COUNT CALLED VARIANTS IN DESJARDINS ASHTON DATASET

Claudia Zirión-Martínez

R Setup

```
library(tidyverse)
library(patchwork)
setwd("/FastData/czirion/Crypto_Diversity_Pipeline/analyses/ploidy/scripts")
```

Desjardins

Number of raw and filtered variants

Using bash count the number of variants called by Freebayes and filter by Snippy.
Snippy filter are:
GT == 1/1
QUAL >= 100
DP >= 10
A0/DP >= 0

```
cd /FastData/czirion/Crypto_Diversity_Pipeline/
tail -n +2 Crypto_Desjardins/config/metadata.csv | cut -d',' -f2 | while read line
do
    raw=$(grep -v "#" Crypto_Desjardins/results/01.Samples/snippy/$line/snps.raw.vcf | wc -l)
    filt=$(grep -v "#" Crypto_Desjardins/results/01.Samples/snippy/$line/snps.filt.vcf | wc -l)
    echo $line,$raw,$filt >> analyses/ploidy/data/processed/snp_counts_desjardins.csv
done
```

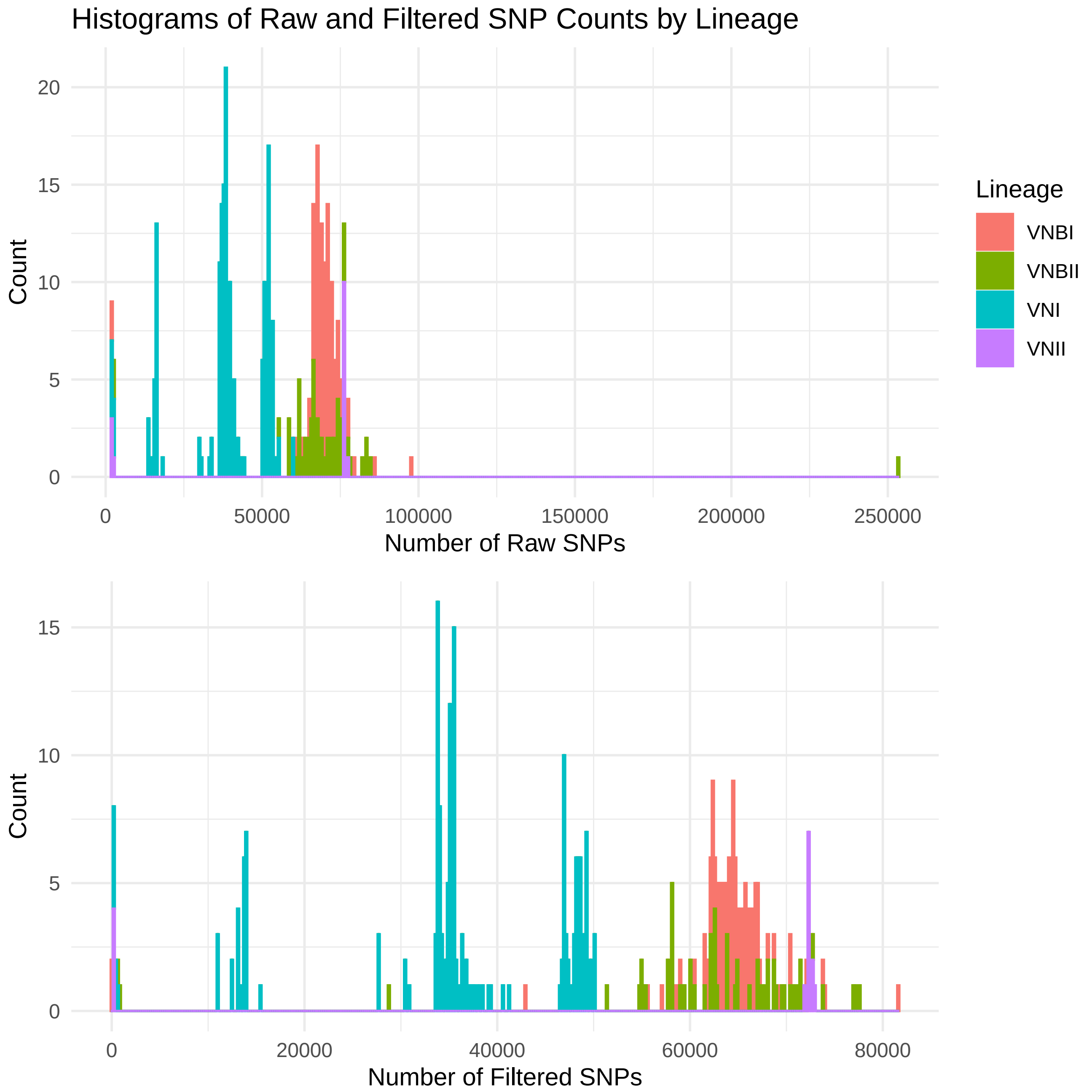
Metadata

```
metadata <- read_csv("../../../Crypto_Desjardins/config/metadata.csv")
metadata <- metadata %>%
    select(sample, strain, lineage)
```

Variant counts

```
snp_counts <- read_csv("../data/processed/snp_counts_desjardins.csv", col_names = c("sample", "n_raw",

snp_counts <- snp_counts %>%
        mutate(n_removed = n_raw - n_filt, percent_filt = (n_filt / n_raw)*100)%>%
        left_join(metadata, by = "sample")
```



Since Snippy removes all the heterozygous variants, the samples with a big number of removed variants might have high heterozygosity.

```
summary <- snp_counts %>%
    summarise(max_raw = max(n_raw),
              max_filtered = max(n_filt),
              max_removed = max(n_removed),
              median_raw = median(n_raw),
              median_filt = median(n_filt),
              median_removed = median(n_removed))
summary
```

max_raw	max_filtered	max_removed	median_raw	median_filt	median_removed
253198	81701	224449	58616	50063	3801

Here are the 30 samples with highest number of likely heterozygous variants.

```
sorted <- snp_counts %>%
    arrange(desc(n_removed))
head(sorted, 30)
```

sample	n_raw	n_filt	n_removed	percent_filt	strain	lineage
SRS409075	253198	28749	224449	11.35436	Bt206	VNBII
SRS885851	97753	42827	54926	43.81144	NRHc5009.ENR	VNBI
SRS881185	83039	68113	14926	82.02531	PMHc1049.THER1.STOR	VNBII
SRS885877	73198	58977	14221	80.57187	NRHc5023.ENR	VNBI
SRS885841	77690	64521	13169	83.04930	NRHc5009.REL.INI	VNBI
SRS404740	59707	47322	12385	79.25704	LP-RSA2296	VNI
SRS881210	74437	62342	12095	83.75136	PMHc1002.ENR	VNBII
SRS881170	60012	49301	10711	82.15190	NRHc5036.ENR	VNI
SRS885847	72614	62087	10527	85.50280	PMHc1047.ENR.CLIN1	VNBII
SRS881166	74028	63591	10437	85.90128	NRHc5032.ENR.CLIN.ISO	VNBI
SRS885853	74969	65041	9928	86.75719	Muc418-1	VNBI
SRS885871	74043	64835	9208	87.56398	Muc367-1	VNBI
SRS885173	72086	63164	8922	87.62312	NRHc5004.ENR	VNBI
SRS885177	67651	58784	8867	86.89302	NRHc5019.ENR	VNBI
SRS881211	70753	61973	8780	87.59063	PMHc1050.ENR.CLIN1	VNBI
SRS881243	68723	60290	8433	87.72900	Muc479-1	VNBI
SRS881239	41914	33509	8405	79.94703	NRHc5048.ENR.ISO	VNI
SRS881237	55044	46643	8401	84.73766	NRHc5025.ENR.CLIN.1	VNI
SRS885186	71332	63099	8233	88.45820	NRHc5020.ENR	VNBI
SRS885863	73581	65376	8205	88.84902	Ftc222-2	VNBI
SRS885893	71277	63128	8149	88.56714	NRHc5045.ENR.CLIN.ISO	VNBI
SRS881140	74717	66659	8058	89.21531	NRHc5030.ENR.CLIN.ISO	VNBI
SRS881151	71899	64146	7753	89.21682	NRHc5029.ENR.CLIN.ISO	VNBI
SRS885192	74609	67155	7454	90.00925	Ftc225-3	VNBI
SRS881201	73326	66018	7308	90.03355	NRHc5027.ENR.CLIN1	VNBI
SRS881236	72758	65455	7303	89.96262	NRHc5041.ENR.STOR	VNBI
SRS881180	84855	77611	7244	91.46308	PMHc1029.ENR.STOR	VNBII
SRS885169	72141	65044	7097	90.16232	Gbc573-1	VNBI
SRS885888	70430	63430	7000	90.06105	NRHc5040.ENR.CLIN.ISO	VNBI
SRS881240	71662	64664	6998	90.23471	NRHc5021.ENR	VNBI

Ashton

Number of raw and filtered variants

Using bash count the number of variants called by Freebayes and filter by Snippy.
Snippy filter are:
GT == 1/1
QUAL >= 100
DP >= 10
A0/DP >= 0

```
cd /FastData/czirion/Crypto_Diversity_Pipeline/
tail -n +2 Crypto_Ashton/config/metadata.csv | cut -d',' -f2 | while read line
```

```
do
    raw=$(grep -v "#" Crypto_Ashton/results/01.Samples/snippy/$line/snps.raw.vcf | wc -l)
    filt=$(grep -v "#" Crypto_Ashton/results/01.Samples/snippy/$line/snps.filt.vcf | wc -l)
    echo $line,$raw,$filt >> analyses/ploidy/data/processed/snp_counts_ashton.csv
done
```

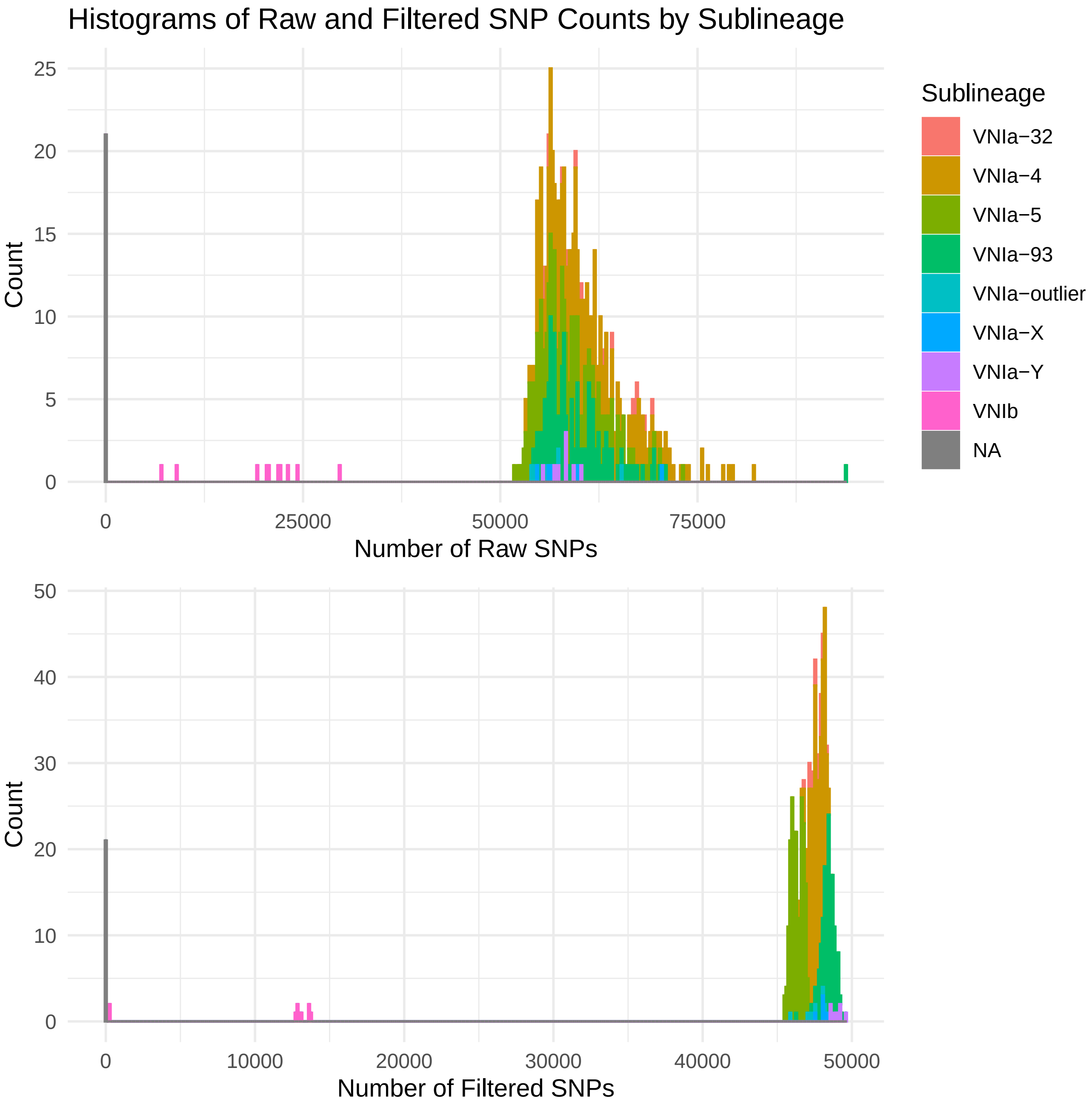
Metadata

```
metadata <- read_csv("../../../Crypto_Ashton/config/metadata.csv")
metadata <- metadata %>%
    select(sample, strain, lineage, VNI_subdivision)
```

Variant counts

```
snp_counts <- read_csv("../data/processed/snp_counts_ashton.csv", col_names = c("sample", "n_raw", "n_
```

```
snp_counts <- snp_counts %>%
        mutate(n_removed = n_raw - n_filt, percent_filt = (n_filt / n_raw)*100)%>%
        left_join(metadata, by = "sample")
```



Since Snippy removes all the heterozygous variants, the samples with a big number of removed variants might have high heterozygosity.

```
summary <- snp_counts %>%
    summarise(max_raw = max(n_raw),
              max_filtered = max(n_filt),
              max_removed = max(n_removed),
              median_raw = median(n_raw),
              median_filt = median(n_filt),
              median_removed = median(n_removed))
summary
```

max_raw	max_filtered	max_removed	median_raw	median_filt	median_removed
93810	49602	46506	58518.5	47508.5	11353.5

Here are the 30 samples with highest number of likely heterozygous variants.

```r
sorted <- snp_counts %>%
    arrange(desc(n_removed))
head(sorted, 30)
```

sample	n_raw	n_filt	n_removed	percent_filt	strain	lineage	VNI_subdivision
ERS2541310	93810	47304	46506	50.42533	04CN-65-080	VNI	VNIa-93
ERS1142849	82051	48000	34051	58.50020	20427_3#55	VNI	VNIa-4
ERS1142823	79490	48031	31459	60.42395	20427_3#32	VNI	VNIa-4
ERS1142830	78989	48106	30883	60.90215	20427_3#37	VNI	VNIa-4
ERS1142841	78184	48071	30113	61.48445	20427_3#47	VNI	VNIa-4
ERS1142842	76362	48139	28223	63.04052	20427_3#48	VNI	VNIa-4
ERS1142818	75626	48112	27514	63.61833	20427_3#29	VNI	VNIa-4
ERS1142815	75531	48143	27388	63.73939	20427_3#26	VNI	VNIa-4
ERS1142795	73258	46719	26539	63.77324	20427_3#15	VNI	VNIa-5
ERS1142843	73818	48121	25697	65.18871	20427_3#49	VNI	VNIa-4
ERS1142762	73415	47975	25440	65.34768	20427_2#65	VNI	VNIa-4
ERS1142698	72884	47828	25056	65.62208	20427_2#7	VNI	VNIa-4
ERS1142790	71992	48122	23870	66.84354	20427_3#11	VNI	VNIa-4
ERS1142712	70344	46506	23838	66.11225	20427_2#18	VNI	VNIa-5
ERS1142826	71471	48190	23281	67.42595	20427_3#34	VNI	VNIa-4
ERS1142794	71342	48210	23132	67.57590	20427_3#14	VNI	VNIa-4
ERS1142733	71067	47950	23117	67.47154	20427_2#38	VNI	VNIa-4
ERS1142797	71199	48159	23040	67.63999	20427_3#16	VNI	VNIa-4
ERS1142724	70970	47986	22984	67.61449	20427_2#29	VNI	VNIa-4
ERS1142793	69665	46754	22911	67.11261	20427_3#13	VNI	VNIa-5
ERS1142700	70627	47928	22699	67.86073	20427_2#9	VNI	VNIa-32
ERS1142751	69091	46518	22573	67.32860	20427_2#54	VNI	VNIa-5
ERS1142825	69453	46885	22568	67.50608	20427_3#33	VNI	VNIa-5
ERS1142699	70437	47910	22527	68.01823	20427_2#8	VNI	VNIa-4
ERS1142759	70990	48734	22256	68.64911	20427_2#62	VNI	VNIa-93
ERS1142701	68745	46506	22239	67.65001	20427_2#10	VNI	VNIa-5
ERS1142833	70563	48415	22148	68.61245	20427_3#40	VNI	VNIa-X
ERS1142816	68914	46781	22133	67.88316	20427_3#27	VNI	VNIa-5
ERS1142704	70777	48787	21990	68.93058	20427_2#13	VNI	VNIa-93
ERS1142848	70133	48149	21984	68.65384	20427_3#54	VNI	VNIa-4