DETECT CHROMOSOMAL DUPLICATIONS

Claudia Zirión-Martínez

Setup

```
library(tidyverse)
library(ggbeeswarm)
library(ggtree)
library(ggtreeExtra)
library(ape)
library(phytools)
library(ggnewscale)
library(RColorBrewer)
setwd("/FastData/czirion/Crypto_Diversity_Pipeline/analyses/tree_duplications/scripts")
```

Metadata

Use the metadata table that has all the samples included in the final Crypto_Desjardins_Ashton dataset (n = 1055) and H99 .

```
metadata <- read.delim(
    "../../data/processed/metadata_ashton_desj_all_fungalpop_H99.csv",
    header=TRUE,
    sep=",",
    stringsAsFactor = TRUE)
```

Select needed columns, remove H99 and get the number of samples per dataset and lineage, per lineage, and total.

```
metadata <- metadata %>%
    select(sample, strain, source, lineage, dataset, vni_subdivision)%>%
    filter(!strain == "H99") %>%
    group_by(dataset, lineage)%>%
    mutate(samples_in_dataset_lineage = n_distinct(sample))%>%
    ungroup() %>%
    group_by(lineage)%>%
    mutate(samples_in_lineage = n_distinct(sample))%>%
    ungroup()%>%
    mutate(total_samples = n_distinct(sample))
```

Chromosome names and length

Get the nice chromosome names

```
chromosome_names = read.delim(
    "../../../Crypto_Desjardins_Ashton/results_joined/02.Dataset/chromosomes.csv",
    header=TRUE, sep=",")
chromosome_names <- chromosome_names %>%
    mutate(chromosome = str_pad(chromosome, 2, pad = "0"))%>%
```

```
    mutate(chromosome = as.factor(chromosome))
levels(chromosome_names$chromosome) <- paste("chr", chromosome_names$chromosome, sep="")
```

Get the chromosome lengths

```
# /usr/bin/bash
tail -n +2 ../../../Crypto_Desjardins/config/chromosomes.csv | \
    cut -d',' -f1 | sort | uniq | while read line \
do
seqkit fx2tab ../../../Crypto_Desjardins/data/references/$line.fasta \
    -l -i -n >> ../../data/processed/chromosome_lengths.tsv
done
```

```
chromosome_lengths = read.delim(
    "../../data/processed/chromosome_lengths.tsv",
    header=FALSE,
    col.names=c("accession", "length"),
    sep="\t")
```

Get metrics of chromosomes from called CNVs

```
cnv_calls <- read.delim(
    "../../../Crypto_Desjardins_Ashton/results_joined/02.Dataset/cnv/cnv_calls.tsv",
    header=TRUE, sep="\t")
```

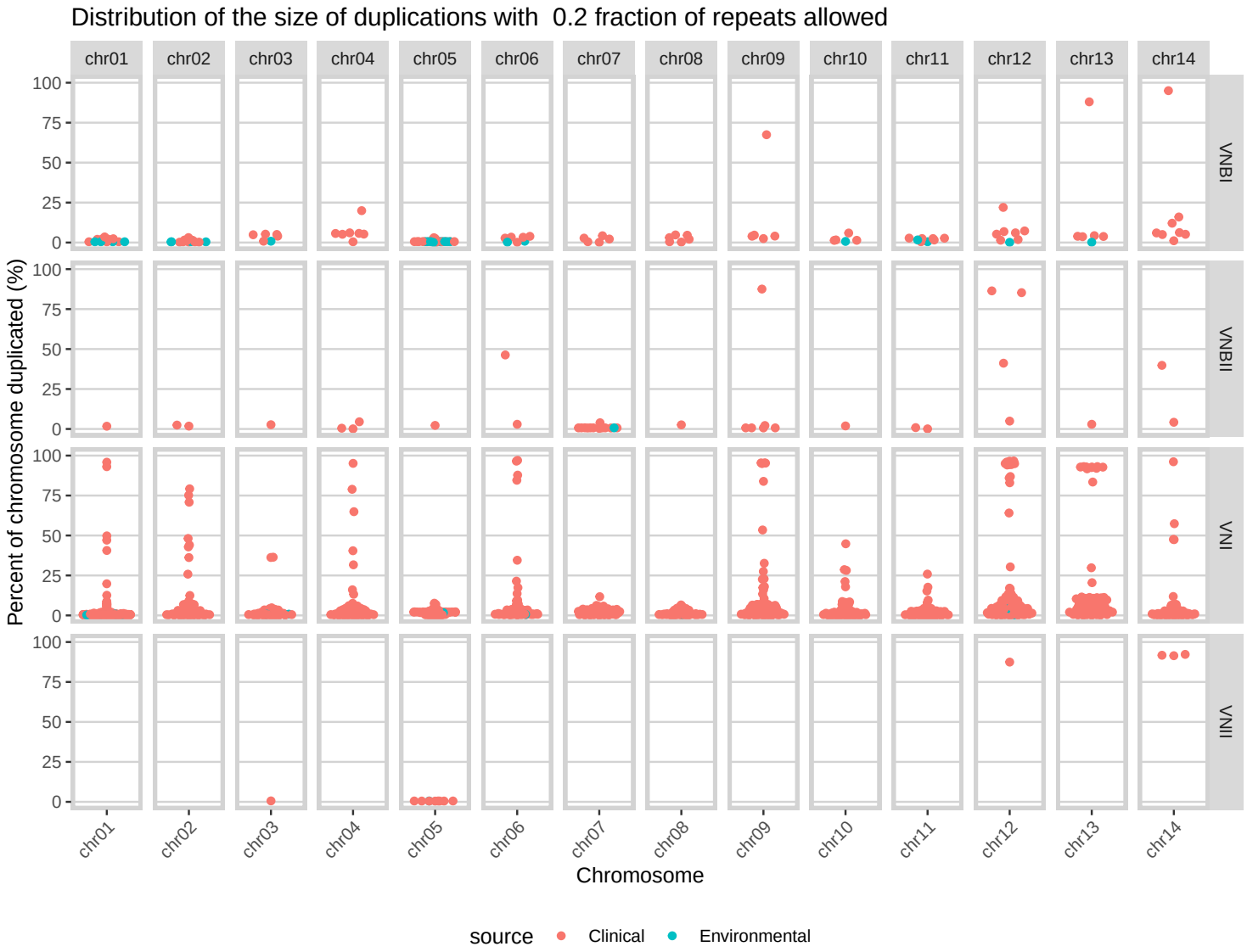
Define a threshold of allowed fraction of repetitive sequences in called CNV regions.

```
repeat_fraction_threshold <- 0.2
```

Get the percentage of the size of each chromosome that is part of called duplications `percent_cnv_size` and the percentage of the size of the chromosome that is in between the leftmost and rightmost duplicated regions `percent_size_covered`.

```
cnv_percent <- cnv_calls %>%
    filter(repeat_fraction < repeat_fraction_threshold) %>%
    filter(cnv == "duplication")%>%
    group_by(accession, sample, cnv) %>%
    summarise(total_cnv_size = sum(region_size),
                n_cnvs = n(),
                first = min(start),
                last = max(end),
                mean_cnv_depth = round(mean(smooth_depth),2)) %>%
    left_join(chromosome_lengths, by="accession") %>%
    left_join(chromosome_names, by="accession") %>%
    left_join(metadata, by=c("sample", "lineage")) %>%
    mutate(percent_cnv_size = round((total_cnv_size / length) * 100, 2),
            size_covered = last - first,
            percent_size_covered = round((size_covered / length) * 100, 2))%>%
    mutate(chromosome = as.factor(chromosome)) %>%
    select(dataset,lineage, samples_in_lineage, samples_in_dataset_lineage,
    total_samples, sample,strain, source, accession, chromosome,
```

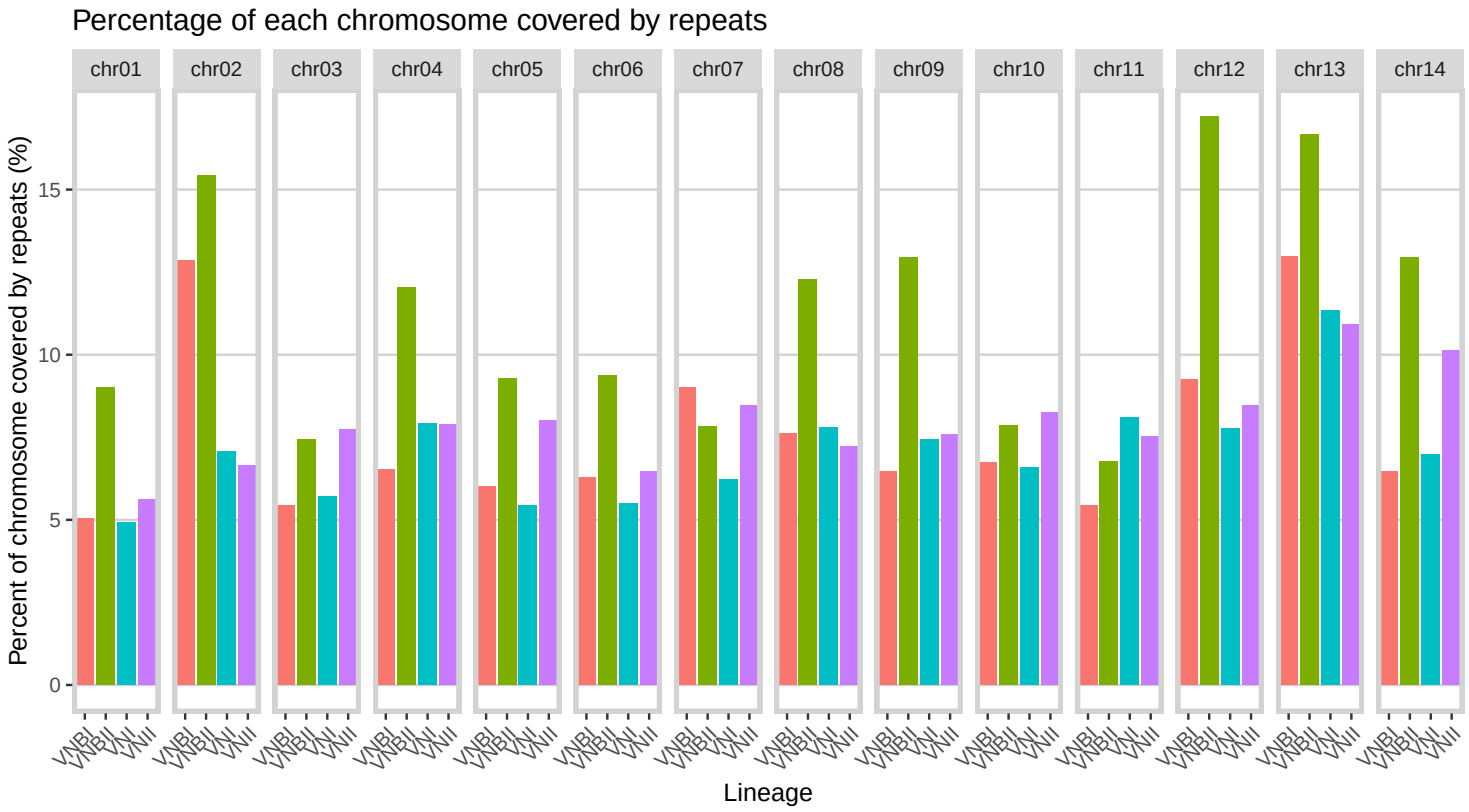
```
    percent_cnv_size, percent_size_covered, mean_cnv_depth, n_cnvs)
```



Get the percentage of each chromosome covered by repeats to know how much of a chromosome might not have reliable CNV calls.

```
lineages <- unique(metadata$lineage)
repeats_all <- data.frame()
for(lineage in lineages){
    repeats_path <-paste(
        "../../../Crypto_Desjardins/results/03.References/",
        lineage, "/", lineage, "_repeats.bed",
        sep ="")
    repeats <- read.delim(repeats_path,
        header=FALSE,
        col.names=c("accession", "start", "end", "repeat_type"),
        sep="\t")
    repeats$lineage <- lineage
    repeats_all <- rbind(repeats_all, repeats)
}
```

```
repeats_percent <- repeats_all %>%
    mutate(repeat_size_each = end - start)%>%
    group_by(accession, lineage) %>%
    summarise(repeat_size = sum(repeat_size_each)) %>%
    left_join(chromosome_lengths, by="accession") %>%
    left_join(chromosome_names, by=c("accession","lineage")) %>%
    mutate(percent_repeat_size = round((repeat_size / length) * 100, 2))%>%
    mutate(chromosome = as.factor(chromosome))%>%
    select(lineage, accession, chromosome, percent_repeat_size)
```



Most chromosomes have repeats in between 5 and 10% of the size. In VNBII it is closer to 15% for some chormosomes.

Get the chromosome median depth as another measure to dectect duplicated chromosomes

```
depth_by_chrom_good_desjardins <- read.delim(
    "../../../Crypto_Desjardins/results/04.Intermediate_files/02.Dataset/depth_quality/depth_by_chrom_
     header=TRUE, sep="\t")
depth_by_chrom_good_ashton <- read.delim(
    "../../../Crypto_Ashton/results/04.Intermediate_files/02.Dataset/depth_quality/depth_by_chrom_good
     header=TRUE, sep="\t")
depth_by_chrom_good <- rbind(depth_by_chrom_good_desjardins, depth_by_chrom_good_ashton)
depth_by_chrom <- depth_by_chrom_good %>%
    select(sample, accession, norm_chrom_median)
```

Join CNV metrics with normalized chromosome depth

```
cnv_and_depth <- left_join(depth_by_chrom, cnv_percent, by = c("sample", "accession" ))
```

Make a threshold for the size of the chromosome that must be in called duplicated regions and a threshold for the normalized chromosome median depth.

```
percent_size_threshold <- 80
depth_threshold <- 1.55
```

Keep, as putative duplications, the chromosomes whose metrics are above any of the thresholds.

```
duplications <- cnv_and_depth %>%
    filter(norm_chrom_median > depth_threshold | percent_cnv_size > percent_size_threshold)%>%
    select(dataset, lineage, samples_in_lineage, samples_in_dataset_lineage,
        total_samples, source, strain, sample, chromosome, norm_chrom_median,
        mean_cnv_depth, percent_cnv_size, percent_size_covered, n_cnvs)
```

```
head(select(duplications, strain, chromosome,
    norm_chrom_median, mean_cnv_depth, percent_cnv_size,
    percent_size_covered, n_cnvs))
```

strain	chromosome	norm_chrom_median	mean_cnv_depth	percent_cnv_size	percent_size_covered	n_cnvs
Bt134	chr12	2.02	1.85	95.86	97.67	2
8-1	chr14	2.03	1.90	92.26	96.56	2
Bt109	chr09	1.97	1.87	87.54	94.24	4
MW-RSA3834	chr12	1.97	1.88	95.15	97.34	2
LP-RSA3042	chr14	1.92	1.82	96.14	98.53	3
WM626	chr14	2.15	2.04	91.72	95.95	2

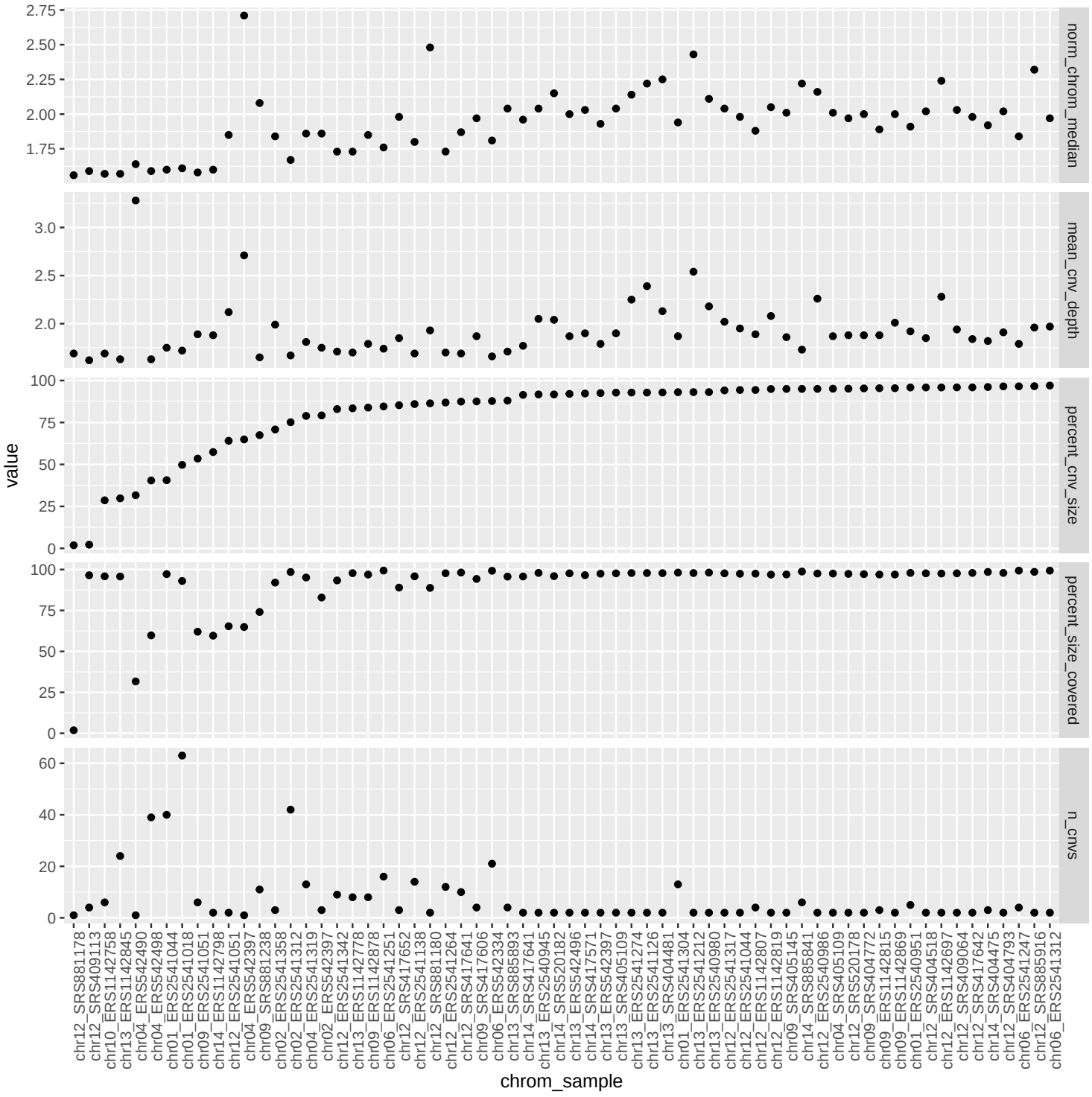
Write the duplications table to a table to use it in the scripts `duplications_gather_plots.xsh` and `duplications_polish.qmd` to polish the results.

```
dir.create("../results/tables/putative", recursive = TRUE)
write_tsv(duplications, "../results/tables/duplications_putative.tsv")
```

Plot all metrics of putative duplications.

```
duplications_long <- duplications  %>%
    arrange(percent_cnv_size) %>%
    mutate(chrom_sample = paste(chromosome, sample, sep="_"))
duplications_long$chrom_sample <- factor(duplications_long$chrom_sample,
    levels = duplications_long$chrom_sample)
duplications_long <- pivot_longer(duplications_long,
    cols = c("norm_chrom_median", "mean_cnv_depth",
        "percent_cnv_size", "percent_size_covered",
        "n_cnvs"),
    names_to = "variable",
```

```
    values_to = "value")
duplications_long$variable <- factor(duplications_long$variable,
    levels = c("norm_chrom_median", "mean_cnv_depth",
        "percent_cnv_size", "percent_size_covered",
        "n_cnvs"))
```



Get multiple summary tables

Number of duplicated chromosomes per sample.

```
dup_sample <- duplications %>%
    group_by(dataset,lineage, sample, strain, source) %>%
    summarise(n_chroms = n_distinct(chromosome),
            chromosomes = paste(chromosome, collapse = ", ")) %>%
    arrange(desc(n_chroms))
dup_sample
```

dataset	lineage	sample	strain	source	n_chroms	chromosomes
Ashton	VNI	ERS542397	14936_1#6	Clinical	3	chr02, chr04, chr13
Ashton	VNI	ERS2541044	04CN-64-074	Clinical	2	chr01, chr12
Ashton	VNI	ERS2541051	04CN-65-032	Clinical	2	chr09, chr12
Ashton	VNI	ERS2541312	04CN-32-011	Clinical	2	chr02, chr06
Desjardins	VNI	SRS405109	Bt92	Clinical	2	chr04, chr13
Desjardins	VNII	SRS417641	C12	Clinical	2	chr12, chr14
Ashton	VNI	ERS1142697	20427_2#6	Clinical	1	chr12
Ashton	VNI	ERS1142758	20427_2#61	Clinical	1	chr10
Ashton	VNI	ERS1142778	20949_2#15	Clinical	1	chr13
Ashton	VNI	ERS1142798	20949_2#21	Clinical	1	chr14
Ashton	VNI	ERS1142807	20427_3#21	Clinical	1	chr12
Ashton	VNI	ERS1142815	20427_3#26	Clinical	1	chr09
Ashton	VNI	ERS1142819	20427_3#30	Clinical	1	chr12
Ashton	VNI	ERS1142845	20427_3#51	Clinical	1	chr13
Ashton	VNI	ERS1142869	20949_2#42	Clinical	1	chr09
Ashton	VNI	ERS1142878	20427_4#13	Clinical	1	chr09
Ashton	VNI	ERS2540945	04CN-65-072	Clinical	1	chr13
Ashton	VNI	ERS2540951	04CN-64-024	Clinical	1	chr01
Ashton	VNI	ERS2540980	04CN-64-011	Clinical	1	chr13
Ashton	VNI	ERS2540986	04CN-65-001	Clinical	1	chr12
Ashton	VNI	ERS2541018	04CN-64-128	Clinical	1	chr01
Ashton	VNI	ERS2541126	BMD3144	Clinical	1	chr13
Ashton	VNI	ERS2541138	04CN-03-053	Clinical	1	chr12
Ashton	VNI	ERS2541212	04CN-03-039	Clinical	1	chr13
Ashton	VNI	ERS2541247	BMD2209	Clinical	1	chr06
Ashton	VNI	ERS2541251	04CN-03-081	Clinical	1	chr06
Ashton	VNI	ERS2541264	BMD3117	Clinical	1	chr12
Ashton	VNI	ERS2541274	BMD761	Clinical	1	chr13
Ashton	VNI	ERS2541304	04CN-65-056	Clinical	1	chr01
Ashton	VNI	ERS2541317	04CN-64-090	Clinical	1	chr12
Ashton	VNI	ERS2541319	04CN-32-042	Clinical	1	chr04
Ashton	VNI	ERS2541342	UI_31647-2	Clinical	1	chr12
Ashton	VNI	ERS2541358	CNS_936	Clinical	1	chr02
Ashton	VNI	ERS542334	14892_1#38	Clinical	1	chr06
Ashton	VNI	ERS542490	14893_1#4	Clinical	1	chr04
Ashton	VNI	ERS542496	14893_1#10	Clinical	1	chr13
Ashton	VNI	ERS542498	14893_1#12	Clinical	1	chr04
Desjardins	VNBI	SRS881178	NRHc5005.ENR	Clinical	1	chr12
Desjardins	VNBI	SRS881238	PMHc1026.ENR	Clinical	1	chr09
Desjardins	VNBI	SRS885841	NRHc5009.REL.INI	Clinical	1	chr14
Desjardins	VNBI	SRS885893	NRHc5045.ENR.CLIN.ISO	Clinical	1	chr13
Desjardins	VNBII	SRS417606	Bt109	Clinical	1	chr09
Desjardins	VNBII	SRS417652	MW-RSA2967	Clinical	1	chr12
Desjardins	VNBII	SRS881180	PMHc1029.ENR.STOR	Clinical	1	chr12

dataset	lineage	sample	strain	source	n_chroms	chromosomes
Desjardins	VNI	SRS404475	LP-RSA3042	Clinical	1	chr14
Desjardins	VNI	SRS404481	Bt139	Clinical	1	chr13
Desjardins	VNI	SRS404518	Bt134	Clinical	1	chr12
Desjardins	VNI	SRS404772	Bt141	Clinical	1	chr09
Desjardins	VNI	SRS404793	MW-RSA6134	Clinical	1	chr12
Desjardins	VNI	SRS405145	Bt117	Clinical	1	chr09
Desjardins	VNI	SRS409064	MW-RSA1955	Clinical	1	chr12
Desjardins	VNI	SRS409113	Br795	Clinical	1	chr12
Desjardins	VNI	SRS417642	In2632	Clinical	1	chr12
Desjardins	VNI	SRS520178	MW-RSA3834	Clinical	1	chr12
Desjardins	VNI	SRS885916	PMHc1031A.ENR.INI.LP	Clinical	1	chr12
Desjardins	VNII	SRS417571	8-1	Clinical	1	chr14
Desjardins	VNII	SRS520182	WM626	Clinical	1	chr14

Number of samples with duplications in each group of dataset-lineage-chromosome.

```
dup_dataset_lineage_chromosome <- duplications %>%
    group_by(dataset,lineage, chromosome) %>%
    summarise(n_samples = n_distinct(sample),
            samples_in_dataset_lineage = first(samples_in_dataset_lineage))%>%
    mutate(percent_samples = round((n_samples / samples_in_dataset_lineage) * 100, 1))%>%
    select(dataset,lineage, chromosome,
            n_samples, samples_in_dataset_lineage, percent_samples)%>%
    arrange(chromosome, desc(lineage), desc(n_samples))
dup_dataset_lineage_chromosome
```

dataset	lineage	chromosome	n_samples	samples_in_dataset_lineage	percent_samples
Ashton	VNI	chr01	4	668	0.6
Ashton	VNI	chr02	3	668	0.4
Ashton	VNI	chr04	4	668	0.6
Desjardins	VNI	chr04	1	185	0.5
Ashton	VNI	chr06	4	668	0.6
Ashton	VNI	chr09	4	668	0.6
Desjardins	VNI	chr09	2	185	1.1
Desjardins	VNBII	chr09	1	64	1.6
Desjardins	VNBI	chr09	1	122	0.8
Ashton	VNI	chr10	1	668	0.1
Desjardins	VNII	chr12	1	16	6.2
Ashton	VNI	chr12	10	668	1.5
Desjardins	VNI	chr12	7	185	3.8
Desjardins	VNBII	chr12	2	64	3.1
Desjardins	VNBI	chr12	1	122	0.8
Ashton	VNI	chr13	9	668	1.3
Desjardins	VNI	chr13	2	185	1.1
Desjardins	VNBI	chr13	1	122	0.8
Desjardins	VNII	chr14	3	16	18.8
Ashton	VNI	chr14	1	668	0.1
Desjardins	VNI	chr14	1	185	0.5
Desjardins	VNBI	chr14	1	122	0.8

Number of samples with duplications in each group of lineage-chromosome.

```
dup_lineage_chromosome <- duplications%>%
    group_by(lineage, chromosome) %>%
    summarise(n_samples = n_distinct(sample), samples_in_lineage = first(samples_in_lineage))%>%
    mutate(percent_samples = round((n_samples / samples_in_lineage) * 100, 1))%>%
    select(lineage, chromosome, n_samples, samples_in_lineage,percent_samples)%>%
    arrange(chromosome, desc(lineage), desc(n_samples))
dup_lineage_chromosome
```

lineage	chromosome	n_samples	samples_in_lineage	percent_samples
VNI	chr01	4	853	0.5
VNI	chr02	3	853	0.4
VNI	chr04	5	853	0.6
VNI	chr06	4	853	0.5
VNI	chr09	6	853	0.7
VNBII	chr09	1	64	1.6
VNBI	chr09	1	122	0.8
VNI	chr10	1	853	0.1
VNII	chr12	1	16	6.2
VNI	chr12	17	853	2.0
VNBII	chr12	2	64	3.1
VNBI	chr12	1	122	0.8
VNI	chr13	11	853	1.3
VNBI	chr13	1	122	0.8
VNII	chr14	3	16	18.8
VNI	chr14	2	853	0.2
VNBI	chr14	1	122	0.8

Number of samples with duplications in each group of lineage-datset.

```
dup_lineage_dataset <- duplications%>%
    group_by(dataset,lineage) %>%
    summarise(n_samples = n_distinct(sample),
        samples_in_dataset_lineage = first(samples_in_dataset_lineage))%>%
    mutate(percent_samples = round((n_samples / samples_in_dataset_lineage) * 100, 1))%>%
    select(lineage, n_samples, samples_in_dataset_lineage, percent_samples)%>%
    arrange(desc(lineage), desc(n_samples))
dup_lineage_dataset
```

dataset	lineage	n_samples	samples_in_dataset_lineage	percent_samples
Desjardins	VNII	3	16	18.8
Ashton	VNI	35	668	5.2
Desjardins	VNI	12	185	6.5
Desjardins	VNBII	3	64	4.7
Desjardins	VNBI	4	122	3.3

Number of samples with duplications in each chromosome.

```r
dup_chromosome <- duplications %>%
    group_by(chromosome) %>%
    summarise(n_samples = n_distinct(sample), total_samples = first(total_samples))%>%
    mutate(percent_samples = round((n_samples / total_samples) * 100, 1))%>%
    select(chromosome, n_samples,total_samples, percent_samples)%>%
    arrange(chromosome, desc(n_samples))
dup_chromosome
```

chromosome	n_samples	total_samples	percent_samples
chr01	4	1055	0.4
chr02	3	1055	0.3
chr04	5	1055	0.5
chr06	4	1055	0.4
chr09	8	1055	0.8
chr10	1	1055	0.1
chr12	21	1055	2.0
chr13	12	1055	1.1
chr14	6	1055	0.6