## The influence of chromosomal inversions 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 represent 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 a 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

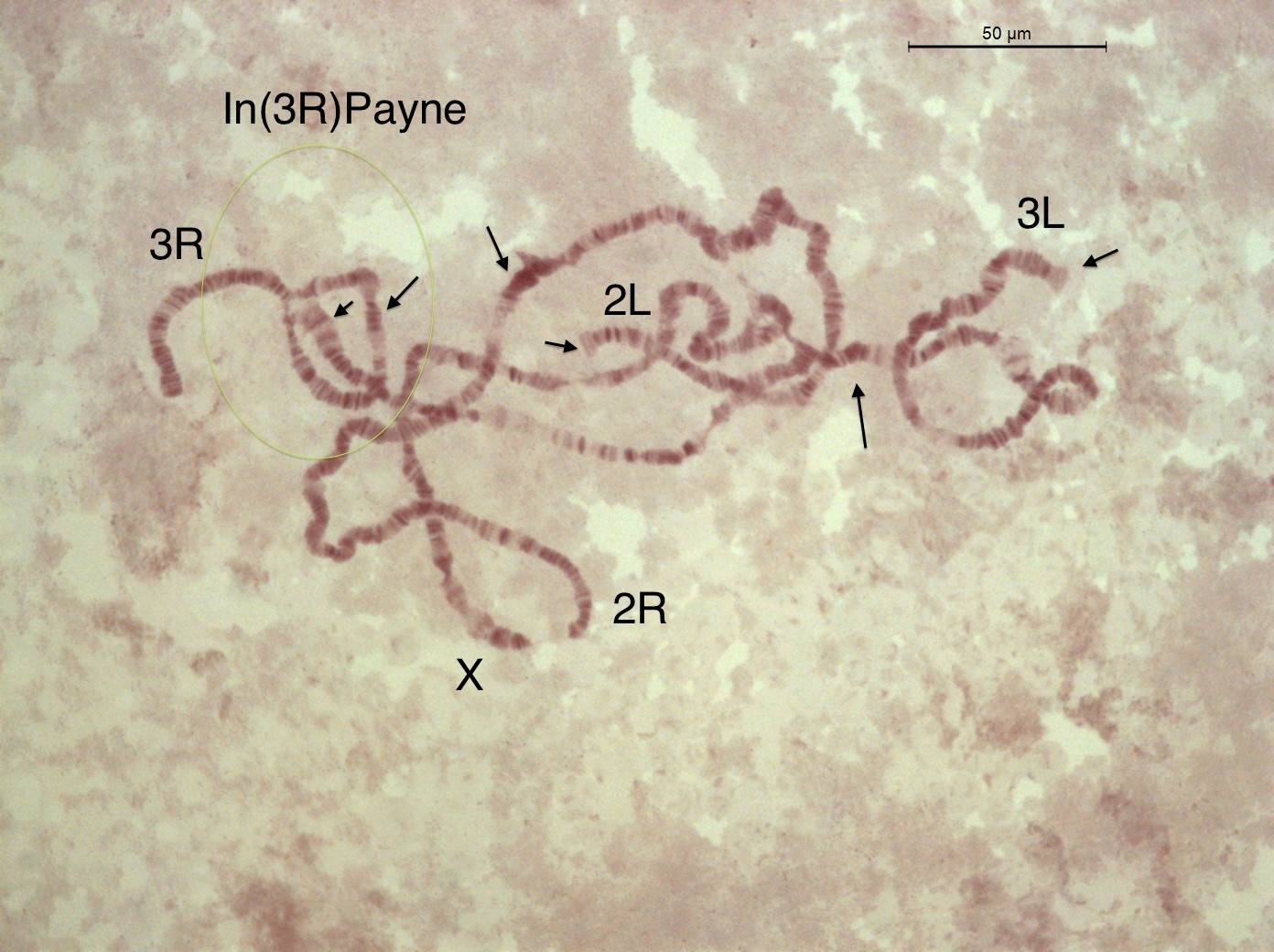


Figure 1

In this book chapter, we will employ a bioinformatics analysis pipeline to assess the influence of inversions 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 of this book chapter 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 fonts and colors that indicate the syntax of the BASH shell scripts, which represents the main scripting language used in this analysis pipeline, besides specific scripts written in th *Python* and *R*. It is possible to copy and paste these code snippets directly from the document and then paste and exceute them in a terminal window on a workstation computer or computer server with a LINUX operation system. However, I would strongly 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 need to download sequencing data from the Short Read Archive (SRA; XXX) to obtain genomic information of individuals which have presviously been characterized for their karyotype. These data will allow us to analyze the influence of inversions on genome-wide patterns of variation and differentiation. We will use the *Drosophila* Nexus dataset and focus on genomic data of haploid individuals collected in Siavonga/Zambia with known karyotypes, which have been sequenced by shotgun sequencing from paired-end libraries with Illumina technology. In a first step, we will obtain data from a metadata-table, which contains the sample names, the corresponding IDs from the SRA database and the inversion status of common inversions. We will use this infomation to select (up to) 20 individuals from each karyotype (INV and ST) for each of the two common cosmopolitan inversions, *In(2L)t* and *In(3R)Payne*. For each of the two inversions we make sure, that INV and ST individuals only carry either the focal but not the other inversion to avoid possible linked effects on genomic variation if both inversions (or only the “wrong” inversion) are present. While such putative effects are indeed also interesting, this is beyond the scope of this book chapter. Here, we want to keep the analysis simple and focused on the individual effects of each inversion. After the samples, that we will use for our downstream analyses, have been selected from the metadata table with a specific *R* script based on the criteria described above, we will download their raw sequencing data 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 information of the genomic position for each of the two 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 raw sequencing data based on base-quality (PHRED score >= 18) and map the trimmed and 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. In brief, this mapping pipeline first uses the program cutadapt to trim away sequencing adapter sequences which are still present in the raw reads and further trims flanking sequences if their base quality encoded as PHRED score is below 18. Since the sequencing data represent paired-end libraries, we will only retain intact read-pairs, where each read has a minimum length of 25 bp for the downstream analyses. We use the program bwa-mem2 for mapping (i.e. aligning) the trimmed read-pairs against the reference genome before sorting the resulting BAM file, which we filter for mapping quality (PHRED >= 20) and correctly oriented read-pairs, by genomic position. Using the program Picard, we then identify and remove PCR duplicates in the dataset, which represents reads with identical mapping position that are particularly problematic in pooled sequencing data as they may bias allele frequency estimatiates. Finally, we use the program GATK to realign reads around indel position to avoid false positive SNP variants due to misalignment. To keep the analysis pipeline concise, I collected all commands for the mapping pipleine in a separate shell script mapping.sh, which can be fond in the shell/ folder.

## 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  
 ### 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 correct 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. To speed this analyses up, I am using GNU parallel with 100 threads.

## SNP calling using freebayes with 100 threads  
for index in ${!DATA[@]}; do  
  
 INVERSION=${DATA[index]}  
 while  
 IFS=',' read -r ID SRR Inv  
 do  
 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 an African population from Zambia.

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 variation 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 for each individual. Then, we will calculate *π* in 200,000 bp windows along the whole genome for the standard and the inverted individuals.

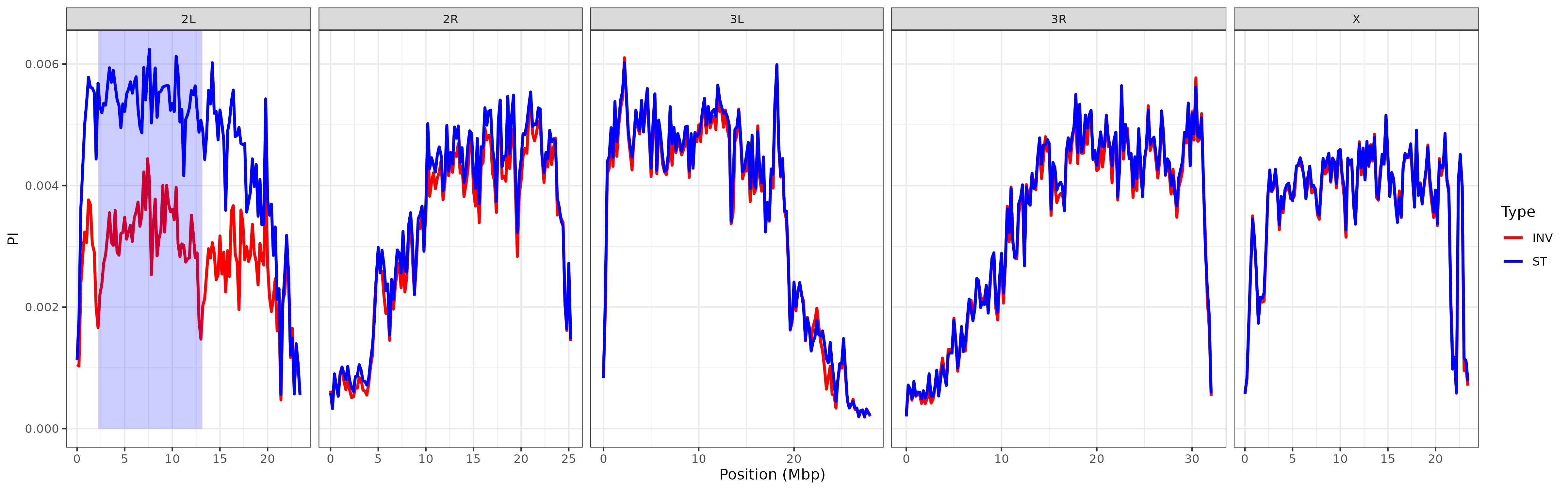
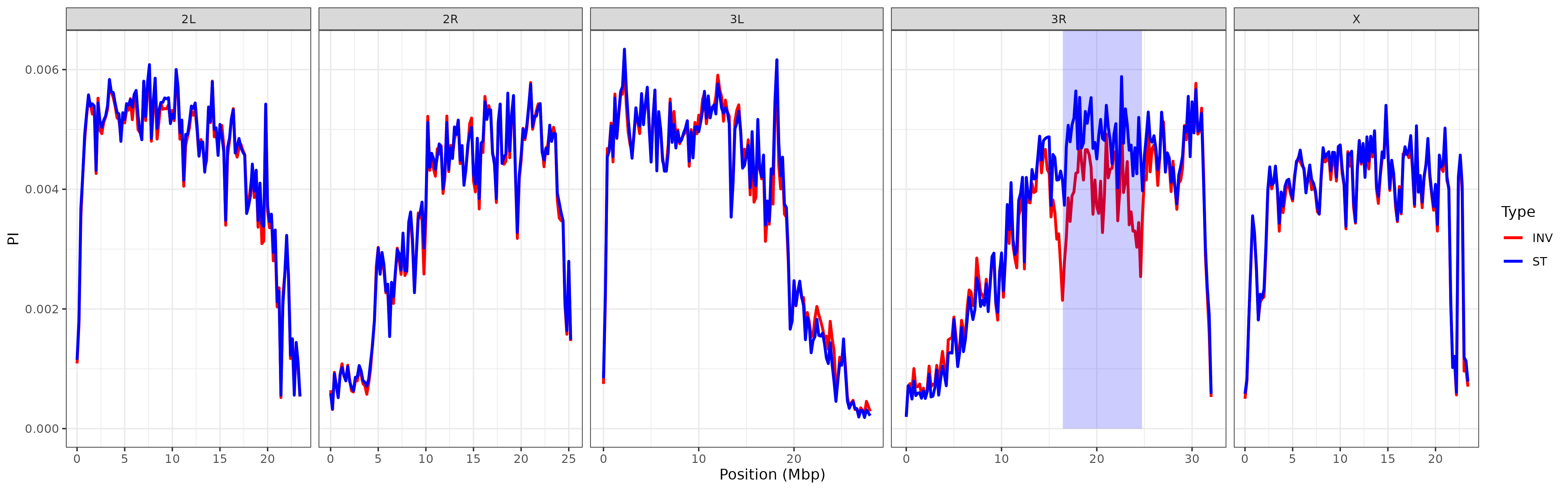
#### (2.1) The influence of inversions on genetic variation

## 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 for each inversion, 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 for 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. These patterns may also indicate that the inversion is not old enough and gene flux and novel mutations did not have enough time yet to reconstitute genetic variation 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

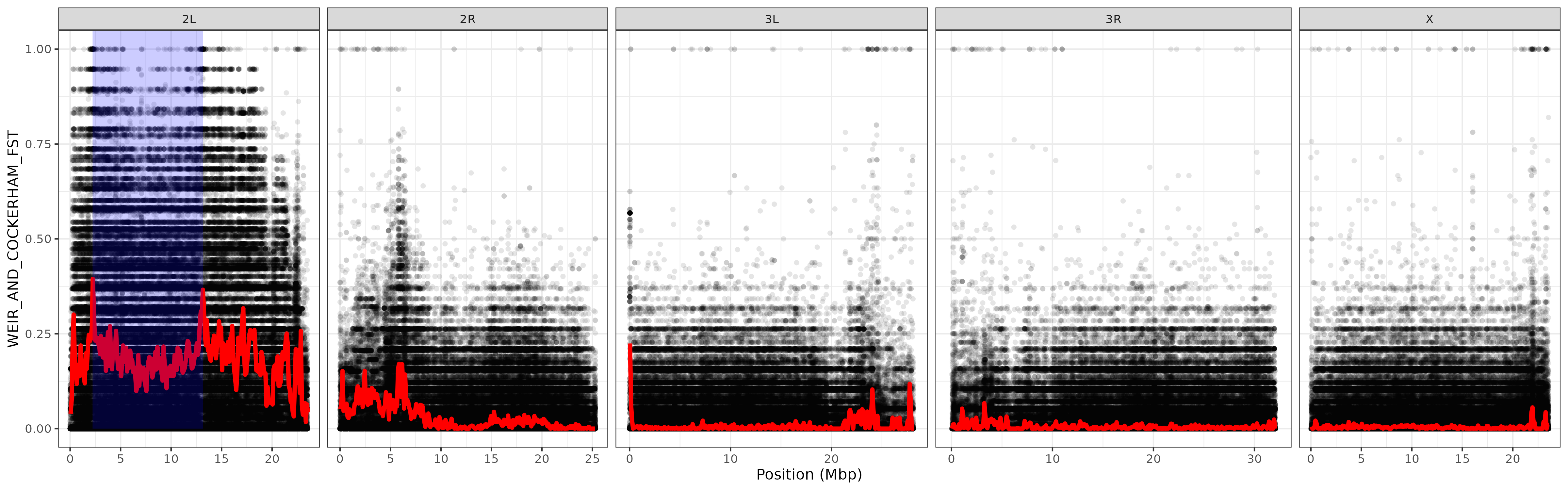
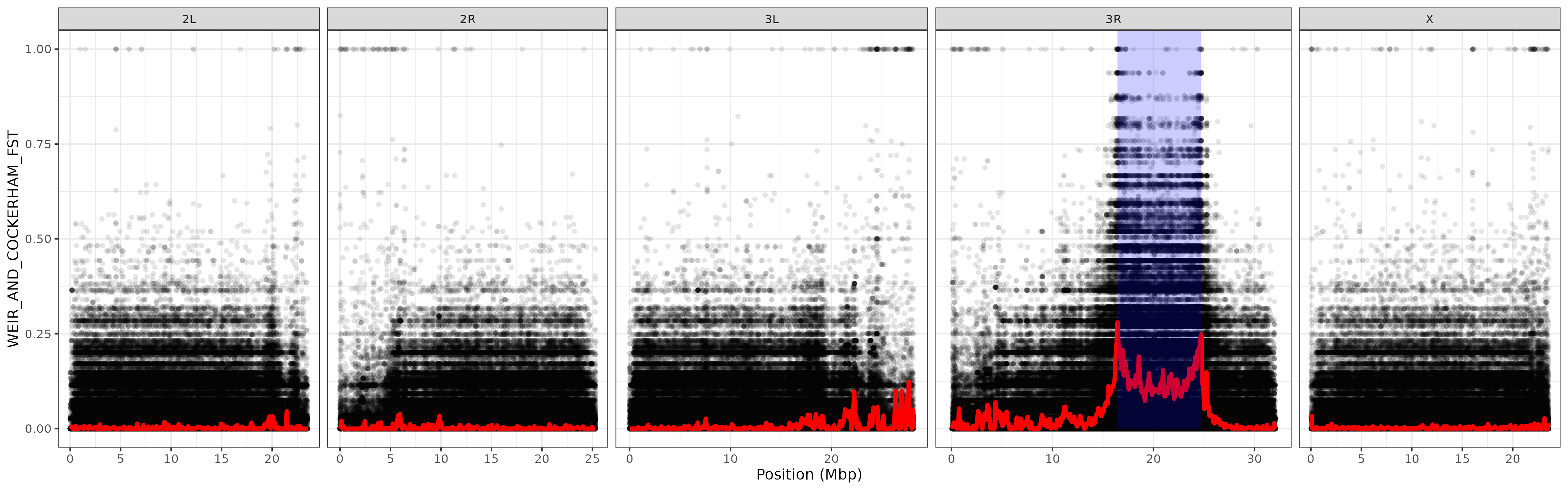
In the next part, we will use the diploid SNP dataset generated above to calculate *F*ST estimates among the INV and ST indivduals for each inversion using the method of Weir & Cockerham as implemented in VCFtools. The fixation index *F*ST summarizes genetic structure and is scaled between 0 (no differentiation) and 1 (complete differentiation). We will use this metric to identify single SNPs, which are strongly differentiated between the karyotpyes. In addition, we will calculate FST averaged in 200,000 bp windows to find genomic regions, where many neighbouring SNPs show similar differentiation patterns.

## 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

This analysis resulted in 62 and 26 diagnostic SNPs for *In(2L)t* and *In(3R)Payne*, respectively. In the following paragraphs, we will use these marker SNPs to indirectly infer inversion frequencies in other genomic datasets, but before that, we will test if inversions influence population structure in *D. melanogaster* population samples from North America and Europe without prior information on inversion frequencies in the corresponding samples.

#### (3.2) The influence of inversions on population structure

To this end, we will use the largest Pool-Seq dataset of natural *D. melanogaster* populations available to date. The DEST v.2.0 dataset combines more than 700 population samples of world-wide fruitflies from different sources that were densely collected through space and time mostly from North American and from European populations. All shotgun sequence data were processed with a standardized trimming and mapping pipeline (as described above) prior to joint SNP calling with the heuristic variant caller PoolSNP. Moreover, DEST v.2.0. also provides rich metadata, including detailed information on the sampling date and location, basic sequencing statistics (such as read depths, SNP counts, etc.) and recommendations based on data quality assessments. With the help of these metadata, we will subset the full data and only consider population samples of high quality and focus on samples collected from North America and Europe. Then, we will apply principal component analyses (PCA) to the allel frequency dataset in order to identify genome-wide differences among the samples in our datasets. In essence, the first few orthogonal PC-axes capture most of genetic variation shared across genome-wide SNPs and reflect the shared evolutionary history of the populations in the dataset. For these reasons, PCA is a very popular model-free method to quantify population structure. In our example, we will compare the results of PCA applied to SNPs either located within the genomic region spanned by each of the inversions or in distance to these genomic regions. These analyses will reveal if the inversions have an influence on population structure in their genomic region and how these patterns differ from genome-wide estimates.

As a first step, we will download both the DEST v.2.0 SNP data in VCF file-format and the corresponding metadata as a comma-separated (CSV) table from the DEST website. In addition, we will download two scripts from the DEST pipeline that are needed for the downstream analaysis.

### download VCF file and metadata for DEST dataset  
cd ${WD}/data  
wget -O DEST.vcf.gz http://berglandlab.uvadcos.io/vcf/dest.all.PoolSNP.001.50.3May2024.ann.vcf.gz  
wget -O meta.csv https://raw.githubusercontent.com/DEST-bio/DESTv2/main/populationInfo/dest\_v2.samps\_3May2024.csv  
  
### download sripts  
cd ${WD}/scripts  
wget https://raw.githubusercontent.com/DEST-bio/DESTv2\_data\_paper/main/16.Inversions/scripts/VCF2sync.py  
wget https://raw.githubusercontent.com/DEST-bio/DESTv2\_data\_paper/main/16.Inversions/scripts/overlap\_in\_SNPs.py

Using the metadata table, we identify all samples that we will include in our continent-wide analyses of populations from North America and Europe. Furthermore, we will exclude samples that did not pass the quality thresholds defined previously for the DEST v.2.0. dataset.

### Split metadata by continent  
  
## remove single quotes from metadata table  
sed -i "s/'//g" ${WD}/data/meta.csv  
  
## split by continent  
awk -F "," '$6 =="Europe" {print $1}' ${WD}/data/meta.csv >${WD}/data/Europe.ids  
awk -F "," '$6 =="North\_America" {print $1}' ${WD}/data/meta.csv >${WD}/data/NorthAmerica.ids  
  
## get data for populationes that did not pass the quality criteria (no PASS and average read depths < 15)  
awk -F "," '$(NF-7) !="Pass" || $(NF-9)<15 {print $1"\t"$(NF-7)"\t"$(NF-9)}' ${WD}/data/meta.csv >${WD}/data/REMOVE.ids

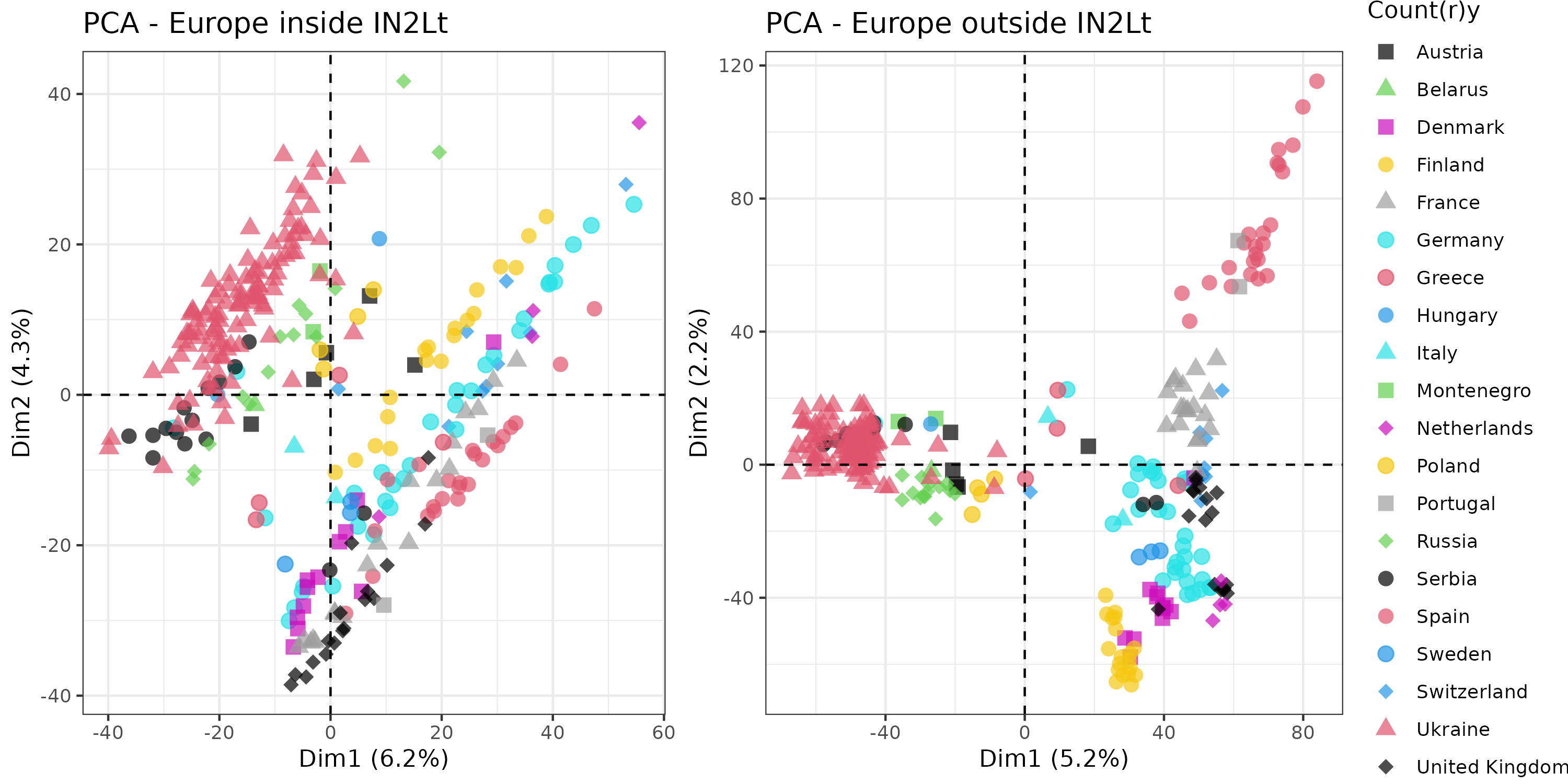
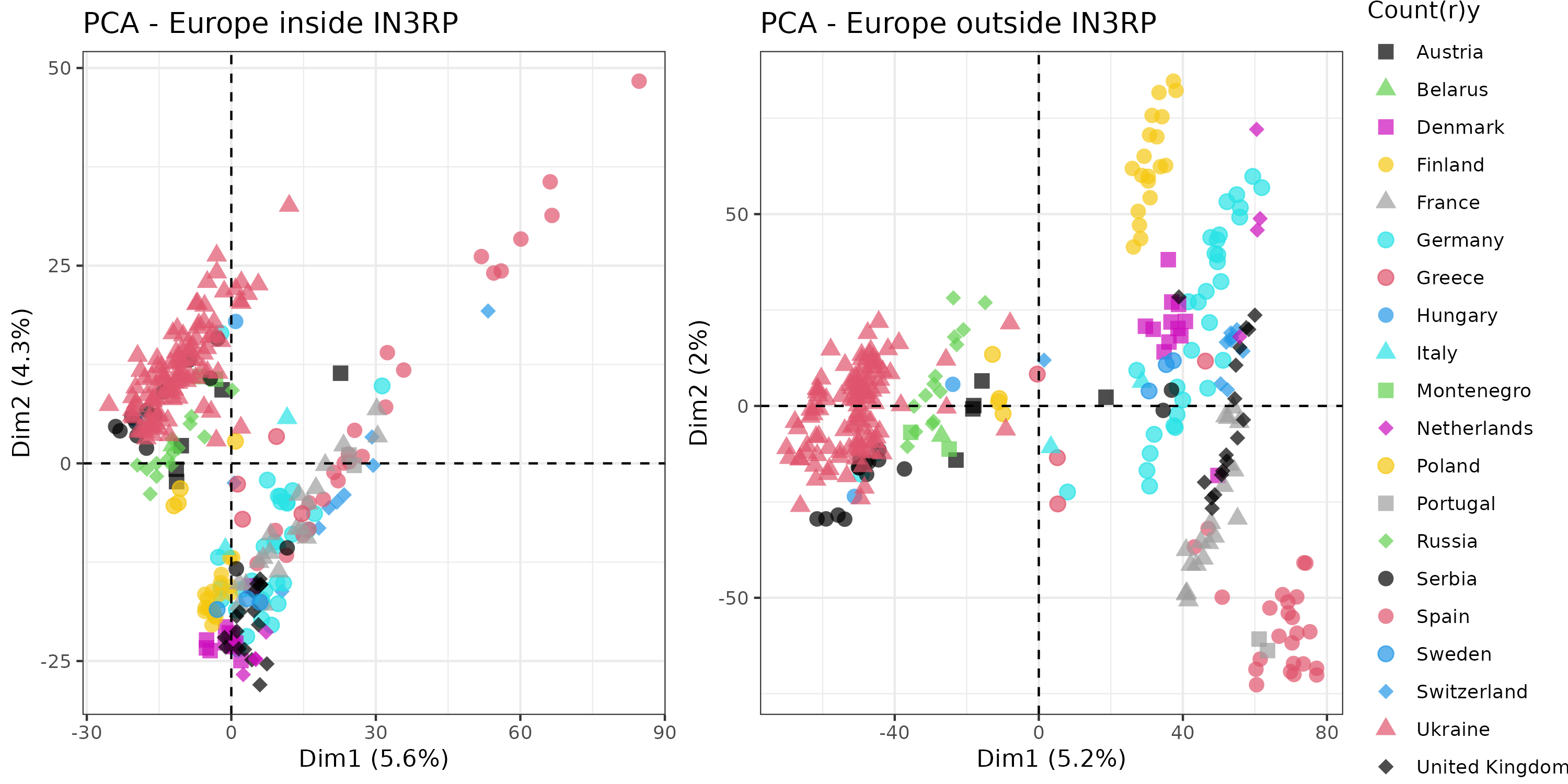
Next, we will apply several filtering steps to the VCF file and based on metadata information, we will construct two continent-specific datasets consisting of allele frequency data that we will use for all downstream analyses. Specifically, we will (1) isolate continent-specific populations, (2) remove problematic samples (based on DEST recommendations), remove (3) populations with < 15-fold average read depth, (4) only retain bilallic SNPs, (5) convert the allele counts to frequencies of the reference allele and obtain (6) read-depths for each position and population sample. Finally, (7) we will restrict our analyses to 50,000 randomly drawn genome-wide SNPs. The final files will represent a two-dimensional matrix of frequencies of the allele that is present in the reference genome, where each row represents one polymorphic genomic position and each column shows the data for one poluation sample.

### subset the VCF file   
  
mkdir ${WD}/results/SNPs  
  
for continent in NorthAmerica Europe; do  
  
 conda activate vcftools  
  
 ## decompress VCF file  
 pigz -dc ${WD}/data/DEST.vcf.gz |  
  
 ## keep header and position with only one alternative allele  
 awk '$0~/^\#/ || length($5)==1' |  
  
 ## keep continental data and remove bad quality samples  
 vcftools --vcf - \  
 --keep ${WD}/data/${continent}.ids \  
 --remove ${WD}/data/REMOVE.ids \  
 --recode \  
 --stdout |  
  
 ## remove rows with missing data  
 grep -v "\./\." |  
  
 ## randomly samples 50,000 SNPs  
 python ${WD}/scripts/SubsampleVCF.py \  
 --input - \  
 --snps 50000 |  
  
 ## convert VCF to allele frequencies and weigths (of the reference allele)  
 python ${WD}/scripts/vcf2af.py \  
 --input - \  
 --output ${WD}/results/SNPs/${continent}  
  
done

Now, we will employ principal component analyses (PCA) test if the genetic variation in the genomic region spanned by an inversion influences the signals of population structure. To this end, we will execute the *R* script PCA\_inv.r in the scripts/ folder to carry out the following analysis steps: At first, we will load the allele frequency datasets generated above and split the data in two subsets, where one subset contains SNPs located within the genomic region of a given inversion and the other contains SNPs from the remaining part of the genome. Then, we will perform PCA on the transposed allele frequency matrices, where columns represent chromosomal positions and rows represent population samples. Then, we will use metadata information to highlight the country/county of origin in scatterplots that show the first two PC-axes for each subset for each of the continents and inversions.

### use PCA to test for patterns inside and outside the genomic region spanned by an inversion  
for index in ${!DATA[@]}; do  
  
 INVERSION=${DATA[index]}  
 St=${Start[index]}  
 En=${End[index]}  
 Ch=${Chrom[index]}  
  
 Rscript ${WD}/scripts/PCA\_Inv.r \  
 ${INVERSION} \  
 ${Ch} \  
 ${St} \  
 ${En} \  
 ${WD}  
  
done

The plots in Figure 4 reveal that the genomic region spanned by the inversions (shown on the left) indeed differs in population structure compared to patterns based on the remaining genome (on the right) as shown in scatterplots for the first two PC-axes, which together explain between 8%-12% of the total genetic variation (of all SNPs). Notably, the PC-scores of genome-wide SNPs in Europe (shown on the right) both cluster populations according to expectations based on geography. Conversely, PC-scores calculated from SNPs inside the breakpoints of *In(2L)t* and *In(3R)Payne* are much more compressed and appear to mostly follow a diagonal rather than clustering according to geography - particularly for PC1. The same analyses for North American samples show a very similar pattern and the corresponding graphs cand be found in the folder output/. We may thus speculate that genetic variation associated with inversion, and thus particularly the inversion frequencies in the investigated populations strongly contribute to these patterns. However, unless the individuals that have been pooled before sequencing were karyotyped prior to DNA extraction, it is not possible to directly investigate the karyotypic status in these population samples. However, there is hope…

### (4) Inversion frequencies in Pool-Seq data

We will now take advantage of the diagnostic marker SNPs that we isolated above and estimate inversion frequencies in each population sample. This will allow us to test more directly how inversions influence genetic variation and population structure and if the two inversions are associated with environmental variation and exhibit clinal variation.

#### (4.1) Estimating inversion frequencies in Pool-Seq data with diagnostic markers

We will now convert the VCF file to the SYNC file format using the Python script VCF2sync.py from the DEST pipeline. The SYNC file format is commonly used to store allele counts in pooled re-sequencing data as colon-separated lists in the form A:T:C:G:N:Del for each population sample and position. We will then obtain allele counts from the SYNC file at the positions of inversion-specific marker SNPs that are present in the DEST dataset using the Python script overlap\_in\_SNPs.py. To speed these calculations up, I am using GNU parallel with 100 threads.

### convert VCF to SYNC file format  
conda activate parallel  
gunzip -c ${WD}/data/DEST.vcf.gz |  
 parallel \  
 --jobs 100 \  
 --pipe \  
 -k \  
 --cat python3 ${WD}/scripts/VCF2sync.py \  
 --input {} |  
 gzip >${WD}/data/DEST.sync.gz  
  
### Get positions at inversion specific marker SNPs  
for index in ${!DATA[@]}; do  
  
 INVERSION=${DATA[index]}  
 gunzip -c ${WD}/data/DEST.sync.gz |  
 parallel \  
 --pipe \  
 --jobs 100 \  
 -k \  
 --cat python3 ${WD}/scripts/overlap\_in\_SNPs.py \  
 --source ${WD}/results/SNPs\_${INVERSION}/${INVERSION}\_diag.txt \  
 --target {} \  
 >${WD}/data/DEST\_${INVERSION}.sync  
done

For each population and for each of the two inversions, we now calculate the median frequency of the inversion-specific alleles across all diagnostic markers to obtain an estimate of the corresponding inversion frequency with a custom Python script. Before that, we need to obtain the names of all samples in the VCF file in the correct order and then output the estimated inversion frequencies as a tab-delimted file.

### get the names of all samples in the VCF file and store as an array  
NAMES=$(gunzip -c ${WD}/data/DEST.vcf.gz | head -150 | awk '/^#C/' | cut -f10- | tr '\t' ',')  
  
# Calculate median frequencies for marker SNPs  
for index in ${!DATA[@]}; do  
  
 INVERSION=${DATA[index]}  
  
 python3 ${WD}/scripts/inversion\_freqs.py \  
 --marker ${WD}/results/SNPs\_${INVERSION}/${INVERSION}\_diag.txt \  
 --input ${WD}/data/DEST\_${INVERSION}.sync \  
 --names $NAMES \  
 --inv ${INVERSION} \  
 >${WD}/results/SNPs\_${INVERSION}/${INVERSION}.af  
  
done

To visually inspect the distribution of inversion-specific alleles across all diagnostic markers for each inversion and population sample, we plot historgrams of all inversion-specific allele frequencies and the actual allele frequencies of all diagnostic SNPs against their genomic position. In addition, we plot the median frequency, which we consider the estimated inversion frequency, as a dashed line atop the frequency histogram. We therefore need to first generate a table with the inversion-specific allele frequencies of the diagnostic SNPs for all population samples in the DEST VCF file. Then, we generate the above-mentioned plots of allele frequencies in *R* using the script Plot\_InvMarker.r in the scripts/ folder.

### generate plots for each population  
for index in ${!DATA[@]}; do  
  
 INVERSION=${DATA[index]}  
 Ch=${Chrom[index]}  
  
 ### convert VCF to allele frequency table for each SNP and population sample  
 gunzip -c ${WD}/data/DEST.vcf.gz |  
 awk -v Ch=${Ch} '$1~/^#/|| $1 == Ch' |  
 python3 ${WD}/scripts/AFbyAllele.py \  
 --input - \  
 --diag ${WD}/results/SNPs\_${INVERSION}/${INVERSION}\_diag.txt \  
 >${WD}/results/SNPs\_${INVERSION}/${INVERSION}\_pos.af  
  
 ### make plots in R  
 Rscript ${WD}/scripts/Plot\_InvMarker.r \  
 ${INVERSION} \  
 ${WD}  
done

The example in Figure 5 below shows the distribution of inversion-specific alleles for *In(2L)t* in a population sample collected in 2015 close to Mautern in the beautiful Wachau area along the Danube in Austria. As you can see in the scatterplot, two sets of SNPs are located around the breakpoints of *In(2L)t* and the frequencies of the invesion-specific alleles span from 0% to more than 60%, which is quite a broad range and presumably the result of sampling error in the PoolSeq data. Assume that even if a diasgnostic SNP is in full linkage disequilibrium with the inversion, it will not necessarly depict the “true” frequency of the inversion in the population due to binomial sampling. When sequencing the pooled DNA from multiple samples with NGS methods such as Illumina, we are usually sampling (i.e., sequencing) 50-100 DNA fragments at every genomic position, which corresponds to a 50-100 fold sequencing depth. This is, of course, only a very small fraction of the total millions of copies of DNA from all body cells of the individuals in the extracted DNA. The resulting sampling error leads to deviations from the expected frequency (i.e., the true allele frequency in all the DNA copies) and these deviations become even bigger the lower the sequencing depths are. However, if we further assume that each SNP is a reliable estimator of the “true” inversion frequency due to perfect linkage disequilibrium, we expect that the inferred frequencies across all marker SNPs roughly follow a binomial distribution, where the sequencing depth corresponds to the number of trials *n* and the expected inversion frequency corresponds to the number of successces *p*. However, other factors, such as sequencing and mapping errors or imperfect linkage disequilibrium of some diagnostic SNPs in certain geographic regions (see above in the introduction to chapter 3) may also influence the distirbution of frequencies. Rather than calculating the mean frequencies across all markers, we use the median to estimate the population inversion frequency and compare inversion patterns in all population samples, since this statistic is more robust to assymetric distributions. In our example in Figure 5 below, the median is shown as a dashed red line at app. 25%.

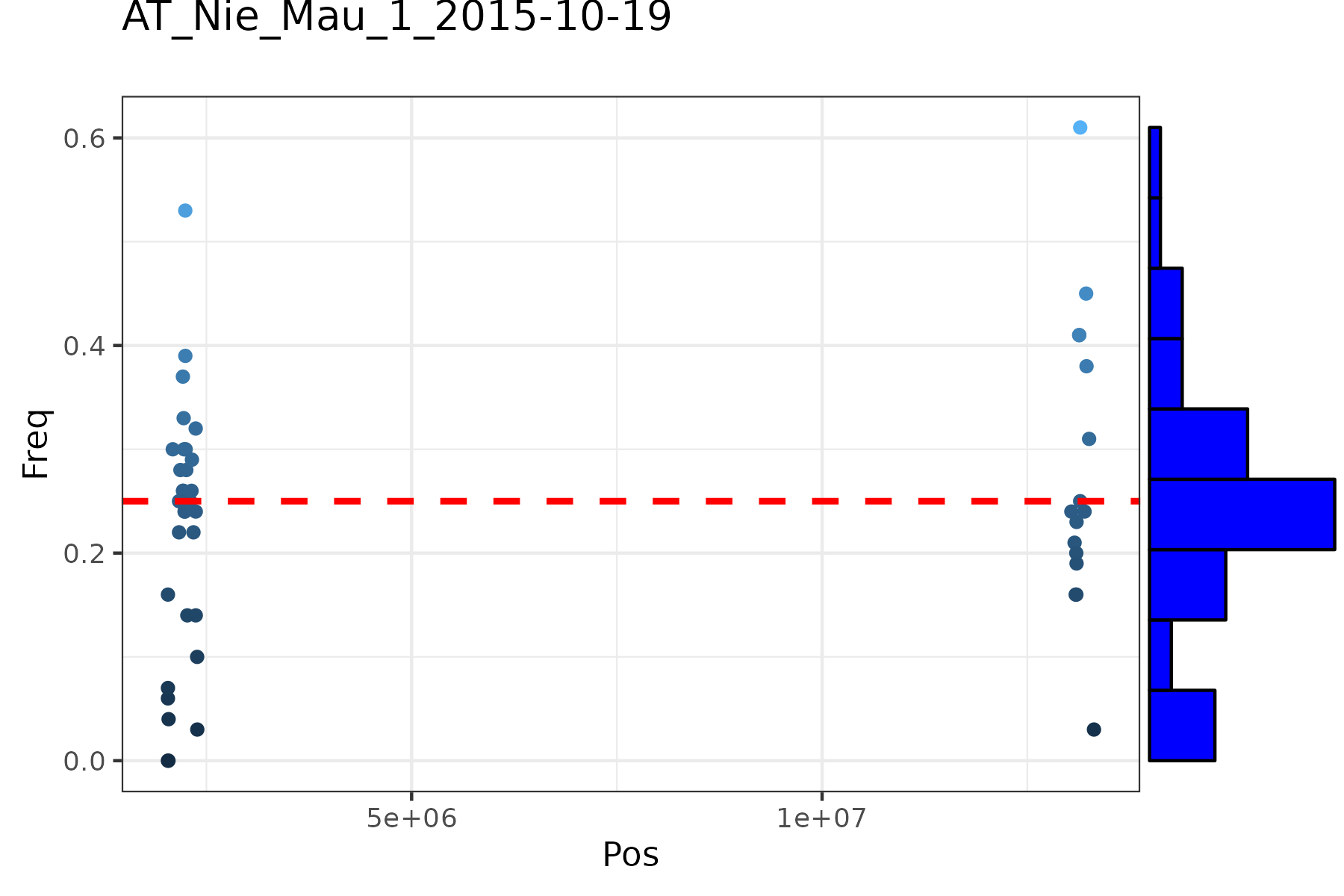


Figure 5

#### (4.2) The influence of inversions on population structure revisited

In paragraph 3.2 we found that the genomic regions spanned by inversions differ in patterns of population structure from the remaining genome. We speculated that inversion frequencies may play an important role. Now, that we have estiated inversion frequencies in all populations based on the dianostic marker SNPs, we can test this hypothesis. We will test for correlations between the scores of PC1 based on SNPs located either inside and outside the genomic region spanned by each inversion and inversion frequencies.

We will execute the *R*-script Plot\_PCAInvFreq.r, which will plot scatterplots based on the PC-scores of the first PC-axis (Dim.1) and the inversion frequency for *In(2L)t* and *In(3R)Payne*, fit a linear regression line to each of the plots and add the adjusted *R*2 value (the determinant of correlation) in the top-right corner of each plot, which describes the propotion of the variance explained by the correlation.

## does the Inv Frequency influence the PCA results?  
for index in ${!DATA[@]}; do  
  
 INVERSION=${DATA[index]}  
 Rscript ${WD}/scripts/Plot\_PCAInvFreq.r \  
 ${INVERSION} \  
 ${WD}  
done

Consistent with our hypothesis, we can see in Figure 6 below, that all plots on the left side, which show scatterplots of inversion frequencies and PC1 are characterized by very strong and highly significant correlations, which explain between

