cause problems - best traits climb to the fitness peak through small incremental steps

Three main methods:

1) QTL mapping (linkage mapping) - assumed that two parents are homozygous (for continuous traits) - F_1 is heterozygous but no recombination - mate the F_1 generation and get one generation of recombination for these traits - **The Idea of QTL** use markers linked to chromosomes, crossing (over many generations)

linkage

| QTN
| o

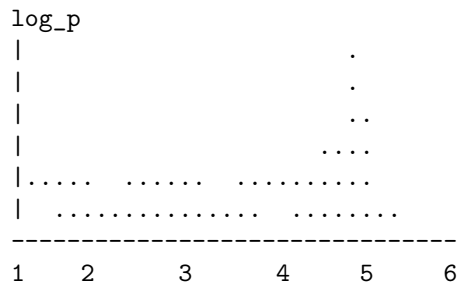


2) GWAS:

$$y_{trait} = \mu_{intercept} + \beta_{effectsize} * SNP_i + covariates_{populationsize}$$

collect data run regressions (millions)

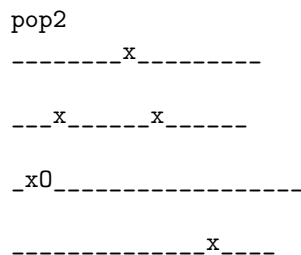
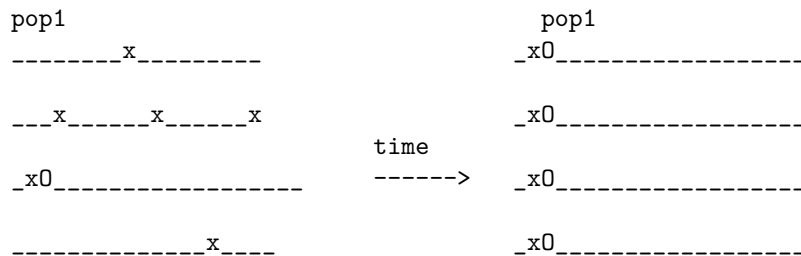
**Manhattan plot



3) Selection SCANS (based on selection sweeps):

x= SNPs

selection favors mutation next to 0



- lose variation as everything favors the mutation

- diversity is higher in pop 2
- FST is populations struction high FST will be for pop2 low for pop 1 after selction
 - we study the few large effect size mutations for QTL and QTN that we study because they are easy to find
 - majority are small but hard to find

PAPER DISCUSSION:

forward vs reverse approach

apply scientific method and don't assume that evrything is representing the sample of arge mutations we have mostly been studying

- 1) large effect allele mutation (mendelian)
- 2) basic theory in terms of gen. theoary neither requires or predicts the importance of large mutations
- 3) making assumption that the little parts we find act the same way the big parts act the sam way and thats probably not true

They will argue that there will be knowledge to found using this information

DISCUSION OF GROUP PROJECT

Notes on Seastar Wasting Disease:

BACKGROUND ON DISEASE

- High mortaility (between 70-100%)
- many species impacted (not species specific)
- loss of legs withing hours or days, turns inside out!
- turn into goo
- Alaska to Baha
- First report in 2012, less severe currently (reported in the 1970s and 80s in isolation)
- study (Hewson et al., 2014) linked Densovirus implicated but not very confidnet
- potentially something in microbiom can lower host immunity and allow for opportunistic infection by pathogen
- Seastar Disease Outbreak Distribution

INFO ON WHAT DATA WE HAVE

- focusing on *Pisaster ercaceous*
- RNA extractions and polly a tail and sequenced mRNA on 3 illumina hiseq lanes
- also pulled our ribosmal 16s structure
- none of samples pos for densavirus (qPCR) positive control
- DNA virus
- is it cyclical
 - 1) never been so extreme before
 - 2) do go through booms and busts though
- All arrived healthy
- put into aquaria with artifial seawater (kept at 12°C)
- From same site
- were no infected; some developed SWD on there own
- first and only time course study of this kind

Hypothesis ideas:

My idea: (Environmental Component)

- Intertidal VS Subtidal
 - genetic differences (local adapt.) related to susceptibility
 - gene expression (immune related genes higher in SUB group)

Other ideas:

- Resistance genes (stayed healthy or recovered)
- w/in Intertidal -> gen. differences between three groups
- both groups, Microbiome -> H vs S (differences (yes or no))
- Microbiome through time
- h^2 of microbiome
 - day as replicate
 - relatedness VS microbiome trait
- immune related genes (Sam)
 - what kind of pathogen based which genes are expressed
 - which are the immune relatedness genes (conserved in other species)
 - potential problems -> viruses that are opportunistic and enter at tail end of infection and immune response genes still expressed (confounded?)