Number of genes in CSD regions

Cyril Matthey-Doret

March 21, 2019

Regions definition

Here, CSD regions are defined as the regions surrounding significant SNPs until a non-significant SNP is encountered. Low SNP density will likely cause an overestimation of region sizes, but this is the most objective definition with available data. There are 11 significant SNPs out of 706 in total, and a total of 8120 annotations found on the 75.4Mb of anchored genome.

This generates multiple identical (overlapping) regions for consecutive significant SNPs (1/SNP) and directly adjacent regions in case there a single non-significant SNP between two significant ones (Figure 2). Overlapping and directly adjacent intervals are merged.

Annotations

The number of annotated genes (i.e. not included isoforms or other features) contained in each (non-overlapping, non-adjacent) region is shown in table 1. The density of annotations is highly variable between regions (Figure 3). In total, there are 381 annotated genes among all 6 regions.

GO terms

Performing a functional enrichment test reveals nothing interesting; the predicted functions of genes in CSD regions are too diverse to compute an enrichment for one particular term (Table 2).

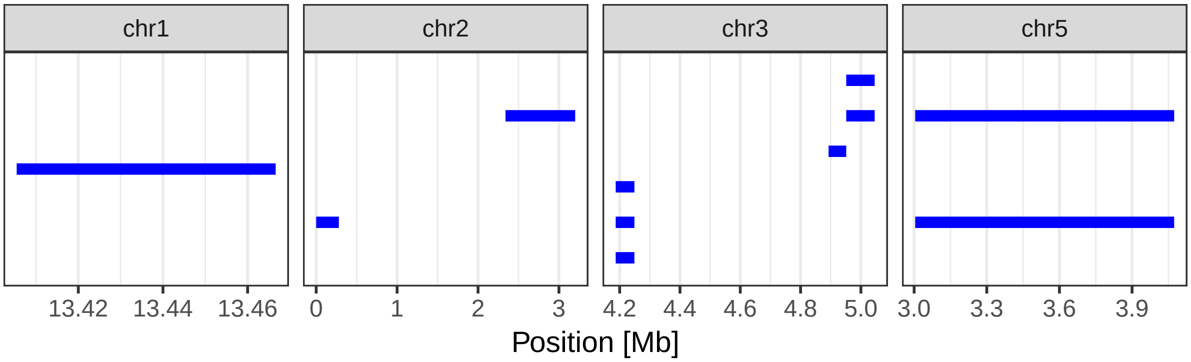

Figure 1: Visual representation of genomic ranges for each region. Each blue interval represent the span of a region. Regions are dodged vertically to show overlap and are separated in different panels by chromosomes.

Table 1: Summary of the different CSD regions after merging overlapping or directly adjacent regions. Note the first region on chromosome 2 starts at 0 since the first SNP on the chromosome is significant

chr	start	end	length_bp	n_genes
chr1	13405438	13466635	6.12e+04	0
chr2	0	279861	2.80e+05	45
chr2	2339662	3200955	8.61e+05	115
chr3	4186852	4248871	6.20e+04	5
chr3	4892861	5045763	1.53e+05	3
chr5	3004402	4073958	1.07e+06	213

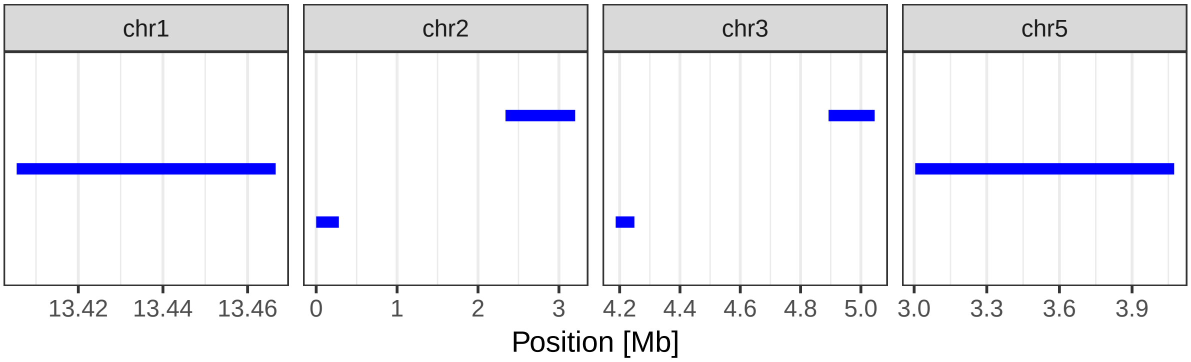

Figure 2: Visual representation of genomic ranges for each region after merging overlapping and adjacent regions. Each blue interval represent the span of a region. Regions are dodged vertically to show overlap and are separated in different panels by chromosomes.

Table 2: Top 10 most significantly enriched GO terms in CSD regions

GO	n_csd	n_genome	q-value	term
NA	199	4715	0.0000014	NA
GO:0000028	1	1	0.4690482	ribosomal small subunit assembly
GO:0000266	1	2	0.4690482	mitochondrial fission
GO:0000288	1	1	0.4690482	nuclear-transcribed mRNA catabolic process, deadenylation-dependent decay
GO:0000289	1	2	0.4690482	nuclear-transcribed mRNA poly(A) tail shortening
GO:0000462	1	1	0.4690482	maturation of SSU-rRNA from tricistronic rRNA transcript (SSU-rRNA, 5.8S rRNA, LSU-r
GO:0000703	1	1	0.4690482	oxidized pyrimidine nucleobase lesion DNA N-glycosylase activity
GO:0000932	1	2	0.4690482	P-body
GO:0001054	1	1	0.4690482	RNA polymerase I activity
GO:0001055	1	1	0.4690482	RNA polymerase II activity

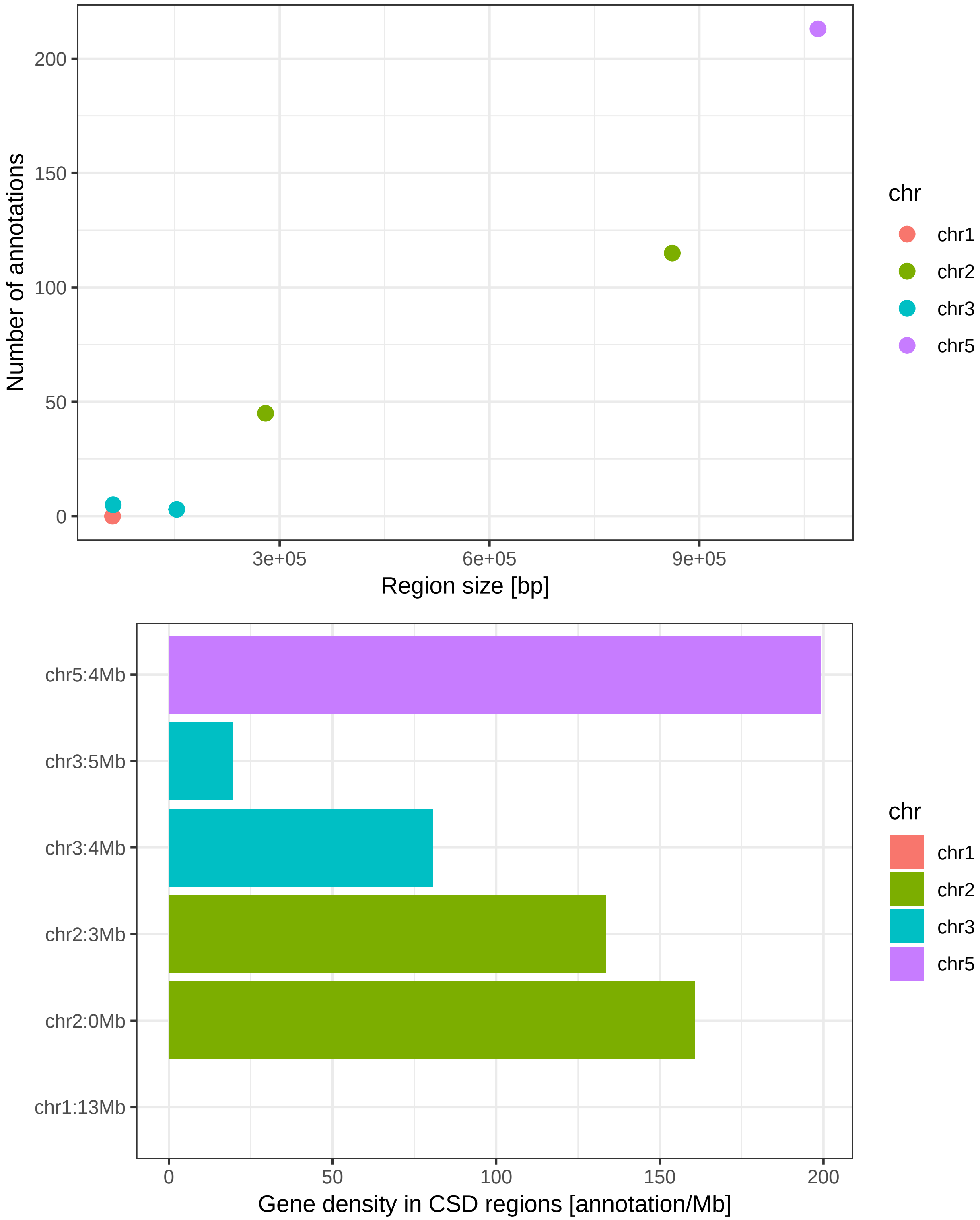

Figure 3: Top: Number of annotated features in each region compared to their size, in basepairs. Bottom: Density of annotations in each CSD region per megaase in each CSD regions. Colors represent chromosomes in both panels.