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

Martin Kapun¹

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

### Introduction

Chromosomal inversions are structural mutations that result in the 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 light microscopes 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. These structures allowed investigating the influence of inversions on recombination patterns in the early days of genetics research almost 100 years ago in the fruit fly *Drosophila melanogaster*, and 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 thousand 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 to varying numbers of TEs in different genomes. 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 the strong suppression of recombination with standard arrangement chromosomes (i.e., in heterokaryotypes) 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 independently. However, the suppression of recombination is not perfect and two processess may lead to rare genetic exchange - so called “gene flux” - across karyotypes: (1) Two recombination events (double recombination) that happen at the same time within the inverted region may results in viable recombinant gamets. However, the synaptonemal complexes, which initiate crossing over, can only form when the homologous chromatids are fully paired. Thus, double recombination events never occur in the proximity of the inversion breakpoints and thus lead to a gradual increase of gene flux probability towards the inversion center. Conversely, (2) gene conversion, which results from the repair of DNA double-strand breaks, can lead to rare genetic exchange across the whole inverted region since this mechanism does not depend on paired chromatids.

When a new inversion arises, it captures a single haplotype of the ancestral standard arrangement. The evolutionary fate of the new inversion is initally determined by genetic drift and by the fitness of the captured haplotype relative to the rest of the population. Subsequently divergence among the karyotypes builds up continuously due to the accumulation of novel mutations. However, geneflux in distance to the breakpoints keeps homogenizing the genetic variation which will lead to a pattern of sequence divergence that resembles a “suspension bridge” patterns in inversions that are sufficiently old to harbour many novel mutations. Accordingly, inversions may strongly influence the patterns of genentic variation in the corresponding genomic region.

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

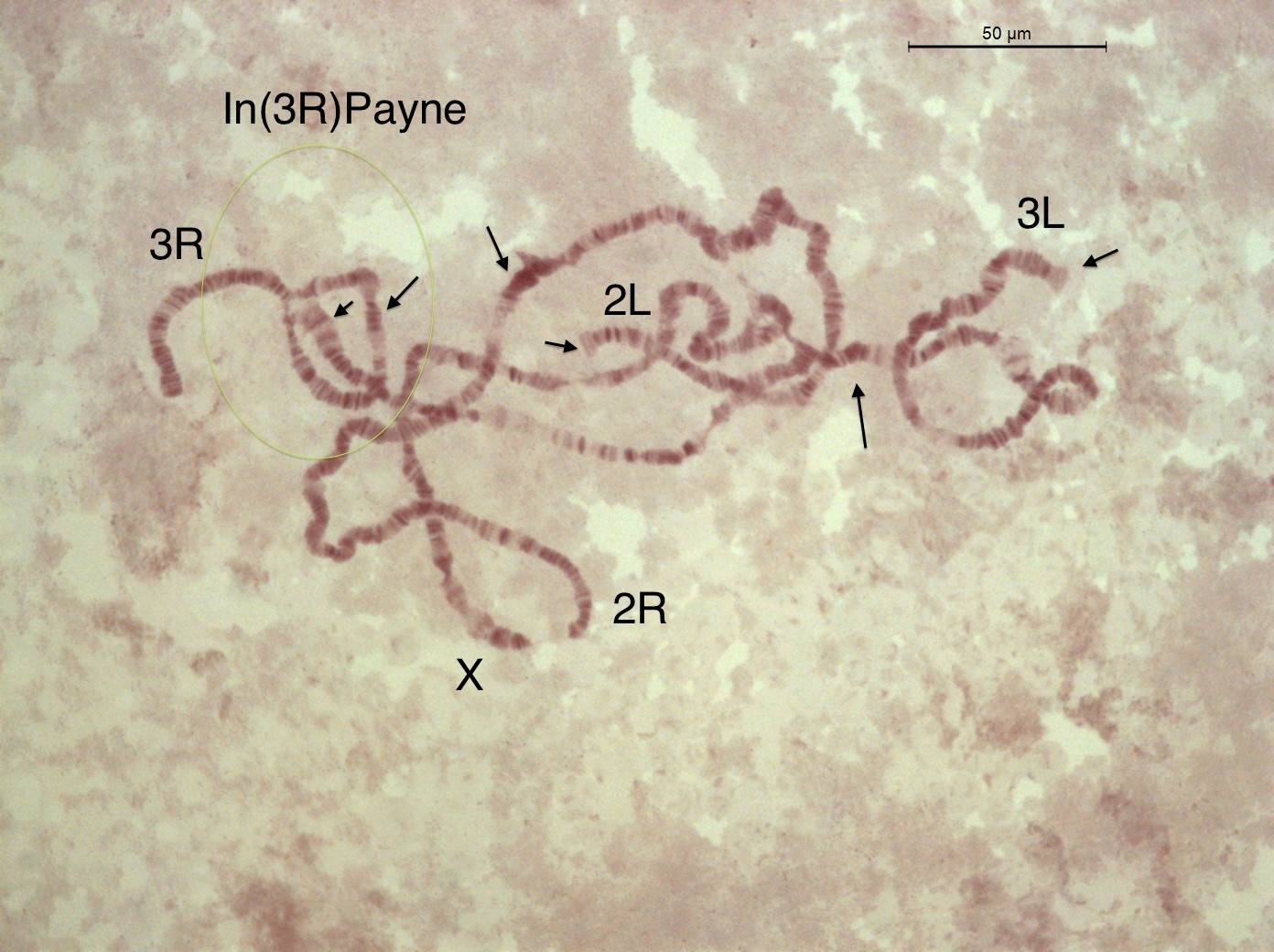


Figure 1

In this book chapter, we will employ a bioinformatics analysis pipeline to assess the influence of inversion on genetic variation in natural populations. We will focus on the fruit fly *Drosophila melanogaster*, which is characterized by seven chromosomal inversions that 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 a shell-script called dependencies.sh which is located in the shell/ folder. Here and throughout this chapter, the code blocks, as shown below, are highlighted by boxes with a different font and colors that highlight the syntax of the BASH scripting language that is throughout for the analysis pipeline. It is possible to copy and paste these code snippets directly from this document and paste it into the terminal window on a workstation computer or computer server with a LINUX operation system. However, I would recommend to open the script main.sh, which is located in the shell/ folder and which contains the whole analysis pipeline shown here, in an integrated development environment (IDE) program such as the [VScode](https://code.visualstudio.com/) editor and and execute the individual commands from the script bit by bit.

### define working directory  
WD=</Github/InvChapter> ## replace with path to the downloaded GitHub repo https://github.com/capoony/InvChapter  
  
## install dependencies  
sh ${WD}/shell/dependencies

Then, we 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. As you will see in the code block below, we will focus on the two inversions *In(2L)t* and *In(3R)Payne* which we abbreviate as *IN2Lt* and *IN3RP* for the sake of simplicity. The arrays DATA, Chrom, Start and End contain the information of the genomic position for both inversions.

## Get information of individual sequencing data and isolate samples with known inversion status  
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 SRA accession numbers for the first 20 individuals with inverted and standard karyotype, respectively.  
Rscript ${WD}/scripts/ReadXLS.r ${WD}  
  
### Define arrays with the inverions names, chromosome, start and end breakpoints; These data will be reused in the whole pipleine for the sequential analysis and visulaization of both focal inversions  
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 ReadXLS.r  
 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](https://flybase.org/). 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/dmel-all-chromosome-r6.57.fasta.gz  
  
## 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 ressources). Note that this step may take quite some time  
 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 FreeBayes  
 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 genome in chuncks of 100,000bps and and use GNU parallel for multithreading. I am using 100 threads. Please adjust to your system.   
 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

### (2) Patterns of genomic variation and differentiation associated with different karyotpes in African populations

In the next step, we will use population genetics statistics to compare genetic variation among Zambian indivduals with inverted and standard arrangement for the two inversions *In(2L)t* and *In(3R)Payne*. We will calulate Nei’s *π* (nucleotide diversity) as an estimator of genetic diversity in a population. Since VCFtools, the program which we use to calculate the population genetic statistics, does not allow calculating these statistics from haploid VCF files, we will first convert the haploid VCF to a diploid version by duplicating the haploid haplotype. Then, we will calculate π in 200,000 bp windows along the whole genome for the standard and the inverted individuals.

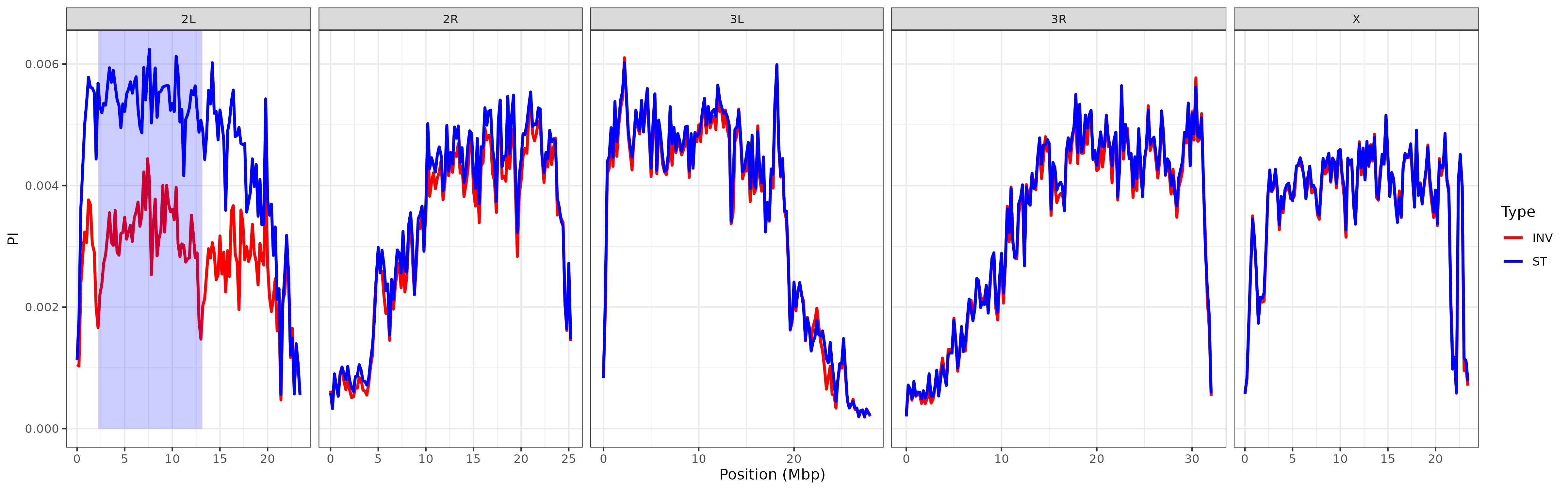
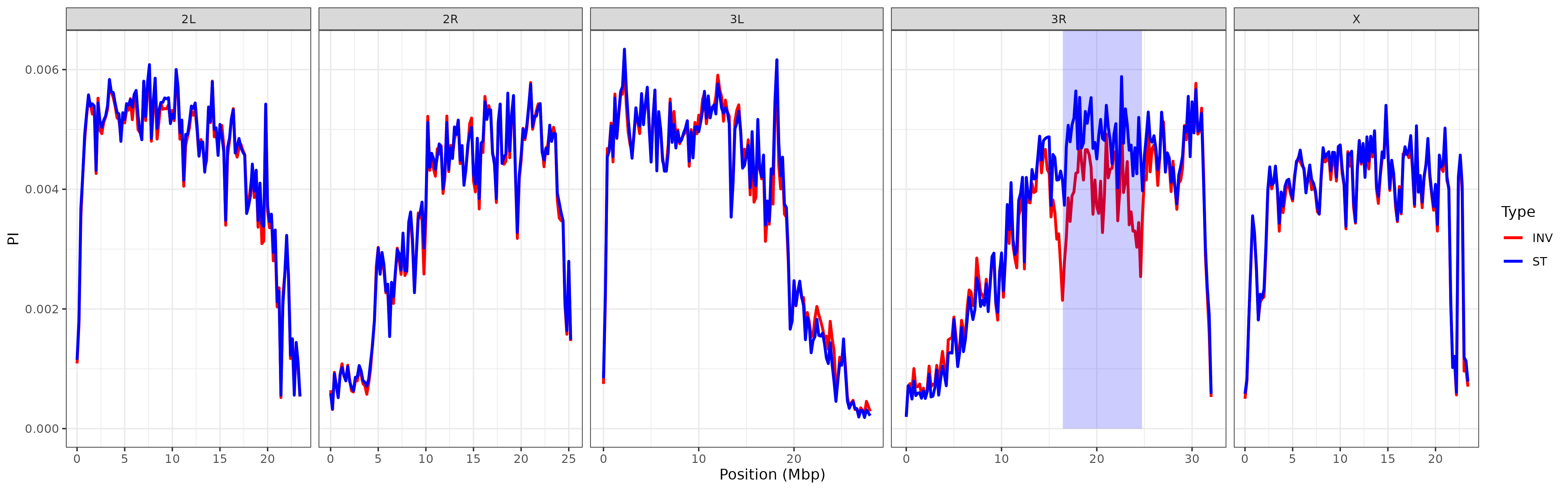
#### (2.1) The influence of inversions on genetic variation

## calculate pi per karyotype  
for index in ${!DATA[@]}; do  
 INVERSION=${DATA[index]}  
  
 mkdir ${WD}/data/${INVERSION}  
 output\_dir=${WD}/data/${INVERSION}  
  
 ### split input file with sample IDs based on Inversions status  
 awk -F',' '  
 {  
 filename = $3 ".csv"  
 filepath = "'$output\_dir'/" filename  
 if (filename == ".csv") next  
 print $1 >> filepath  
 }  
 ' ${WD}/data/${INVERSION}.txt  
  
 ### filter VCF for biallelic SNPs  
 conda activate vcftools  
  
 vcftools --gzvcf ${WD}/results/SNPs\_${INVERSION}/SNPs\_${INVERSION}.vcf.gz \  
 --min-alleles 2 \  
 --max-alleles 2 \  
 --remove-indels \  
 --recode \  
 --out ${WD}/results/SNPs\_${INVERSION}/SNPs\_${INVERSION}  
  
 gzip ${WD}/results/SNPs\_${INVERSION}/SNPs\_${INVERSION}.recode.vcf  
  
 ### convert haploid VCF to diploid  
 python ${WD}/scripts/hap2dip.py \  
 --input ${WD}/results/SNPs\_${INVERSION}/SNPs\_${INVERSION}.recode.vcf.gz \  
 --output ${WD}/results/SNPs\_${INVERSION}/SNPs\_${INVERSION}.recode\_dip.vcf.gz  
  
 for karyo in INV ST; do  
  
 ### calculate pi in 200kbp windows  
 vcftools --gzvcf ${WD}/results/SNPs\_${INVERSION}/SNPs\_${INVERSION}.recode\_dip.vcf.gz \  
 --keep ${WD}/data/${INVERSION}/${karyo}.csv \  
 --window-pi 200000 \  
 --out ${WD}/results/SNPs\_${INVERSION}/${INVERSION}\_${karyo}\_pi  
 done  
  
 ## combine pi of INV and ST chromosomes  
 awk 'NR ==1 {print $0"\tType"}' ${WD}/results/SNPs\_${INVERSION}/${INVERSION}\_INV\_pi.windowed.pi >${WD}/results/SNPs\_${INVERSION}/${INVERSION}\_pi.tsv  
 awk 'NR>1 {print $0"\tINV"}' ${WD}/results/SNPs\_${INVERSION}/${INVERSION}\_INV\_pi.windowed.pi >>${WD}/results/SNPs\_${INVERSION}/${INVERSION}\_pi.tsv  
 awk 'NR>1 {print $0"\tST"}' ${WD}/results/SNPs\_${INVERSION}/${INVERSION}\_ST\_pi.windowed.pi >>${WD}/results/SNPs\_${INVERSION}/${INVERSION}\_pi.tsv  
  
done

After we merged all *π* estimates for both karyotypes in a single file, we plot the genome-wide patterns in a line-plot for each chromosome and each karyotype. Here and in the rest of this pipeline, we will use *ggplot* from the *tidyverse* package in *R* for plotting. Note, for the sake of brevity, I collected all *R* code in individual *R*-scripts, which can all be found in the scripts/ folder.

### plot PI as line plot  
for index in ${!DATA[@]}; do  
  
 INVERSION=${DATA[index]}  
 St=${Start[index]}  
 En=${End[index]}  
 Ch=${Chrom[index]}  
  
 Rscript ${WD}/scripts/Plot\_pi.r \  
 ${INVERSION} \  
 ${Ch} \  
 ${St} \  
 ${En} \  
 ${WD}  
  
done

With the exception of the genomic region spanned either by *In(2L)t* (Figure 2; top) or *In(3R)Payne* (Figure 2; bottom), we find that *π*-values are very similar across INV (red) and ST (blue) samples across the whole genome. In contrast, we see that genetic variation within the inverted regions is markedly reduced in INV individuals which suggest that the population size of the inversion is smaller than the standard arrangement or that the inversion is not old enough and gene flux and novel mutations did not have enough time yet, to reconstitute genetic varaition similar to the ancestral arrangement. Moreover, we see that the reduction of genetic variation is not equal across the whole inverted region. Rather, the reduction is strongest close to the breakpoints. This indicates that gene-flux in the center of the inversion has, at least partilally, shifted genetic varaition from the standard arrangement into the inverted chromosomes. In the case of *In(2L)t*, we further find that the effect of the inversion on genetic variation is not confined to the region spanned by the inversion but rather spreads millions of basepairs beyond the breakpoints (Figure 2; top). Finally, we also observe that the distribution of nucleotide diversity varies along each chromosome and is strongly reduced close to the centro- and telomers (Figure 2). These particular regions are characterized by reduced recombination rates, which influences the extend of backround selection and leads to reduced variation in these chromosomal regions.

#### (2.2) The influence of inversions on genetic differentiation

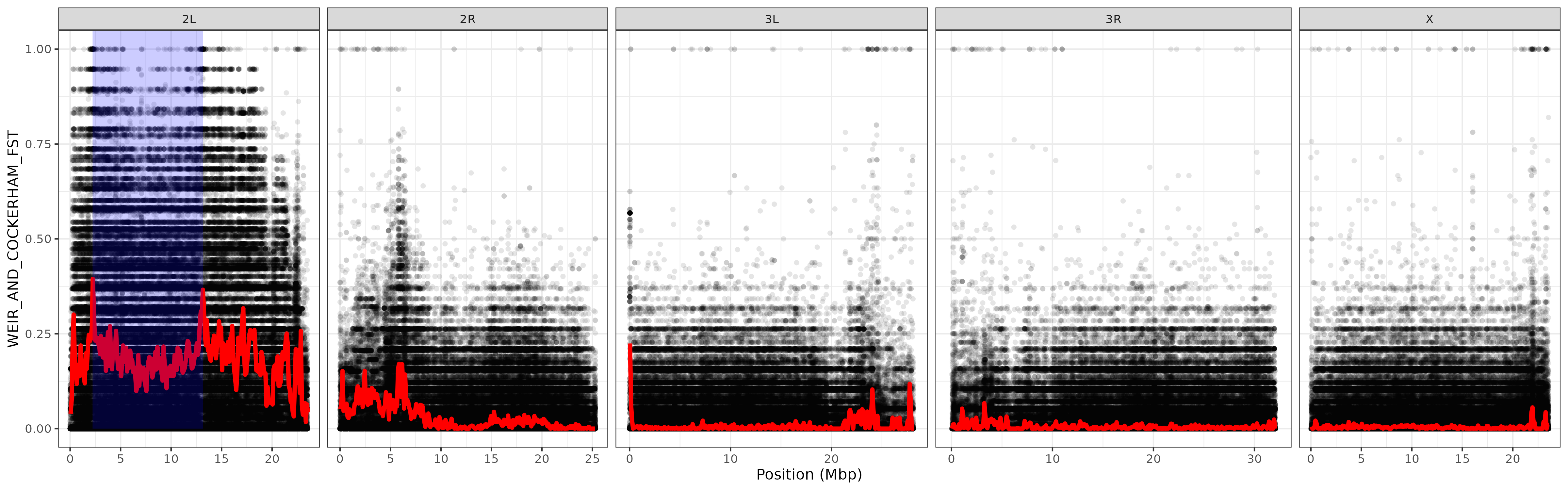
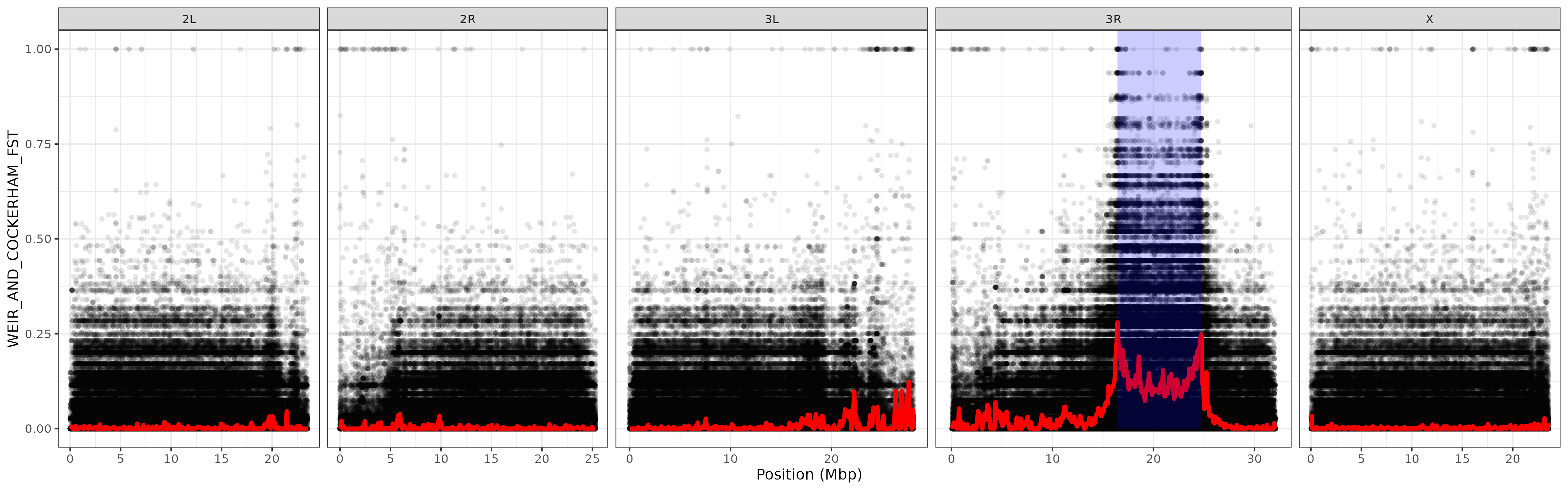
In the next step, we will use the diploid SNP dataset generated above to calculate *F*ST estimates among the INV and ST indivduals for each inversion using the method of Weir & Cockerham as implemented in VCFtools. The fixation index *F*ST summarizes genetic structure and is scaled between 0 (no differentiation) and 1 (complete differentiation).

## calculate FST between karyotypes  
for index in ${!DATA[@]}; do  
 INVERSION=${DATA[index]}  
  
 conda activate vcftools  
  
 ## calculate FST per SNP  
 vcftools --gzvcf ${WD}/results/SNPs\_${INVERSION}/SNPs\_${INVERSION}.recode\_dip.vcf.gz \  
 --weir-fst-pop ${WD}/data/${INVERSION}/INV.csv \  
 --weir-fst-pop ${WD}/data/${INVERSION}/ST.csv \  
 --out ${WD}/results/SNPs\_${INVERSION}/${INVERSION}.fst  
  
 ## calculate FST in 200kbp windows  
 vcftools --gzvcf ${WD}/results/SNPs\_${INVERSION}/SNPs\_${INVERSION}.recode\_dip.vcf.gz \  
 --weir-fst-pop ${WD}/data/${INVERSION}/INV.csv \  
 --weir-fst-pop ${WD}/data/${INVERSION}/ST.csv \  
 --fst-window-size 200000 \  
 --out ${WD}/results/SNPs\_${INVERSION}/${INVERSION}\_window.fst  
  
 conda deactivate  
done

Now, we plot both SNP-wise *F*ST as well as *F*ST - values in 200kbp windows. These type of plots are so-called Manhattan plots, where each dot respresents a polymorphic genomic position on the x-axis and the corresponding *F*ST - value on the y-axis. On top, we are plotting the window-wise *F*ST as a line and highlight the region of the corresponding inversion by a transparent blue box.

for index in ${!DATA[@]}; do  
  
 INVERSION=${DATA[index]}  
 St=${Start[index]}  
 En=${End[index]}  
 Ch=${Chrom[index]}  
  
 ### plot FST as Manhattan Plots  
 Rscript ${WD}/scripts/Plot\_fst.r \  
 ${INVERSION} \  
 ${Ch} \  
 ${St} \  
 ${En} \  
 ${WD}  
done

As you can see in Figure 3 for *In(2l)t* (top) and *In(3R)Payne* (bottom), genetic differentiation is elevated among the karyotypes within the inversion and particularly at and around the inversion breakpoints. These patterns suggest, that novel mutations building up over time in the proximity of the inversion breakpoints result in strong differentiation. Conistent with theory, the suppression of recombination prevents genetic homogenization among the karyotypes across the whole inverted region, but specifically at the breakpoints. Similar to Figure 2, we also see that patterns of differentation may spread way beyond the inversion breakpoint, as shown for *In(2L)t*, which further emphasizes the genome-wide impact of inversions on genetic variation.

### (3) SNPs in strong linkage disequilibrium with inversions

Several SNPs in the Manhattan plots of Figure 3 that are clustered at the inversion breakpoints show an *F*ST - value of one, which indicates complete fixation for different alleles among the two karyotpes. We therefore assume that these SNPs are in complete linkage disequilibrium (LD) with the inversion - at least in the particular Zambian population sample that we investigate here. This means that one allele is associated with the inverted karyotype and the other with the standard arrangement. Thus, it is possible to use these SNPs as diagnostic markers that allow to (1) estimate if the sequencing data of an individual with unknown karyotype is carrying the inversion simply by tracing for the inversion-specific allele at the correpsonding diagnostic markers. Furthermore, it is possible to estimate the frequency of inverted chromosomes in pooled sequencing data, where multiple individuals are pooled prior to DNA extraction and the pool of DNA is then sequenced jointly. In the latter type of datasets, it is assumed that the frequency of an allele in the pool corresponds to the actual frequency of the allele in the population from which the pooled individuals were randomly sampled. Thus, the median frequency of the inversion-specific alleles in the pooled dataset should roughly correspond to the inversion frequency given that these SNPs have been found to be in tight LD with the inversion. However, I need to caution here, that these markers should - at best - only be applied to sequencing data from samples collected in the same broader geographic region, or that diagnostic maker SNPs are defined using a mixed samples of individuals with known karyotype from all areas where the corresponding inversion occurs. The evolutionary history of inversions with a broad geographic distribution may be very complex and characterized by the emergence and fixation of different SNPs within the inversion in different geographic regions.

#### (3.1) Inversion-specific diagnostic marker SNPs

In the following, we will isolate SNPs located within 200kbp distance to each of the breakpoints that are in full LD with either of the two focal inversoin and obtain their alleles that are fixed within the inverted chromosomes. We will use a custom script that searches the inversion-specifc VCF files for SNPs with fixed differences among the INV and ST individuals as defined above within 200kbp around each inversion breakpoint.

## obtain diagnostic SNPs for each inversion  
for index in ${!DATA[@]}; do  
  
 INVERSION=${DATA[index]}  
 St=${Start[index]}  
 En=${End[index]}  
 Ch=${Chrom[index]}  
   
 ### store the chormosome, start and endpoints of each inversion as a comma-separated string  
 BP="${Ch},${St},${En}"  
  
 ### only retain the header and the rows on the "correct" chromosome and focus on the focal individuals that are either INV or ST  
 gunzip -c ${WD}/results/SNPs\_${INVERSION}/SNPs\_${INVERSION}.recode.vcf.gz |  
 awk -v Ch=${Ch} '$1~/^#/|| $1 == Ch' |  
 python ${WD}/scripts/DiagnosticSNPs.py \  
 --range 200000 \  
 --breakpoints ${BP} \  
 --input - \  
 --output ${WD}/results/SNPs\_${INVERSION}/${INVERSION} \  
 --MinCov 10 \  
 --Variant ${WD}/data/${INVERSION}.txt  
done

#### (3.2) Estimating inversion frequencies in Pool-Seq data

This analysis resulted in 62 and 26 diagnostic SNPs for *In(2L)t* and *In(3R)Payne*, respectively. In the next part, we will apply this diganostic maker SNPs to the largest Pool-Seq dataset of natural *D. melanogaster* populations available. The DEST v.2.0 dataset combines more than 700 population samples of world-wide fruitfly populations from different sources that were densely collected through space and time and