CellPress

Review

Dynamic Quantitative Trait Locus Analysis of Plant Phenomic Data

Zitong Li[1,2] and Mikko J. Sillanpää[1,2,*]

Advanced platforms have recently become available for automatic and systematic quantification of plant growth and development. These new techniques can efficiently produce multiple measurements of phenotypes over time, and introduce time as an extra dimension to quantitative trait locus (QTL) studies. Functional mapping utilizes a class of statistical models for identifying QTLs associated with the growth characteristics of interest. A major benefit of functional mapping is that it integrates information over multiple timepoints, and therefore could increase the statistical power for QTL detection. We review the current development of computationally efficient functional mapping methods which provide invaluable tools for analyzing large-scale timecourse data that are readily available in our post-genome era.

QTL Mapping and High-Throughput Phenotyping

In plant genetics, **quantitative trait locus** (QTL) **mapping** (see Glossary) is often used to identify QTLs or causal genes associated with phenotypes of interest [1]. QTL mapping is a crucial step in **marker-assisted selection** (MAS), which has been successfully applied in many plant breeding programs [2]. Recent advances in **next-generation sequencing** (NGS) techniques have provided fast and inexpensive access to genomic information on a large scale, which allows the execution of the QTL and/or association mapping based on genome-wide marker data [3,4] (**genome-wide association mapping**). In addition to the genotypic information, QTL mapping also requires high-quality phenotype data. Intuitively, increasing the sample size in a QTL analysis can improve the power to correctly identify QTLs. From another perspective, it is also beneficial to perform sample collection of plants under similar or exchangeable microenvironmental conditions to ensure that environmental variance and noise are minimized. However, traditional plant phenotyping approaches largely utilize manual laboratory experiments and visual scoring by experts, and these practices are often time-consuming and it is difficult to arrange desirable growth conditions for individual phenotypes or repeats. Consequently, the development of high-throughput and/or automated phenotype platforms [5–11], involving both automated recording and screening of phenotypes by various imaging techniques, and effectively allocating and monitoring environmental conditions, has started to gain more and more attention. This advance greatly eases the measurement of timecourse phenotype data and will provide new insight into genetic studies of plant growth. Accordingly, this review describes functional mapping – a class of statistical methods designed to efficiently integrate temporal information and identify QTLs associated with phenotypic dynamics. Computational efficiency is an important consideration in some recently developed approaches to meet the new challenge from high-dimensional genotype and phenotype data. The review also discusses the limitations of the current methods and highlights future research directions.

Trends

High-throughput imaging techniques are capable of measuring time-series of plant phenotypes, which may potentially facilitate the QTL analysis of developmental and growth related traits.

A major benefit of functional mapping is that it integrates information over multiple timepoints, and therefore could increase the statistical power for QTL detection.

To handle high-dimensional genotyping and phenotyping data, computational efficiency is the focus of the novel statistical methods for dynamic QTL analysis.

[1]Biocenter Oulu, Oulu, Finland
[2]Department of Mathematical Sciences and Department of Biology, University of Oulu, 90014 Oulu, Finland

*Correspondence:
Mikko.Sillanpaa@oulu.fi
(M.J. Sillanpää).

High-Throughput Phenotyping

A high-throughput phenotyping platform (HTPP) may either be set up in controlled environments, such as growth chambers and greenhouses, or in the field. Imaging techniques [12,13] are core facilities of a HTPP and are utilized to record phenotypes as pictures or videos, including for example visible and/or digital imaging for traits such as height, shoot biomass, yield traits, root architecture and morphology, fluorescence imaging for traits such as photosynthetic status, and 3D imaging for traits such as root structure. These 2D and/or 3D images obtained by HTPPs may be stored on a high-performance computing infrastructure and be read and processed by mathematical and computational image-analysis tools to extract various traits. Therefore, HTPPs do not only rely on advanced imaging and remote sensing techniques but also require high-performance computational tools for processing and managing image data [8].

HTPPs have been successfully developed in some controlled environments (e.g., Australian Plant Phenomics Facility, www.plantphenomics.org.au), where microenvironmental conditions such as light and water are automatically adjusted, and the positions of plants can also be relocated to minimize the environmental heterogeneity between individuals and repeats. Such environmental homogeneity may also be achieved by efficient use of experimental design and analysis within the HTPP framework (e.g., choosing suitable block size and/or using appropriate statistical models for the analysis) with less cost than relocating the plants [14]. However, it has been reported that the QTLs identified in controlled environments may not contribute to crop improvement in the field [6]. By contrast, many HTPPs which have been directly developed in the field cannot adequately monitor environmental factors, such as the temporal effects of climate and atmospheric variation, or the spatial effect caused by soil variation [8]. The use of these platforms in the field will require improvements to permit better monitoring of environmental factors, and application of an appropriate experimental design would also be useful to maintain a sufficient degree of environmental homogeneity within a field. Another issue with HTPPs is that they are currently only available for a limited number of plant species, and more generic phenotyping platforms that are applicable for multiple species will be needed in the future.

High-Throughput Phenotyping Facilitates the Measurement of Developmental Traits

Studying the developmental process (e.g., growth) of the traits is often interesting. Analyzing developmental behavior of a trait is only possible if there are repeated measurements of individual phenotypes over time. Monitoring trait development by traditional phenotyping approaches is far from simple work. For example, obtaining repeated measurements of some traits such as root architecture is not possible using conventional methods because these methods would necessitate destroying the plants. By contrast, some HTPPs, relying on various imaging techniques, are able to more conveniently monitor the dynamic growth of the traits without damaging the plant [15–20]. Therefore, HTPPs can efficiently bring time as an extra dimension to the phenotype data, which may potentially facilitate QTL analysis of developmental and growth-related traits. To efficiently utilize timecourse data generated by HTPPs, advanced statistical methods are needed [18,19]. Ideally such methods should integrate the phenotypic information over multiple timepoints, map the dynamic phenotype–genotype relationship, and account for possible random errors introduced by temporal and/or spatial environmental factors.

Functional QTL Mapping

Analysis of quantitative trait loci involves modeling, estimation, and hypothesis testing. Statistical approaches for analyzing a quantitative trait and/or multiple correlated traits at a single timepoint (Box 1) have been well established [21–23]. When phenotype records at multiple timepoints are available, one may analyze each single timepoint separately and identify QTLs associated with phenotypes at that particular timepoint. This approach ignores the dependency between repeated phenotypic measurements. For example, one may expect that the two phenotypic measurements at neighboring timepoints should have closer values than the two at a greater

Glossary

False discovery rate (FDR): in multiple hypothesis testing, FDR is the expected portion of falsely rejected null hypotheses among the rejected hypotheses.

Family-wise error rate (FWER): the probability of having one incorrectly rejected null hypothesis among all the hypotheses.

Genome-wide association (GWA) mapping: identifies SNPs that are significantly associated with a quantitative trait among a genome-scale SNP set based on population data. Because a high-density SNP panel is applied, the detected significant markers should be in high linkage disequilibrium with QTLs, and can be used to proxy the QTL positions.

Marker-assisted selection (MAS): a molecular strategy to indirectly improve economically relevant traits during their early development stages, based on selection targeted on a trait-associated set of markers.

Multiple-split test: a hypothesis-testing method for variable selection. The data are repeatedly and randomly divided into two parts, the first part is used to perform variable selection and the second is used to construct a test statistic. The *P* value of each SNP is averaged over multiple replicates to reduce the uncertainty. This method can be used to control FWER.

Next-generation sequencing (NGS): utilizes efficient parallel sequencing and imaging techniques to simultaneously produce thousands to millions of reads with low cost. Advances in NGS facilitate the genotyping of high-density SNP panels to be used later in genetic studies.

Permutation test: hundreds of datasets are generated by randomly shuffling phenotypes into a different order and destroying phenotype–genotype relationships. Each shuffled dataset is analyzed to construct an empirical distribution of SNP test statistics (i.e., null distribution). The observed SNP test statistic is then tested against this distribution.

Quantitative trait locus (QTL): a segment of a DNA sequence which contributes to the variation of a quantitative trait by containing or being linked to the genes determining that trait.