QTL-seq 分析报告

Written by 王鹏飞

Email: wangpf0608@126.com

一 QTL-seq 原理

QTL-seq[1]是一种将 Bulked-segregant analysis (BSA)[2,3]和高通量测序相结合，快速定位 QTL 的方法。目标性状有差异的双亲构建的分离群体中分别选取极端表型个体进行等量混合构建两个极端表型 bulked DNA pool 并进行测序。随后进行变异分析筛选出双亲间 SNP 位点并分别计算两个 bulked DNA pool 中每个 SNP 位点上某一亲本基因型 read 覆盖深度占该位点总 read 深度的比值，即 SNP index，通过两个 bulked DNA pool 的 SNP index 相减即得到 ΔSNP index。在基因组所有区域中，目标基因及其连锁的区域由于根据表型受到相反的选择在两个 bulked DNA pool 中表现出不同的趋势，因此 ΔSNP index 会显著偏离 0 附近；另一方面，于目标形状无关的区域则两个 bulked DNA pool 则表现为相似的变化趋势，因此 ΔSNP index 会在 0 附近波动（图 1）。

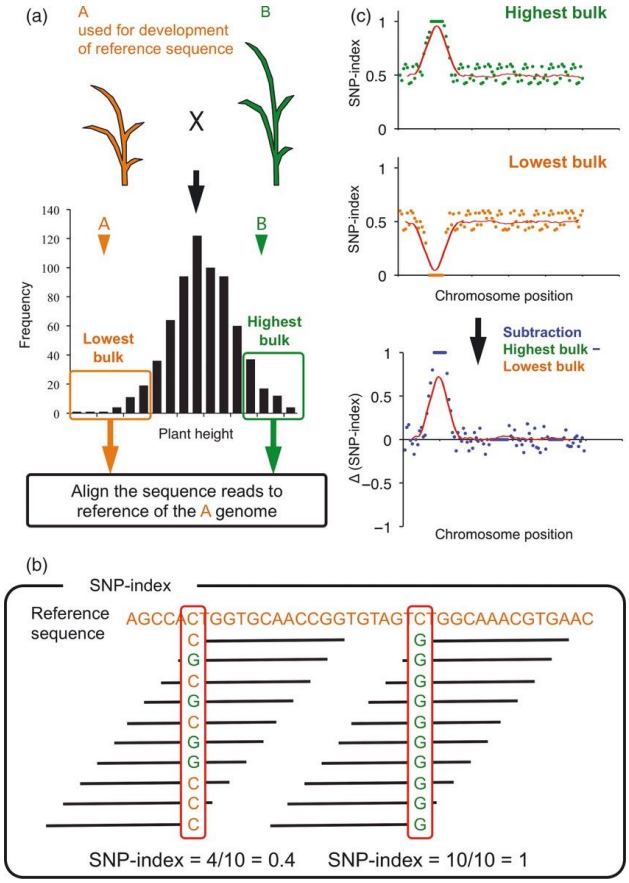
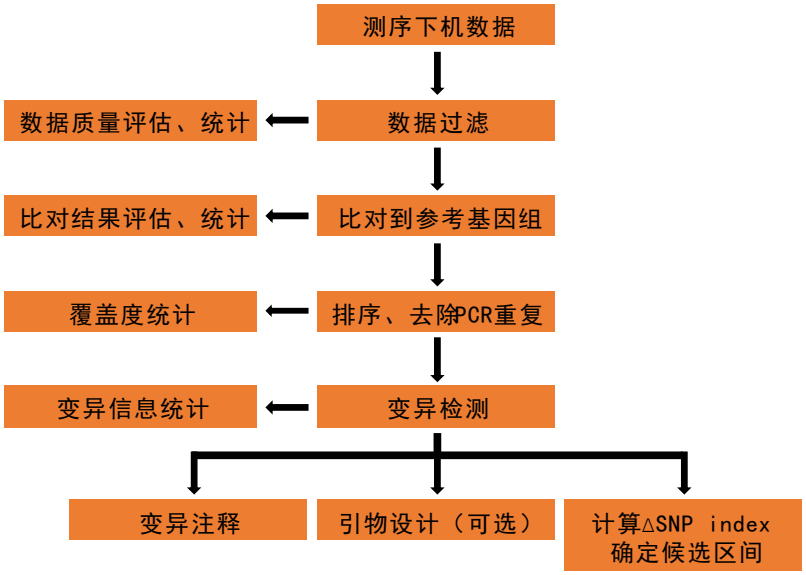


图 1 QTL-seq 原理示意图

二 QTL-seq 流程图



三 QTL-seq 分析流程

1 取样、提取 DNA、建库和测序

选取分离群体中极端表型个体各 30-50 株以及双亲取幼嫩组织，分别提取 DNA 并将分离群体极端表型个体按照等量原则混合构建 bulked DNA pools，然后进行建库和高通量测序（混池 DNA 深度宜与单株个数相同）。（具体流程应参考实验设计以及公司测序报告）

2 数据过滤

使用 fastp[4]（version: 0.20.0）对 raw data 进行过滤得到 clean data。统计过滤前后 total bases、total reads、Q30、Q20、GC content 以及有效数据比率（data_stat.csv/txt），同时使用 FastQC（version: 0.11.9）对过滤前后的数据进行质量评估（QC/sample_fastqc.html）。

3 比对到参考基因组

使用 Bowtie2[5]（version: 2.4.1）软件将 clean read 比对到甘蓝型油菜 ZS11 参

考基因组[6]，得到 SAM（Sequence Alignment/Map）格式文件并对比对结果进行统计（01.Mapping/align_stat.csv），随后使用 SAMtools（version: 1.9）对比对结果（SAM 文件）按照染色体和位置进行排序并转换为 BAM（Binary Alignment/Map）格式文件[7]，使用 Picard tools[8]（version: 2.23.2）去除建库过程中产生的 PCR 重复并统计样本基因组覆盖率（01.Mapping/cov_stat.txt）。

4 变异分析

变异分析使用 The Genome Analysis Toolkit，GATK[9]（version: 3.8-0-ge9d806836）完成，首先使用 GATK 的 HaplotypeCaller 功能对样本单独分析再使用 CombineGVCFs 功能合并，随后使用 GenotypeGVCFs 功能得到 SNP 和 INDEL 信息，最后使用 VariantFiltration 功能过滤原始的变异位点的到可靠的变异信息。

5 QTL-seq 分析

在变异分析得到的可靠 SNP 位点中，筛选亲本内纯和并且亲本间不同的位点，并且进一步过滤掉双亲和两各混池中低覆盖深度的位点，此时统计剩余 SNP 位点在基因组上的分布情况并绘图。利用最终筛选出的 SNP 进行下一步的分析，首先计算每个混池的 SNP index 值，随后计算 ΔSNP index 值并绘图，其中点图按照 500kb 窗口大小、250kb 步长进行滑窗统计，折线图是由 R 软件（version: 4.0.2）扩展包 QTLseqr[10]（version: 0.7.5.2）按照 2Mb 窗口大小统计得到，置信区间按照 Takagi 等（2013）[1]描述方法计算得到，超出 95 或 99%置信区间的区域视为目标性状候选 QTL。

6 引物设计（若需要请单独备注）

为进行 QTL 验证和进一步精细定位，根据变异分析中得到的可靠 INDEL 位点进行全基因组范围的引物设计。首先根据 INDEL 位置提取上下游各 250 bp 基因组序列，随后使用 primer3[11]（version: 2.5.0）进行引物设计（引物退火温度、长度、GC 含量和扩增片段长短等偏好需提前说明，同时需根据实验室电泳仪器能区分的 INDEL 大小挑选合适的位点，引物根据参考基因组设计，其真实位置和特异性需根据序列比对和实验结果进行验证）。

参考文献

1 Hiroki Takagi, Akira Abe, Kentaro Yoshida, Shunichi Kosugi, Satoshi Natsume, Chikako Mitsuoka, Aiko Uemura, Hiroe Utsushi, Muluneh Tamiru, Shohei Takuno, Hideki Innan, Liliana M. Cano, Sophien Kamoun, Ryohei Terauchi. QTL -seq: rapid mapping of quantitative trait loci in rice by whole genome resequencing of DNA from two bulked populations[J]. The Plant Journal, 2013, 74(1).

2 Giovannoni James J., Wing Rod A., Ganal Martin W., Tanksley Steven D.. Isolation of molecular markers from specific chromosomal intervals using DNA pools from existing mapping populations[J]. Narnia, 1991, 19(23).

3 R. W. Michelmore, I. Paran,R. V. Kesseli. Identification of Markers Linked to Disease-Resistance Genes by Bulked Segregant Analysis: A Rapid Method to Detect Markers in Specific Genomic Regions by Using Segregating Populations[J]. Proceedings of the National Academy of Sciences of the United States of America, 1991, 88(21).

4 Shifu Chen, Yanqing Zhou, Yaru Chen, Jia Gu. fastp: an ultra-fast all-in-one FASTQ preprocessor[J]. Bioinformatics, 2018, 34(17).

5 Ben Langmead, Steven L Salzberg. Fast gapped-read alignment with Bowtie 2[J]. Nature Methods: Techniques for life scientists and chemists, 2012, 9(4).

6 Jia-Ming Song, Zhilin Guan, Jianlin Hu, Chaocheng Guo, Zhiquan Yang, Shuo Wang, Dongxu Liu, Bo Wang, Shaoping Lu, Run Zhou, Wen-Zhao Xie, Yuanfang Cheng, Yuting Zhang, Kede Liu, Qing-Yong Yang, Ling-Ling Chen, Liang Guo. Eight high-quality genomes reveal pan-genome architecture and ecotype differentiation of Brassica napus[J]. Nature Plants, 2020, 6(1).

7 Li Heng, Handsaker Bob, Wysoker Alec, Fennell Tim, Ruan Jue, Homer Nils, Marth Gabor, Abecasis Goncalo, Durbin Richard. The Sequence Alignment/Map format and SAMtools.[J]. Bioinformatics (Oxford, England), 2009, 25(16).

8 “Picard Toolkit.” 2019. Broad Institute, GitHub Repository. http://broadinstitute.github.io/picard/; Broad Institute

9 Aaron McKenna, Matthew Hanna, Eric Banks, Andrey Sivachenko, Kristian

Cibulskis, Andrew Kernytsky, Kiran Garimella, David Altshuler, Stacey Gabriel, Mark Daly, Mark A. DePristo. The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data[J]. Cold Spring Harbor Laboratory Press, 2010, 20(9).

10 Ben N. Mansfeld, Rebecca Grumet. QTLseqr: An R Package for Bulk Segregant Analysis with Next-Generation Sequencing[J]. The Plant Genome, 2018, 11(2).

11 Untergasser Andreas, Cutcutache Ioana, Koressaar Triinu, Ye Jian, Faircloth Brant C., Remm Maido, Rozen Steven G.. Primer3—new capabilities and interfaces[J]. Narnia, 2012, 40(15).