Correlation Trait Loci (CTL) mapping: phenotype network inference subject to genotype

Danny Arends[1,2], Yang Li[2,3], Gudrun A. Brockmann[1], Ritsert C. Jansen[2], Robert W. Williams[4], and Pjotr Prins[2,4,5]

[1]Humboldt University, Berlin, Germany,
[2]University of Groningen, The Netherlands
[3]University Medical Center Groningen, The Netherlands
[4]University of Tennessee Health Science Center, USA
[5]University Medical Center Utrecht, The Netherlands

20 September 2016

Paper DOI: http://dx.doi.org/10.21105/joss.00087
Software Repository: https://github.com/DannyArends/CTLmapping
Software Archive: https://dx.doi.org/10.5281/zenodo.163510

Summary

The CTLmapping repository contains an implementation of the Correlation Trait Loci (CTL) algorithm first presented in (Danny Arends 2014). CTL mapping allows geneticists to pursue network inference by discovering genetic loci associated with correlation difference between phenotypes.

CTL mapping is complementary to the proven quantitative trait locus (QTL) mapping method which maps/associates each separately observed phenotype against genotype. CTL mapping, in contrast, associates correlation differences observed *between* two phenotypes at a time, subject to the genotype. In other words, QTL mapping treats phenotypes independently while CTL mapping connects phenotypes. CTL show very similar profiles to QTL, but get interesting when they differ (see figure 1).

CTL differs from mediation [Chick:2016], for example, where the goal is to use covariates tied to genomic position, i.e., mRNA expression, and to find change in strength of QTL signal. The CTL method does not require phenotype tied to genomic location and provides an unbiased method to look for those genomic loci which control correlation differences between phenotypes.

The CTL method is somewhat related to ANCOVA for QTL mapping which is often used to determine changes in correlation. ANCOVA is used in QTL mapping to adjust for covariates when searching for QTL and is used to improve mapping power, because variance in the dependent variable is absorbed by covariates. CTL, in contrast, are calculated without knowledge of QTL.

By comparing differences between QTL and CTL and by connecting phenotypes CTL mapping provides a mechanism for inference and discovery of causality (see chapter 4 (Danny Arends 2014)). This is particularly of interest when phenotype correlations change with conditions, for example in pathways with highly correlated gene expression patterns (see figure 1). CTL mapping differs from existing correlation methods, such as set test methods (e.g., (K. Wang, Li, and Hakonarson 2010)) in that CTL mapping does not require prior information on sets (e.g., pathways) and uses (existing) QTL information for inference.

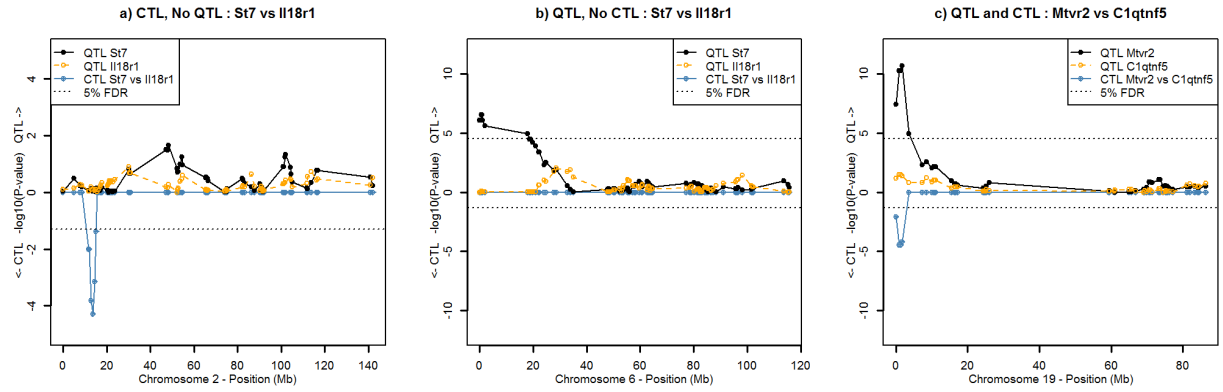

Figure 1: Examples of colocated CTL and QTL profiles, as found in GeneNetwork dataset GN207 (BXD mouse eye mRNA). (a) CTL without a colocalizing QTL between the expression *St7* and *Il18r1* genes, i.e., CTL changes at ~ 15 Mb at chromosome 2 from -0.39 B locus, to 0.86 D locus while both genes do not show a difference in mean expression. (b) *St7* gene shows a QTL at chromosome 6 and no CTL are detected between *St7* and *Il18r1* (possibly implying that the expression of this gene is regulated by some variant at this locus). (c) Expression variation of *Mtvr2* is linked to a CTL with *C1qtnf5* (0.85 B locus to -0.46 D locus) and to a conventional QTL, both of which map to proximal chromosome 19.

CTL analysis can be performed on combined phenotypes obtained from the whole spectrum of data types: i.e., from classical phenotypes, such as yield and disease susceptibility, to high-throughput experimental data, such as micro-arrays, RNA-seq and/or protein abundance measurements. This is especially useful in combined datasets, e.g. a combination of: classical phenotypes, protein abundance and gene expression (see figure 2).

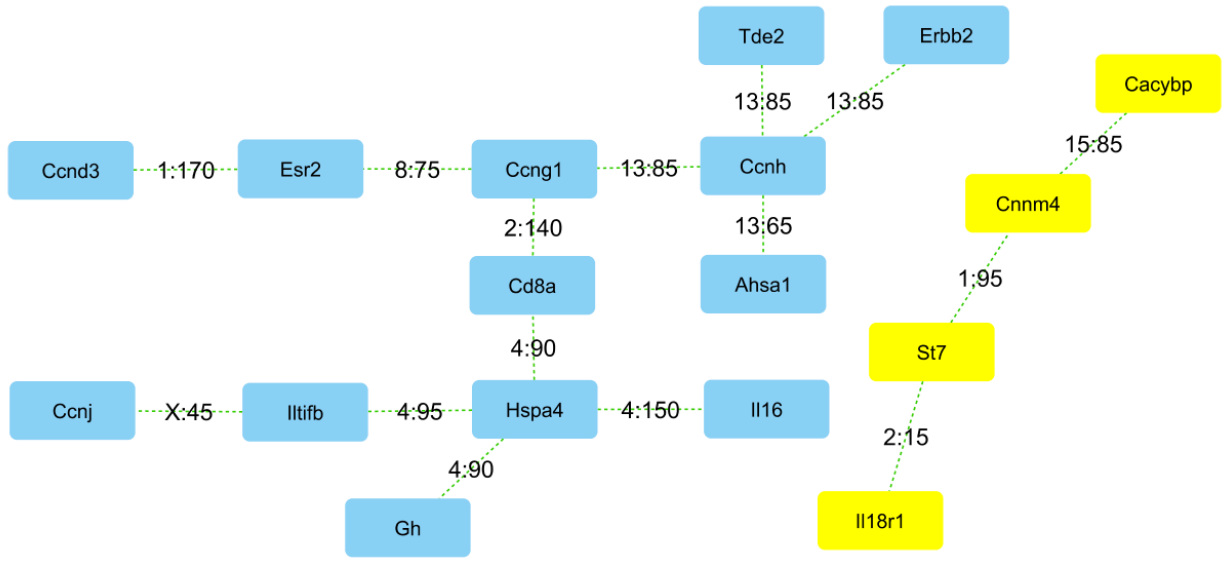

Figure 2: By network inference CTL discover the genetic wiring of classical phenotypes and identify known and new key players in the genetic / protein network underlying classical phenotypes using CTL and QTL information; as discovered without prior information in GeneNetwork BXD mouse datasets (again from GN207). Here we visualize significantly changed correlation between genes in different studies as edges between the genes (nodes) with links showing the location (Chr:Position) at which correlation was lost.

CTL mapping can be applied in model organism experimental crosses, such as mouse and the plant *Arabidopsis thaliana* (see example datasets below); as well as outbred-crosses, such as the Mouse diversity outbred cross (Mouse-DO), the Rat heterogeneous stock (Rat HS) and in A. *thaliana* MAGIC (Koning and McIntyre 2014);

as well as in natural populations, such as human. For statistical power, in general: the more individuals the better. But, as a rule of thumb it is about the same as for QTL, i.e., about 100 individuals for a recombinant inbred line (RIL), and 1,000 individuals for genome-wide association (GWA) in human. This rule of thumb was determined by performing power analysis using genome wide simulations with varying effect sizes and minor allele frequencies. (Danny Arends 2014).

The CTL mapping software is provided as a free and open source (FOSS) package for the R Project for Statistical Computing (Team 2005). Data structures of the CTL mapping R package have been harmonized with the popular R/qtl package (D Arends et al. 2010), allowing users to quickly and efficiently re-analyze previous (R/)QTL experiments. Additional advantages of close integration with R/qtl are the many input formats supported by R/qtl, and access to all plot and helper functions provided by R/qtl.

The core CTL mapping algorithm is written in standalone C making it easy to integrate the CTL mapping algorithm into other languages that support bindings to C functions. As a proof of concept the CTL repository provides bindings for the D programming language.

CTL has been integrated into GeneNetwork (GN), a FOSS framework for web-based genetics that can be deployed anywhere (Sloan et al. 2016). This allows results from CTL mapping to be interactively explored using the GeneNetwork web interface. Additionally results from CTL mapping can be visualized by plotting routines provided by the R package and results can be exported to external tools (such as Cytoscape (Shannon et al. 2003)) for visualization and interactive exploration.

Example datasets

CTL mapping comes with several example datasets (in Rdata format) for the user to explore:

- 301 gene expression traits measured on 109 *Saccharomyces cerevisia* (Brem et al. 2002)
- 9 Metabolite expression traits measured on 403 *Arabidopsis Thaliana* (Blair, Kliebenstein, and Churchill 2012)
- 24 Metabolite expression traits measured on 162 *Arabidopsis Thaliana* (Keurentjes et al. 2006)

(instructions can be found in the README).

Future work

CTL is computationally very intensive, phenotypes *O(n2)*, both in terms of RAM use and CPU. Future work includes research into improving the CTL algorithm for large scale correlations and inference, including the use of GPU/supercomputing. In the context of GeneNetwork we are also working on adding exploratory interactive visualization (such as Cytoscape and D3 interactive graphics).

References

Arends, D, P Prins, R C Jansen, and K W Broman. 2010. "R/qtl: high-throughput multiple QTL mapping." *Bioinformatics (Oxford, England)* 26 (23). Groningen Bioinformatics Centre, University of Groningen, Groningen, The Netherlands.: 2990–2. doi:10.1093/bioinformatics/btq565.

Arends, Danny. 2014. "High-Throughput Computational Methods and Software for Quantitative Trait Locus (QTL) Mapping." PhD thesis. http://hdl.handle.net/11370/29cf3cc5-6596-4d5a-a87c-490b31676a96.

Blair, R H, D J Kliebenstein, and G A Churchill. 2012. "What can causal networks tell us about metabolic pathways?" *PLoS Computational Biology* 8 (4): e1002458. doi:10.1371/journal.pcbi.1002458.

Brem, R B, G Yvert, R Clinton, and L Kruglyak. 2002. "Genetic dissection of transcriptional regulation in

budding yeast." *Science (New York, N.Y.)* 296 (5568): 752–5. doi:10.1126/science.1069516.

Keurentjes, J J B, J Fu, R C H de Vos, A Lommen, R D Hall, R J Bino, L H W van der Plas, R C Jansen, D Vreugdenhil, and M Koornneef. 2006. "The genetics of plant metabolism." *Nature Genetics* 38 (7): 842–9. doi:10.1038/ng1815.

Koning, D.J. de, and Lauren M McIntyre. 2014. "GENETICS and G3: Community-Driven Science, Community-Driven Journals." *G3: Genes|Genomes|Genetics* 4 (9). Genetics Society of America: 1567–8. doi:10.1534/g3.114.013680.

Shannon, P, A Markiel, O Ozier, N S Baliga, J T Wang, D Ramage, N Amin, B Schwikowski, and T Ideker. 2003. "Cytoscape: a software environment for integrated models of biomolecular interaction networks." *Genome Research* 13 (11): 2498–2504. doi:10.1101/gr.1239303.

Sloan, Zachary, Danny Arends, Karl W. Broman, Arthur Centeno, Nicholas Furlotte, Harm Nijveen, Lei Yan, Xiang Zhou, Robert W. Williams, and Pjotr Prins. 2016. "GeneNetwork: Framework for Web-Based Genetics." *JOSS* 1 (2). The Open Journal. doi:10.21105/joss.00025.

Team, RDC. 2005. "R: A language and environment for statistical computing." *R Foundation for Statistical Computing* 1. http://r-project.kr/sites/default/files/2/%EA/%B0/%95/%EA/%B0/%95/%EC/%A2/%8C/%EC/%86/%8C/%EA/%B0/%9C/_/%EC/%8B/%A0/%EC/%A2/%85/%ED/%99/%94.pdf.

Wang, K., M. Li, and H. Hakonarson. 2010. "Analysing biological pathways in genome-wide association studies." *Nat Rev Genet* 11 (12): 843–54. doi:10.1038/nrg2884.