Natural genetic variation impacts expression levels of coding, non-coding and antisense transcripts in fission yeast

Mathieu Clément-Ziza[1,2], Francesc X. Marsellach[3], Sandra Codlin[3], Manos A. Papadakis[4], Susanne Reinhardt[1], Maria Rodriguez-Lopez[3], Stuart Martin[3], Samuel Marguerat[3], Alexander Schmidt[5], Eunhye Lee[3], Christopher T. Workman[4], Jürg Bähler[3] , and Andreas Beyer [1,2]

[1] Biotechnology Centre, Technische Universität Dresden, Germany. [2] CECAD, University of Cologne, Köln, Germany. [3] University College London, London, United Kingdom. [4] Center for Biological Sequence Analysis, Technical university of Denmark, Denmark. [5] Biozentrum, University of Basel, Switzerland

mathieu.clement-ziza@uni-koeln.de

Abstract

Our current understanding of how natural genetic variation affects gene expression beyond well-annotated coding genes is still limited. The use of deep sequencing technologies for the study of expression quantitative trait loci (eQTLs) has the potential to close this gap. Here, we generated the first recombinant strain library for fission yeast and conducted an RNA-seq-based QTL study of the coding, non-coding, and antisense transcriptomes. We show that the frequency of distal effects (*trans*-eQTLs) greatly exceeds the number of local effects (*cis*-eQTLs) and that non-coding RNAs are as likely to be affected by eQTLs as protein-coding RNAs. We identified a genetic variation of *swc5* that modifies the levels of 871 RNAs, with effects on both sense and antisense transcription, and show that this effect most likely goes through a compromised deposition of the histone variant H2A.Z. The strains, methods, and datasets generated here provide a rich resource for future studies [CZMC$^+$14].

Introduction

Variation in gene expression, which in turn is often caused by natural genetic variation, is a major factor causing intra-species phenotypic differences. Hence, investigating the influence of genetic variation on gene expression has been the focus of intense research. The identification of expression quantitative trait loci (eQTLs), that is, genomic regions that are linked with the expression of a specific transcript, has primarily been conducted using DNA microarrays [BYCK02]. High-throughput sequencing of cDNA (RNA-seq) has great potential to provide qualitatively and quantitatively new insights beyond mere mRNA quantification.

Previous RNA-seq based eQTL studies have focused on measuring new traits, and have been limited to the detection of local eQTLs, so called *cis*-eQTLs [LHW$^+$11, MP11, PMP$^+$10], or have identified a only a relatively small proportion of distant eQTLs [BMZ$^+$13]. *cis*-eQTLs are located at or close to the genes whose expression they directly affect; while *trans*-eQTLs are remote from the genes whose expression they affect. Further, sequence variation information contained in the RNA-seq data has not been exploited in the framework of eQTL mapping.

Here, we have conducted an expression QTL study characterized by a design enabling a high statistical power for association detection and by a broad investigation of pervasive expression beyond well-annotated coding genes (i.e. non-coding and antisense transcripts). The high statistical power contributed to the improved discovery of *trans*-eQTLs, suggesting that previous studies may have been overestimating the fraction of *cis*-eQTLs. First, we generated a recombinant strain library for fission yeast (*Schizosaccharomyces pombe*) suitable for powerful QTL studies, which was subsequently subjected to high-resolution measurements of growth kinetics and strand-specific RNA-seq. Whereas microarray probes rely on a fixed reference genome, RNA-seq allows for the individualized quantification of transcripts taking genomic variation into account. We show that our approach, which explicitly includes individual genomes, reduces the potential for false-positive eQTLs. Further, because RNA-seq measures the actual transcript sequences of a given strain, it can also be used for genotyping the strain library. We developed a computational framework for the robust genotyping of recombinant strains using RNA-seq data, which eliminates the need for separate genotyping experiments. Finally, RNA-seq makes no assumptions about the structure of genomic features. In the context of QTL studies, it can thus be used to identify genetic variants affecting non-annotated features. Here, we present a striking example of a variation affecting antisense transcription of hundreds of *S. pombe* genes detected in this study.

Results and methods

Generation and phenotyping of a recombinant fission yeast strain library

We generated the first recombinant strain library for *Schizosaccharomyces pombe* suitable for QTL studies. In order to enable the detection of association at a high statistical power, we selected closely related parental strains to reduce the genetic complexity of the library (0.05% divergence, comparable to the average divergence between two humans). This cross was subsequently subjected to high-resolution measurements of growth kinetics and deep strand-specific RNA-seq (average effective depth 37.5x).

Genotyping of the strain library by RNA-seq

eQTL studies require both the genotypes and the expression profiles of each individual of the studied population. We developed a strategy enabling the genotyping of a recombinant strain library through RNA-seq. Thus, separate genotyping experiments of the segregant strains were not needed. First we sequenced the genome of the parental strains a great depth in order to detect potentially all genomic variants. Then we used RNAseq data of the segregants to detect genomic variation at the sites polymorphic when comparing the progenitors. On average half of the sites could be directly genotyped after conservative filtering. It was sufficient to identify haplotype blocks and thus infer genotypes at the remaining sites. This lead to the genotyping of the whole library at 4,481 sites.

Accounting for individual genomes improves transcript quantification

In microarray-based expression studies, sequence variation in probe regions can affect the hybridization efficiency. Because this leads to an allele-specific signal bias, sequence variation can inflate the number of false *cis*-eQTL calls [ATL+07]. Notably, RNA-seq studies are neither immune to such artifacts [DMP+09] as transcript quantification usually involves the mapping of sequence reads to a reference genome. In this study, gene expression quantification was performed by aligning reads to strain-specific genomes in order to minimize this bias. Using both simulations and real data, we compared this strategy to reference genome mapping. Results show that aligning RNA-seq data against individualized genomes marginally improves transcript quantification, while ignoring individual sequence variation can inflate the number of falsely detected *cis*-eQTLs.

trans-eQTLs greatly exceed cis-eQTLs in abundance

After mapping the QTLs using a Random Forest based approach [MASB10, PCZL+13], one of the most surprising results of this study was the small fraction of *cis*-eQTL that were detected. It is generally assumed that *cis*-eQTLs can be detected more easily than *trans*-eQTLs [ASWB13, HLB+11, SMD+03]: (i) direct effects are stronger that distant indirect ones, and (ii) searching for *trans* linkages involves testing a much larger number of hypotheses. Strikingly we detected a much higher fraction of trans-eQTLs (˜90%) than *cis*-eQTL. We attribute this to the high statistical power of our study, which could be due to several factors: the genetic similarity of the parental strains reducing the complexity, the use of deep sequencing reducing trait noise, and/or the advanced methods we used to analyze these data.

Non-coding and expression are strongly affected by genetic variation

RNA-seq enables the quantification of entire transcriptomes, including non-coding RNAs (ncRNAs). The high sequencing depth used here and the extensive annotation of the fission yeast genome enabled us to quantify transcript levels for 1,428 annotated ncRNAs. Thus, this analysis presents the first comparative eQTL mapping for coding versus non-coding transcript levels at a genomic scale. We showed that that the expression of non-coding RNAs is at least as much affected by genetic variation as the expression of protein-coding RNAs. To further investigate the importance of non-coding RNAs as effectors of eQTLs, we predicted the most likely causal gene for each eQTL. Our result suggests that non-coding RNAs substantially contribute as effectors of the genetic variation of gene expression.

A frameshift in swc5 causes major eQTL hotspot, reduces H2A.Z deposition increase antisense transcription

We identified an eQTL hotspot (locus regulating numerous genes) affecting the sense expression of 817 genes and the anti-sense expression of 1,384 traits. This QTL hotspot shows more widespread gene expression effects than any other hotspot reported so far. Because of its extraordinary strength, we wanted to unravel its molecular basis. We identified a frame-shift polymorphism in the gene *swc5* as being the molecular regulator at this locus. Swc5 is a component of the Swr1 protein complex controlling the chromosomal deposition of the histone variant H2A.Z. H2A.Z has been associated with the control of antisense transcription if fission yeast [ZFZ+09]. We showed that the effect of *swc5* hotspot most likely goes through a compromised deposition of the histone variant H2A.Z, which consequently leads to an increase of read-through antisense transcription. We performed numerous experiments and analyses that all corroborated this hypothesis. Notably we studied expression changes in strains deleted for *swc5*, and we analyzed the H2A.Z occupancy via ChIP-seq.

Conclusion

Several methodological aspects have been developed in this study, for instance RNA-seq based genotyping or strain specific genome mapping. Moreover, the high statistical power to detect eQTLs characterizing this study led to interesting findings regarding the genetic control of non-coding expression and the relative importance of *trans* effect. The detailed experimental and analytic validation on one of the casual genes (*swc5*) offers new insights on how a QTL could modulate its target genes. This study has been published in *Molecular System Biology* [CZMC+14].

Presentation outline

The presentation will first motivate the need of studying the genetic basis of molecular traits. Then the concepts of eQTL mapping will be presented. The main part of the presentation will focus on both methodological aspects (RNA-seq based genotyping or random forest based QTL mapping), and on the result highlights (the importance of the regulation of non-coding RNA, the proportion of *cis*/*trans*). Finally the *swc5* eQTL hotspot will be briefly presented.

References

[ASWB13] M. Ackermann, W. Sikora-Wohlfeld, and A. Beyer. Impact of Natural Genetic Variation on Gene Expression Dynamics. *PLoS Genet*, 9(6):e1003514, June 2013.

[ATL$^+$07] Rudi Alberts, Peter Terpstra, Yang Li, Rainer Breitling, Jan-Peter Nap, and Ritsert C. Jansen. Sequence Polymorphisms Cause Many False cis eQTLs. *PLoS ONE*, 2(7):e622, July 2007.

[BMZ$^+$13] A. Battle, S. Mostafavi, X. Zhu, J. B Potash, C. Weissman, M. M .and McCormick, C. D Hauden-schild, K. B Beckman, J. Shi, R. Mei, A. E Urban, S. B Montgomery, Douglas F Levinson, and D. Koller. Characterizing the genetic basis of transcriptome diversity through RNA-sequencing of 922 individuals. *Genome research*, October 2013.

[BYCK02] R. B. Brem, G. Yvert, R. Clinton, and L. Kruglyak. Genetic dissection of transcriptional regulation in budding yeast. *Science (New York, N.Y.)*, 296(5568):752–755, April 2002.

[CZMC$^+$14] M. Clément-Ziza, F. X. Marsellach, S. Codlin, M. A. Papadakis, S. Reinhardt, M. Rodriguez-Lopez, S. Martin, S. Marguerat, A. Schmidt, E. Lee, C. T. Workman, J. Bahler, and A. Beyer. Natural genetic variation impacts expression levels of coding, noncoding, and antisense transcripts in fission yeast. *Molecular Systems Biology*, 10(11):764, November 2014.

[DMP$^+$09] J. F Degner, J. C Marioni, A. A Pai, J. K Pickrell, E. Nkadori, Y. Gilad, and J. K Pritchard. Effect of read-mapping biases on detecting allele-specific expression from RNA-sequencing data. *Bioinformatics (Oxford, England)*, 25(24):3207–3212, December 2009.

[HLB$^+$11] B. Holloway, S. Luck, M. Beatty, J-A Rafalski, and B. Li. Genome-wide expression quantitative trait loci (eQTL) analysis in maize. *BMC Genomics*, 12(1):336, June 2011.

[LHW$^+$11] E. Lalonde, K. C.H. Ha, Z. Wang, A. Bemmo, C. L. Kleinman, T. Kwan, T. Pastinen, and J. Majewski. RNA sequencing reveals the role of splicing polymorphisms in regulating human gene expression. *Genome Research*, 21(4):545–554, April 2011.

[MASB10] J.J. Michaelson, R. Alberts, K. Schughart, and A. Beyer. Data-driven assessment of eQTL mapping methods. *BMC Genomics*, 11(1):502, 2010.

[MP11] Jacek Majewski and Tomi Pastinen. The study of eQTL variations by RNA-seq: from SNPs to phenotypes. *Trends in Genetics*, 27(2):72–79, February 2011.

[PCZL$^+$13] P. Picotti, M. Clément-Ziza, H. Lam, D. S. Campbell, A. Schmidt, E. W. Deutsch, H. Rst, Z. Sun, O. Rinner, L. Reiter, Q. Shen, J. J. Michaelson, A. Frei, S. Alberti, U. Kusebauch, B. Wollscheid, R. L. Moritz, A. Beyer, and R. Aebersold. A complete mass-spectrometric map of the yeast proteome applied to quantitative trait analysis. *Nature*, 494(7436):266–270, February 2013.

[PMP$^+$10] JK. Pickrell, JC. Marioni, AA. Pai, JF. Degner, BE. Engelhardt, E. Nkadori, J-B. Veyrieras, M. Stephens, Y. Gilad, and JK. Pritchard. Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature*, 464(7289):768–772, March 2010.

[SMD$^+$03] E. E. Schadt, S. A. Monks, T. A. Drake, A. J. Lusis, N. Che, V. Colinayo, T. G. Ruff, S. B. Milligan, J. R. Lamb, G. Cavet, P. S. Linsley, M. Mao, R. B. Stoughton, and S. H. Friend. Genetics of gene expression surveyed in maize, mouse and man. *Nature*, 422(6929):297–302, March 2003.

[ZFZ$^+$09] M Zofall, T. Fischer, K. Zhang, M. Zhou, B. Cui, T. D. Veenstra, and S.I.S. Grewal. Histone H2A.Z cooperates with RNAi and heterochromatin factors to suppress antisense RNAs. *Nature*, 461(7262):419–422, August 2009.