DETECT CHROMOSOMAL DUPLICATIONS

Claudia Zirión-Martínez

Setup

```r
setwd("/FastData/czirion/Crypto_Diversity_Pipeline/analyses/tree_duplications/scripts")
library(tidyverse)
```

Detect partial duplications

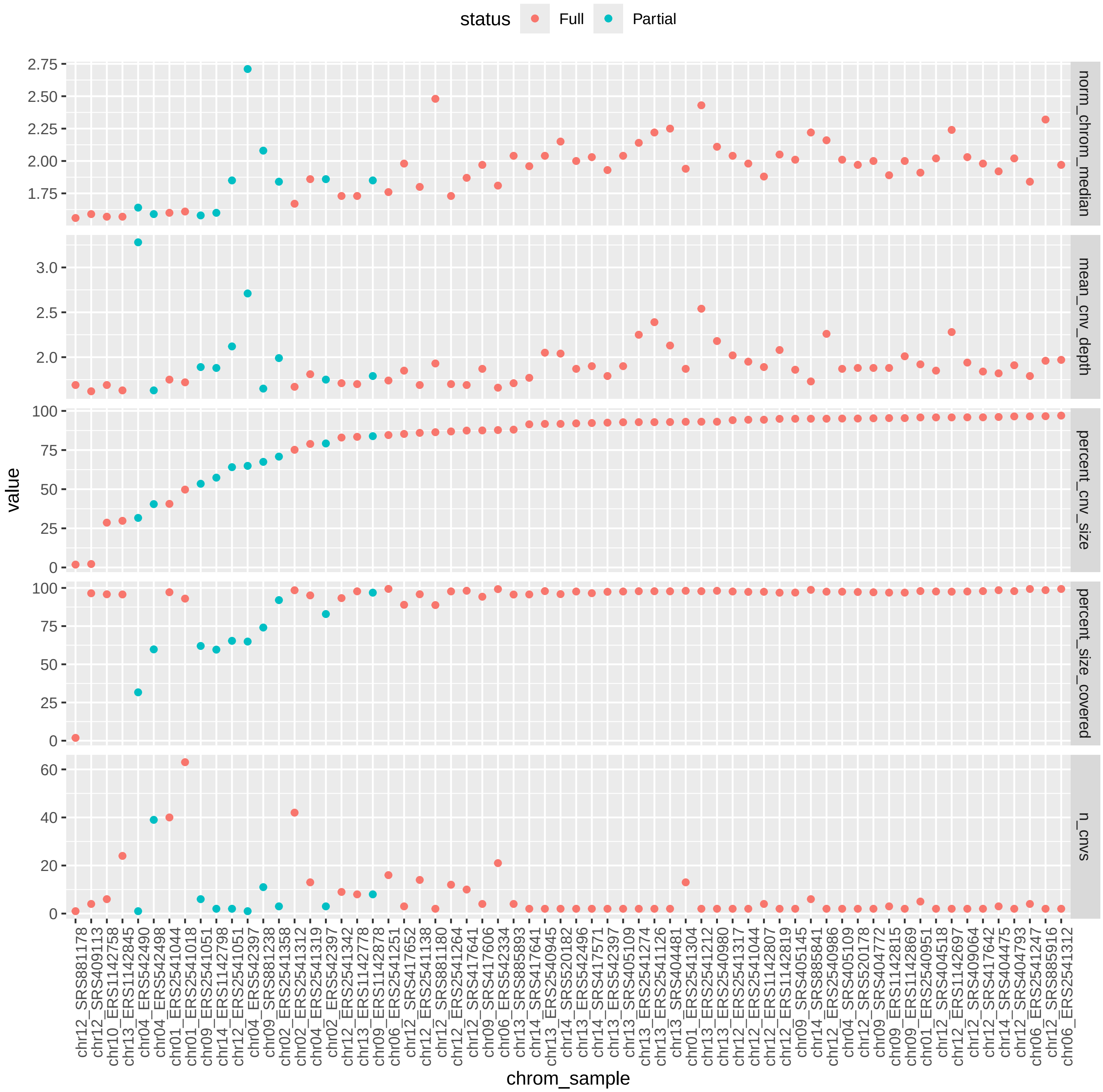
After manual inspection of plots (see scripts `duplications_detect.qmd` `duplications_gather_plots.xsh`) I found that the following samples have partial duplications instead of full chromosome duplications.

```r
partial <- data.frame(sample=
                            c("ERS2541051",
                            "ERS2541051",
                            "ERS542397",
                            "ERS1142798",
                            "ERS542490",
                            "ERS1142878",
                            "ERS2541358",
                            "SRS881238",
                            "ERS542397",
                            "ERS542498"),
                    chromosome =
                            c("chr09",
                            "chr12",
                            "chr02",
                            "chr14",
                            "chr04",
                            "chr09",
                            "chr02",
                            "chr09",
                            "chr04",
                            "chr04"))

partial$sample_chromosome <- paste(partial$sample, partial$chromosome, sep="_")
```

```r
duplications_putative <- read_tsv("../results/tables/duplications_putative.tsv")

duplications_putative <- duplications_putative %>%
    mutate(sample_chromosome = paste(sample, chromosome, sep="_"),
            status = ifelse(sample_chromosome %in% partial$sample_chromosome, "Partial", "Full"))
```

```r
duplications_long <- duplications_putative  %>%
    arrange(percent_cnv_size) %>%
    mutate(chrom_sample = paste(chromosome, sample, sep="_"))
```

```r
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"))
```



Remove partial duplications

```
duplications_full <- duplications_putative %>%
    filter(status == "Full")

duplications_full_strain <- duplications_full %>%
    select(dataset, lineage, sample, strain,
            source, chromosome,
            norm_chrom_median, percent_cnv_size)
```

```
write_tsv(duplications_full_strain,
    "../results/tables/duplications_polished.tsv")
```

Get multiple summary tables

Number of duplicated chromosomes per sample.

```
dup_sample <- duplications_full %>%
    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	ERS2541044	04CN-64-074	Clinical	2	chr01, 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	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	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	ERS542334	14892_1#38	Clinical	1	chr06
Ashton	VNI	ERS542397	14936_1#6	Clinical	1	chr13
Ashton	VNI	ERS542496	14893_1#10	Clinical	1	chr13
Desjardins	VNBI	SRS881178	NRHc5005.ENR	Clinical	1	chr12
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

dataset	lineage	sample	strain	source	n_chroms	chromosomes
Desjardins	VNBII	SRS881180	PMHc1029.ENR.STOR	Clinical	1	chr12
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_full %>%
    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	1	668	0.1
Ashton	VNI	chr04	1	668	0.1
Desjardins	VNI	chr04	1	185	0.5
Ashton	VNI	chr06	4	668	0.6
Ashton	VNI	chr09	2	668	0.3
Desjardins	VNI	chr09	2	185	1.1
Desjardins	VNBII	chr09	1	64	1.6
Ashton	VNI	chr10	1	668	0.1
Desjardins	VNII	chr12	1	16	6.2
Ashton	VNI	chr12	9	668	1.3
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
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_full%>%
    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	1	853	0.1
VNI	chr04	2	853	0.2
VNI	chr06	4	853	0.5
VNI	chr09	4	853	0.5
VNBII	chr09	1	64	1.6
VNI	chr10	1	853	0.1
VNII	chr12	1	16	6.2
VNI	chr12	16	853	1.9
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	1	853	0.1
VNBI	chr14	1	122	0.8

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

```
dup_lineage_dataset <- duplications_full%>%
    group_by(dataset,lineage) %>%
    summarise(n_samples = n_distinct(sample), samples_in_dataset_lineage = first(samples_in_dataset_li
    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	29	668	4.3
Desjardins	VNI	12	185	6.5
Desjardins	VNBII	3	64	4.7
Desjardins	VNBI	3	122	2.5

Number of samples with duplications in each chromosome.

```
dup_chromosome <- duplications_full %>%
    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	1	1055	0.1
chr04	2	1055	0.2
chr06	4	1055	0.4
chr09	5	1055	0.5
chr10	1	1055	0.1
chr12	20	1055	1.9
chr13	12	1055	1.1
chr14	5	1055	0.5