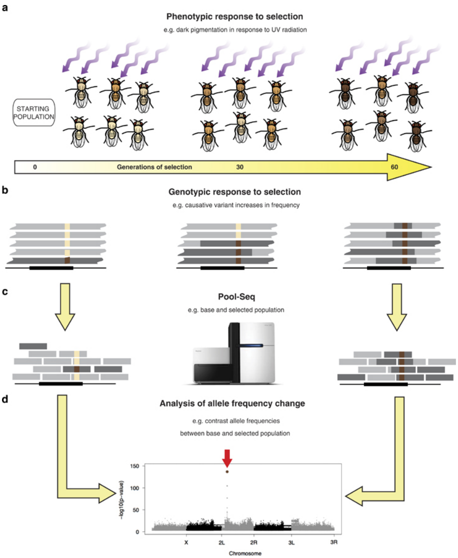
Guppy ddRAD data analysis overview

Alexander Jueterbock

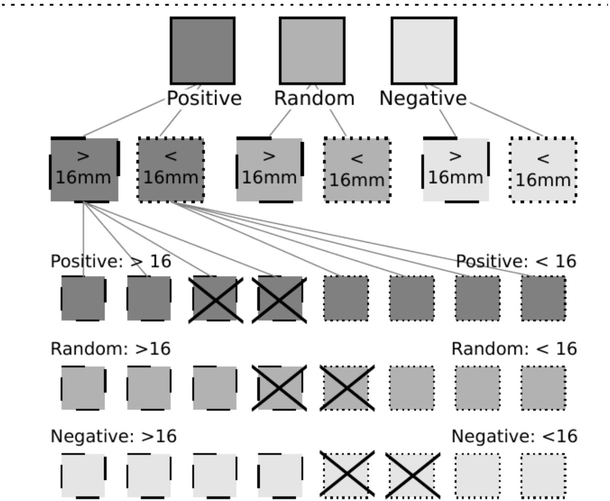
Nov 2014

Evolve and Resequence (E&R) studies


a
Phenotypic response to selection
e.g. dark pigmentation in response to UV radiation
STARTING POPULATION
0
Generations of selection
30
60
b
Genotypic response to selection
e.g. causative variant increases in frequency
c
Pool-Seq
e.g. base and selected population
d
Analysis of allele frequency change
e.g. contrast allele frequencies
between base and selected population
150
100
50
0
−log10(p-value)
X
2L
2R
3L
3R
Chromosome

Review in Schlotterer et al. (2014) Heredity

Experiment overivew



Pooling

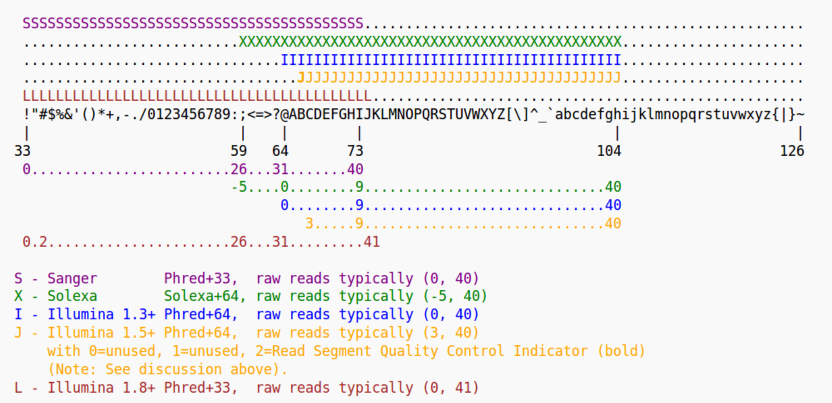
- Critical: equimolar concentrations of individuals expected
- Best to pool individuals at the very latest step
- Recommended: >40 individuals/pool
  - Higher numbers decrease
    - sampling error
    - unequal representation of individuals in the pool
  - But more difficult to discriminate minor allele frequencies from sequencing errors

Fastq output

One example sequence
(Quality scores are ASCII encoded)

```
@SEQ_ID
GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTC
+
!’’*((((***+))%%%++)(%%%%).1***-+*’’))**55CCF>>>>>>CC
```

Phred quality scores in Fastq



```
 SSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSS.....................................................
 ..........................XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX......................
 ...............................IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII......................
 .................................JJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJ......................
 LLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLL....................................................
 !"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrstuvwxyz{|}~
 |                         |    |        |                              |                     |
33                        59   64       73                            104                   126
 0........................26...31.......40
                          -5....0........9.............................40
                                0........9.............................40
                                   3.....9.............................40
 0.2......................26...31.........41

S - Sanger        Phred+33,  raw reads typically (0, 40)
X - Solexa        Solexa+64, raw reads typically (-5, 40)
I - Illumina 1.3+ Phred+64,  raw reads typically (0, 40)
J - Illumina 1.5+ Phred+64,  raw reads typically (3, 40)
    with 0=unused, 1=unused, 2=Read Segment Quality Control Indicator (bold)
    (Note: See discussion above).
L - Illumina 1.8+ Phred+33,  raw reads typically (0, 41)
```

Figure : Quality overview of raw reads

Quick and dirty analysis

- DDocent Pipeline (not for pooled and replicted data)
- STACKS (not tried, not targeted for pooled and replicated data)

Demultiplexing by barcode

- 'process radtags' from STACKS did not work well on our data
- DDemux used
- Unpaired reads are discarded
- Barcodes are removed

Paired read:

ADAPTER1 AATTAAATTCNNNNCCG ADAPTER2
ADAPTER1 TTAATTTAAGNNNNGGC ADAPTER2

First read: AATTCNNN
Second read: CGGNNN

Barcode
Restriction enzyme overhang (EcoRI and MspI)
Target sequence

Trimming 1

- I use TrimGalore!
  - Uses 'cutadapt' for adapter trimming
  - Can handle paired end reads
  - removes orphan reads (reads without a pair)
- Removing internal adapters (0.1%-0.2% of reads)
  - Can have deviating internal barcodes

Before: AATTCNNADAPTER1AATTAAATTCNNN

After: AATTCNN

- Minimum read length? 20bp default, I set 50bp (single sequence) as the lower limit, so 100bp paired end

Trimming 2

- Cut off the Restriction enzyme overhangs (5bp from read 1 and 3bp from read2)

First read: AATTC NNN

Second read: CGG NNN

- Trim bases with a Phred quality score <20 (99% base call accurracy, Phred+33 encoding)

Phred Score	Probability of incorrect base	Base call accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1000	99.9%

Read qualities before trimming

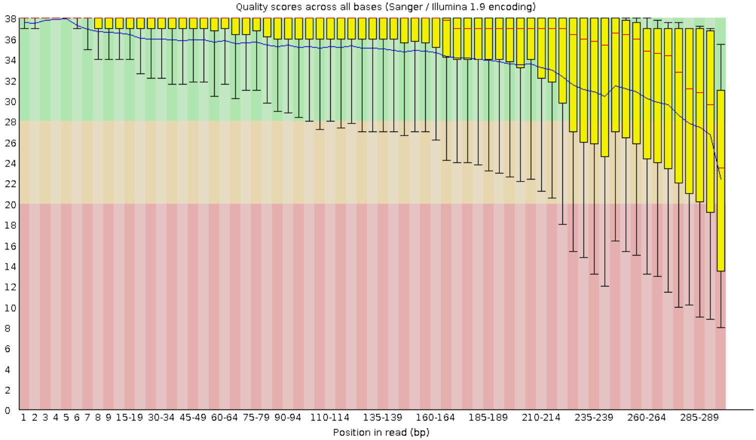


Figure : Quality before trimming

Read qualities after trimming

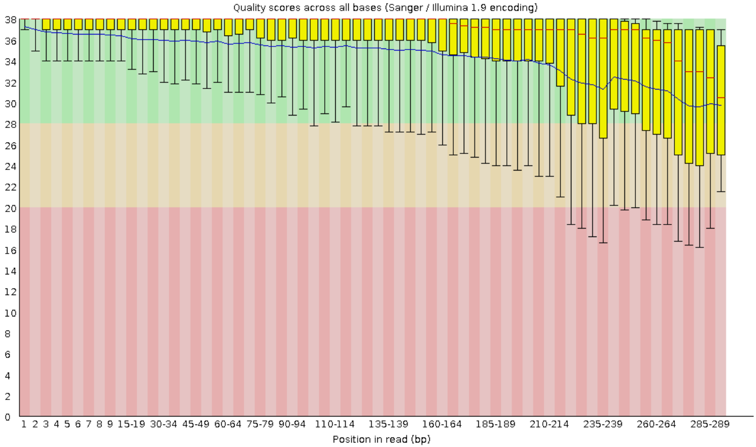


Figure : Quality after trimming

Number of Sequences

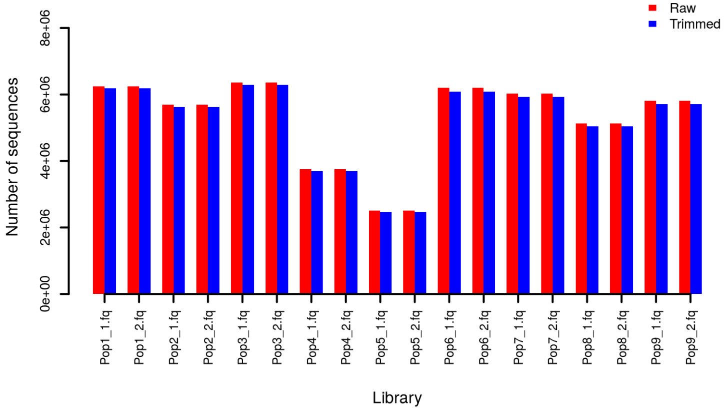


Figure : Number of sequences

CG Percentage

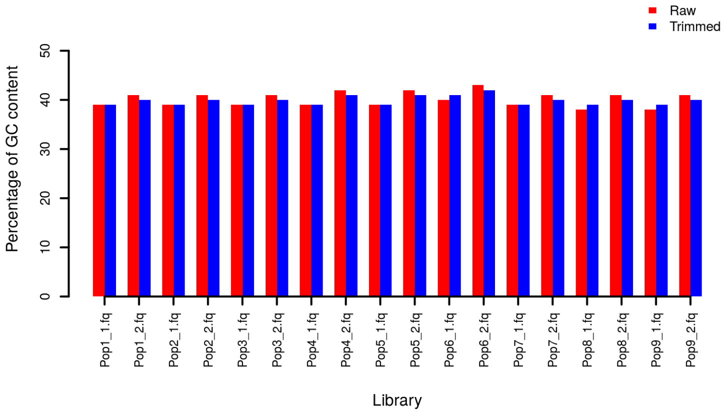


Figure : Percentage of GC

Quality of raw reads

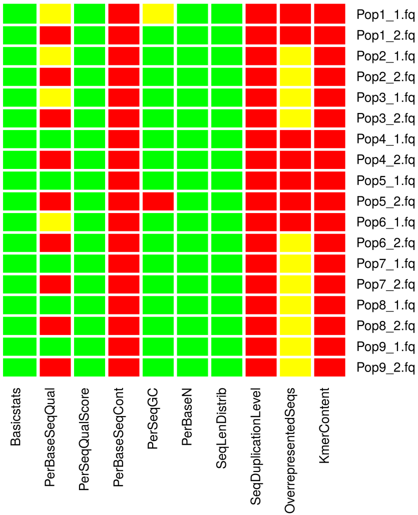
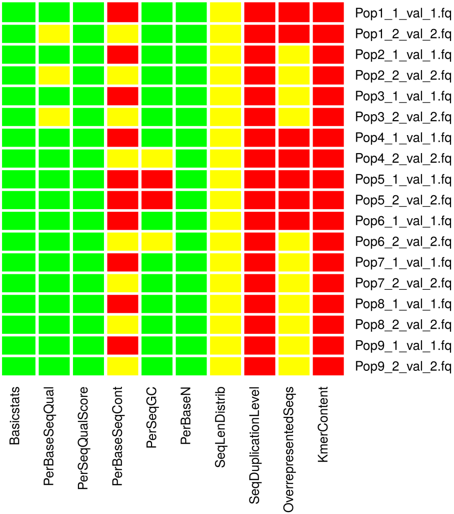


Figure : Quality overview of raw reads

Quality of trimmed reads



	Basicstats	PerBaseSeqQual	PerSeqQualScore	PerBaseSeqCont	PerSeqGC	PerBaseN	SeqLenDistrib	SeqDuplicationLevel	OverrepresentedSeqs	KmerContent
Pop1_1_val_1.fq	🟩	🟩	🟩	🟥	🟩	🟩	🟨	🟥	🟥	🟥
Pop1_2_val_2.fq	🟩	🟨	🟩	🟨	🟩	🟩	🟨	🟥	🟥	🟥
Pop2_1_val_1.fq	🟩	🟩	🟩	🟥	🟩	🟩	🟨	🟥	🟨	🟥
Pop2_2_val_2.fq	🟩	🟨	🟩	🟨	🟩	🟩	🟨	🟥	🟨	🟥
Pop3_1_val_1.fq	🟩	🟩	🟩	🟥	🟩	🟩	🟨	🟥	🟨	🟥
Pop3_2_val_2.fq	🟩	🟨	🟩	🟨	🟩	🟩	🟨	🟥	🟨	🟥
Pop4_1_val_1.fq	🟩	🟩	🟩	🟥	🟩	🟩	🟨	🟥	🟥	🟥
Pop4_2_val_2.fq	🟩	🟩	🟩	🟨	🟨	🟩	🟨	🟥	🟥	🟥
Pop5_1_val_1.fq	🟩	🟩	🟩	🟥	🟥	🟩	🟨	🟥	🟥	🟥
Pop5_2_val_2.fq	🟩	🟩	🟩	🟥	🟥	🟩	🟨	🟥	🟥	🟥
Pop6_1_val_1.fq	🟩	🟩	🟩	🟥	🟩	🟩	🟨	🟥	🟥	🟥
Pop6_2_val_2.fq	🟩	🟩	🟩	🟨	🟨	🟩	🟨	🟥	🟨	🟥
Pop7_1_val_1.fq	🟩	🟩	🟩	🟥	🟩	🟩	🟨	🟥	🟨	🟥
Pop7_2_val_2.fq	🟩	🟩	🟩	🟨	🟩	🟩	🟨	🟥	🟨	🟥
Pop8_1_val_1.fq	🟩	🟩	🟩	🟥	🟩	🟩	🟨	🟥	🟨	🟥
Pop8_2_val_2.fq	🟩	🟩	🟩	🟨	🟩	🟩	🟨	🟥	🟨	🟥
Pop9_1_val_1.fq	🟩	🟩	🟩	🟥	🟩	🟩	🟨	🟥	🟨	🟥
Pop9_2_val_2.fq	🟩	🟩	🟩	🟨	🟩	🟩	🟨	🟥	🟨	🟥

Figure : Quality overview of trimmed reads

Kmer content - overrepresented sequences at the beginning

Sequence	Count	Obs/Exp Max
CCCTAGC	5615	195.27312
CTAGCCC	5660	193.232
GCGTTGG	6795	180.89699
CCTAGCC	6130	178.19095

- Overrepresented sequences due to ddRAD stacks (high duplication levels)

Duplication Percentage

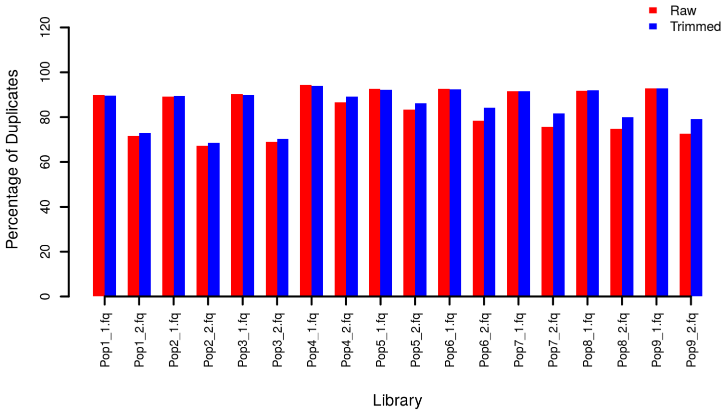
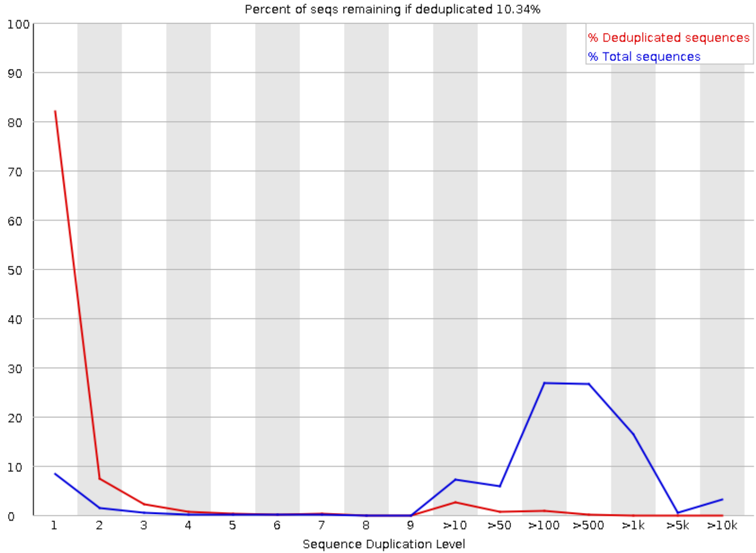


Figure : Percentage of Duplicates

Duplication levels



PCR duplicates and ddRADs can not be discriminated

- RAD: equal only at one end
- ddRAD: equal at both ends, makes identification of PCR duplicates impossible
- Tin2014 introduces four random bp to allow for duplicate detection in ddRAD reads

Mapping (Present state)

- Recommended:
  - Avoid as it can cause allele frequency biases:
    - Seeding (mapping of read subsets); reason: it discriminates against diverged reads?!
    - Local alignment and soft-clipping (removing of terminal mismatches)
- Allow gaps as ungapped alignment leads to false positive SNPs and mapping inaccuracy
- Map only proper pairs, discard broken pairs
- I use Bowtie2 (one of the recommendations in Schlotterer2014)
  - Uses semi-global but multi-seed alignment
  - Alternative1: BWA aln and sampe but only optimal for up to 100bp reads
  - Alternative 2: BWA mem (used in the DDocent pipeline, but uses seeding and local alignment)

Realign around Indels

- reads around indels: generally misaligned, results in false SNPS
- istead of realigning: ignore regions around indels
- Programs Dindel or GATK

Filtering

- Filter out broken pairs
- Filter out ambiguously mapped reads? Schlotterer2014 sees mapping quality as more important than targeting only unqiuely mapped reads.
- Mapping quality above 20

Coverage

- 50 recommended
- 20 as the absolute minimum. What do we aim for?
- less than 50: Sliding window analysis recommended. Only possible for species with genome or transcriptome reference
- Upper limit (too high coverage could result from copy number variations)
    - twice the mean coverage
    - remove top 2% coverages
    - mean coverage plus two standard deviations
- Coverage heterogeneity has to be taken into account in subsequent analyses. Or subsample to a homogenous coverage over the entire genome.

Variant detection 1

- Major problem: discriminate raw variants from sequencing errors
    - Some set threshold for minimum allele count
    - Some remove all multiallelic SNPS Beissinger2014
    - Inappropriate for pooled data Raineri2012
- Use algorithm that takes strand bias into account (only accept SNPs that are occurring at similar frequencies on both strands)

Variant detection 2

- Consensus approach (need to check in how far they are taking coverage variation into account):
  - Pool-hmm (corrects for site coverage variation)
  - SNVER (takes strand bias into account)
  - CRISP (takes strand bias into accunt)
  - snape (takes strand bias into accunt)
  - samtools mpileup followed by popoolation (gives a p-value for strand bias)
- Biological replicates (3):
  - Take only SNPs that are identified in all three replicates Robasky2013

Nucleotide diversity

- popoolation allows calculation of Tajima's D and Watterson's theta
- Using a sliding window approach (window size between 5 and 50 kbp)

How Tajima's D, Watterson's theta and Fst were used in our cod paper

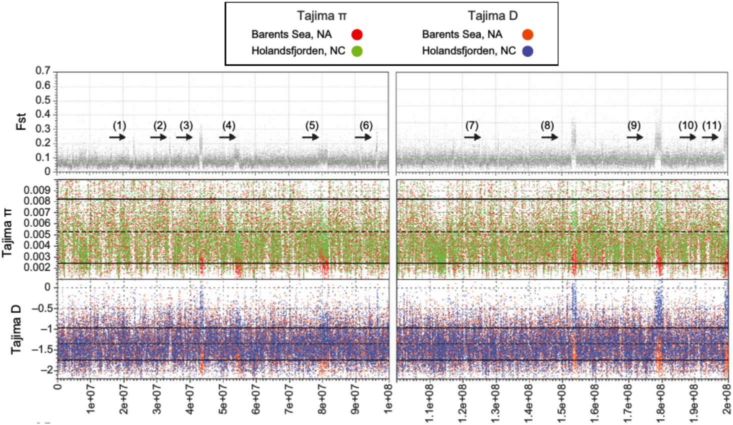


Figure : Population genetic parameters of cod populations

Adaptive differentiation

- Consensus approach
  - popoolation2 (Fisher's exact test and CMH (Cochran-Mantel-Haenzel) test (takes biological replicates into account, used also by Huang2014)
  - pool-hmm (detection of selective sweeps)
  - Bayenv (calculates environmental correlation, so this might not be the right approach for E&R studies)
  - SelEstim (detects and quantifies selection)
- Replicates:
  - In the CMH test taken into account
  - Do several pairwise comparison and identify overlapping outliers
  - Pool the allele frequencies before applying the test
  - Calculate a composite p-value or log-likelihood from the single tests
- Do we compare the selected with the non-selected populations or also the selected with each other?

Expected results

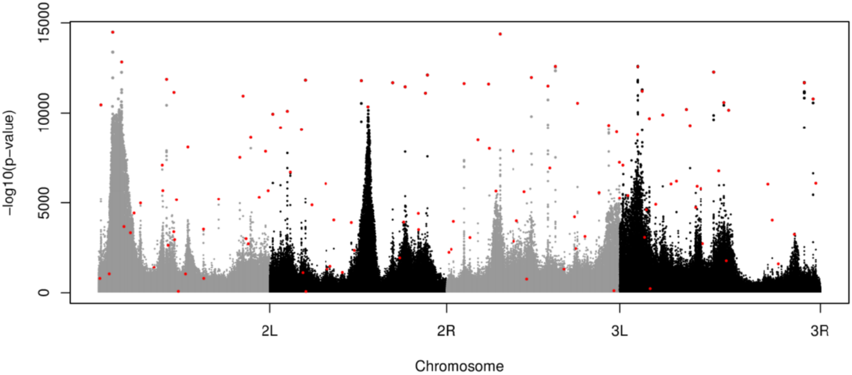


Figure : Manhatten plot with p-values from CMH test

Annotation

- Have to see what annotation of the genome is already available
- Programs for SNP annotation: SNPeff or AnnoVar