Building the genetic map of *Tribolium castaneum* from individual-level RNA-seq data

Charles Rocabert

Contents

1 Introduction

Individual RNA-seq data is derived from a population of *Tribolium castaneum*, at the 1st and 21st generations of an experimental evolution protocol. The experiment comprises two environmental conditions (Control and Hot-Dry) with 4 replicates each (figure 1). For each of the four replicates, approximately 60 individuals have been sampled at generation 1, and approximately 16 at generation 21, leading to 614 samples (figure 2).

SNP detection was performed with GATK software on all RNA-seq samples.

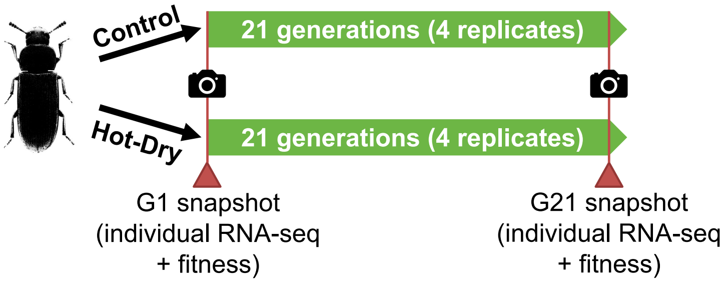


Figure 1 – A population of *Tribolium castaneum* is placed under selection in two different environments (with 4 independent replicates each): Control, and Hot-Dry. At generations 1 and 21, individual RNA-seq samples and fitness measurements are performed.

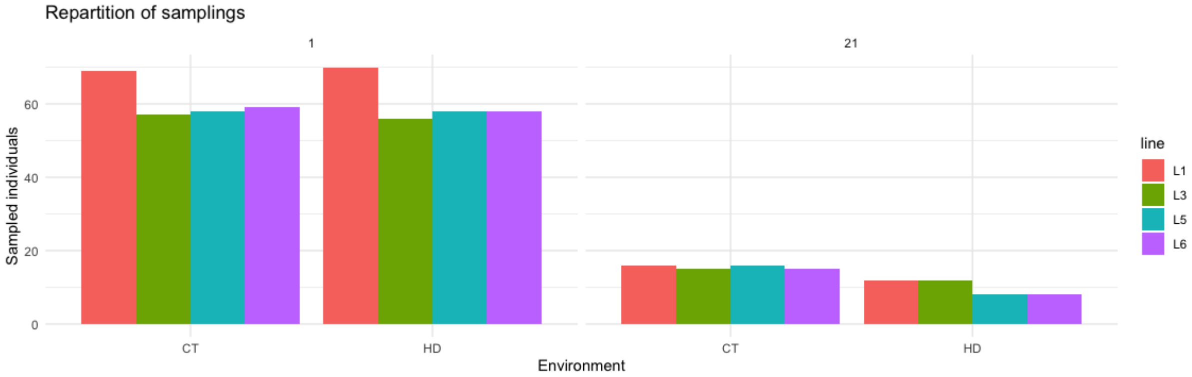


Figure 2 – Repartition of samples by environment, line and generation.

2 Pipeline

2.1 Pre-processing

Before calling Lep-MAP3 functions, a filtered VCF has been produced from the raw SNP call (DP >= 3, F_MISSING <= 0.5, MAF > 0.0), and a pedigree file containing father/mother relationships is produced (fathers and mothers are not part of the sampling).

2.2 ParentCall2

The function `ParentCall2` is called with the following parameters:

— `halfSibs = 1`;

— `removeNonInformative = 1`;

2.3 Filtering2

The function `Filtering2` is called with the following parameters:
— `removeNonInformative = 1`;
— `dataTolerance = 0.01`;

2.4 LOD limit exploration to find the best chromosome separation

The function `SeparateChromosomes2` is called with the following parameters:
— `lodLimit` $\in \{5, 20\}$;
— `distortionLod = 1`;
— `numThreads = 10`;

To find the LOD limit leading to the best chromosome separation, `LODlimit` values have been explored in the range $\{5, 20\}$. *Tribolium castaneum* has 10 chromosomes, thus the most accurate chromosome separation is obtained with `LODlimit = 12` (figure 2).

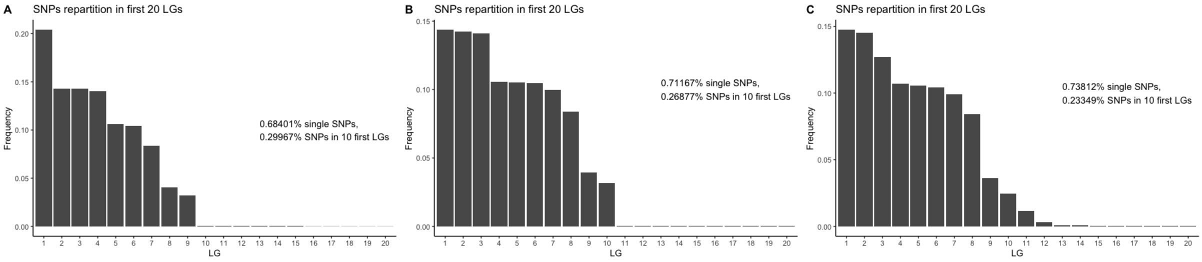


Figure 3 – Repartition of SNPs in the 20 first linkage groups. **(A)** `LODlimit = 11`. **(B)** `LODlimit = 12`. **(C)** `LODlimit = 13`

For `LODlimit = 12`, the repartition of SNPs in the 10 first linkage groups accurately matches the known repartition of SNPs in chromosomes (figure 3).

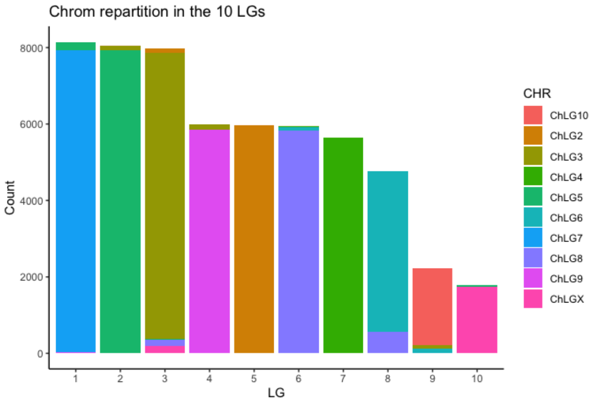


Figure 4 – Repartition of SNPs in the 10 first linkage groups. SNPs are colored by their known repartition in the 10 chromosomes of *Tribolium castaneum*.

2.5 Manual reset of additional linkage groups

All SNPs belonging to linkage groups > 10 are re-labelled to zero with an *ad hoc* script (a SNP labelles to zero does not belong to any linkage group).

2.6 JoinSingles2All

The function `JoinSingles2All` is called with the following parameters:
— `lodLimit = 8`;
— `distortionLod = 1`;
— `iterate = 1`;
— `numThreads = 10`;

This step adds 29,235 single SNPs to the first 10 linkage groups (51% increase), leading to a total of 85,727 SNPs (figure 4).

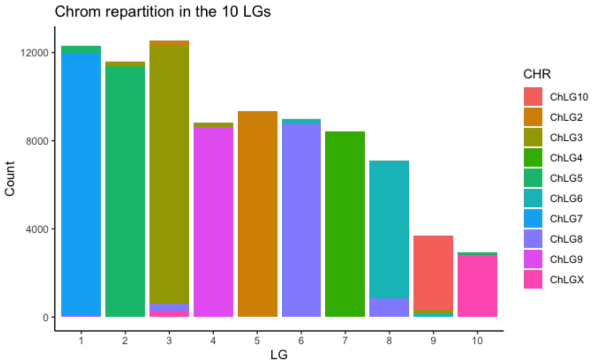
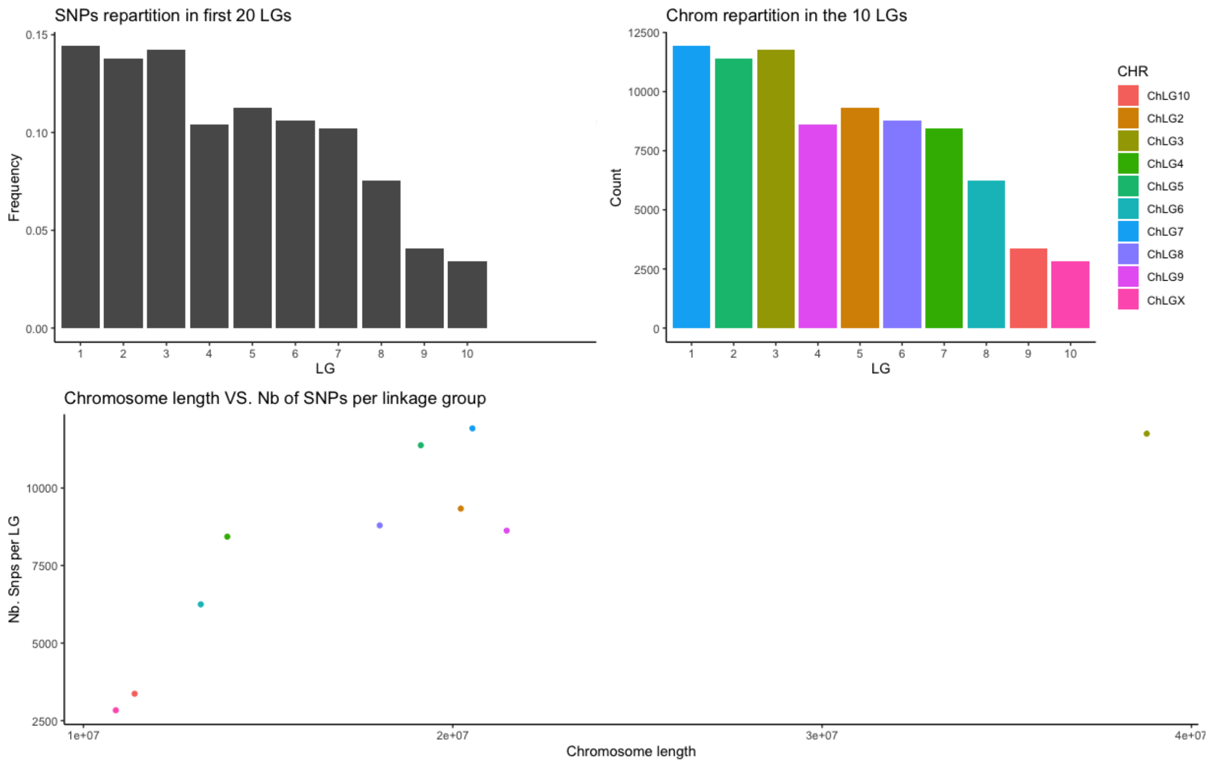


Figure 5 – Repartition of SNPs in the 10 first linkage groups after running Lep-MAP3 function `JoinSingles2All`. SNPs are colored by the known repartition in the 10 chromosomes of *Tribolium castaneum*.

2.7 Manual cleaning of the map

All single SNPs, as well as SNPs attributed to a linkage group but not belonging to the major chromosome of this group are removed with an *ad hoc* script (for example in figure 4, the major chromosome of linkage group 1 is ChLG7).

After this step, the map is ready for markers sorting. The number of SNPs per linkage group correlates to known chromosome lengths (figure 5, adjusted $R^2 = 0.46$, p-value $=$ 0.02).



2.8 OrderMarkers2

The function `OrderMarkers2` is called with the following parameters:

— `outputPhasedData = 1`;

— `sexAveraged = 1`;

— `numThreads = 10`;

2.9 Manual cut of map ends for each chromosome

Genetic distance inflation is observed at both ends of each chromosome map (figure 6).

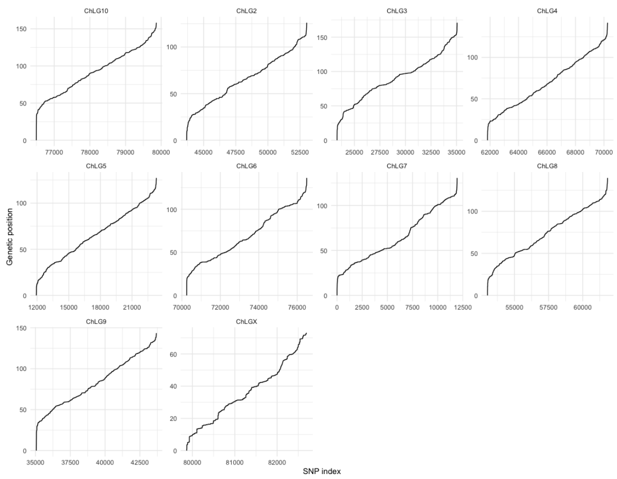


Table 1 – List of cutoff values at map ends for each chromosome.

Chromosome	Genetic distance cutoffs (cM)	Physical distance cutoffs (bp)
ChLGX	$\varnothing$	$\varnothing$
ChLG2	$29 < d < 87$	$3 \times 10^6 < d < 2 \times 10^7$
ChLG3	$50 < d < 145$	$\varnothing$
ChLG4	$42 < d < 110$	$d < 1.2 \times 10^7$
ChLG5	$36 < d < 101$	$\varnothing$
ChLG6	$37 < d < 100$	$\varnothing$
ChLG7	$37 < d < 100$	$\varnothing$
ChLG8	$48 < d < 100$	$\varnothing$
ChLG9	$55 < d < 120$	$5 \times 10^6 < d$
ChLG10	$60 < d < 125$	$2.5 \times 10^6 < d$

Figure 6 – Raw genetic map with genetic distance inflation at chromosome ends.

Thresholds are applied on each chromosome based on the visual interpretation of Marey maps (figure 7). Cutoff thresholds are applied on physical and genetic distances, following the table 1. The resulting Marey map is shown on figure 8, and indicates that physical positions from the reference genome correspond well to predicted genetic distances (figure 8).

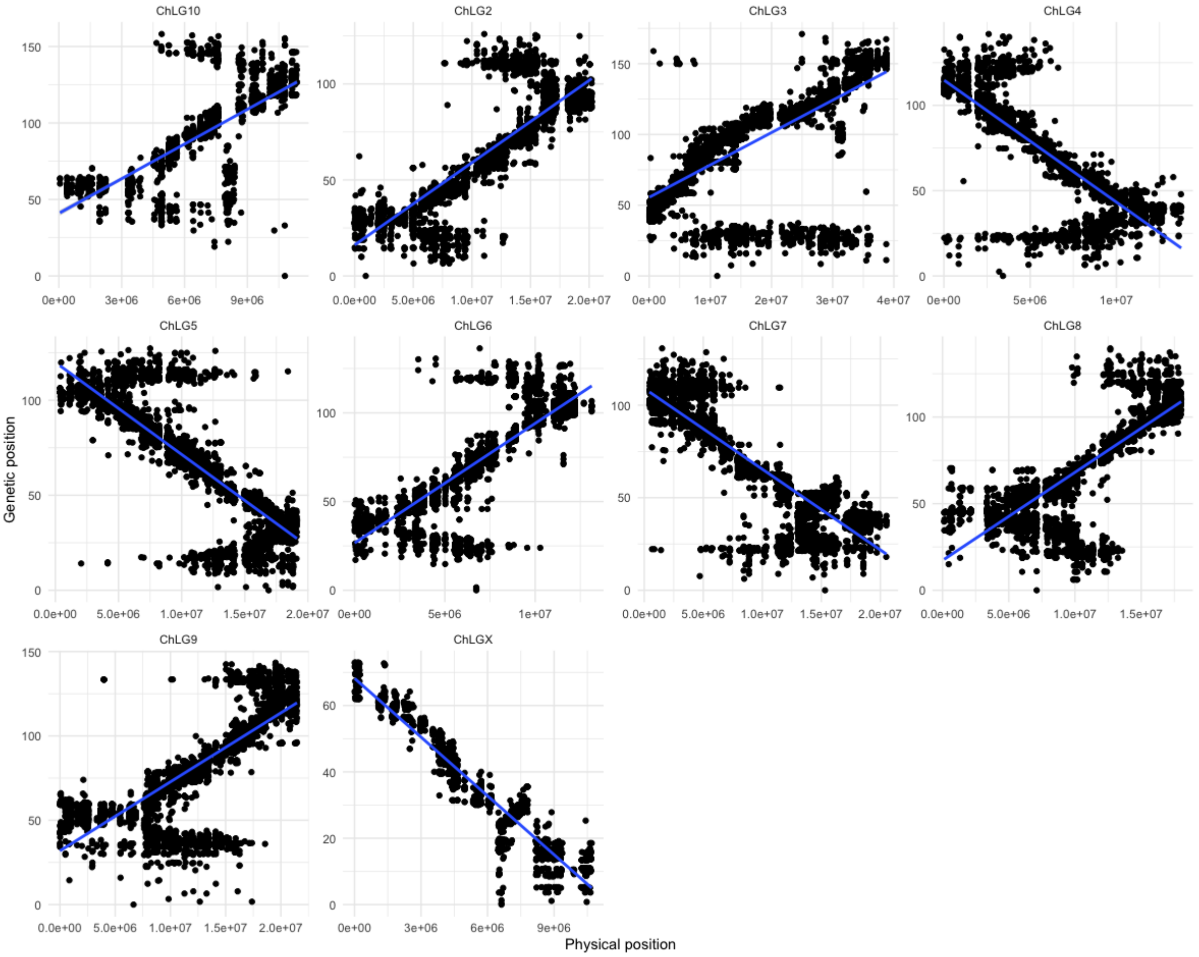


Figure 7 – Marey maps before cutting chromosome ends.

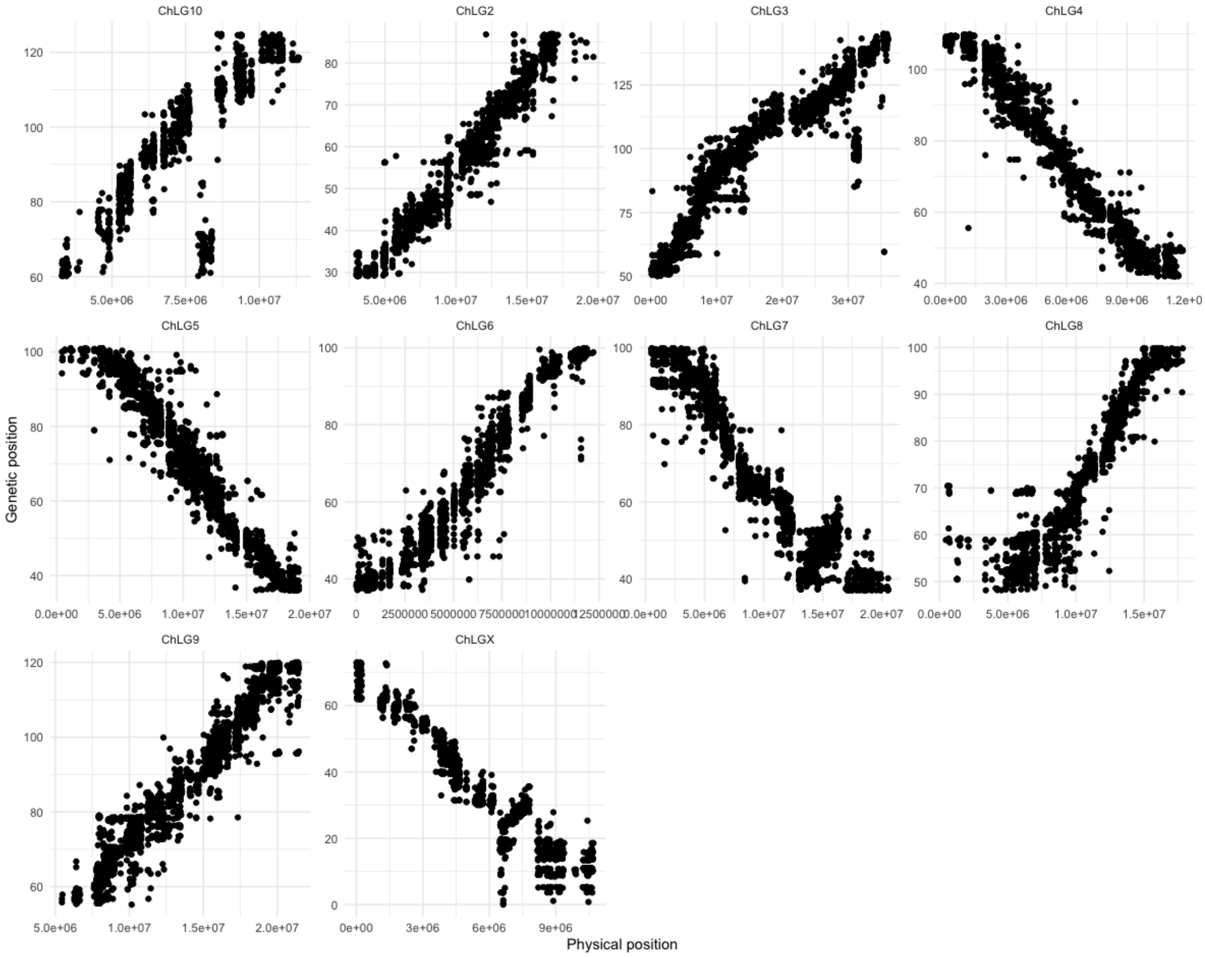


Figure 8 – Marey maps after cutting chromosome ends.