Map-Reduce programming model for QTL applications

Behrang Mahjani, Salman Toor, Sverker Holmgren
Dept. IT, Div. Scientific Computing, Uppsala University, Uppsala, Sweden

Introduction

The motivation for this work is the computational challenges involved in identifying **multiple QTL in an experimental cross**. In a linear regression model for detection of quantitative trait loci, the LOD score can be computed as $\mathrm{LOD} = n/2 \; log_{10}(RSS_0/RSS)$ [1]. In this definition, LOD is linearly related to the $log_{10}(RSS)$, the logarithm of the residual sum of squared errors. Finding one or multiple interacting QTLs in the genome is then a matter of finding a point maximizing the LOD, or minimizing the RSS.

In optimization problems, knowledge of the optimization landscape can be used to improve the search procedure. For this problem, we present another log transformation of the RSS as the objective function in the optimization. This **alternative objective function** is simple to handle and makes the ideal peak shape around a QTL linear with a known slope, when plotted on a linkage map. This new objective function increases the **performance** and the **accuracy** of the **PruneDIRECT** search algorithm in finding multiple QTL.

The aim is to explore **Map-Reduce** programming model for permutation testing of QTL positions. Map-Reduce is well known model for managing several independent tasks of an application which makes it a good candidate for permutation testing.

According to the preliminary studies, we are hopeful that Map-Reduce model will be helpful in finding more than three QTL.

The R statistical software is widely used by the biologist and statisticians community. In order to make a familiar environment settings for the biologists, we have chosen the **R-Hadoop** framework that allows the transparent use of Hadoop within the R programming environment.

Research group: Hadoop - PruneDIRECT

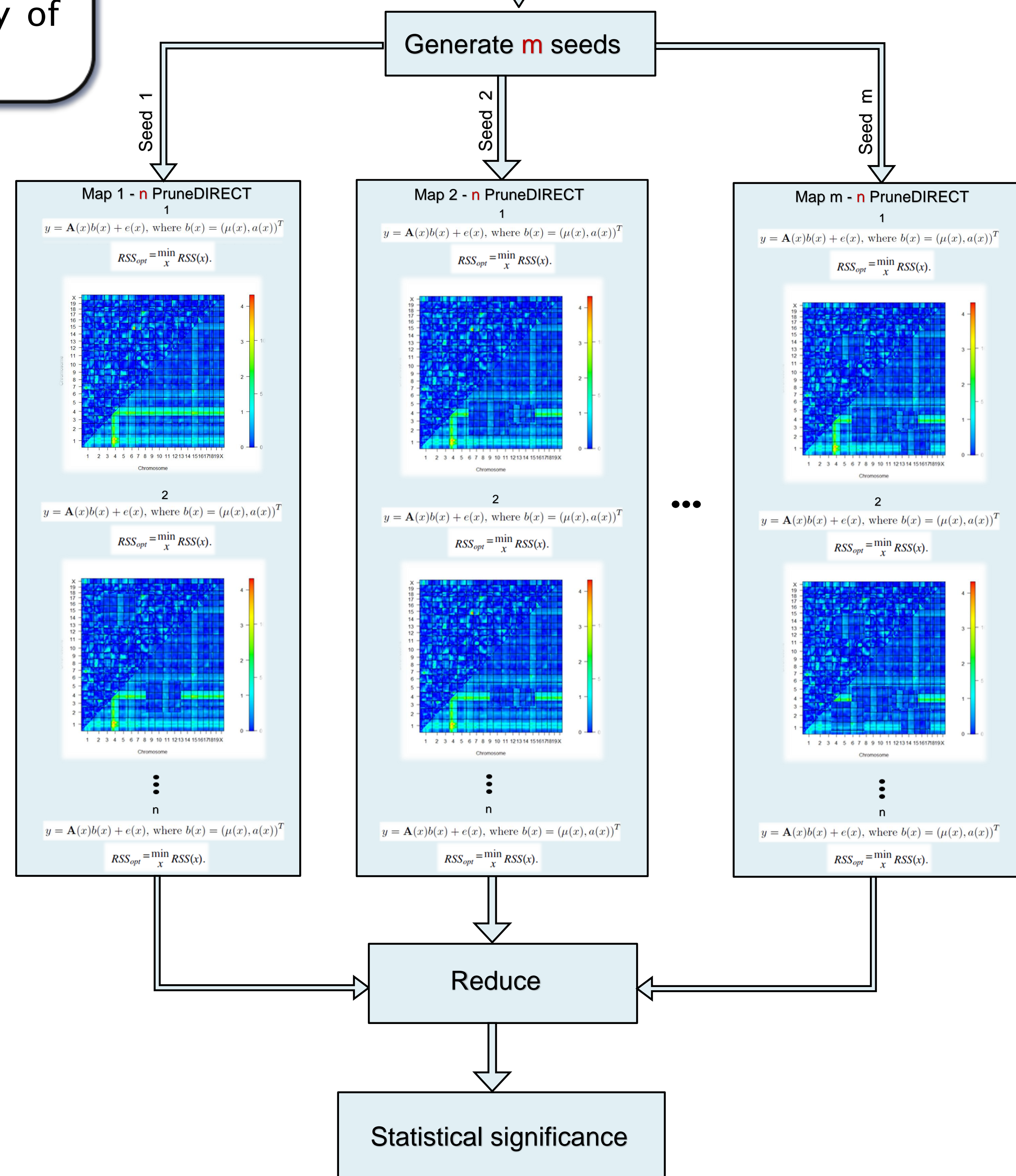Dept. IT, Div. Scientific Computing, Uppsala University, Uppsala, Sweden
Web site: http://www.it.uu.se/research/group/ccac/bioinformatics
Participants:

Sverker Holmgren   (Professor)   *E-mail*: sverker.holmgren@it.uu.se
Behrang Mahjani   (PhD student)   *E-mail*: behrang.mahjani@it.uu.se
Salman Toor   (Post-Doc)   *E-mail*: salman.toor@it.uu.se

References:
[1] Broman, K.W and Speed, T.P. A review of methods for identifying QTLs in experimental crosses. In: Seillier-Moiseiwitsch, F., ed., Statistics in Molecular Biology and Genetics. Vol. 33 of IMS Lecture Notes Monograph Series, pp. 114-42, 1999.

[2] Simultaneous search for multiple QTL using the global optimization algorithm DIRECT. Kajsa Ljungberg, Sverker Holmgren, and Örjan Carlborg. In Bioinformatics, volume 20, pp 1887-1895, 2004.

[3] Fast and accurate detection of multiple quantitative trait loci. Carl Nettelblad, Behrang Mahjani, and Sverker Holmgren. In Journal of Computational Biology, volume 20, pp 687-702, 2013.