Population genomics using Pool-Seq

Dr. Robert Kofler

October 17, 2014

Availability of slides

http://drrobertkofler.wikispaces.com/NGSandEELecture

SEQUENCING STRATEGIES

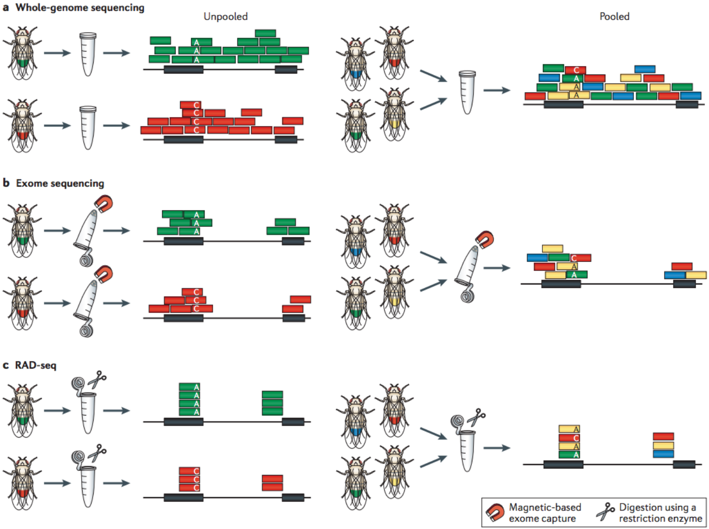

a Whole-genome sequencing
Unpooled
Pooled
A
C
b Exome sequencing
c RAD-seq
Magnetic-based exome capture
Digestion using a restriction enzyme

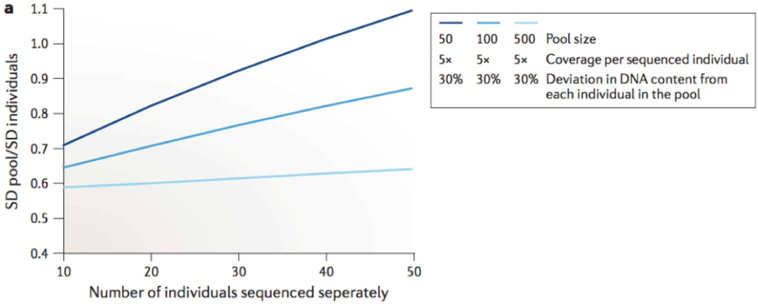
Figure 2 | **Comparison of sequencing strategies.** Three different sequencing approaches — whole-genome sequencing (part **a**), exome sequencing (part **b**) and restriction-site-associated DNA sequencing (RAD-seq; part **c**) — are compared, and sequencing of individuals (left panel) is contrasted with sequencing of pools of ...ed to reflect the individual from

enriched for exonic sequences (part **b**). RAD-seq only determines the sequence next to restriction sites, which results in stacked sequence reads (part **c**). Both exome sequencing and RAD-seq direct the sequencing efforts to targeted regions. This reduction in genome coverage allows a higher read count at a given genomic position and thus a more accurate allele frequency estimate at the covered genomic regions than

POOLING: PROS AND CONS

- +++ cost effective; a single Illumina lane may be sufficient to estimate allele frequencies in a population
- + bioinformatics analysis is, in my opinion, simpler; Especially now, since many software tools have been developed
- - haplotype information is lost
- - distinguishing between minor alleles and sequencing errors may be difficult

Major advantage cost



- with the same sequencing effort (i.e.: number of reads) the allele frequency in a population can be more accurately estimated with Pool-Seq than with sequencing individuals
- Pool-seq performs especially well with large pools (i.e.: many individuals entering the pools)

Schlötterer, Tobler, Kofler, Nolte (2014) Nat. Rev. Genetics

ANOTHER ADVANTAGE: SOFTWARE

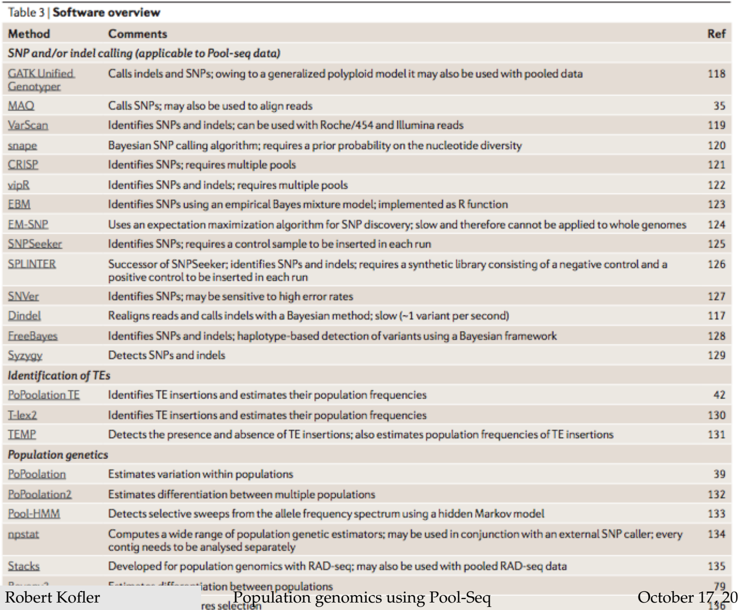


Table 3 | **Software overview**

Method	Comments	Ref
SNP and/or indel calling (applicable to Pool-seq data)		
GATK Unified Genotyper	Calls indels and SNPs; owing to a generalized polyploid model it may also be used with pooled data	118
MAQ	Calls SNPs; may also be used to align reads	35
VarScan	Identifies SNPs and indels; can be used with Roche/454 and Illumina reads	119
snape	Bayesian SNP calling algorithm; requires a prior probability on the nucleotide diversity	120
CRISP	Identifies SNPs; requires multiple pools	121
vipR	Identifies SNPs and indels; requires multiple pools	122
EBM	Identifies SNPs using an empirical Bayes mixture model; implemented as R function	123
EM-SNP	Uses an expectation maximization algorithm for SNP discovery; slow and therefore cannot be applied to whole genomes	124
SNPSeeker	Identifies SNPs; requires a control sample to be inserted in each run	125
SPLINTER	Successor of SNPSeeker; identifies SNPs and indels; requires a synthetic library consisting of a negative control and a positive control to be inserted in each run	126
SNVer	Identifies SNPs; may be sensitive to high error rates	127
Dindel	Realigns reads and calls indels with a Bayesian method; slow (~1 variant per second)	117
FreeBayes	Identifies SNPs and indels; haplotype-based detection of variants using a Bayesian framework	128
Syzygy	Detects SNPs and indels	129
Identification of TEs		
PoPoolation TE	Identifies TE insertions and estimates their population frequencies	42
T-lex2	Identifies TE insertions and estimates their population frequencies	130
TEMP	Detects the presence and absence of TE insertions; also estimates population frequencies of TE insertions	131
Population genetics		
PoPoolation	Estimates variation within populations	39
PoPoolation2	Estimates differentiation between multiple populations	132
Pool-HMM	Detects selective sweeps from the allele frequency spectrum using a hidden Markov model	133
npstat	Computes a wide range of population genetic estimators; may be used in conjunction with an external SNP caller; every contig needs to be analysed separately	134
Stacks	Developed for population genomics with RAD-seq; may also be used with pooled RAD-seq data	135
[illegible]	Estimates differentiation between populations	79
[illegible]	res selection	136

WHAT COULD BE DONE WITH POOL-SEQ?

- Genotype-phenotype mapping: Pool-GWAS, E&R, mapping induced mutations,..
- Reverse Ecology, i.e.: the use of genomics to study ecology; For example to identify loci responsible for adaptation to some environments
- Domestication; Identify the genomic basis of artificially selected traits
- Genome evolution; Transposable element activity, polymorphism, selective sweeps
- Trajectories of selected alleles; Experimental evolution, clonal interference, dynamics of clonal populations
- Study cancer progression

Schlötterer, Tobler, Kofler, Nolte (2014) Nat. Rev. Genetics

WHEN NOT TO POOL

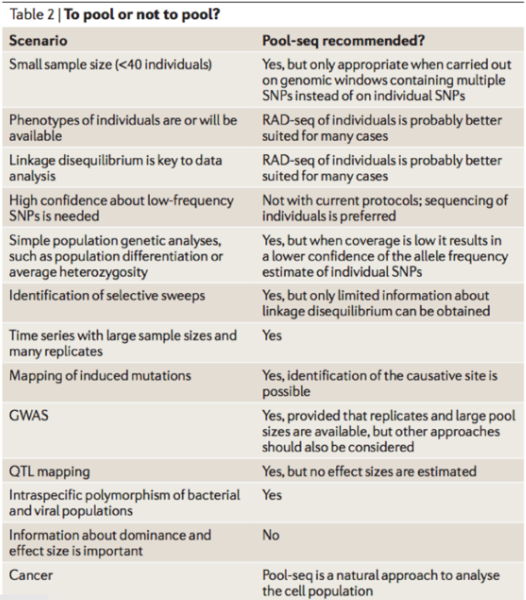


Table 2 | **To pool or not to pool?**

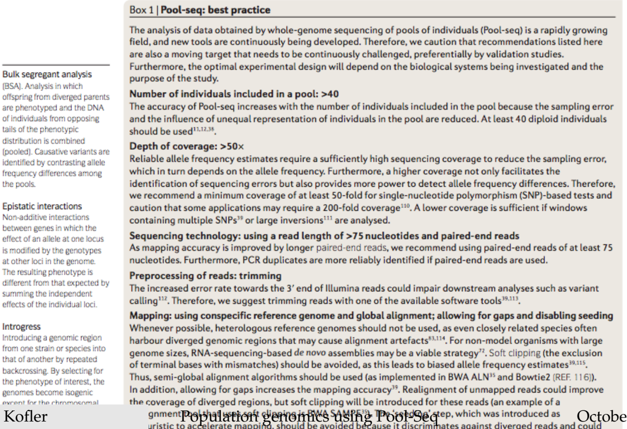
	Scenario	**Pool-seq recommended?**
	Small sample size (<40 individuals)	Yes, but only appropriate when carried out on genomic windows containing multiple SNPs instead of on individual SNPs
NO	Phenotypes of individuals are or will be available	RAD-seq of individuals is probably better suited for many cases
NO	Linkage disequilibrium is key to data analysis	RAD-seq of individuals is probably better suited for many cases
NO	High confidence about low-frequency SNPs is needed	Not with current protocols; sequencing of individuals is preferred
	Simple population genetic analyses, such as population differentiation or average heterozygosity	Yes, but when coverage is low it results in a lower confidence of the allele frequency estimate of individual SNPs
	Identification of selective sweeps	Yes, but only limited information about linkage disequilibrium can be obtained
	Time series with large sample sizes and many replicates	Yes
	Mapping of induced mutations	Yes, identification of the causative site is possible
	GWAS	Yes, provided that replicates and large pool sizes are available, but other approaches should also be considered
	QTL mapping	Yes, but no effect sizes are estimated
	Intraspecific polymorphism of bacterial and viral populations	Yes
NO	Information about dominance and effect size is important	No
	Cancer	Pool-seq is a natural approach to analyse the cell population

Section 2

Aim of this lecture

Best practice analysis of Pool-seq data

Learn the pipeline for analysing Pool-seq data adhering to the best practices as suggested in our recent review [Schlötterer, Tobler, Kofler, Nolte (2014) Nat. Rev. Genetics]. As an example application we will analyse *Drosophila* data with PoPoolation (Kofler et al. (2011) PLoS One).



Bulk segregant analysis
(BSA). Analysis in which offspring from diverged parents are phenotyped and the DNA of individuals from opposing tails of the phenotypic distribution is combined (pooled). Causative variants are identified by contrasting allele frequency differences among the pools.

Epistatic interactions
Non-additive interactions between genes in which the effect of an allele at one locus is modified by the genotypes at other loci in the genome. The resulting phenotype is different from that expected by summing the independent effects of the individual loci.

Introgress
Introducing a genomic region from one strain or species into that of another by repeated backcrossing. By selecting for the phenotype of interest, the genomes become isogenic except for the chromosomal

Box 1 | **Pool-seq: best practice**

The analysis of data obtained by whole-genome sequencing of pools of individuals (Pool-seq) is a rapidly growing field, and new tools are continuously being developed. Therefore, we caution that recommendations listed here are also a moving target that needs to be continuously challenged, preferentially by validation studies. Furthermore, the optimal experimental design will depend on the biological systems being investigated and the purpose of the study.

Number of individuals included in a pool: >40
The accuracy of Pool-seq increases with the number of individuals included in the pool because the sampling error and the influence of unequal representation of individuals in the pool are reduced. At least 40 diploid individuals should be used[11,12,38].

Depth of coverage: >50×
Reliable allele frequency estimates require a sufficiently high sequencing coverage to reduce the sampling error, which in turn depends on the allele frequency. Furthermore, a higher coverage not only facilitates the identification of sequencing errors but also provides more power to detect allele frequency differences. Therefore, we recommend a minimum coverage of at least 50-fold for single-nucleotide polymorphism (SNP)-based tests and caution that some applications may require a 200-fold coverage[110]. A lower coverage is sufficient if windows containing multiple SNPs[39] or large inversions[111] are analysed.

Sequencing technology: using a read length of >75 nucleotides and paired-end reads
As mapping accuracy is improved by longer paired-end reads, we recommend using paired-end reads of at least 75 nucleotides. Furthermore, PCR duplicates are more reliably identified if paired-end reads are used.

Preprocessing of reads: trimming
The increased error rate towards the 3′ end of Illumina reads could impair downstream analyses such as variant calling[112]. Therefore, we suggest trimming reads with one of the available software tools[39,113].

Mapping: using conspecific reference genome and global alignment; allowing for gaps and disabling seeding
Whenever possible, heterologous reference genomes should not be used, as even closely related species often harbour diverged genomic regions that may cause alignment artefacts[83,114]. For non-model organisms with large genome sizes, RNA-sequencing-based *de novo* assemblies may be a viable strategy[72]. Soft clipping (the exclusion of terminal bases with mismatches) should be avoided, as this leads to biased allele frequency estimates[39,115]. Thus, semi-global alignment algorithms should be used (as implemented in BWA ALN[35] and Bowtie2 (REF. 116)). In addition, allowing for gaps increases the mapping accuracy[39]. Realignment of unmapped reads could improve the coverage of diverged regions, but soft clipping will be introduced for these reads (an example of a gnment tool that uses soft clipping is BWA SAMPE[35]). The 'seeding' step, which was introduced as uristic to accelerate mapping, should be avoided because it discriminates against diverged reads and could

PoPoolation 1

OPEN ACCESS Freely available online

PLoS one

PoPoolation: A Toolbox for Population Genetic Analysis of Next Generation Sequencing Data from Pooled Individuals

Robert Kofler[1], Pablo Orozco-terWengel[1], Nicola De Maio[1], Ram Vinay Pandey[1], Viola Nolte[1], Andreas Futschik[2], Carolin Kosiol[1], Christian Schlötterer[1]*

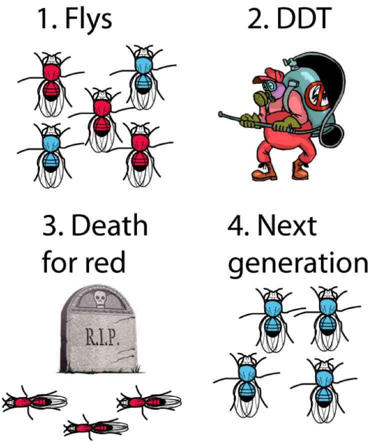
[1] Institute of Population Genetics, Vetmeduni Vienna, Vienna, Austria, [2] Department of Statistics, University of Vienna, Vienna, Austria

WHAT CAN YOU DO WITH POPOOLATION?

- Perform genome-wide scans for positively selected regions in populations sequenced as pools
- Obtain genome-wide estimates of natural variation
- You may just estimate natural variation at synonymous or non-synonymous sites
- trim fastq-reads

Positive selection?

Individuals with beneficial mutations (e.g.: resistance to DDT) will produce more progeny, and thus the beneficial allele rises in frequency.



1. Flys

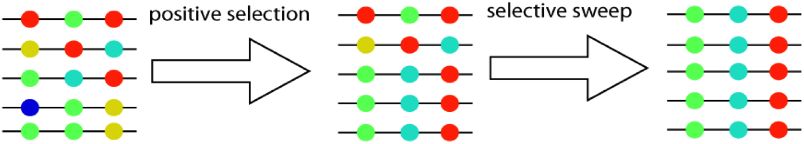
2. DDT

3. Death for red

4. Next generation

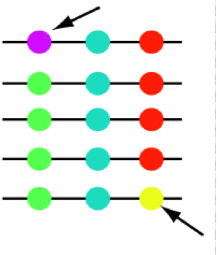
GENOMIC SIGNATURE OF POSITIVE SELECTION #1

When an allele increases in its population frequency, nearby variants also increase in its frequency ⇒ Hitchhiking. This leads to a selective sweep which erases variation around a positively selected allele.



Genomic signature of positive selection #2

After the sweep new mutations appear and restore diversity, but they appear very slowly (mutations are rare) and they are initially of low frequency



Selective sweeps thus lead to regions with reduced variability in the neighbourhood of the selected sites, i.e.: few SNPs having low population frequencies.

Example: KEL-locus in humans

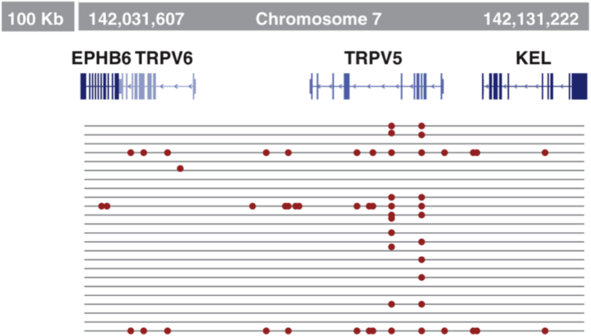


Fig. 3. Low diversity and many rare alleles at the Kell blood antigen cluster. On the basis of three different statistical tests, the 115-kb region (containing four genes) shows evidence of a selective sweep in Europeans (*28*).

KEL extends for 115kb and is the largest locus of positive selection described in humans. These genes are important determinants of blood type.

TAJIMAS π

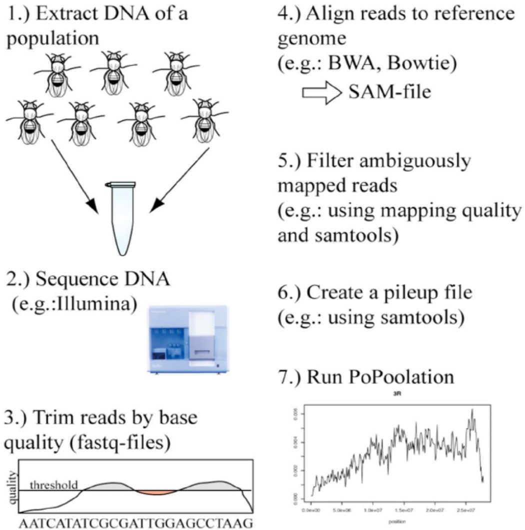
One measure of variability within populations is Tajima's π, which is defined as the average pairwise difference between randomly chosen individuals. Can be calculated for SNPs, windows, genes, etc...

$$\pi = \frac{\sum\limits_{SNPs}(1 - \sum\limits_{alleles} f_a^2)}{L}$$

- f_a frequency of the given allele
- L length of the investigated window

Value of Tajima's π varies between $0 - 1$, where a low π indicates no natural variation (few SNPs with low population frequencies) and a high π a large amount of natural variation (many SNPs with very balanced allele frequencies). In this Walkthrough I will show how to obtain genome-wide estimates of Tajima's π with PoPoolation.

WORKFLOW WITH POPOOLATION



1.) Extract DNA of a population

2.) Sequence DNA (e.g.:Illumina)

3.) Trim reads by base quality (fastq-files)

4.) Align reads to reference genome (e.g.: BWA, Bowtie) ⇨ SAM-file

5.) Filter ambiguously mapped reads (e.g.: using mapping quality and samtools)

6.) Create a pileup file (e.g.: using samtools)

7.) Run PoPoolation

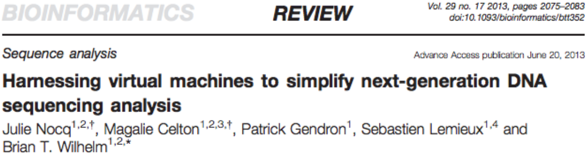
BIOINFORMATICS WORKFLOW: BEST PRACTICES

Currently, we recommend to use the following pipeline:

- trimming of reads
- mapping of reads; semi-global alignments, no seeding
- remove duplicates
- remove ambiguously mapped reads and broken pairs
- converting to mpileup
- subsampling to uniform coverage
- remove regions around indels
- PoPoolation

Virtual Machine

The course will be held on a Virtual Machine. Advantages?



BIOINFORMATICS **REVIEW** *Vol. 29 no. 17 2013, pages 2075–2083 doi:10.1093/bioinformatics/btt352*

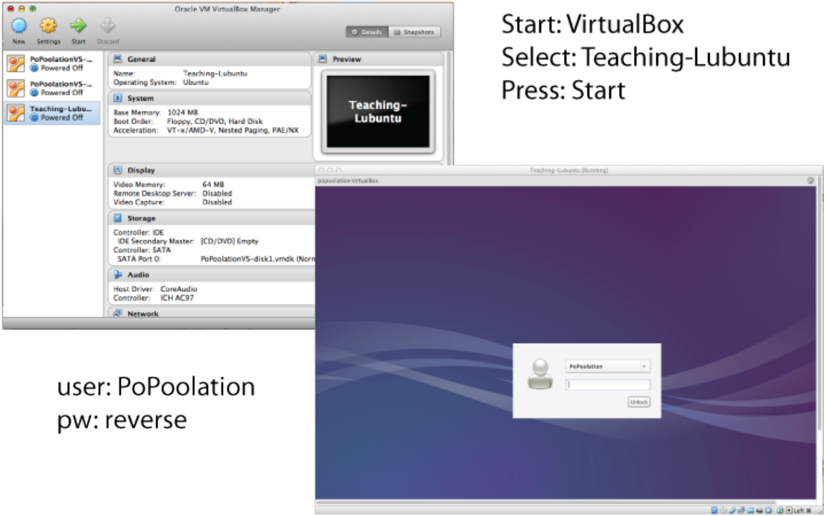
Sequence analysis Advance Access publication June 20, 2013

Harnessing virtual machines to simplify next-generation DNA sequencing analysis

Julie Nocq[1,2,†], Magalie Celton[1,2,3,†], Patrick Gendron[1], Sebastien Lemieux[1,4] and Brian T. Wilhelm[1,2,*]

- correct versions of software and all required libraries are preinstalled so you may actually be able to repeat the demonstrated analysis
- reproducibility of analysis is increased; data may be shared together with the software necessary for analysing them
- more stable pipelines; entire pipelines may be shared
- this comes at a cost: performance loss of about 25%

Open the virtual machine



General
Name: Teaching-Lubuntu
Operating System: Ubuntu

System
Base Memory: 1024 MB
Boot Order: Floppy, CD/DVD, Hard Disk
Acceleration: VT-x/AMD-V, Nested Paging, PAE/NX

Display
Video Memory: 64 MB
Remote Desktop Server: Disabled
Video Capture: Disabled

Storage
Controller: IDE
IDE Secondary Master: [CD/DVD] Empty
Controller: SATA
SATA Port 0: PoPoolationVS-disk1.vmdk (Norn

Audio
Host Driver: CoreAudio
Controller: ICH AC97

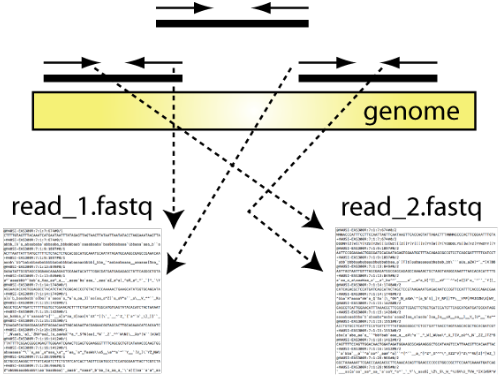
Network

Start: VirtualBox
Select: Teaching-Lubuntu
Press: Start

user: PoPoolation
pw: reverse

PAIRED END READS

We start with paired end reads. With the Illumina technology you will get two fastq-files for paired end reads. Reads are always provided in the same order in both fastq-files. The two reads of one pair can therefore be recognized by having the same index in the both fastq-files.



Note: the first read (read_1.fastq) is not necessarily the 5′ read. Assignment as the first read is a stochastic process (therefore usually 50% 5′ reads and 50% 3′ reads).

REMEMBER: FIRST STEP

ALWAYS make sure your data are complete. Quick and dirty

```
cd Desktop/popoolation1
wc read_*
>  209288  209288 11279782 read_1.fastq
>  209288  209288 11279782 read_2.fastq
>  418576  418576 22559564 total
```

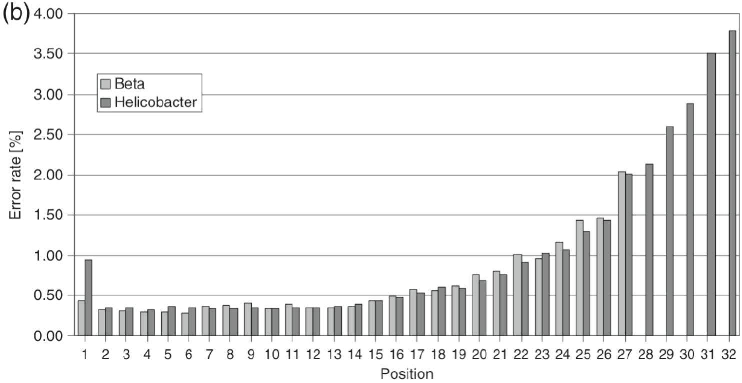
more professional

```
md5sum read_*
>MD5 (read_1.fastq) = fd8fdfce336391e106fdc84ee60dd622
>MD5 (read_2.fastq) = 18d01db158ea29334d90b68880d9f6bb
```

The wordcount or the md5 sum should be compared to the values provided by the sequencing facility (e.g.: BGI).

TRIMMING OF READS; WHY?

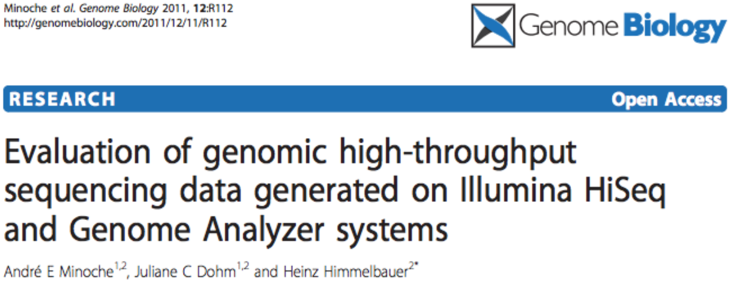
Before we map the paired end reads we need to deal with a problem: Error rate of the reads is increasing with the length. Also remember your FastQC results, the base quality is decreasing with the length of the reads.



Source: Dohm J. (2008) Substan)al biases in ultrashort read data sets from

Trimming of reads; Why?

A recent study clearly showed that trimming of the reads at low quality is the single quality filtering step that most dramatically improves the results (reduces analysis artefacts).



Minoche *et al. Genome Biology* 2011, **12**:R112
http://genomebiology.com/2011/12/11/R112

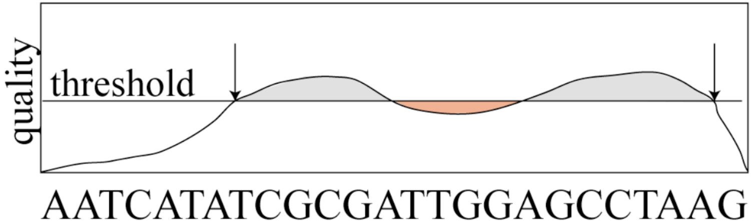
Genome **Biology**

RESEARCH **Open Access**

Evaluation of genomic high-throughput sequencing data generated on Illumina HiSeq and Genome Analyzer systems

André E Minoche[1,2], Juliane C Dohm[1,2] and Heinz Himmelbauer[2*]

TRIMMING ALGORITHM OF POPOOLATION



- Given some arbitrary quality threshold (usually base quality of 20) the algorithm finds the highest scoring substring of the read.
- some fraction of the bases may be below the quality threshold, as long as a new high score can be achieved.
- the algorithm is very similar to dynamic programming (Smith-Waterman)
- handles single end reads as well as paired end reads

Trimming

```
mkdir pe

# Trimming
perl ~/programs/popoolation/basic-pipeline/trim-fastq.pl --
    disable-zipped-output --input1 read_1.fastq --input2 read_2.
    fastq --min-length 50 --no-5p-trim --quality-threshold 20 --
    fastq-type illumina --output1 pe/read_1.tr.fastq --output2
    pe/read_2.tr.fastq --outputse pe/read_se.tr.fastq
```

- –input the input files
- –output.. the output files; this will create three files with the extensions _1 _2 _SE;
- –min-length discard reads that are after trimming smaller than this threshold; Note this step may create orphan reads, i.e.: reads who lost their mate :(
- –no-5p-trim only trim reads at the 3′ end; this is necessary for the removal of duplicates
- –quality-threshold reads should on average have a score higher than this threshold
- –fastq-type is the encoding of the base quality in sanger or illumina (remember offset)
- –disable-zipped-output in the newest versions of PoPoolation the output of the fastq files is per default zipped. Here we disable this feature

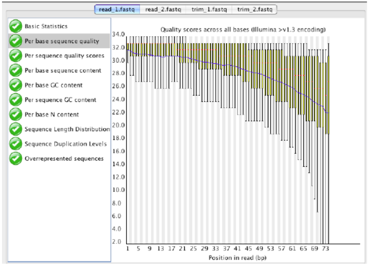
Trim statistic

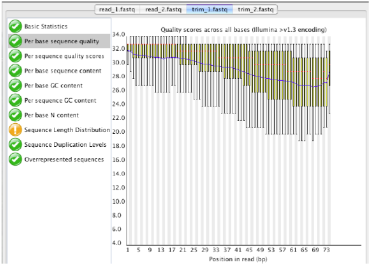
```
Read-pairs processed: 52322
Read-pairs trimmed in pairs: 52322
Read-pairs trimmed as singles: 0

FIRST READ STATISTICS
First reads passing: 52322
5p poly-N sequences trimmed: 0
3p poly-N sequences trimmed: 124
Reads discarded during 'remaining N filtering': 0
Reads discarded during length filtering: 0
Count sequences trimmed during quality filtering: 19928

Read length distribution first read
length count
50 322
51 327
52 351
53 359
54 358
55 381
56 366
```

Effect of trimming on quality





Exercise:

- open FastQC
- load read_1.fastq
- load pe/read_1.tr.fastq
- compare the base quality between trimmed and untrimmed
- compare the sequence length distribution between trimmed and untrimmed

PREPARING THE REFERENCE SEQUENCE FOR MAPPING

```
mkdir wg
awk ’{print $1}’ dmel-2R-chromosome-r5.22.fasta
    > wg/dmel-2R-short.fasta
bwa index wg/dmel-2R-short.fasta
```

Remember: With this command we are removing the description of the fasta entry. This step is strongly recommended as unnecessarily long fasta identifiers may lead to problems in downstream analysis.

Paired end mapping

```
bwa aln -I -m 100000 -o 1 -n 0.01 -l 200 -e 12 -d 12 -t 2 wg/dmel
    -2R-short.fasta pe/read_1.tr.fastq > pe/read_1.sai
bwa aln -I -m 100000 -o 1 -n 0.01 -l 200 -e 12 -d 12 -t 2 wg/dmel
    -2R-short.fasta pe/read_2.tr.fastq > pe/read_2.sai
bwa sampe wg/dmel-2R-short.fasta pe/read_1.sai pe/read_2.sai pe/
    read_1.tr.fastq pe/read_2.tr.fastq > pe/pe.sam
```

- -I input is in Illumina encoding (offset 64); do not provide this when input is in sanger! Very important parameter!
- -m not important; just telling bwa to process smaller amounts of reads at once
- -l 200 seed size (needs to be longer than the read length to disable seeding)
- -e 12 -d 12 gap length (for insertions and deletions)
- -o 1 maximum number of gaps
- -n 0.01 the number of allowed mismatches, in terms of probability of missing the read. In general the lower the value the more mismatches are allowed. The exact translation is shown at the beginning of the mapping
- -t 2 number of threads, the more the faster

CONVERTING SAM TO BAM

- sam.. Sequence Alignment Map format ⇒ optimized for humans
- bam.. binary sam ⇒ optimized for computers

It is easily possible to convert a sam to bam and vice versa a bam to sam. In the following we convert a sam into a bam and finally sort the bam file

```
samtools view -Sb pe/pe.sam > pe/pe.bam
```

- -S input is sam
- -b output is bam (-S may be merged with -b to -Sb)
- ’sort - outpufile’ input for sorting is the pipe (rather than a file)

SORTING WITH PICARD

Here we use Picard to sort reads, which is a bit more complicated than sorting with samtools. Sorting with Picard is however necessary as otherwise the downstream analysis (MarkDuplicates) would not work.

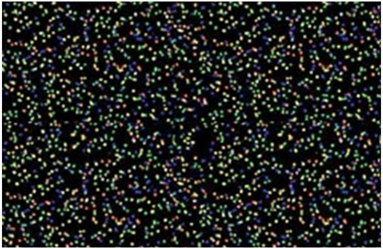
```
java -Xmx2g -jar ~pic/SortSam.jar I= pe/pe.bam O= pe/pe.sort.bam
     VALIDATION_STRINGENCY=SILENT SO=coordinate
```

- Picard runs with Java
- -Xmx2g give Java 2 Gb of memory
- -jar SortSam use the Java software SortSam
- I= input
- O= output
- SO= sort order; sort by coordinate
- VALIDATION_STRINGENCY= Picard is like a Princess that is constantly complaining about every small deviation of our sam file from the most stringent requirements. I have never found a sam file satisfying all of Picards demands ⇒ 'shut up Picard'

DUPLICATES

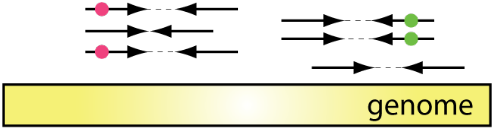
There are two sources of duplicates with the Illumina technology

- Optical duplicates: they are occurring during the sequencing step; The algorithm responsible for identifying isolated clusters wrongly identifies a single cluster as two (see picture below).
- PCR duplicates; occurs during sample preparation where a PCR step is necessary to amplify the amount of DNA for the sequencing.



HOW TO RECOGNIZE DUPLICATES

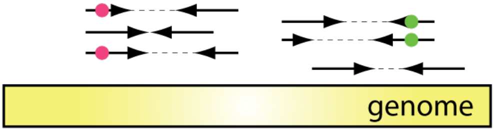
One common proxy is to use the mapping positions of PE reads to recognize duplicates where reads having exactly identical positions are marked as duplicates. This has the advantage that it also allows for sequencing errors within duplicated reads. The chances that two PE reads accidentally have identical positions are minimal.



Duplicates are marked by dots having identical colors.

DUPLICATES WITH TRIMMED READS

During trimming reads may be truncated at sequences having a low quality, thus duplicated reads may end up having different lengths. Fortunately, removal of duplicates can still be performed if only the 3′ ends of reads are trimmed (remember that the quality deteriorates mostly at the 3′ end of reads). This is because Picard recognizes duplicates by PE reads having identical 5′ positions (5′ of the read not the genome).



Duplicates are marked by dots having identical colors.

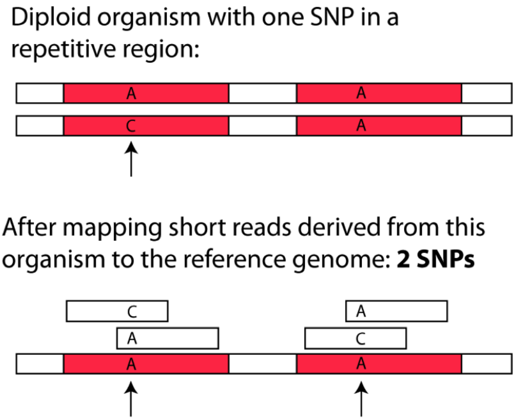
Removing duplicates

```
# the following only accepts a sam file sorted by Picard.
java -Xmx2g -jar ~pic/MarkDuplicates.jar I= pe/pe.sort.bam O= pe/pe.rmd.sort.bam M= pe/dupstat.txt VALIDATION_STRINGENCY=SILENT REMOVE_DUPLICATES=true
```

- I= input file
- O= output file for reads
- M= output file of statistics (how many identified duplicates)
- REMOVE_DUPLICATES= remove duplicates from the output file rather than just marking them (remember flag in sam-file 0x400)

Another problem, ambiguous mapping

Ambiguously mapped reads can lead to wrong SNPs.



Diploid organism with one SNP in a repetitive region:

After mapping short reads derived from this organism to the reference genome: **2 SNPs**

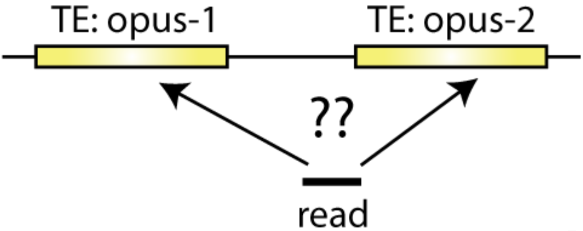
⇒ therefore exclude ambiguously mapped reads

AMBIGUOUS MAPPING POSITION AND MAPPING QUALITY

Remember column 5 of the sam file contains the mapping quality. Similarly to the base quality, the mapping quality is the log scaled probability that the position of the read is wrong.

- 20.. one out of 100 reads is wrongly mapped
- 30.. one out of 1000 reads is wrongly mapped
- 0.. every read is wrongly mapped

A major cause for incorrect or ambiguous mapping positions are repetitive regions in the genome



REMOVE LOW QUALITY ALIGNMENTS

The following command ensures that we remove ambiguously mapped reads and only retain PE reads where both mates align with the reference genome

```
samtools view -q 20 -f 0x0002 -F 0x0004 -F 0x0008 -b
    pe/pe.rmd.sort.bam > pe/pe.qf.rmd.sort.bam
```

- -q 20 only keep reads with a mapping quality higher than 20 (remove ambiguously aligned reads)
- -f 0x0002 only keep proper pairs (remember flags from sam file)
- -F 0x0004 remove reads that are not mapped
- -F 0x0008 remove reads with an un-mapped mate

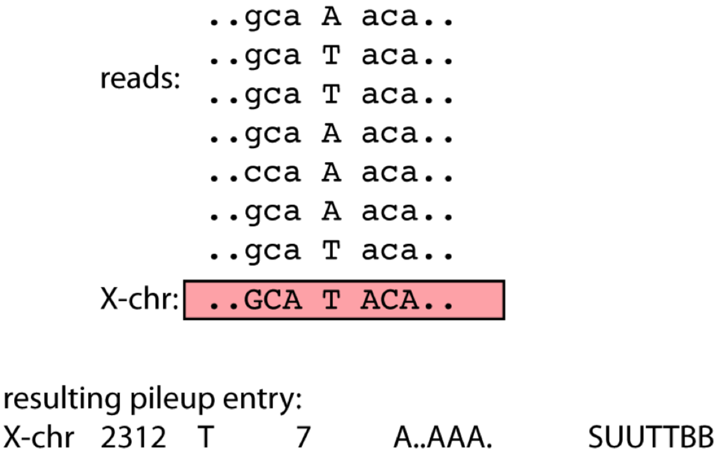
Note ′-f′ means only keep reads having the given flag and ′-F′ discard all reads having the given flag.

CREATING A MPILEUP FILE

```
samtools mpileup -B -Q 0 -f wg/dmel-2R-short.fasta pe/pe.qf.rmd.sort.bam > pe/pe.mpileup
less pe/pe.mpileup
```

- -B disable BAQ computation (base alignment quality)
- -Q skip bases with base quality smaller than the given value
- -f path to reference sequence

WHAT IS A PILEUP?



```
reads:  ..gca A aca..
        ..gca T aca..
        ..gca T aca..
        ..gca A aca..
        ..cca A aca..
        ..gca A aca..
        ..gca T aca..
X-chr:  ..GCA T ACA..
```

resulting pileup entry:

```
X-chr   2312   T     7      A..AAA.      SUUTTBB
```

MPILEUP FILE

```
2R 7809811 C 18 .............,..,. B=9>=C<BBB<@A4BC6C
2R 7809812 A 18 .............,..,. @B;66:7ABB@7A8CB;B
2R 7809813 G 18 .............,..,. AA9/:C<<?B@@B?BBAC
```

- col1 reference chromosome
- col2 position
- col3 reference character
- col4 coverage
- col5 bases for the given position ('.' identical to reference character - forward strand; ',' identical to reference - reverse strand)
- col6 quality for the bases

Filtering indels

```
perl ~/programs/popoolation/basic-pipeline/identify-
    genomic-indel-regions.pl --indel-window 5 --min-
    count 2 --input pe/pe.mpileup --output pe/indels
    .gtf
```

- –indel-window how many bases surrounding indels should be ignored
- –min-count minimum count for calling an indel. Note that indels may be sequencing errors as well

```
perl ~/programs/popoolation/basic-pipeline/filter-
    pileup-by-gtf.pl --input pe/pe.mpileup --gtf pe/
    indels.gtf --output pe/pe.idf.mpileup
```

Note: the filter-pileup script could also be used to remove entries overlapping with transposable elements (RepeatMasker produces a gtf as well).

SUBSAMPLING TO UNIFORM COVERAGE

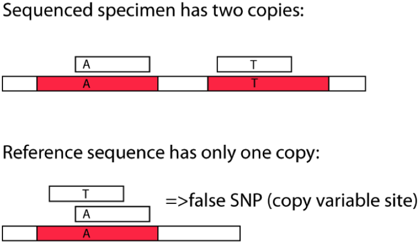
Several population genetic estimators are sensitive to sequencing errors. For example a very low Tajima's D, usually indicative of a selective sweep, may be, as an artifact, frequently be found in highly covered regions because these regions have just more sequencing errors. To avoid these kinds of biases we recommend to subsample to an uniform coverage.

```
perl ~/programs/popoolation/basic-pipeline/subsample-pileup.pl --
    min-qual 20 --method withoutreplace --max-coverage 50 --
    fastq-type sanger --target-coverage 10 --input pe/pe.idf.
    mpileup --output pe/pe.ss10.idf.mpileup
```

- –min-qual minimum base quality
- –method method for subsampling, we recommend without replacement
- –target-coverage which coverage should the resulting mpileup file have
- –max-coverage the maximum allowed coverage, regions having higher coverages will be ignored (they may be copy number variations and lead to wrong SNPs)
- –fastq-type (sanger means offset 33)

Copy number variations may lead to wrong SNPs

We introduced a maximum coverage because copy number variations could lead to wrong SNPs.


Sequenced specimen has two copies:
A
A
T
T
Reference sequence has only one copy:
T
A
A
=>false SNP (copy variable site)

The signatures of these artifactual SNPs are:

- fairly balanced allele frequencies (e.g.: 50% A and 50% T)
- high coverage ← targeted by PoPoolation

INSPECTING THE SUB-SAMPLED PILEUP

```
less pe/pe.ss10.idf.mpileup
```

```
2R 7799887 A 10 AAAAAAAAAA 5555555555
2R 7799889 G 10 GGGGGGGGGG 5555555555
2R 7799890 G 10 GGGGGGGGGG 5555555555
2R 7799892 C 10 CCCCCCCCCC 5555555555
2R 7799893 A 10 AAAAAAAAAA 5555555555
```

Note that the quality has been uniformly set to the ’–min-qual’

BIOINFORMATICS WORKFLOW: FINALLY POPOOLATION1

After these many steps we can finally proceed and estimate the polymorphism in the population with PoPoolation. Now it's also time to apologize for the length of this pipeline. When we started with PoPoolation this pipeline was much shorter. But gradually we eliminated possible confounding factors (e.g.: duplicates, subsampling) and the pipeline grew. We also condensed this information into our recent review [Schlötterer, Tobler, Kofler, Nolte (2014) Nat. Rev. Genetics]

- trimming of reads
- mapping of reads; semi-global alignments, no seeding
- remove duplicates
- remove ambiguously mapped reads and broken pairs
- converting to mpileup
- subsampling to uniform coverage
- remove regions around indels
- PoPoolation ←

GETTING HELP

You can get help for any PoPoolation script with the option '–help'.

```
perl ~/programs/popoolation/Variance-sliding.pl
    --help
```

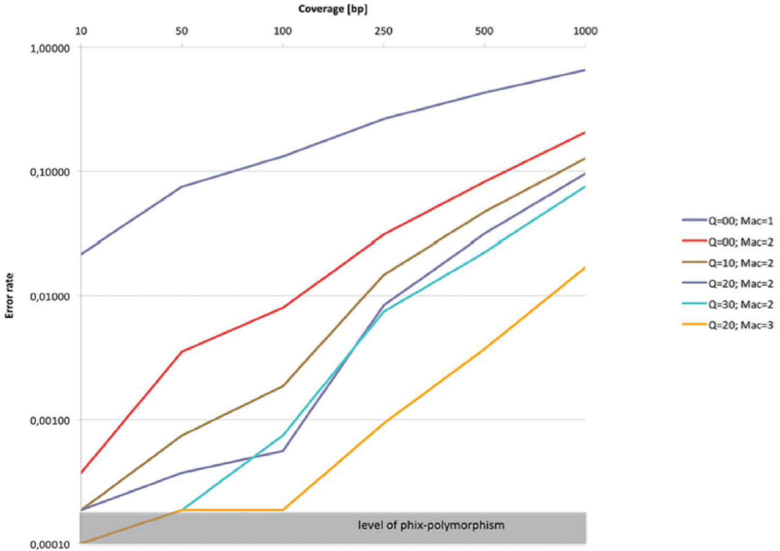
Exit help with pressing 'q'

Calculating Tajima's π

```
perl ~/programs/popoolation/Variance-sliding.pl --fastq-type
    sanger --measure pi --input pe/pe.ss10.idf.mpileup --min-
    count 2 --min-coverage 4 --max-coverage 10 --min-covered-
    fraction 0.5 --pool-size 500 --window-size 1000 --step-size
    1000 --region 2R:7800000-8300000 --output pe/cyp6g1.pi --snp
    -output pe/cyp6g1.snps
```

- –min-coverage –max-coverage: for subsampled files not important; should contain target coverage, i.e.: 10
- –min-covered-fraction minimum percentage of sites having sufficient coverage in the given window
- –min-count minimum occurrence of allele for calling a SNP
- –measure which population genetics measure should be computed (pi/theta/D)
- –pool-size number of chromosomes (thus number of diploids times two)
- –region compute the measure only for a small region; default is the whole genome
- –output a file containing the measure (π) for the windows
- –snp-output a file containing for every window the SNPs that have been used for computing the measure (e.g. π)
- –window-size –step-size control behaviour of sliding window; if step size is smaller than window size than the windows will be overlapping.

How to choose a minimum count threshold?



Output

```
less pe/cyp6g1.pi
```

```
2R 7800500 0 0.218 na
2R 7801500 6 0.683 0.004936240
2R 7802500 13 0.916 0.008076347
2R 7803500 3 0.782 0.002411416
2R 7804500 6 0.599 0.006439348
```

- col 1: reference chromosome
- col 2: position of window (mean value)
- col 3: number of SNPs in the given window
- col 4: fraction of sites in the window having sufficient coverage ($min \leq x \leq max$)
- col 5: measure for the window (π)

SNP Output

```
less pe/cyp6g1.snps
```

The file contains the SNP found in each window

```
>2R:7801500 2R:7801000-7802000 snps:6
2R 7801059 T 10 2 8 0 0 0
2R 7801066 G 10 6 0 0 4 0
```

- col 1: reference chromosome
- col 2: position of SNP
- col 3: reference character
- col 4: coverage
- col 5-9: counts of A, T, C, G, N respectively

Visualize in IGV

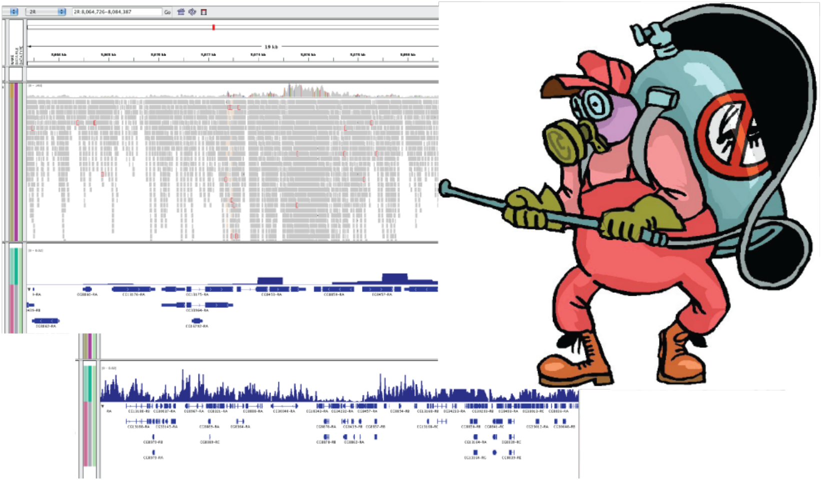
```
perl ~/programs/popoolation/VarSliding2Wiggle.pl --input pe/
    cyp6g1.pi --trackname "pi" --output pe/cyp6g1.wig
samtools index pe/pe.qf.rmd.sort.bam
java -Xmx2g -jar ~/programs/IGV_2.3.26/igv.jar
```

Than:

- create a new genome; load wg/dmel-2R-short.fasta
- store new genome in Desktop/popoolation1
- load the paired end reads (pe/pe.qf.rmd.sort.bam)
- load the annotation (cyp6g1.gtf)
- load the variation (pe/cyp6g1.wig)
- search gene CG8453 (= Cyp6g1)

Background: some variants around Cyp6g1 have recently swept to fixation in *D. melanogaster* as the gene confers resistance to DDT. So we would expect an extreme dip in variability in the neighbourhood of this gene. However, the situation is complex as there has also been a lot of copy number variations (CNVs e.g.: duplications). Zoom in and inspect the coverage to see CNV regions.

Biology, here I come..



ADDITIONAL FEATURES

PoPoolation also allows to:

- Calculate Tajima's D, Wattersons Θ
- Calculate the measure (π, D, Θ) for genes (instead of windows)
- Calculate the measure for synonymous and non-synonymous sites
- Compute the divergence between two species using a Mauve alignment

What next? E.g.: GO analysis with Gowinda (Kofler and Schlötterer, 2012, Bioinformatics)

- migrane inducing AMJ-complex
- AMJ GO-polymerase
- pink poodle inbreedase