The influence of chromosomal inversion on genetic variation and clinal patterns in genomic data

Martin Kapun[1]

[1] *Natural History Museum of Vienna, Vienna, Austria*

Introduction

Chromosomal inversions are structural mutations that result in a reorientation of the gene order in the affected genomic region. The reversal of synteny impedes homologous pairing in heterokaryotypic chromosomes and leads to loop structures at the chromosomal region spanned by the inversion. These characteristic inversion loops (Figure 1) can be examined under the microscope in giant polytene chromosomes, which are thousand-fold replicated chromatids within the nucleus that can be found, for example, in the salivary glands of many Drosophilid larvae. This allowed Alfred Sturtevant, a student of Thomas Hunt Morgan, to investigate the influence of inversions on recombination patterns more than 100 years ago in the fruit fly *Drosophila melannogaster*, which makes these structural polymorphisms one of the first mutations ever to be directly studied. Inversion are considered to be the result of ectopic recombination among repetitive and palindromic sequences as found in tRNAs, ribosomal genes and transposable elements (TEs). Accordingly, the breakpoints of inversion polymorphisms, which can range from less than 1,000 bp to several million basepairs in length, are often enriched for these repetitive sequences. The prevalence of inversions that either include or exclude the centromere (pericentric vs. paracentric inversions) in genomes can vary dramatically, even among closely related taxa, which may be linked varying numbers of genomic TEs. For example, in contrast to the fruit fly *D. melanogaster* which contains many inversions that are pervasive and common in many worldwide populations, its sister taxa *D. mauritiana* and *D. sechellia* are basically inversion-free.

The primary evolutionary effect of inversions is a strong suppression of recombination (but not gene conversion) with standard arrangement chromosomes since crossing-over within the inverted region results in unbalanced gametes that are non-viable. While recombination with paracentric inversions results in acentric and dicentric gametes, crossing-over with pericentric inversions can cause large-scale duplications and deletions in the recombination products. As a result, both the ancestral standard (ST) and the derived inverted (INV) karyotype evolve largely independent, resulting in increased divergence. Since most inversions likely result from unique mutation events, the newly formed inversion evolves from a single haplotype of the ancestral standard arrangement and diverges over time by accumulating novel mutations. Accordingly, inversions may strongly influence the genentic variation in the corresponding genomic region and even beyond their breakpoints. Large chromosomal inversions are generally considered deleterious since they lead to (1) inviable recombination products in heterozygous

state, (2) may results in pseudogenization in the breakpoint regions and (3) shift genes to other genomic regions, which may perturb their expression patterns. However, inveresions may also provide beneficial effects. In particular, when

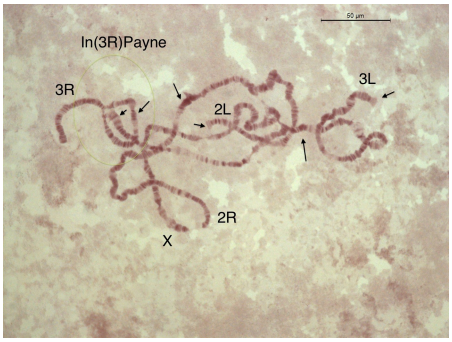

Figure 1: Figure 1

In this book chapter, we will focus on the fruit fly *Drosophila melanogaster*, which is characterized by seven chromosomal inversions, which are commonly found in most world-wide populations. Using genomic data from different sources and a broad range of bioinformatics analyses tools, we will study two common cosmopolitan inversions, *In(2L)t* and *In(3R)Payne*, in their ancestral African origin and investigate their effect on genetic variation and differentiation. We will identify single nucleotide polymorphisms (SNPs) in the proximity of the inversion breakpoints which are fixed for different alleles in the inverted and standard chromosomal arrangements. Using these SNPs as diganostic markers, we will subsequently estimate inversion frequencies in pooled resequencing (PoolSeq) data, where individuals with uncertain inversion status are pooled prior to DNA sequencing. In particular, we will utilize the DEST v.2.0 dataset, which is a collection of whole-genome pooled sequencing data from more than 700 world-wide *Drosophila melanogaster* population samples, densely collected through time and space. Using the inversion-specific marker SNPs, we will estimate the inversion frequencies of our two focal inversions in the PoolSeq data of each population sample and test how inversions influence genome-wide linkage disequilibrium and population structure. Furthermore, we will test for clinal patterns of the inversions in European and North American populations and investigate if these patterns can be explained by demography alone.

(1) Preparing the bioinformatics analyses pipeline

The full analysis pipeline including specific *Python* and *R* scripts can be found at https://github.com/capoony/InvChapter. As a first step, all necessary software needs to be installed. This information can be found in the shell-script `dependencies.sh` which is located in the `shell/` folder.

```
### define working directory
WD=</Github/InvChapter> ## replace with path to the downloaded GitHub repo https://github.co

## install dependencies
sh ${WD}/shell/dependencies
```

Then, we need to download genomic data from the Short Read Archive (SRA; XXX). We will use the *Drosophila* Nexus dataset and focus on genomic data of haploid individuals collected in Siavonga/Zambia with known karyotypes. In a first step, we will use a metadata-table, which contains the sample ID's, the corresponing ID's from the SRA database and the inversion status of common inversions, to select (up to) 20 individuals from each karyotype (INV and ST) for each of the two focal inversions. Finally, we will download the raw sequencing data for these samples from SRA

```
## Get information of individual sequencing data and isolate samples with known inversion st
mkdir ${WD}/data
cd ${WD}/data

### download metadata Excel table for Drosophila Nexus dataset
wget http://johnpool.net/TableS1_individuals.xls

### process table and generate input files for downstream analyses, i.e., pick the ID's and
Rscript ${WD}/scripts/ReadXLS.r ${WD}

### Define arrays with the inverions names, chromosome, start and end breakpoints; These dat
DATA=("IN2Lt" "IN3RP")
Chrom=("2L" "3R")
Start=(2225744 16432209)
End=(13154180 24744010)

## Get read data from SRA
mkdir ${WD}/data/reads
mkdir ${WD}/shell/reads
conda activate sra-tools

### loop over both inversions
for index in ${!DATA[@]}; do
    INVERSION=${DATA[index]}

    ## read info from input file {WD}/data/${INVERSION}.txt that was generated above with Re
    while
        IFS=',' read -r ID SRR Inv
    do
        if [[ -f ${WD}/data/reads/${ID}_1.fastq.gz ]]; then
            continue
```

```
        fi

        echo """
        ## download reads and convert to FASTQ files
        fasterq-dump \
            --split-3 \
            -o ${ID} \
            -O ${WD}/data/reads \
            -e 8 \
            -f \
            -p \
            ${SRR}
        ## compress data
        gzip ${WD}/data/reads/${ID}*
        """ >${WD}/shell/reads/${ID}.sh
        sh ${WD}/shell/reads/${ID}.sh
    done <${WD}/data/${INVERSION}.txt
done
```

In the next step, we will first trim the reads based on base-quality and map the filtered datasets against the *D. melanogaster* reference genome (v.6.57), which we will download from FlyBase. We will use a modified mapping pipeline from Kapun et al. (2020), which further filters for PCR duplicates and improves the alignment of nucleotides around indels.

```
### obtain D. melanogaster reference genome from FlyBase
cd ${WD}/data
wget -O dmel-6.57.fa.gz http://ftp.flybase.net/genomes/Drosophila_melanogaster/current/fasta

### index the reference genome for the mapping pipeline
conda activate bwa-mem2
bwa-mem2 index dmel-6.57.fa.gz
gunzip -c dmel-6.57.fa.gz >dmel-6.57.fa
samtools faidx dmel-6.57.fa
samtools dict dmel-6.57.fa >dmel-6.57.dict
conda deactivate

### trim & map & sort & remove duplicates & realign around indels
for index in ${!DATA[@]}; do
    INVERSION=${DATA[index]}
    while
        IFS=',' read -r ID SRR Inv
    do
        ## ignore header or continue if mapped dataset already exists
        if [[ ${ID} == "Stock ID" || -f ${WD}/mapping/${ID}_RG.bam ]]; then
```

```
            continue
        fi
        ## run the mapping pipeline with 100 threads (modify to adjust to your system resou
        sh ${WD}/shell/mapping.sh \
            ${WD}/data/reads/${ID}_1.fastq.gz \
            ${WD}/data/reads/${ID}_2.fastq.gz \
            ${ID} \
            ${WD}/mapping \
            ${WD}/data/dmel-6.57 \
            100 \
            ${WD}/scripts/gatk/GenomeAnalysisTK.jar
    done <${WD}/data/${INVERSION}.txt
done
```

Using the mapping pipeline, we aligend all reads against the *Drosophila melanogaster* reference genome. Thus, we can now obtain the allelic information for each sample at every position in the reference genome, which is stored in the final BAM files. Since the sequencing data was generated from haploid embryos, we assume that there is only one allele present in each sample at a given genomic position. We will now identify polymorphisms using the FreeBayes variant calling software and store the SNP information across all samples per inversion in a VCF file.

```
## SNP calling using freebayes with 100 threads
for index in ${!DATA[@]}; do

    INVERSION=${DATA[index]}
    while
        IFS=',' read -r ID SRR Inv
    do
        if [[ ${ID} == "Stock ID" ]]; then
            continue
        fi

        mkdir -p ${WD}/results/SNPs_${INVERSION}

        ### store the PATHs to all BAM files in a text, which will be used as the input for
        echo ${WD}/mapping/${ID}_RG.bam >>${WD}/mapping/BAMlist_${INVERSION}.txt

    done <${WD}/data/${INVERSION}.txt

    conda activate freebayes

    ### We assume ploidy = 1 and run FreeBayes in parallel by splitting the reference genom
    freebayes-parallel \
```

```
        <(fasta_generate_regions.py \
            ${WD}/data/dmel-6.57.fa.fai \
            100000) \
        100 \
        -f ${WD}/data/dmel-6.57.fa \
        -L ${WD}/mapping/BAMlist_${INVERSION}.txt \
        --ploidy 1 |
        gzip >${WD}/results/SNPs_${INVERSION}/SNPs_${INVERSION}.vcf.gz
    conda deactivate
done
```

(1) Patterns of genomic variation associated with different karyotpes in African populations

Once a inversions originates and persists in a population, novel mutations will appear and bulid up in frequency

(2) SNPs in strong linkage disequilibrium with inversions