Background: Quantitative Trait Loci (QTL) mapping using bulk segregants is an effective approach for identifying genetic variants associated with phenotypes of interest in model organisms. By exploiting next-generation sequencing technology, the QTL mapping accuracy can be improved significantly, providing a valuable means to annotate new genetic variants. However, setting up a comprehensive analysis framework for this purpose is a time-consuming and error prone task, posing many challenges for scientists with limited experience in this domain.

Findings: Here, we present BSA4Yeast, a comprehensive web-application for QTL mapping via bulk segregant analysis of yeast sequencing data. The software provides an automated and efficiency-optimized data processing, up-to-date functional annotations, and an interactive web-interface to explore identified QTLs.

Conclusion: BSA4Yeast enables researchers to identify plausible candidate genes in QTL regions efficiently in order to experimentally validate causative genetic variations for a phenotype of interest. BSA4Yeast is freely available at https://bsa4yeast.lcsb.uni.lu.

Corresponding Author: Enrico Glaab
University of Luxembourg
Esch-sur-Alzette, Capellen LUXEMBOURG

First Author: Zhi Zhang

Order of Authors: Zhi Zhang
Paul P. Jung
Valentin Grouès
Patrick May
Carole Linster
Enrico Glaab
| Question                                                                 | Answer |
|-------------------------------------------------------------------------|--------|
| Are you submitting this manuscript to a special series or article collection? | No     |
| **Experimental design and statistics**                                   | Yes    |
| Full details of the experimental design and statistical methods used should be given in the Methods section, as detailed in our Minimum Standards Reporting Checklist. Information essential to interpreting the data presented should be made available in the figure legends. |        |
| Have you included all the information requested in your manuscript?      | Yes    |
| **Resources**                                                            | Yes    |
| A description of all resources used, including antibodies, cell lines, animals and software tools, with enough information to allow them to be uniquely identified, should be included in the Methods section. Authors are strongly encouraged to cite Research Resource Identifiers (RRIDs) for antibodies, model organisms and tools, where possible. |        |
| Have you included the information requested as detailed in our Minimum Standards Reporting Checklist? | Yes |
| **Availability of data and materials**                                   | Yes    |
| All datasets and code on which the conclusions of the paper rely must be either included in your submission or deposited in publicly available repositories (where available and ethically appropriate), referencing such data using a unique identifier in the references and in the "Availability of Data and Materials" section of your manuscript. |        |
| Have you have met the above requirement as detailed in our Minimum | Yes |
TECHNICAL NOTE

BSA4Yeast: Web-based QTL linkage analysis and bulk segregant analysis of yeast sequencing data

Zhi Zhang\(^1\), Paul P Jung\(^1\), Valentin Grouès\(^1\), Patrick May\(^1\), Carole Linster\(^1\) and Enrico Glaab\(^1,*\)

\(^1\)Luxembourg Centre for Systems Biomedicine (LCSB), University of Luxembourg, Esch-sur-Alzette, Luxembourg

\(^*\)enrico.glaab@uni.lu

Abstract

**Background:** Quantitative Trait Loci (QTL) mapping using bulk segregants is an effective approach for identifying genetic variants associated with phenotypes of interest in model organisms. By exploiting next-generation sequencing technology, the QTL mapping accuracy can be improved significantly, providing a valuable means to annotate new genetic variants. However, setting up a comprehensive analysis framework for this purpose is a time-consuming and error prone task, posing many challenges for scientists with limited experience in this domain. 

**Findings:** Here, we present BSA4Yeast, a comprehensive web-application for QTL mapping via bulk segregant analysis of yeast sequencing data. The software provides an automated and efficiency-optimized data processing, up-to-date functional annotations, and an interactive web-interface to explore identified QTLs. 

**Conclusion:** BSA4Yeast enables researchers to identify plausible candidate genes in QTL regions efficiently in order to validate their genetic variations experimentally as causative for a phenotype of interest. BSA4Yeast is freely available at https://bsa4yeast.lcsb.uni.lu.

**Key words:** QTLs; BSA; mapping

Background

Deciphering the genetic basis of diseases or complex traits is a major task in biomedical and basic biological research and is a key first step towards a better understanding of the molecular mechanisms behind disorders with genetic components. As a forward genetic approach, linkage analysis of Quantitative Trait Loci (QTLs) using bulk segregant analysis (BSA) in model organisms, such as yeast, is an efficient method for identifying novel genetic variants responsible for heritable phenotypic variability \([1, 2]\). By exploiting the capacity of next-generation sequencing (NGS) technologies to assess large numbers of genetic markers efficiently and integrating NGS analysis with linkage mapping, the precision of QTL mapping can be improved significantly as compared to traditional approaches. In order to perform a linkage analysis using BSA in practice, relevant software packages, such as the baseq python package \([3]\), and web-based software, such as EXPLoRA-web [4], have been made available in recent years. However, these tools require researchers to first determine genetic markers of interest from sequencing data. Moreover, they only provide limited annotations for the discovered QTLs (i.e. only the QTL coordinates) and do not support the interactive exploration and visualization of detailed QTL annotations in a web-browser. Since the analysis of NGS data involves several different command-line software tools and is a time-consuming and laborious task, a more efficient, automated analysis framework that supports annotation-based result interpretation would greatly facilitate NGS-based bulk segregant analysis (NGS-BSA).

For this purpose, we have developed BSA4Yeast, a comprehensive web-based analysis software for QTL mapping via bulk segregant analysis of yeast sequencing data (Fig. 1). BSA4Yeast provides the following main new benefits and features:
Figure 1. Overview of the analysis workflow for the BSA4Yeast web-application. The experimental design and other parameters can be specified on the web-interface. Representative results shown at the bottom include (from left to right): allele frequency, G’ statistic values and functional annotations. SacCer3: the reference genome of *Saccharomyces cerevisiae*.

Figure 2. Overview of the three main phases of the software workflow and the individual analysis steps they include. From left to right the three phases cover the following tasks: (I) pre-processing and alignment of the short reads against the yeast genome; (II) identifying genetic markers between two parental lines; and (III) performing QTL analyses and comprehensively annotating the results.

Materials and Methods

Functionality and Workflow

The BSA4Yeast framework for QTL mapping via bulk segregant analysis of yeast sequencing data is built on custom scripts and open-source bioinformatics software (Fig. 2). The software workflow covers three major functionalities: 1) Pre-processing and aligning short reads (Illumina format) against an up-to-date version of the yeast genome; 2) identifying genetic markers between two parental lines; and 3) performing QTL analyses and comprehensively annotating the results using public data in an automated fashion (Fig. 2). For different types of experiments the design of the experiment (DOE) can be adjusted appropriately using flexible parameter settings. The web-application supports one- or two-bulk designs and designs with multiple biological replicates in each bulk, as well as three file formats as input (.fastq, .bam or .map; both paired-end (PE) or single-end (SE) DNA sequencing data is accepted). For .map input files, BSA4Yeast can be used to identify QTLs for any species of interest, and for *S. cerevisiae* dedicated annotations are generated additionally. After the pre-processing and alignment computations in the first step of the workflow, genetic markers will be identified automatically in the second phase. Optionally, the user can adjust the trade-off between the stringency and coverage of the marker identification by specifying a custom DNA sequencing depth of coverage. For the QTL analyses in the third and final step, the user can adjust the type and width of the used smoothing kernel and has the option to download intermediate results, such as allele frequency files, bam files or map files, for further independent analyses. The QTL peaks, QTL regions and corresponding empirically estimated p-values are determined using the G’ statistic [3]. To facilitate the results interpretation, BSA4Yeast computes various dedicated statistics, such as the allele counts on each chromosome and a summary of the type of mutations in each parental line (e.g. stop gain, stop loss, frameshift or non-synonymous mutations), as well as SNAP scores to evaluate deleteriousness [5]. Additionally, comprehensive annotations for the QTLs and the genome of the parental lines are provided, and all results can be downloaded from the website. The software does not require any registration, but users can optionally create an anonymous account to store results (8 GB) for a longer time to conduct further analyses with different parameters. Overall, the software is designed to enable scientists with limited background knowledge in bioinformatics to run all analysis steps with minimal manual effort, only needing to provide adequately formatted input DNA sequencing files (bam or map files) through a web-browser, and avoiding time-consuming installation and configuration steps on the local computer.

Implementation

The BSA4Yeast web-application has been developed in Python 2.7 using the Flask micro-framework (Fig. 3) [6]. Flask is an extensible web micro-framework, written in Python and therefore fully compatible with the bsa-seq package used for QTL calculations [3] (implemented in Python 2.7). All analyses run as flask asynchronous background tasks using Celery, a task queue/job queue system based on distributed messaging [7], and Redis, an open source (BSD license) message broker between the web-application and the celery worker (Fig. 3).
When a job is complete or users decide via an email message. Moreover, to avoid blocking of the main framework was applied successfully to investigate cellular [10].

Representative runtimes for the example input files are -1 min, ~1.5 hours and ~3.5 hours for .map, .bam and .fastq files. Example parameter settings for fastq analysis are shown in Table 2 (further example settings for other file types are provided on the BSA4Yeast website).

Discussion and possible future extensions

Bulk Segregant Analysis (BSA) based QTL mapping using next generation sequencing technologies is a valuable new approach to identify genes associated with a phenotype of interest. However, the complexity of the software tools, parameter settings and the underlying algorithms used to process the data may prevent a wider application of the computational methods developed for this purpose. Moreover, setting up a comprehensive and efficient analysis pipeline is a laborious, error-prone and time-consuming task, which requires prior experience in bioinformatics. To address these problems, speed up and greatly facilitate bulk segregant analyses for yeast DNA sequencing data, BSA4Yeast was developed as a dedicated web-application for BSA-QTL analysis. Instead of the conventional approach for QTL mapping, which investigates the allele frequency distribution across the chromosomes, BSA4Yeast uses a variant of the G'-statistic [3], which provides multiple advantages over classical allele frequency analyses. Firstly, the G'-statistic is expected to decrease more rapidly around the causal site, providing narrow QTL candidate intervals; and secondly, the G'-statistic takes into account the strength of the evidence, which is estimated using the sample size. However, certain characteristics of the G'-statistic can also complicate analyses, e.g. the variance in read depth strongly influences the variance of the G'-statistic over small spatial scales. The G'-statistic, a smooth version of the G-statistic, previously developed by Paul Magwene et al. [3], is designed to address this limitation and provides a robust framework to analyze BSA sequencing data. It is computed in an automated fashion within BSA4Yeast and has been employed successfully for several biological applications, e.g. to identify genes involved in yeast biofilm formation or chronological aging [2, 16].

Example application

In a first proof-of-concept study, the BSA4Yeast analysis framework was applied successfully to investigate cellular aging in baker’s yeast (S. cerevisiae), detecting two significant QTLs associated with chronological life span regulation [2]. Specifically, a DNA sequencing dataset consisting of paired end (PE) sequenced parental strains and the single end (SE) sequenced segregant bulks was investigated with the software. The web-application was applied on three types of input data from yeast BSA-based QTL studies, representing different experimental designs, and two different types of sequencing methods (PE and SE). Summary statistics for the three input files used for the example analyses are shown in Table 1.

The web-application displays the QTL region annotation, the G’ statistics for each chromosome and the summarized mutation types for the two parental lines. Since the parental DNA data is not available when using map files as input, only the QTL coordinates can be obtained for this input type, whereas the full annotations are generated for bam-file analyses. All of the example datasets are available on the BSA4Yeast server for downloading and testing (https://bsa4yeast.lcsb.uni.lu/).
prior domain knowledge in bioinformatics. Moreover, web-based workflow implementations do not only have advantages over classical software package installations in terms of the simplicity of usage, but further benefits arise from the platform-independence of the software (BSA4Yeast runs on any operating system that supports modern web-browsers) and the fully reproducible analyses, independent of software updates on the client’s computer. Finally, the optional password-protected access to a user account enables users to access data and results from different locations and to share them with trusted collaborators.

Since the BSA4Yeast workflow is implemented in a modular fashion, it can be extended and adjusted, e.g. to cover further annotations and reference genomes for other model organisms used to perform linkage QTL studies (e.g. fruit flies or mice). Moreover, the software can be interlinked with other public internet databases and repositories, which contain further information on identified genes with a phenotype association of interest. The BSA4Yeast source code has been made available on GitLab (https://git-r3lab.uni.lu/zhi.zhang/bsa4yeast) to allow other users to explore, modify or further extend the software.

Summary

In summary, BSA4Yeast was designed and implemented in order to enable users to perform comprehensive NGS-BSA studies efficiently without requiring prior bioinformatics knowledge. The overall software workflow (Fig. 1) covers the following data processing, quality control and analysis tasks:

i. Mapping short reads (PE or SE, Illumina format) to the standard Saccharomyces cerevisiae reference genome (UCSC release SacCer3; BWA, version 0.7.12) [17] and generating alignment files (.bam format) for both parental and bulk samples;

ii. Defining genetic markers between two parental lines given a user-defined coverage threshold (default: 5 ×);

iii. Calculating G’ statistical values [3] for each genetic marker in the bulk pools;

iv. Annotating exonic variants between the two parental lines and within each QTL region with ANNOVAR (version: Mon, 1 Feb 2016) [18];

v. Scoring the functional impact of non-synonymous variants with SNAP (version: 2.0) [5].
Table 2. Input parameters for analysis of fastq files.

| Parameters                     | Value                           |
|--------------------------------|---------------------------------|
| Input file type                | Fastq                           |
| Number of biological replicates| 1                               |
| Title for the result           | myresult                        |
| One- or two-tailed bulk design | 2                               |
| The P1 Fastq file              | ['sake_strain_P1_1.fastq', 'sake_strain_P1_2.fastq'] |
| The P2 Fastq file              | ['white_tec_p1_strain_P2_1.fastq', 'white_tec_p1_strain_P2_2.fastq'] |
| The Bulk H.fastq file          | ['Bulk H.fastq']                 |
| The Bulk L.fastq file          | ['Bulk L.fastq']                 |
| The number of the depth of coverage | 10                             |
| The type of the smoothing kernel | tricube                         |
| The width of the smoothing kernel (bp) | 33750                          |
| The chromosome number to draw  | all                             |
| Whether to draw raw G’ values or not | no                             |

The final results generated by the software, including the annotations, variant functional impact predictions and statistical results for each determined QTL region, can be explored interactively and downloaded using current standard web-browsers (tested on Chrome, Firefox and Safari).

Availability of source code and requirements

- Project name: BSA4Yeast
- Project home page: https://git-r3lab.uni.lu/zhi.zhang/bsa4yeast
- Operating system(s): CentOS 7.2
- Programming language: Python 2.7
- Other requirements: Flask 0.12.2, Celery 4.1.0, Redis 2.10.6
- License: GNU GPL

Availability of supporting data and materials

The used example datasets are available at https://bsa4yeast.lcsb.uni.lu.

Declarations

List of abbreviations

- BSA: Bulk Segregant Analysis
- DOE: Design of Experiment
- NGS: Next Generation Sequencing

Figure 5. The analysis result of the fastq test file. a) A partial annotation on QTL regions. b) A summary of the G’ values for each chromosome, c) A summary type of mutations from the two parental lines.
• PE: Paired-end
• QTL: Quantitative Trait Locus
• SE: Single-end
• VM: Virtual Machine
• web-app: web application
• WSGI: Web Server Gateway Interface

Consent for publication
Not applicable

Competing Interests
The authors declare(s) that they have no competing interests.

Funding
Acknowledgement is made for support by the Fonds Nationale de la Recherche (FNR) Luxembourg, through the National Centre of Excellence in Research (NCER) on Parkinson’s disease (1R–BIC–PFN–15NCER), and as part of the grant project MiRisk (C17/BM/11676395).

Author’s Contributions
ZZ, EG designed the project. ZZ wrote the code of project. ZZ, PJ, EG wrote the manuscript. VG contributed his knowledge to the infrastructure and deployment. PM provided the SNAP2.0 value for mutations. PJ, CL contributed the sequencing data sets for testing. All authors read the manuscript and provided feedback.

Acknowledgements
We thank Yohan Jarosz, Roland Krause and Ursula Heins-Marroquin for advice and helpful discussions, Venkata Satagopam for the calculated SNAP scores and Paul Magwene for kindly providing source code used for the BSA implementation. The bioinformatics analyses presented in this paper were partly conducted using the HPC facilities at the University of Luxembourg (see http://hpc.uni.lu).

References
1. Wilkening S, Lin G, Fritsch ES, Tekkedil MM, Anders S, Kuehn R, et al. An evaluation of high-throughput approaches to QTL mapping in Saccharomyces cerevisiae. Genetics 2014 Mar;196:853–865.
2. Jung PP, Zhang Z, Paczia N, Jaeger C, Ignac T, May P, et al. Natural variation of chronological aging in the Saccharomyces cerevisiae species reveals diet-dependent mechanisms of life span control. NPJ Aging and Mechanisms of Disease 2018;4:3.
3. Magwene PM, Willis JH, Kelly JK. The statistics of bulk segregant analysis using next generation sequencing. PLoS Computational Biology 2011 Nov;7:e1002255.
4. Pulido-Tamayo S, Duitama J, Marchal K. EXPloRA-web: linkage analysis of quantitative trait loci using bulk segregant analysis. Nucleic Acids Research 2016 Jul;44:W142–W146.
5. Bromberg Y, Rost B. SNAP: predict effect of non-synonymous polymorphisms on function. Nucleic Acids Research 2007;35:3823–3835.
6. Grinberg M. Flask web development. O’Reilly Media; 2018.
The result_P1_P2_combined_exonic_variant_function.txt table annotation is:

| chr | coord | ref | P1_alt | P2_alt | P1_variant_type | gene_id_P1 | exon_P1 | nt_P1 | protein_P1 | P1_SNAP | gene_id_link_P1 |
|-----|------|-----|--------|--------|-----------------|------------|---------|------|------------|---------|-----------------|
| 1   | 349  | C   | T      |        | synonymous SNV  | YAL068W    | exom     | c.215T | p.N85N      |         | YAL068W         |
| 1   | 358  | G   | A      |        | synonymous SNV  | YAL068W    | exom     | c.204A | p.W68V      |         | YAL068W         |
|     | 381  |     |        |        |                 |            |          |       |            |         |                 |
| 1   | 424  | GCA | A      |        | nonframedhift insertion | YAL068W | exon | c.903delGCA | p.T300delTA |         | YAL068W         |
|     | 475  | AGTT| AGTT   |        |                 |            |          |       |            |         |                 |
| 1   | 480  | C   | C      |        |                 |            |          |       |            |         |                 |
| 1   | 509  | G   | G      |        |                 |            |          |       |            |         |                 |
| 1   | 519  | C   | T      |        |                 |            |          |       |            |         |                 |
| 1   | 595  | C   | G      |        | nonynymous SNV  | YAL068W-A / YAL068W | exon | c.580G>c.261G | p.H320> p.I87M |         | YAL068W         |
| 1   | 623  | C   | C      |        |                 |            |          |       |            |         |                 |
| 1   | 650  | C   | C      |        |                 |            |          |       |            |         |                 |
| 1   | 653  | C   | C      |        |                 |            |          |       |            |         |                 |
| 1   | 658  | A   | A      |        |                 |            |          |       |            |         |                 |
| 1   | 682  | C   | C      |        |                 |            |          |       |            |         |                 |
| 1   | 683  | G   | G      |        |                 |            |          |       |            |         |                 |
| 1   | 703  | T   | T      |        |                 |            |          |       |            |         |                 |
| 1   | 709  | T   | T      |        |                 |            |          |       |            |         |                 |
| 1   | 738  | C   | A      |        | synonymous SNV  | YAL068W-A | exon | c.220A | p.S76L      |         | YAL068W-A         |
| 1   | 760  | C   | A      |        | nonynymous SNV  | YAL068W-A |
| 1   | 743  | A   | T      |        | nonynymous SNV  | YAL068W-A |
| 1   | 759  | T   | T      |        |                 |            |          |       |            |         |                 |
| 1   | 768  | T   | A      |        |                 |            |          |       |            |         |                 |
| 1   | 772  | C   | C      |        |                 |            |          |       |            |         |                 |

Showing 1 to 25 of 54,616 entries

---

G' Chromosome

---

(c) (d)
