coriandR: Panel of Normals

Report

16-07-2023

Generating PON: "pon"

This report documents the creation of the Panel of Normals (PON) from sequencing runs of healthy individuals, which is later used in Copy Number Variantion analysis of single patient data on the same sequencing platform with `coriandR`.

This PON has to be generated only once, mapping at least three healthy sequencing libraries against the human genome using `bowtie2` and counted with `featureCounts` as specified in the accompanying documentation.

Raw Library Counts

Per default raw read counts per library are counted with `featureCounts` for the entire PON at once and stored in the pon.`fc.tsv` file by the user.

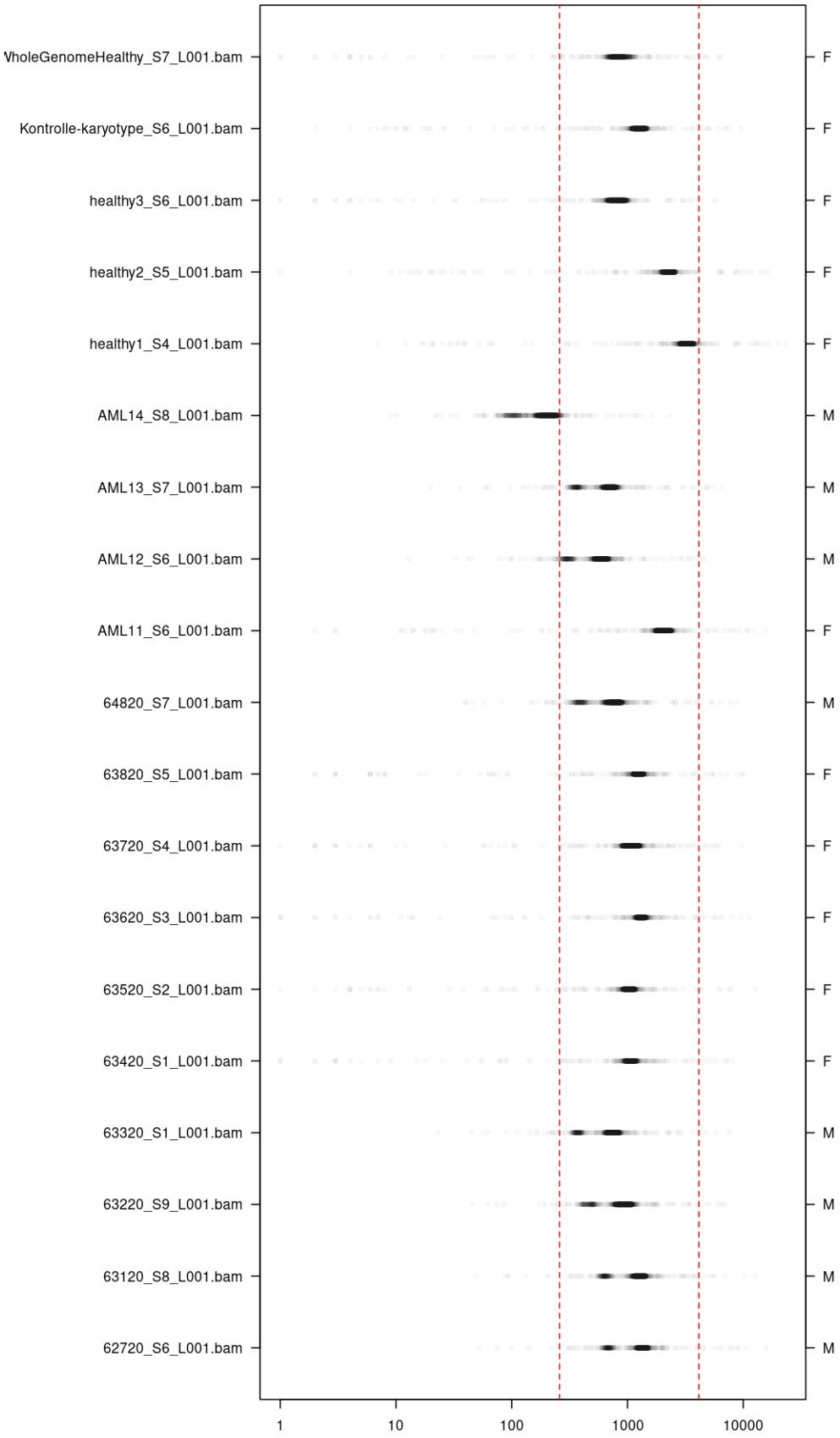
Meta data about the samples (`sample` and `gender`) are specified in the pon.`meta.tsv` file separated by comma (`,`) by the user.

Sequencing depth of individual PON subjects

The following plot shows the sequencing depth of different PON libraries on a logarithmic scale per megabase.

Optimally all samples should be well above 100 and within the red dotted lines, representing 1/4 and 4 times the **median** sequencing depth.

Gender annotation of the samples should match distributions with one single cluster in females (F) or two distinctive clusters for autosomes and single copy X and Y chromosomes in males (M).


WholeGenomeHealthy_S7_L001.bam
Kontrolle-karyotype_S6_L001.bam
healthy3_S6_L001.bam
healthy2_S5_L001.bam
healthy1_S4_L001.bam
AML14_S8_L001.bam
AML13_S7_L001.bam
AML12_S6_L001.bam
AML11_S6_L001.bam
64820_S7_L001.bam
63820_S5_L001.bam
63720_S4_L001.bam
63620_S3_L001.bam
63520_S2_L001.bam
63420_S1_L001.bam
63320_S1_L001.bam
63220_S9_L001.bam
63120_S8_L001.bam
62720_S6_L001.bam
F
F
F
F
F
M
M
M
F
M
F
F
F
F
F
M
M
M
M
1
10
100
1000
10000

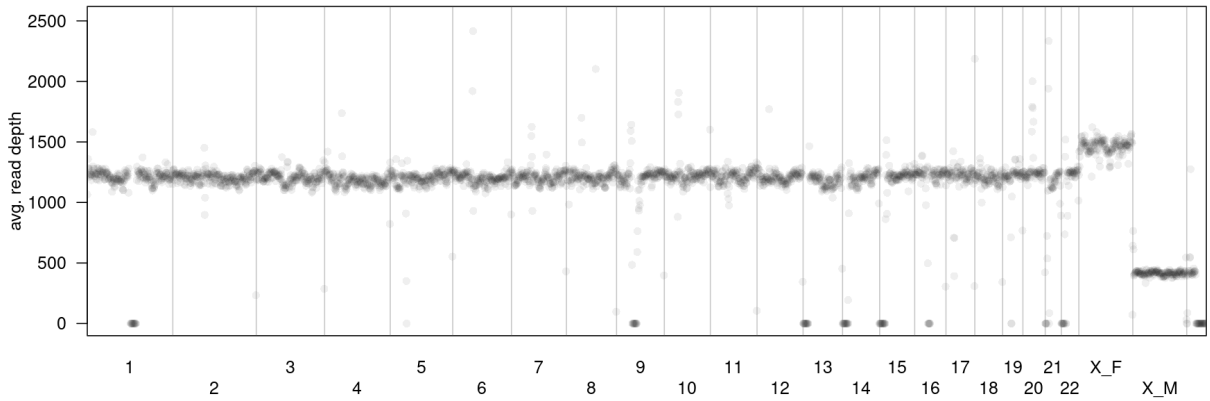
Cumulative PON

For later statistics the read counts of the PON are summed up over each bin.

Ideally, autosomes show up as a straight line of dots with minimal noise well above a read depth of at least 1000 reads per megabase.

Bins with counts outside of 99% of the normal distribution are marked in red. These should probably be masked in the end.

The values of gonosomes for men and women are represented: (from left to right) X male, X female, Y male, Y female.

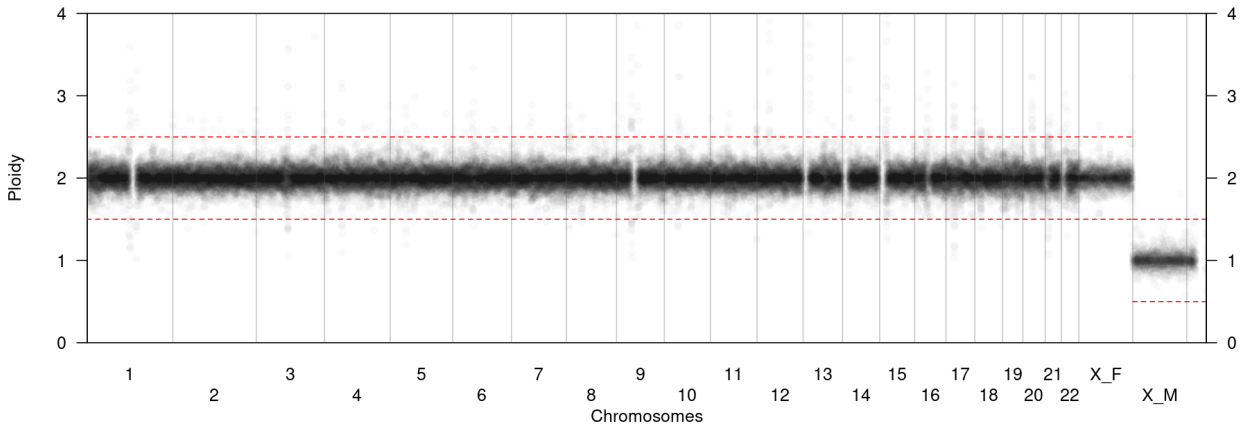


Normalizing PON by library size and per megabase

In order to compare the quality of multiple PONs, read depth is normalized by overall library size and then by the median for each megabase.

The genome plot should result in a tight distribution arount the ploidy of 2.

Gonosomes are shifted to match a ploidy of 2 in males to facilitate comparability.

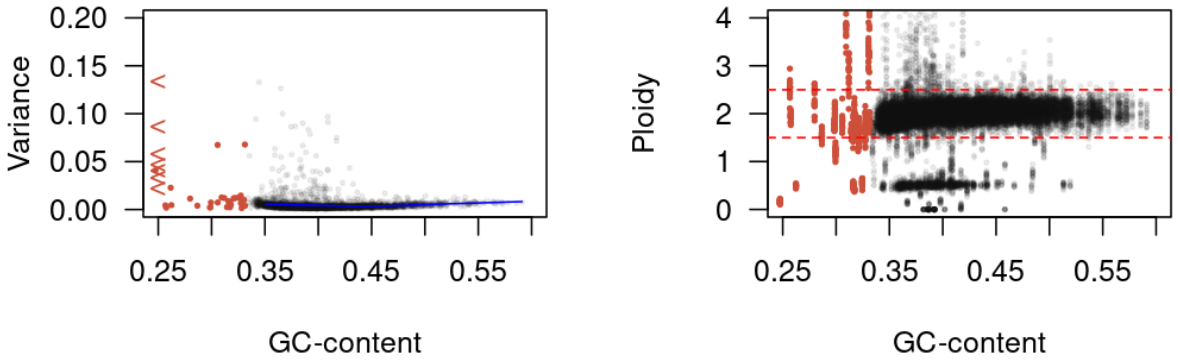


Masking bins with extreme GC-content and high variance

Some of the variance in bin counts might be due to uneven GC-content per bin. GC-bias based on bedtools nucl statistics for same genome and same bins.

Plotting the GC-content vs normalized bin counts: We see hardly any influence of GC content on the data. If GC content has little influence on variance, a large cigar-shaped point cloud appears at 2. If the GC content has a high influence on the reads, a diagonal or a banana-shaped curve can be seen.

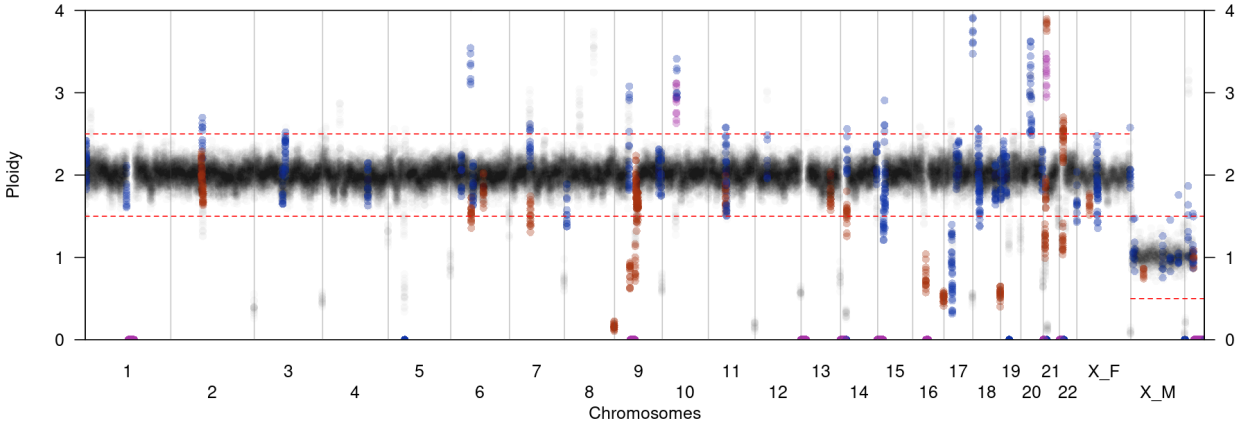
Quinlan A. R. (2014). BEDTools: The Swiss-Army Tool for Genome Feature Analysis. Current protocols in bioinformatics, 47, 11.12.1–11.12.34. doi: 10.1002/0471250953.bi1112s47.

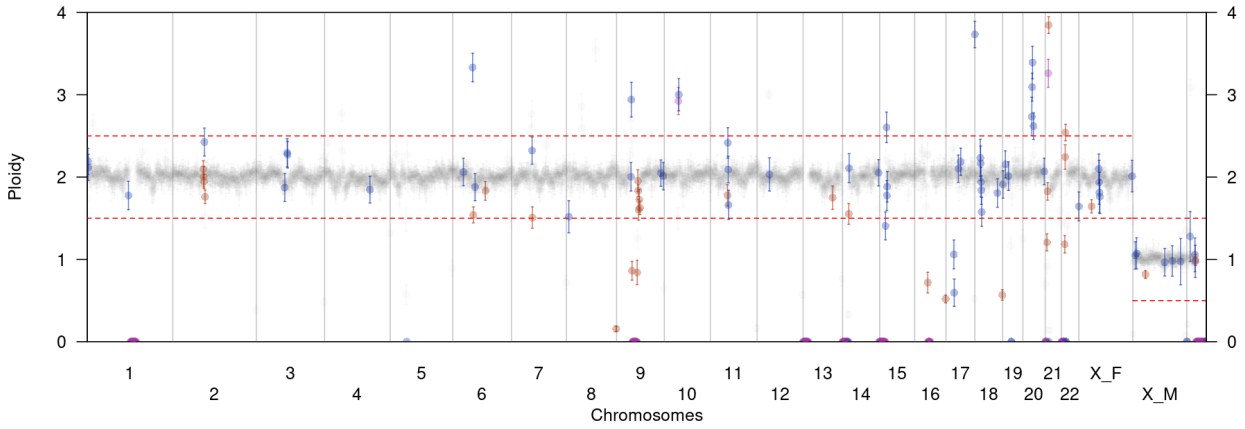


Bins with extreme GC-content are marked in red.

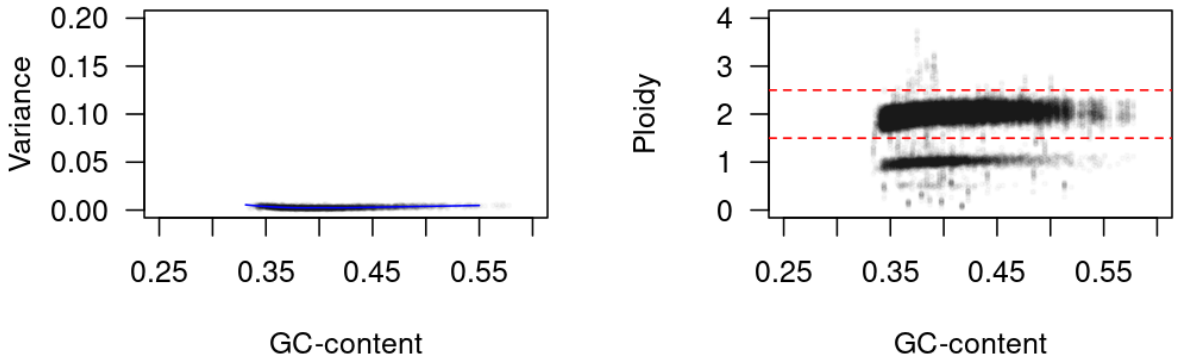
Masking high variance bins

The bins that were excluded from the statistical calculations are shown in color: due to very low GC-content (red), due to high variance (blue), or both (purple).




Ploidy
0
1
2
3
4
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
X_F
X_M
Chromosomes

Variance and normalized abundance after filtering



Filtered PON chromosomal overview

The final filtered PON should exhibit a tight distribution around 2 for all chromosomes and should not exceed the red boundaries, in order to reliably call monoallelic deletions or insertions.