

Title

Full title:

RecombineX: a generalized computational framework for automatic high-throughput gamete genotyping and tetrad-based recombination analysis

Short title:

Tetrad analysis and meiotic recombination landscape

Jing Li^{1,2}, Bertrand Llorente³, Gianni Liti^{2*}, Jia-Xing Yue^{1,2*}

1. State Key Laboratory of Oncology in South China, Collaborative Innovation Center for Cancer Medicine, Sun Yat-sen University Cancer Center, Guangzhou, China.
2. Université Côte d'Azur, CNRS, INSERM, IRCAN, Nice, France.
3. Aix-Marseille Université, CNRS, INSERM, CRCM, Institut Paoli-Calmettes, Marseille, France

Corresponding author:

Gianni Liti < gianni.liti@unice.fr >

Jia-Xing Yue < yuejiaxing@gmail.com >

ORCID:

Jing Li: 0000-0001-5697-9527

Bertrand Llorente:

Gianni Liti: 0000-0002-2318-0775

Jia-Xing Yue: 0000-0002-2122-9221

Running title: tetrad analysis and meiotic recombination landscape

Abstract

Meiotic recombination is an essential biological process that ensures faithful chromosome segregation and promotes parental allele shuffling. Tetrad analysis is a powerful approach to quantify the genetic makeups and recombination landscapes of meiotic products. Here we present RecombineX (<https://github.com/yjx1217/RecombineX>), a generalized computational framework that automates the full workflow of marker identification, gamete genotyping, and tetrad-based recombination profiling based on any organisms and genetic backgrounds with batch processing capability. Aside from conventional reference-based analysis, RecombineX can also perform analysis based on parental genome assemblies, which enables analyzing meiotic recombination landscapes in their native genomic contexts. Additional features such as copy number variation profiling and missing genotype inference further enhance downstream analysis. RecombineX also includes a dedicated module for simulating the genomes and reads of recombinant tetrads, which enables fine-tuned simulation-based hypothesis testing. This simulation module revealed the power and accuracy of RecombineX even when analyzing tetrads with very low sequencing depths (e.g., 1-2X). Tetrad sequencing data from the budding yeast *Saccharomyces cerevisiae* and green alga *Chlamydomonas reinhardtii* were further used to demonstrate the accuracy and robustness of RecombineX for organisms with both small and large genomes, manifesting RecombineX as an all-around one stop solution for future tetrad analysis.

Author Summary

Meiosis is a fundamental cellular process that ensures faithful chromosome segregation and promotes allele shuffling. Tetrad analysis, which isolates and genotypes all four meiotic products (i.e., tetrad) derived from a single meiosis, remains the most straightforward and powerful way of studying meiotic recombination and its modulators at fine scales. The wide application of tetrad analysis in yeasts, filamentous fungi, green algae, and land plants have substantially expand our understanding of meiotic recombination in terms of both genome-wide landscapes and molecular mechanisms. Here we described the first generalized computational framework named RecombineX that automates the full workflow of tetrad analysis based on any organisms and genetic backgrounds. In addition, aside from conventional reference-based analysis, RecombineX can also perform analysis based on parental genome assemblies, which enables analyzing meiotic recombination landscapes in their native genomic contexts. Using both simulated and real tetrad-sequencing data, we further demonstrated RecombineX's trustable performance, versatile usage, and batch-processing capability, manifesting RecombineX as an all-around one stop solution for tetrad analysis. Especially considering that meiotic gamete genome sequencing from different natural and mutant backgrounds can now be acquired, we expect RecombineX to become a popular tool that empowers future tetrad analysis across different genetic backgrounds and species.

Introduction

Meiosis is a fundamental cellular process in eukaryotes, through which sexually reproducing organisms generate their gametes via two successive rounds of cell division. In the first round (meiosis I), homologous chromosomes duplicate, pair and swap genetic materials, and then segregate into two daughter cells. In the next round (meiosis II), the two sets of sister chromatids in each daughter cell further separate into different gamete cells to reduce the total chromosome number by half. The four gamete cells resulting from these two rounds of cell division are collectively referred as a tetrad. In most species, accurate homologous chromosome segregation at meiosis I relies on sister chromatid cohesion in combination with meiotic crossovers (CO) that are reciprocal exchanges of chromosome arms. These meiotic COs result from the repair by homologous recombination of meiotic prophase-induced DNA double strand breaks (DSBs). In addition to COs, DSB repair by meiotic recombination also produces recombinants without reciprocal exchange of chromosome arms called non-crossovers (NCOs). Both COs and NCOs are intrinsically associated with a tract of gene conversion (GC), which detection relies on suitably positioned markers. Both COs and NCOs shuffle parental genetic materials, which promotes the genetic robustness and phenotypic potential of the offspring gene pool.

Given the vital role of meiotic recombination, different methodologies have been developed to characterize its underlying mechanisms and evolutionary implications. For example, meiotic recombination can be indirectly analyzed by examining linkage disequilibrium and haplotype structure with population genomics data (Ptak et al. 2005; Coop et al. 2008; Rasmussen et al. 2014; Spence and Song 2019). While powerful statistical inferences regarding recombination can be made this way with existing genomic data, additional factors such as demographic histories and selection schemes might perplex the result interpretation. In contrast, meiotic recombination can also be studied by directly examining the makeup of gamete genotypes in terms of parental genetic backgrounds. One way of doing this is to perform bulk genotyping analysis for random gametes (Wang et al. 2012, 2012; Hou et al. 2013; Kirkness et al. 2013; Hinch et al. 2019). Although this approach allows for detailed delineation of cumulative recombination landscapes across a large number of gametes, it lacks the power and resolution for inspecting individual meiosis event, which prevents an in-depth view of the meiotic recombination process. Alternatively, at least for a selection of model systems, it is feasible to isolate and genotype all four meiotic products (i.e., tetrad) derived from a single meiosis. This approach is called “tetrad analysis”, which remains the most straightforward and powerful way of studying meiotic recombination and its modulators at fine scales. For instance, a landmark study of this kind was performed on the budding yeast *Saccharomyces cerevisiae*, which led to the first high-resolution meiotic recombination map for eukaryotes (Mancera et al. 2008). Thereafter, similar tetrad-

based genome analysis have been carried out across multiple organisms and genetic backgrounds (including mutants) (Qi et al. 2009; Martini et al. 2011; Lu et al. 2012; Wijnker et al. 2013; Li et al. 2015; Brion et al. 2017; Liu et al. 2018; Marsolier-Kergoat et al. 2018; Liu et al. 2019), which altogether substantially advanced our understanding of meiotic recombination and its genetic modulators.

In contrast to the broad application of tetrad analysis, there is a lack of dedicated computational framework for corresponding data analysis. To our knowledge, ReCombine (Anderson et al. 2011) is the only tool developed for such purpose so far. ReCombine represents an important step towards automated and standardized tetrad analysis, but it was designed in the early days of next-generation sequencing and understandably appears somewhat constrained to cover today's use scenarios in terms of functionality, versatility, and customizability. For example, ReCombine is hardcoded based on the *S. cerevisiae* reference genome and expects the reference-based S288C strain to be one of the two crossing parents. Also, manual configuration and curation are normally needed on a tetrad-by-tetrad basis, making ReCombine less suitable for processing large numbers of tetrads.

Therefore, a new generation of computational solution for high-throughput and high-quality tetrad analysis is much needed. Here we introduce RecombineX, a generalized computational framework that automates the full workflow of gamete sequencing data analysis, especially for organisms whose tetrad can be isolated. Equipped with dedicated modules for polymorphic markers identification, gametes genotyping, recombination profiling, and recombinant tetrad simulation, RecombineX shines in its trustable performance, versatile usage, and batch-processing capability. Our tests based on both simulated and real data further demonstrated its consistent power and accuracy in multiple application scenarios, manifesting RecombineX as an all-around one stop solution for future tetrad analyses.

Results and Discussion

The general design of RecombineX

RecombineX is a Linux-based computational framework designed for automated high-throughput tetrad analysis. It is self-contained by design and can be automatically installed and configured via a pre-shipped installer script. RecombineX comes with a series of task-specific modules handling different workflow phases: genomes and reads preparation -> parental markers identification -> gametes reads mapping and genotyping -> tetrad-based recombination events profiling (Figure 1). Depending on the available input data, RecombineX can be executed in two modes: 1) the reference-based mode and 2) the parent-based mode. For each mode, we numbered the corresponding modules based on their execution orders, ensuring a well-organized data analysis workflow (Figure 2). Within each module, a task-specific executable bash script is provided for invoking the corresponding

module. A template directory with these modules pre-configured is further provided as a testing example, which can be easily adapted for users' own project (Figure S1).

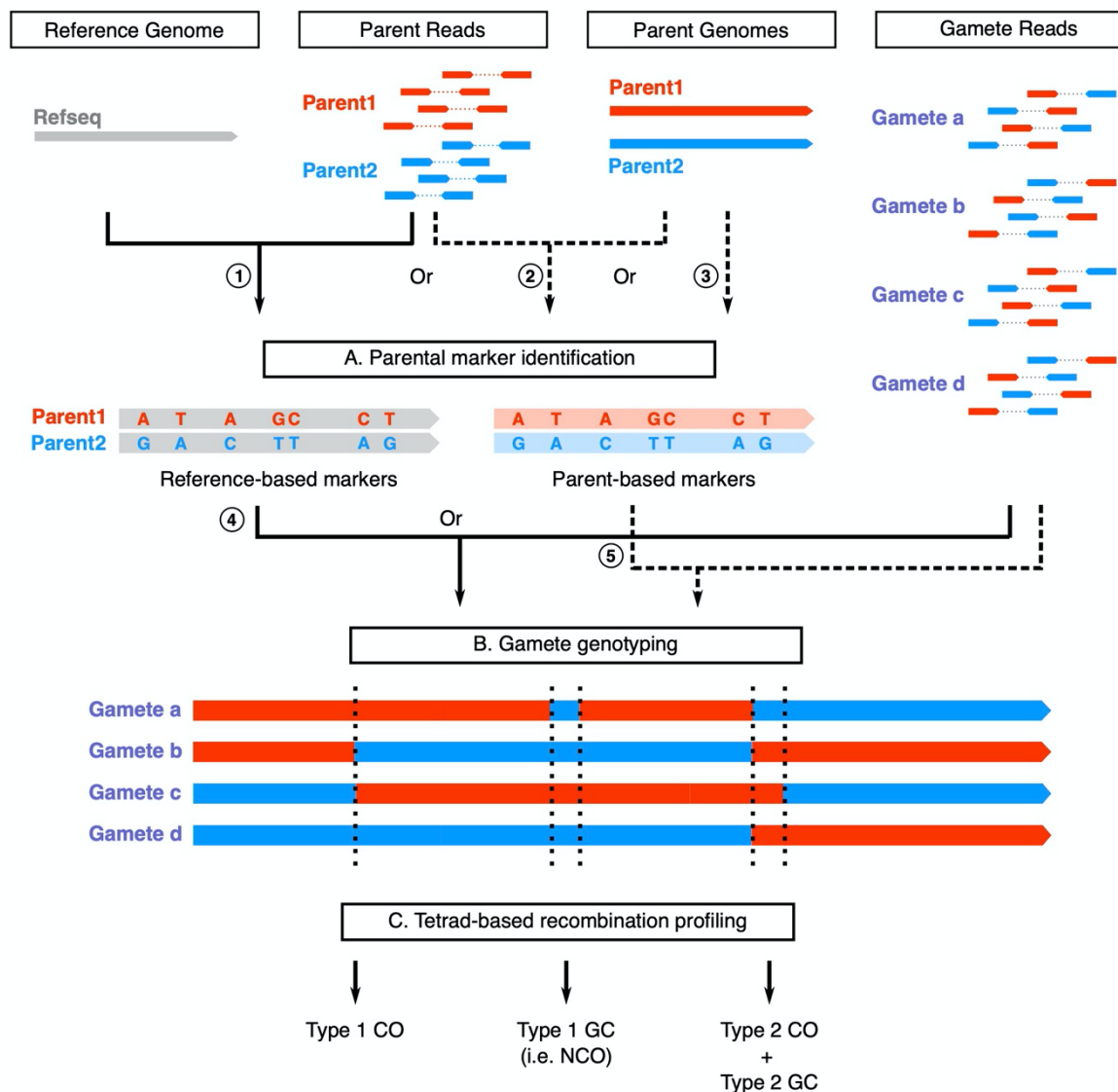


Figure 1. An overview of the RecombineX framework. RecombineX conduct sequencing-based tetrad analysis in three major phases: A) parental marker identification, B) tetrad genotyping, and C) recombination event profiling. Depending on the available input data, users can run RecombineX in either reference-based mode (denoted by solid arrows) or parent-based mode (denoted by dashed arrows). In the reference-based mode, parent reads are mapped to the reference genome for reference-based parental marker identification (①), based on which gamete genotyping is further performed by evaluating the gamete-to-reference read mapping support at each marker position (④). The resulting genotyping assignments across the four gametes from the same tetrad are jointly evaluated for profiling recombination events based on the reference genome coordinate system. In the parent-based mode, whole genome alignment is firstly constructed based on the native genome assemblies of the two crossing parents, upon which parent-based markers are identified accordingly (③). Optionally,

parent-based markers obtained from whole genome alignment can be further leveraged by reciprocal parent-based read mapping (②). In either case, gamete genotyping is performed by evaluating the gamete-to-parent read-mapping at each marker position(⑤). The resulting genotyping assignments across the four gametes from the same tetrad are jointly evaluated for profiling recombination events based on the coordinate systems of the two parental genome assemblies.

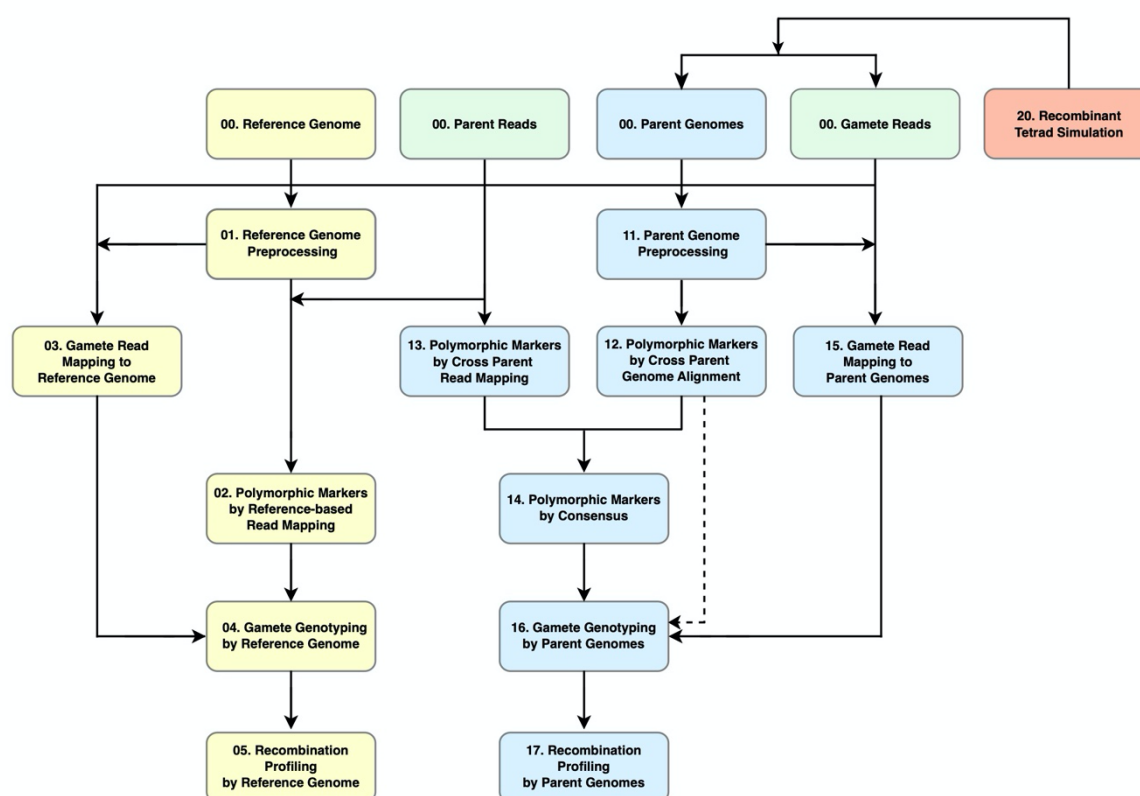


Figure 2. The modular workflow design of RecombineX. RecombineX consists of seventeen task-specific modules, with six modules dedicated for the reference-based mode (colored in yellow) and eight modules dedicated for the parent-based mode (colored in blue). As for the three remaining modules, two are designed for both reference-based and parent-based modes (colored in green), with the last one for simulation analysis (colored in red).

In the reference-based mode, the sequencing reads of the two crossing parents are mapped to the reference genome to identify single nucleotide variants (SNVs) between the two parental backgrounds (Figure S2). The gamete reads are subsequently mapped to the reference genome, upon which RecombineX computes the best supported genotype at each marker position for each gamete (Figure S3). A genotype purity filter is further employed at this step to cull out markers with clear admixed genotype signals. Such admixed genotype signals normally come from ambiguous mapping and therefore it is reasonable to filter them out in normal tetrad analysis. However, admixed genotype signals can also reveal post-meiotic segregation (PMS) of unrepaired

heteroduplex DNA (mismatches) formed during recombination which frequency massively increases after inactivation of the mismatch repair machinery. PMS is important for dissecting the detailed molecular mechanisms of recombination. PMS is directly detectable in filamentous fungi that naturally form octads or after micromanipulation of yeast tetrads (Martini et al. 2011; Yeadon et al. 2016; Marsolier-Kergoat et al. 2018). To allow PMS identification, we added an option that enables RecombineX to report all marker sites with admixed genotype signals. Also, when needed, RecombineX can infer marker-specific missing genotypes by assuming a tetrad-wide 2:2 segregation ratio between the two parental backgrounds, which could come handy to recover genotypes that are otherwise inaccessible. By jointly analyzing genotype segregation and switching patterns within the same tetrad across different marker positions, RecombineX further identifies and classifies recombination events into different categories. In general, profiled recombination events are expected to fall in four major categories: CO without associated GC (referred as Type 1 CO thereafter), NCO (referred as Type 1 GC thereafter), CO with associated GC (referred as Type 2 CO thereafter), and the Type 2 CO associated GC (referred as Type 2 GC thereafter), although more complex cases could be encountered occasionally (Figure S4 and S5). It is worth pointing out that a CO event always associates with a GC tract and therefore Type 1 and Type 2 CO events are biologically equivalent. It is due to lacking available markers that makes the CO-associated GC tracts undetectable in practice. For all profiled recombination events, RecombineX generates detailed reports on their genomic coordinates, marker supports, genotype segregation patterns, and the associated linkage blocks for downstream analysis.

In addition to the reference-based tetrad analysis described above, RecombineX also supports performing tetrad analysis directly based on the genome assemblies of the two crossing parents, which could be especially valuable for analyzing recombination events in their native parental genome contexts. In this parent-based mode, RecombineX builds the whole genome alignment of the two parental genome assemblies and identifies parental markers accordingly (Figure S2). When available, sequencing reads of the two parents can be further provided to derive consensus markers by further leveraging cross-parent read mapping results (Figure S2). Upon the obtaining of parental markers, the downstream analyses such as gametes reads mapping, gametes genotyping, and tetrad-based recombination profiling are performed based on both parental genome coordinate systems (Figure S3), with the corresponding results also reported in two mirrored copies. In this way, users can easily check for association between identified recombination events and various parental genome features (e.g., gene densities, repetitive sequence abundance, GC% contents, parental divergence, DSB hotspots, etc.) based on the same genome coordinate system.

Simulation-based validation for RecombineX's parental marker identification modules

Accurate and robust parental markers identification is a prerequisite for high-quality tetrad analysis. Several studies demonstrated that downstream recombination analysis can be severely compromised when relying on ambiguous markers, which are often derived from genomic regions associated with repetitive sequences or copy number variants (CNVs) (Wijnker et al. 2013; Qi et al. 2014). Therefore, with RecombineX, we designed and implemented multiple filters to effectively culling out markers falling in repetitive and CNV regions (Figure S2; See Materials and Methods for details). As a simulation-based test, we let RecombineX to identify markers segregated between two hypothetical parental genomes: the *S. cerevisiae* reference genome and a simulated *S. cerevisiae* genome with 60,000 SNVs, 6,000 INDELs, and 6 CNVs, mimicking a typical 0.5% genomic divergence between two *S. cerevisiae* natural isolates (Peter et al. 2018). Among the 60,000 simulated SNV sites, 8,087 sites fell in either repetitive or CNV regions, leaving the remaining 51,913 sites as valid targets for marker identification (Figure 3 and Table S1). Based on these two hypothetical parental genomes, we evaluated RecombineX's marker identification performance in both reference-based and parent-based modes. In reference-based mode, RecombineX solely relies on the input parent reads to identify markers. While the power of marker identification positively correlates with parent sequencing depth, it quickly enters diminishing return with sequencing depth >30X (Figure 3 and Table S1). Therefore, as a rule of thumb, we generally recommend using $\geq 30X$ parental reads for marker identification with RecombineX. Based on our simulation, RecombineX is able to recover >90% valid marker targets (47,268 out of 51,913) with 30X parent reads. A close examination of these markers shows no false positive calling was made and ambiguous SNV sites from repeat-/CNV-associated regions have been effectively filtered out, proving the high reliability of RecombineX's marker identification. The only CNV-associated SNV site that escaped from RecombineX's CNV-filter locates near the boundary of a simulated CNV with no detectable per-site mapping depth deviation from the chromosome-wide median. In parent-based mode, RecombineX identified 49,669 markers based on parental genome alignment alone. In comparison, less consensus markers were identified due to the additional filter applied based on reciprocal parent read mapping. The final consensus marker count scales with the sequencing depth of parent reads, while the improvement become marginal with sequencing depth > 50X. Although RecombineX has implemented a unique-alignment-based CNV filter for genome-alignment-based marker identification, we found 38 out of 8,087 CNV-associated markers escaped from this filter. Nearly all of them (37 out of 38) were further filtered out when calling consensus markers, during which an additional mapping-depth-based CNV filter is applied. Therefore, the consensus marker identification protocol appears more robust against ambiguous markers from CNV regions. The only marker that escaped from both CNV filters is the same one as mentioned above, which shows very weak CNV signal in both reference-based and parent-based

modes. According to this result, we recommend opting for consensus marker identification strategy when running RecombineX in parent-based mode, if parent reads are available.

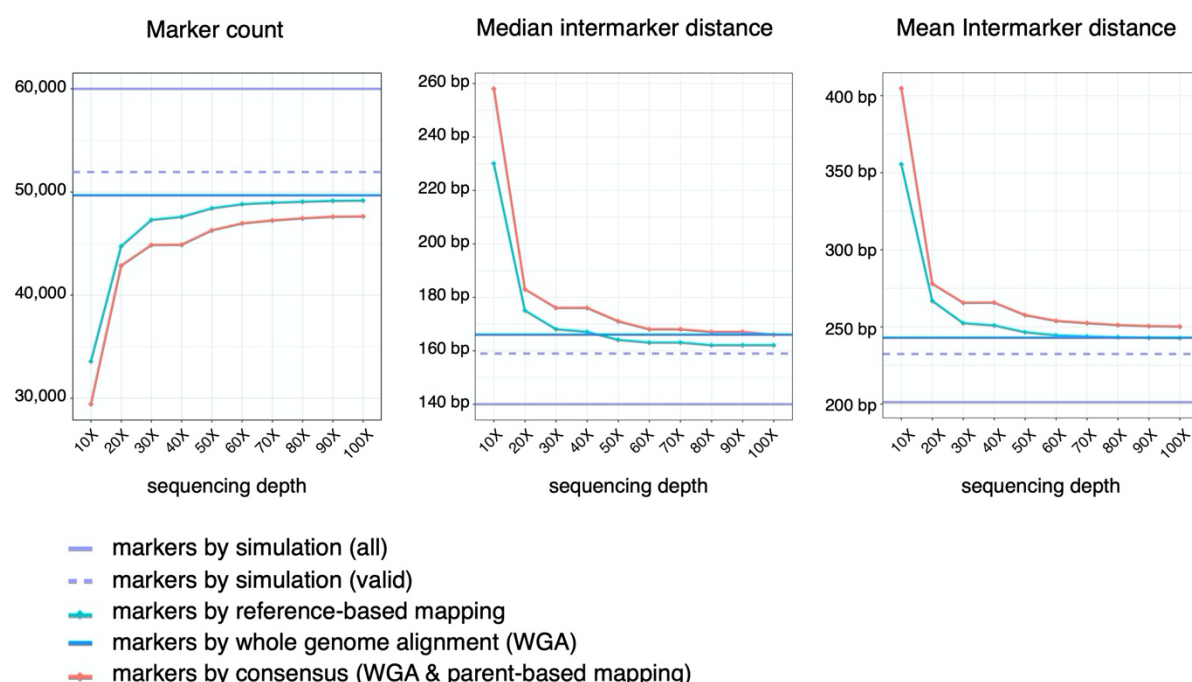


Figure 3. Performance of polymorphic marker identification with RecombineX. A total of 60,000 SNV markers (denoted as “all”) together with 6,000 Indels and 6 CNVs were simulated for two hypothetical parent genomes, among which 51,913 of them are considered as valid marker discovery targets as they are not associated with repetitive regions nor CNV regions. We gauged RecombineX’s performance for marker identification based on these two hypothetical parent genomes and their reads (simulated sequencing coverage: 10X, 20X, 30X, ..., 100X) using different marker identification protocols implemented in RecombineX: reference-based mapping (green), whole genome alignment (blue), and consensus between whole genome alignment and parent-based read mapping (red). For each case, the total marker counts as well as the median and mean inter-marker distance were plotted respectively.

Simulation-based validation for RecombineX’s gamete genotyping module

To assess RecombineX’s gamete genotyping accuracy, we used RecombineX’s built-in simulation module to simulate one recombinant tetrad derived from two hypothetical crossing parents: P1 and P2. The inputs of this simulation module include a reference or parent genome to set up the coordinate system and a list of parental markers projected to the same coordinate system. Here we used the *S. cerevisiae* reference genome as the coordinate system and the above identified 49,153 reference-based markers based on 100X parent reads (Figure 3 and Table S1). For the simulated recombinant tetrads, we introduced CO and NCO events based on the count and size parameters estimated from real yeast

tetrads (Mancera et al. 2008) (Figure 4A). For the four resulting recombinant gametes, sequencing reads of varied depths (1X, 2X, 4X, 8X, 16X, 32X, 64X) were further simulated. With this simulated dataset, we evaluated RecombineX's genotyping performance for mimicked scenarios of both fully and partially viable tetrads. For the scenario of fully viable tetrad, the simulated reads of all four gametes were used for genotyping (Figure 4B). For the scenario of partially viable tetrad, the simulated reads of only 3 gametes were used, leaving the remaining one to be inferred by RecombineX (Figure 4C and 4D). RecombineX infers missing genotypes by assuming a tetrad-wide 2:2 segregation ratio between the two parental genotypes across the whole genome. While such ratio can deviate from 2:2 in genomic regions with GC tracts, the cumulative size of all GC tracts is typically orders of magnitude smaller in comparison to the genome size, making this assumption holds in general. Finally, it is worth emphasizing that although a pre-specified gamete-tetrad correspondence enables extra features such as missing genotyping inference, RecombineX's raw genotyping function do not require any tetrad information as the *priori* when performing raw genotyping. Therefore, users can also use RecombineX to perform plain genotyping analysis for random gametes derived from known parents. Moreover, additional tools are available for reconstructing the gamete-tetrad correspondence map from random gametes (Sakhanenko et al. 2019), which could be used in combination with RecombineX when processing sequencing data from randomly collected gametes.

For the scenario of fully viable tetrad, while the power of genotyping positively correlates with tetrad sequencing depth, a high accuracy is consistently maintained even with very limited tetrad sequencing data available. For example, the upper bound of RecombineX's raw and inferred genotyping error rates are estimated as 7×10^{-4} and 2.64×10^{-3} respectively for 1X-sequenced tetrad (Figure 4E and Table S2). Aside from sequencing depth, Q_{net} cutoff is another important parameter for RecombineX's genotyping analysis, which specifies the cumulative genotyping-supporting score leveraged over all mapped reads at a given marker site. While a higher Q_{net} cutoff helps to filter out ambiguous genotyping signals, setting it too high will compromise RecombineX's genotyping power for shallowly sequenced tetrads, leaving the genotypes of many markers as undetermined (Figure 4E and Table S2). Therefore, in general, lenient Q_{net} cutoffs such as 10 or 20 are recommended for tetrad with very shallow sequencing depth (e.g., 1X). Our simulation shows that RecombineX is able to deliver highly accurate genotyping results even with such lenient Q_{net} cutoffs.

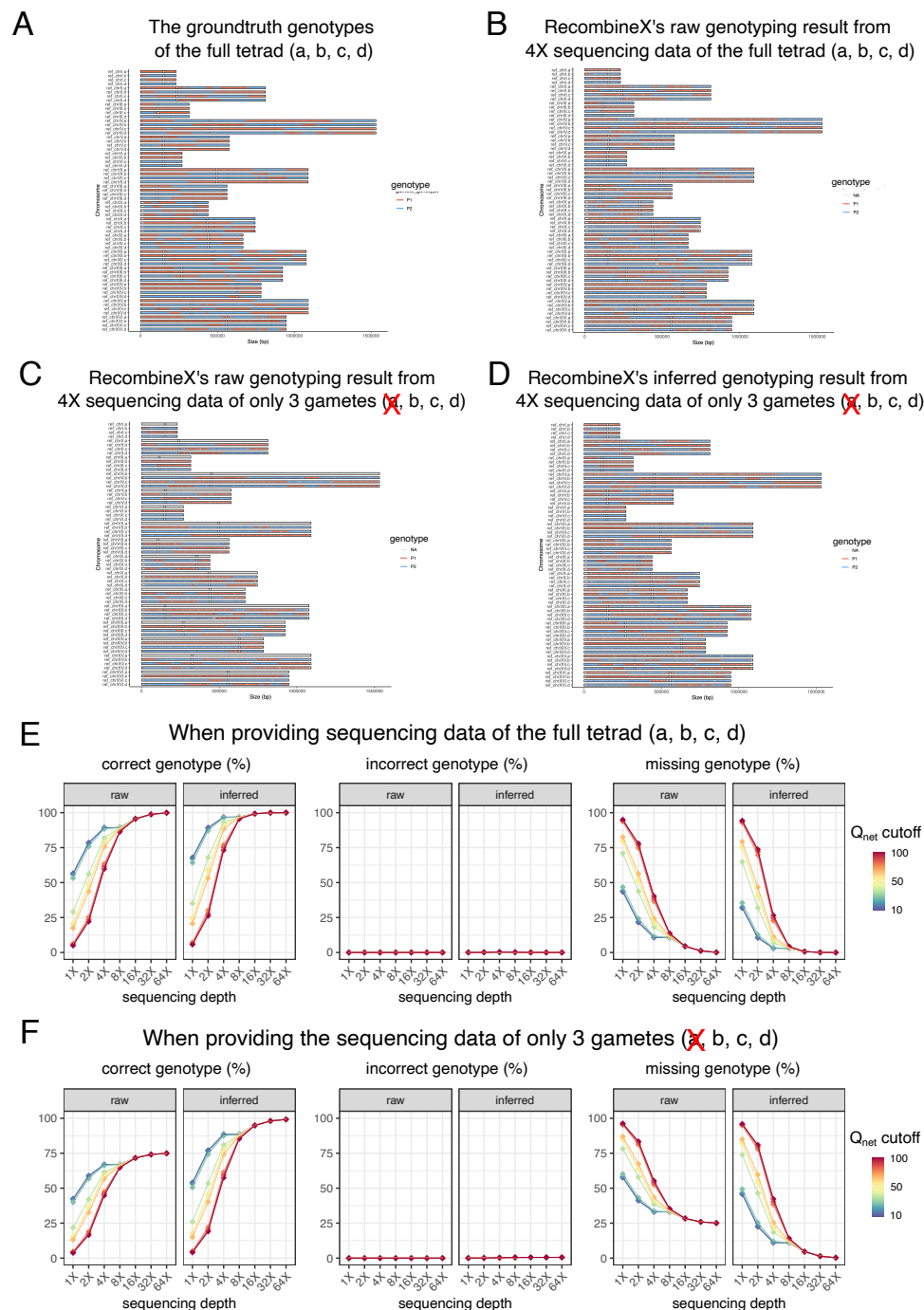


Figure 4. Performance of raw and inferred tetrad genotyping with RecombineX. (A) The ground truth genotypes of the simulated tetrad (a, b, c, d). (B) RecombineX's raw genotyping result for the simulated tetrad based on 4X sequencing data with a Q_{net} cutoff of 30. (C) The raw genotyping result for the simulated tetrad based on 4X sequencing data of only 3 gametes (b, c, d) with a Q_{net} cutoff of 30. (D) The inferred genotyping result for the simulated tetrad based on 4X sequencing data of only 3 gametes (b, c, d) with a Q_{net} cutoff of 30. (E) The calculated percentage of corrected, incorrect, and missing genotypes based on RecombineX's tetrad genotyping results in comparison to the simulated ground truth when the sequencing data of all four gametes (a, b, c, d) are provided. (F) The calculated

percentage of corrected, incorrect, and missing genotypes based on RecombineX's genotyping results in comparison to the simulated ground truth when the sequencing data of only three gametes (b, c, d) are provided. For E and F, the tested sequencing depths are 1X, 2X, 4X, 8X, 16X, 32X and 64X and the tested Q_{net} cutoffs are 10, 20, 30, 40, 50, 60, 70, 80, 90, 100.

For the scenario of partially viable tetrad, the error rate of our inferred tetrad genotypes ranges from 6×10^{-5} to 5.5×10^{-3} across different sequencing depth and Q_{net} cutoff combinations (Figure 4F and Table S3). As demonstrated in our simulation, RecombineX can almost completely recover the true genotypes of the mimicked inviable gamete with 4X sequencing reads from the other three viable gametes (Figure 4D and 4F), which highlights the power of such missing genotype inference. In terms of application value, the inferred missing genotypes from the inviable gametes can be potentially used to map the genetic basis of gamete lethality. Moreover, such missing genotype inference enables a more cost-effective design of trait-mapping experiments by making better use of shallowly sequenced samples. To better support these potential applications, RecombineX reports both raw and inferred genotyping results with rich graphical and textual outputs, making them highly amiable for further integration with genetic mapping tools such as R-qt1 (Broman et al. 2003).

Simulation-based validation for RecombineX's recombination profiling module

High quality genotyping performance of RecombineX lays the foundation for accurate recombination event identification and classification. Here we assessed RecombineX's recombination profiling performance with the genotyping results obtained with different gamete sequencing depths and Q_{net} cutoffs (Figure 5 and Table S4). Our simulated tetrad has been introduced with 90 COs (each with an associated GC tract) and 65 NCOs, which translates into 90 Type 2 COs, 65 Type 1 GCs, and 90 Type 2 GCs (Figure S4 and S5). It is worth mentioning that the power and accuracy of all genotype-based recombination analyses are ultimately linked to the density and distribution of available markers, which needs to be taken into consideration when interpreting the results of such analyses. For example, among our simulated recombination events, one CO event (which locates very close to the chromosome end), and five Type 2 GC events are inherently nondetectable due to the lack of available markers in the corresponding genomic regions. We excluded these undetectable events from our downstream analyses, which left a total of 89 COs, 65 Type 1 GCs, and 85 Type 2 GCs to be identified by RecombineX.

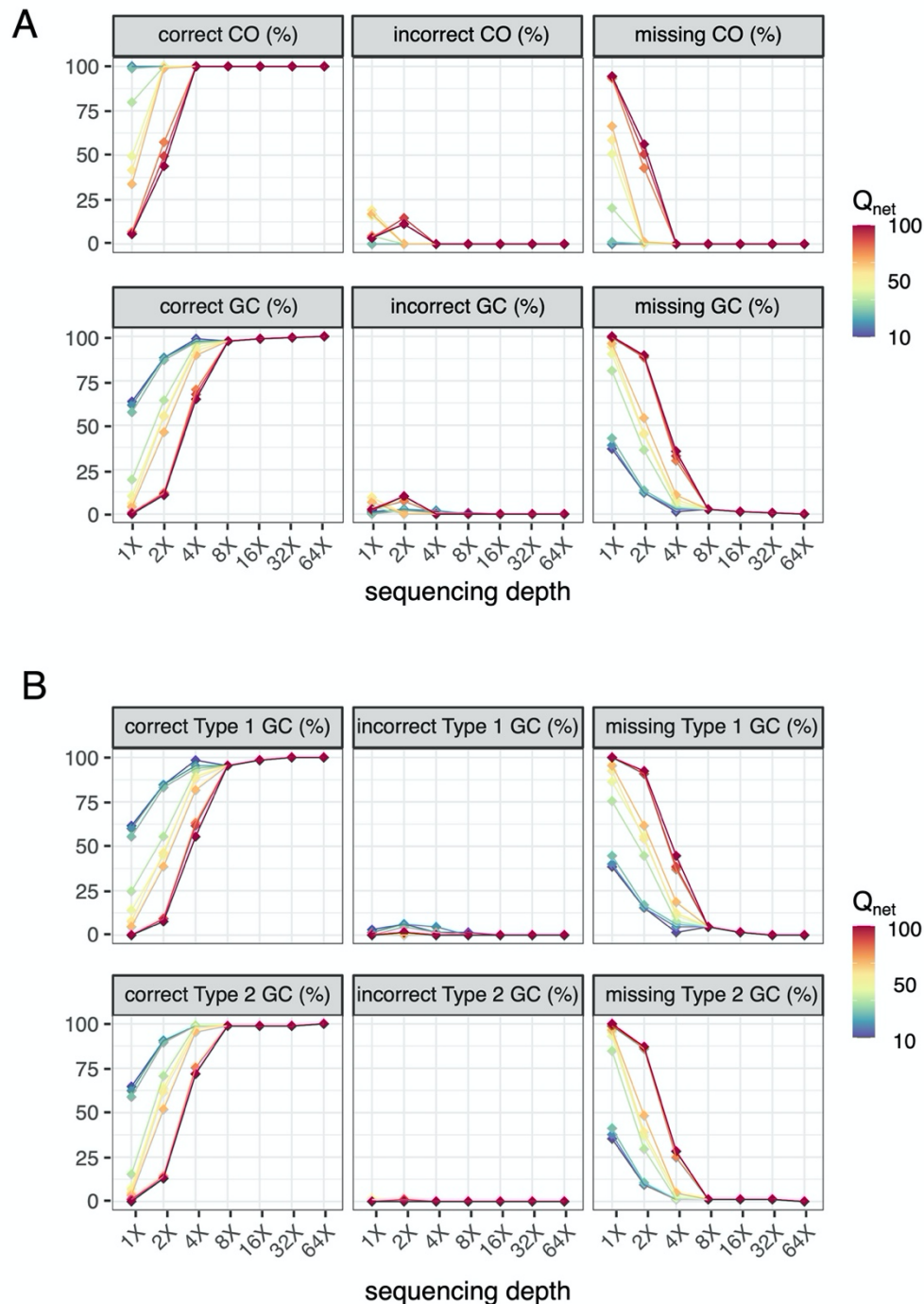


Figure 5. Performance of recombination profiling with RecombineX based on simulated data. The RecombineX-identified CO and GC events were compared with the simulated ground truth. (A) The percentage of corrected, incorrect, and missed total CO and GC events identified by RecombineX given different gamete sequencing coverage and Q_{net} cutoffs. (B) The percentage of corrected, incorrect, and missed Type 1 GC (NCO) and Type 2 GC (CO-associated with GC) events identified by RecombineX given different gamete sequencing coverage and Q_{net} cutoffs. Tested sequencing depth: 1X, 2X, 4X, 8X, 16X, 32X and 64X. Tested Q_{net} cutoff: 10, 20, 30, 40, 50, 60, 70, 80, 90, 100.

Echoing with what we found for genotyping, the performance of recombination profiling positively correlates with tetrad sequencing depth. With moderate sequencing depth (e.g., depth $\geq 8X$), RecombineX can recover all simulated CO and GC events, regardless of the specific Q_{net} cutoff value used for genotyping. For more shallowly sequenced tetrads, the performance of RecombineX's recombination profiling began to be compromised due to the reduction of available markers with strong-enough genotype signals. This is more evident for GC events than for CO events due to the much smaller genomic footprint of GC tracts. Therefore, lenient Q_{net} cutoffs such as 10 or 20 are recommended for recombination profiling on shallowly sequenced tetrads, which helps to maintain a relatively high sensitivity of event calling without severe compromise in specificity.

Applying RecombineX to real tetrad sequencing data

After systematically characterizing RecombineX's module-by-module performance with simulated data, we further applied it to real budding yeast (*Saccharomyces cerevisiae*) and green alga (*Chlamydomonas reinhardtii*) tetrads retrieved from previous studies for further demonstration (Callender et al. 2016; Liu et al. 2018). For the budding yeast example, the sampled tetrads are derived from a cross between S288C and YPS128 strains, for which the native genome assemblies of both parents are available (Yue et al. 2017). Therefore, we performed RecombineX analysis in both reference-based and parent-based modes for this case. As for the green alga example, the sampled tetrads were derived from a cross between CC408 and CC2936 ecotypes, for which no native parental genome assembly is available. Therefore, we only executed RecombineX in the reference-based mode here. For both examples, we compared all recombination events automatically profiled by RecombineX against the curated recombination event lists reported by the respective original studies.

For the yeast example, 50,199 reference-based markers (mean intermarker distance = 234.35 bp) and 48,558 parent-based consensus markers (mean intermarker distance = 240.64 bp) were identified respectively (Table S5). Our genotyping and recombination profiling analysis based on these markers shows a good concordance between RecombineX and the original study, with RecombineX that completely recovered almost all previously reported CO and GC events (Figure 6A and Table S6). There are a few events that were only called by RecombineX, which were further verified in our manual IGV inspection. As for those only called by the original study, we found most of them were filtered out by RecombineX at either marker identification or genotyping stage due to CNV-association or ambiguous genotypes. For instance, by default, RecombineX requires a minimal genotype purity of 90%, meaning at least 90% of the mapped reads should support the same genotype signal at the corresponding marker position. By this standard, some of event-defining markers (and therefore the corresponding event) from the original study will be disregarded by RecombineX (Figure S6). While we found such a strong genotype purity filter is generally beneficial for preventing

the inclusion of suspicious markers potentially derived from unreliable read mapping, users can easily adjust its stringency cutoff with RecombineX when needed.

For the green alga example, we identified 412,210 reference-based markers with an average intermarker distance of 260.28 bp (Table S5). Comparing to the yeast example, we noticed a lower level of concordance of called recombination events between RecombineX and the original study for the green alga example, especially for GC events (Figure 6B and Table S7). While RecombineX successfully recovered all CO events reported by the original study, it also identified a substantial number of CO events that were not reported before. By manually examining the local read mapping profiles of these events in IGV (Robinson et al. 2011), we verified these RecombineX-only CO events as legitimate CO events (Figure S7). We did notice that some of these events span over assembly gaps and potentially were filtered out in the original study for this reason. As for discrepant GC calls, our IGV inspection suggests that most of them could be explained by different stringency criteria employed by the two studies during marker identification and gamete genotyping, especially in complex genomic regions (Figure S7). Also, It is worth noting that the size of GC tracts in green alga are substantially smaller (median size = 73 bp and 364 bp for Type 1 and Type 2 GC respectively) when compared with yeast (median size = 1681 bp and 1841 bp for Type 1 and Type 2 GC respectively)(Liu et al. 2018). This means that the inclusion or exclusion of a single marker makes a big difference in GC event calling for green alga.

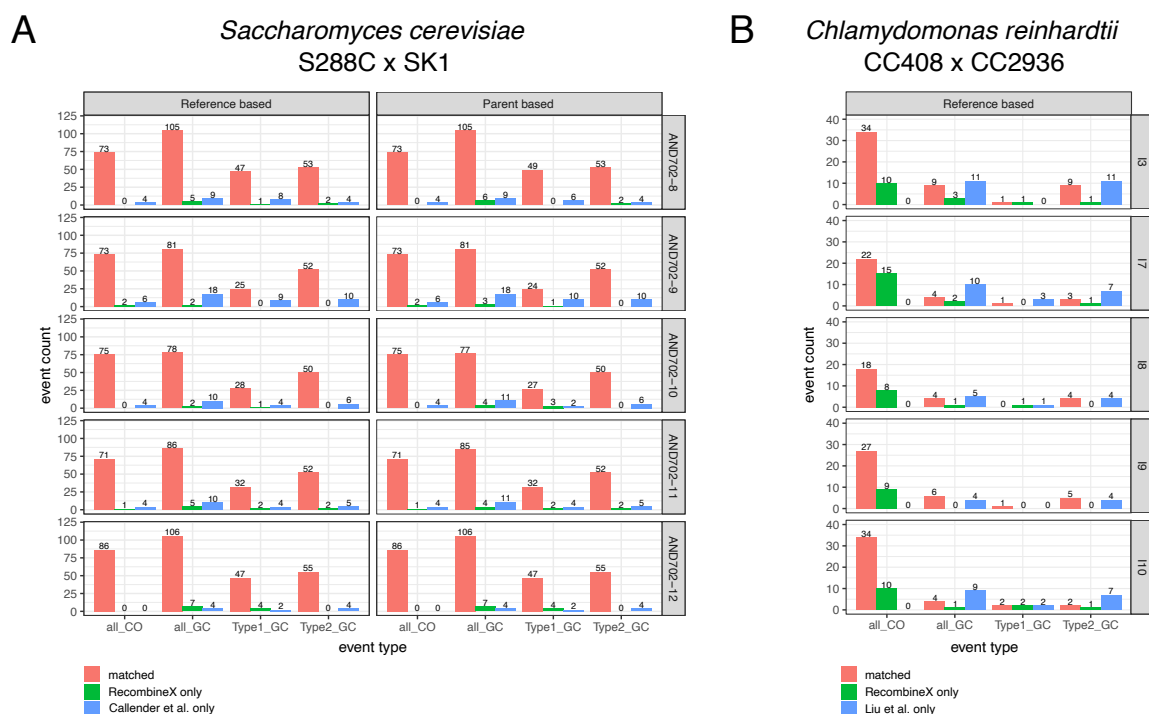


Figure 6. Applying RecombineX to real yeast and green alga tetrads for recombination profiling. Previously sequenced tetrads from yeast (AND702-8, AND702-9, AND702-10, AND702-11,

AND702-12) and green alga (I3, I7, I8, I9, I10) were used for re-processing with RecombineX, the results of which were further compared with the original studies (Callender et al. for the yeast example and Liu et al. for the green alga example). (A) Matched and unmatched total CO and GC event numbers identified by RecombineX (in reference-based and parent-based modes respectively) and the original studies. (B) Matched and unmatched Type 1 GC (NCO) and Type 2 GC (CO-associated GC) events identified by RecombineX and the original studies.

Last but not least, meiotic structural rearrangement could occur due to the numerous DSBs triggered during meiosis (Murakami and Keeney 2008; Turner et al. 2008), which could make significant impacts on the genotypes of the affected gametes. RecombineX's bonus feature for CNV-profiling comes especially helpful in discovering such gamete-specific structural rearrangements. When analyzing the five yeast S288C-SK1 tetrads using RecombineX, we found two interesting cases of gamete-specific structural rearrangement that have not been noticed before (Figure 7). One such structural rearrangement is a large segmental duplication on chromosome IV (chrIV) of the gamete AND1702-8:a (i.e. the gamete a of tetrad AND1702-8) (Figure 7A). By design, RecombineX automatically flagged CNV regions like this and set the corresponding genotypes to "NA" as a conservative measure. To reveal the exact genomic arrangement of this duplication, we retrieved the monosporic isolate of this gamete and performed long-read-based genome sequencing and assembly. A joined comparison between the resulting *de novo* AND1702-8:a assembly and the S288C and SK1 genomes unraveled the intriguing nature of this gamete-specific rearrangement: a tandem duplication with the duplicated copies inherited from both parental backgrounds (Figure 7B-7C). The breakpoints of this tandem duplication are associated with Spo11 DSB hotspots and Ty-related repetitive sequences annotated along the S288C and SK1 genomes, which echoes similar observations made in mouse recently (Lukaszewicz et al. 2021). Comparatively, the rearrangement that RecombineX identified in the gamete AND1702-12:a (i.e. the gamete a of tetrad AND1702-12) appears more complex, in which both chromosome VI (chrVI) and chromosome IX (chrIX) are involved (Figure 7D). Here we also applied long-read sequencing and assembly to illuminate this complex rearrangement, which suggests both tandem and dispersed duplications have contributed to this complex rearrangement (Figure 7E). Again, parental genomic features such as Spo11 DSB hotspots and Ty-related repetitive sequences are associated with the breakpoints, hinting their roles in triggering meiotic DSBs and driving gamete genome rearrangements (Figure 7E). These two cases of gamete-specific rearrangements demonstrated the power of resolving and understanding of tetrad formation and meiotic recombination within the native contexts of their parental genomes. As high-quality genome sequencing and assembly become increasingly affordable, future tetrad analyses are highly likely to shift away from the current reference-based convention and to embrace the parent-based new paradigm instead. In this sense, RecombineX with its built-in support for conducting analysis in parental genome space is expected to greatly facilitate such parent-based tetrad analysis.

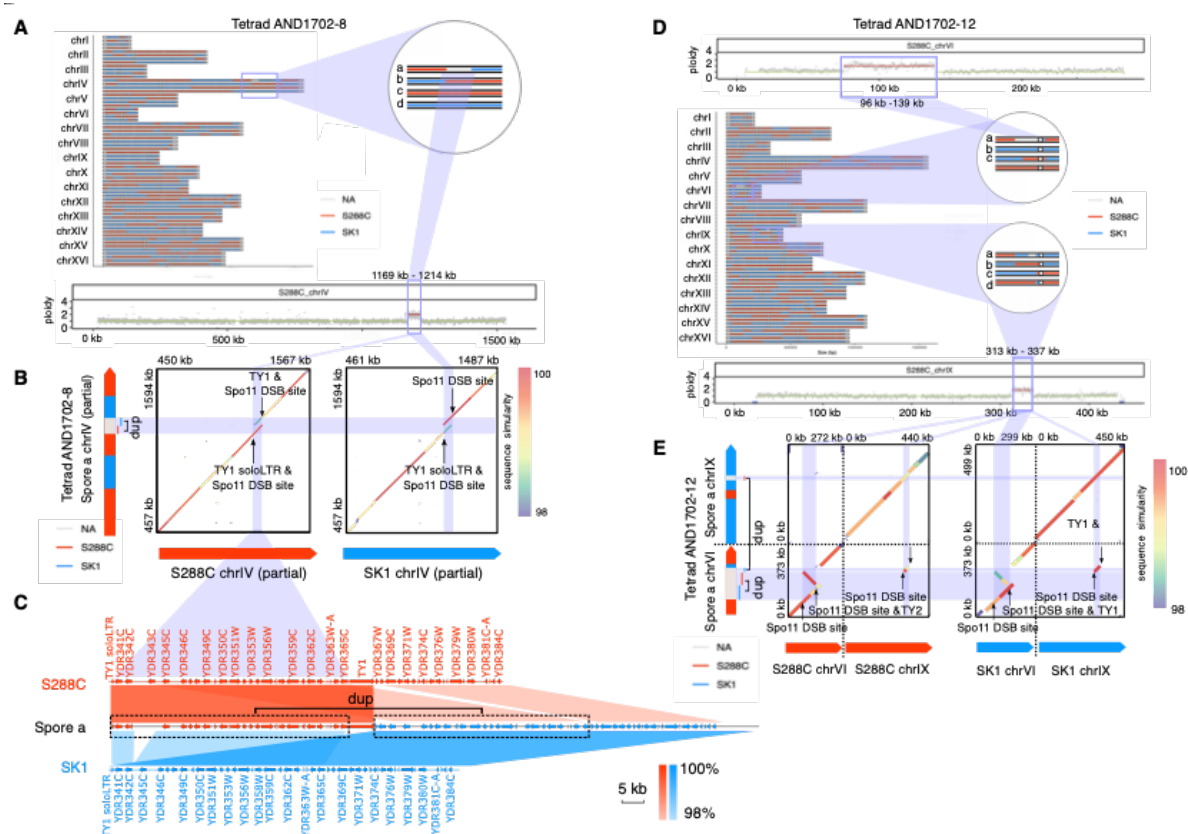


Figure 7. RecombineX enables the discovery of complex structural rearrangements in gametes. (A) A 45-kb CNV identified by RecombineX in gamete AND1702-8:a. (B) Genome sequence comparison between AND1702-8:a and its crossing parents (S288C and SK1), with differences in sequence similarity depicted by rainbow colors. (C) The gene synteny correspondence between AND1702-8:a and its crossing parents (S288C in red and SK1 in blue), with the red and blue shades representing different degrees of sequence similarity. (D) A more complex multi-chromosome-involved CNV identified by RecombineX in gamete AND1702-12:a. (E) Genome sequence comparison between the AND1702-12:a and its crossing parents (S288C and SK1), with differences in sequence similarity depicted by rainbow colors.

In summary, we developed RecombineX as a generalized computational framework that automates the full workflow of marker identification, gamete genotyping, and tetrad-based recombination profiling in a high-throughput fashion, capable of processing hundreds of tetrads in a single batch. Aside from conventional reference-based analysis, RecombineX can also perform analysis based on parental genome assemblies, which enables analyzing meiotic recombination landscapes in their native genomic contexts. Additional features such as copy number variation profiling and missing genotype inference further extends its usage for various downstream analyses. Finally, RecombineX also ships with a dedicate module for simulating the genomes and reads of recombinant tetrads, which

enables fine-tuned simulation-based hypothesis testing. This simulation module revealed the power and accuracy of RecombineX even when analyzing tetrads with very low sequencing depths (e.g., 1-2X). Tetrad sequencing data from the budding yeast *Saccharomyces cerevisiae* and green alga *Chlamydomonas reinhardtii* were further used to demonstrate the accuracy and robustness of RecombineX in tetrad analysis for organisms with both small and large genomes. As demonstrated in these examples, RecombineX unifies different functional modules under an integrated framework and provides a generalized one-stop solution for tetrad analysis. At the frontend, RecombineX shines in its modular design and parameter-rich customizability, making it highly amiable to different model systems and use scenarios. Behind the scenes, RecombineX implements thoughtful and rigorous algorithms, delivering trustable performance against biological and technical noises. The combination of these merits and the extended capacities of parent-based mode support, CNV profiling, missing genotype inference, batch processing, and tetrad simulation, together makes RecombineX a comprehensive platform for high-performance tetrad analysis. Especially considering that meiotic gamete genome sequencing from different natural and mutant backgrounds can now be acquired, we expect RecombineX to become a popular tool that empowers future tetrad analysis across different genetic backgrounds and species.

Materials and Methods

Software implementation

Preprocessing for reference and parental genomes

For both reference based and parent-based modes, RecombineX preprocesses the input genome(s) to generate necessary intermediate files for downstream analysis. These preprocessing steps include: cleaning up and relabeling the input genome file (by pre-shipped Perl scripts), indexing the input genome files by samtools (Li et al. 2009) (version: 1.9; options: faidx), profiling repetitive sequences by windowmasker (Morgulis et al. 2006) (version: 1.0.0; options: -checkdup true -mk_counts), profiling GC content by bedtools (Quinlan and Hall 2010) (version: 2.27.1 ; options: makewindows -w 250), profiling mappability by gemtools (Marco-Sola et al. 2012) (version: 1.7.1; options: -m 0.02 -e 0.02).

Reference-based parental marker identification

The parental Illumina reads are processed with trimmomatic (Bolger et al. 2014) (version: 0.38; options: PE -phred 33 ILLUMINACLIP:TruSeq3-PE-2.fa:2:30:10 SLIDINGWINDOW:5:20 MINLEN:36) to trim off adapter contamination and low quality bases. The trimmed reads are mapped to the preprocessed reference genome with bwa (Li and Durbin 2009) (version: 0.7.17; option: mem). The resulting read mapping bam file is further filtered with samtools (Li et al. 2009) (options: -q 30 -F

3340 -f 2) to only retain uniquely aligned and properly paired reads. Based on the filtered bam file, Picard Tools (<https://broadinstitute.github.io/picard/>) (version: 2.19.0) is used for sorting, mate information fixing, PCR duplicates removal, and indexing. Afterwards, GATK3 (McKenna et al. 2010) (version: 3.6-6) is used for performing local read mapping around Indels for better accuracy. Based on the GATK3-realigned bam file, samtools is used again for mpileup file production (options: mpileup -C 0 -q 30), mapping depth calculation (options: depth -aa), and mapping summary statistics calculation (option: flagstat). After read mapping, FREEC (Boeva et al. 2011) (version: 11.4) is used to perform sliding-window-based CNV profiling accordingly. Typically, FREEC requires a long list of customized parameters tuned for the input genomes. In RecombineX, the suitable parameters are automatically estimated based on the preprocessed genome file. Freebayes (Garrison and Marth 2012) (version: 1.3.4) is used for SNV and Indel calling. The called variants are further processed by vt (Tan et al. 2015) (version: 0.57721) and vcflib (Garrison et al. 2021) (version: 1.0.1) for variant decomposition (vt option: decompose_blocksub -a), normalization, annotation, and filtering (vcflib option: vcfilter -f QUAL > 30 & QUAL / AO > 1 & SAF > 0 & SAR > 0 & RPR > 1 & RPL > 1). In addition to typical variant calling filtering (quality score ≥ 30), SNVs that fall in repetitive regions, CNV regions, or immediate Indel-flanking regions (10 bp) are further filtered out. As an additional CNV filtering, SNV sites with per-site mapping depths strongly deviating from the chromosome-wide median (e.g., $> 1.5X$ chromosome-wide median or < 0.5 chromosome-wide median) are filtered out from the candidate marker list. The filtered SNVs derived from both parents are compared to each other to generate a candidate reference-based marker set comprising SNVs segregated between the two parents. Such candidate marker set is further leveraged with the mpileup file generated during read mapping to control for potential false negative and false positive from SNV calling. Finally, RecombineX pre-shipped with several plotting scripts in R to generate graphic reports on CNV and marker distribution along the reference genome coordinates.

Parent-based parental marker identification

Parent-based parental marker identification can be performed in two strategies: the genome-alignment-based marker identification and consensus-based marker identification. For the genome-alignment-based marker identification, RecombineX employs mummer3 (Kurtz et al. 2004) (version: 3.23) for genome alignment building (options: nucmer -g 90 -l 20 -c 65), filtering (options: delta-filter -1), and SNV extraction (options: show-snps -Clr). The extracted SNVs are further processed by vt (Tan et al. 2015) and vcflib (Garrison et al. 2021) for decomposition, normalization, and annotation. Afterwards these marker candidates are filtered based on the repetitive profiling results generated at the parent-genome preprocessing step to remove repetitive-region-associated markers. Also, as the specific choice of query and target genome assemblies at the genome alignment step potentially could lead to directional bias, a reciprocal filter is further applied to only retain the reciprocal SNV calls recovered in both comparison directions (i.e., A to B and B to A). The corresponding filtered SNV set

is defined as genome alignment based parental marker set. As for the consensus-based marker identification strategy, there are several extra steps to be taken. First, cross-parent read mapping is performed by mapping the Illumina reads of one parent to the genome assembly of the other parent. Accordingly, mapping based SNV, Indel, and CNV calling are carried out by freebayes (Garrison and Marth 2012) and FREEEC (Boeva et al. 2011). Subsequently, the mapping based SNVs are further filtered by repetitive sequences and CNVs identified along the parental genomes. The detailed read mapping, variant calling, and variant filtering protocols are the same as those mentioned above, except for that both parental genome assemblies rather than a single reference genome is used in this case. Afterwards, the resulting mapping-based SNV calling sets are intersected with the genome-alignment-based SNV marker sets to derive a consensus marker set, upon which a final reciprocal filter is further applied to make sure the final consensus set is strictly symmetrical relative to the two crossing parents. For both genome-alignment-based and consensus-based marker set, RecombineX will also plot their respective marker distributions along the genome coordinates of both parental genomes.

Reference-based gamete read mapping and genotyping

RecombineX uses a strictly defined master sample table to document the metadata for each sequenced gamete sample, its Illumina reads, and its corresponding tetrad. According to this file, RecombineX will automatically perform reference-based read mapping and CNV profiling for each defined gamete sample based on the same protocols adopted for parental marker identification. As for genotyping, RecombineX takes the inputs from the mpileup file generated by gamete read mapping and the reference-based parental marker list generated by marker identification to evaluate marker-specific reads support from each gamete across all marker sites. At each marker site, two quality control parameters are calculated. A Q_{net} score is calculated as the cumulative sequencing score difference between the major allele (i.e., the best supported base) and all minor alleles (if any). In the meantime, a base purity score is calculated as the proportion of reads supporting the major allele at the corresponding marker site. A genotype is tentatively assigned only when both Q_{net} and base purity meet their pre-defined cutoffs (50 for Q_{net} and 0.9 for base purity by default). The tentatively assigned genotypes are further filtered based on the gamete specific CNV profiles generated by FREEEC (Boeva et al. 2011), during which the genotypes of markers falling in gamete specific CNV regions will be set as “NA”. In addition to this raw genotyping result, RecombineX will generate another copy of genotyping result (labeled by the “inferred” tag in its file name) by further inferring the possible missing genotypes based on a general 2:2 parental allele segregation ratio across the tetrad. For both raw and inferred genotyping results, RecombineX will make both tetrad-based and batch-based genotyping plots with pre-shipped R scripts to visualize the composition and segregation of two parental genetic backgrounds across the genome.

Parent-based gamete read mapping and genotyping

The protocol used by RecombineX for parent-based gamete read mapping and genotyping is largely the same as that used for reference-based gamete read mapping and genotyping, except for now the analyses are separately performed based on both parental genome assemblies. Therefore, after reads mapping, two genotyping results will be obtained separately, each based on the genome space of a single parent. These two genotyping calls will be cross validated with each other. Those markers with conflicted genotyping calls will be ignored for downstream analysis (i.e., their genotypes will be reassigned to “NA”). Like reference-based analysis, features such as gamete CNV profiling and missing genotype inference are fully supported for parent-based analysis.

Reference- and parent-based recombination profiling

The recombination profiling procedures in the reference-based and parent-based modes are essentially the same, except that the parent-based mode will perform the analysis twice, each based on the genome space of a single parent. RecombineX implemented a modified version of the original recombination event profiling algorithms used by ReCombine (Anderson et al. 2011) to cover all foreseeable scenarios (Figure S4 and S5). Briefly, RecombineX scans through the tetrad-wide genotypes at every marker site to classify them into different categories based on the segregation ratio of the parental alleles (e.g., 2:2 or 3:1 or 1:3 or 4:0 or 0:4). Consecutive markers with identical segregation ratios are grouped together, based on which preliminary linkage blocks are identified. By ignoring the remaining markers with missing genotypes, the calculated preliminary linkage blocks are further extended to form final linkage blocks. The outer bounds of each final linkage block are defined by the midpoint of the outermost markers of this linkage block and their immediate flanking markers. Users can restrict such linkage block identification operation by modifying the minimal number of supporting markers (default: 1 marker) and minimal block size (default: 1 bp) parameters. According to the identified final linkage blocks, RecombineX systematically examines the genotype switch patterns and the number of gametes involved in genotype switches between each pair of adjacent linkage blocks to classify the local recombination events. Those recombination events that are in close adjacency (controlled by the “merging range” parameter) can be further merged when needed. Upon the completion of the calculation, detailed tabular reports on the marker-wide parental allele segregation ratio, preliminary and final linkage blocks, as well as lists of recombination events will be reported.

Recombinant tetrad genome and reads simulation

RecombineX performs recombinant tetrad genome simulation based on an input genome assembly and a list of clearly defined parental markers. Both reference assembly and native parental genome assembly can be used as the input here, as long as the accompanying parental marker list is based on the same genome coordinates. With these inputs, RecombineX first simulate the two parental

genomes by projecting parental markers to the input genome assembly. Based on the simulated parental genomes, CO and GC events are further simulated based on various user-specified parameters, which include the number of CO and GC events, the ratio of CO and GC events, the mean and standard deviation of GC tracts, etc. These recombination events are randomly placed into the four resulting gametes. For both simulated parental and gamete genomes, paired-end Illumina reads are further simulated by ART (Huang et al. 2012) with user-specified sequencing depth.

Simulation based analysis

Genome and reads simulation for parental marker identification

To evaluate the performance of parental marker identification, a pair of hypothetical crossing parents, P1 and P2, were simulated for this study. The genome of P1 is an exact copy of the budding yeast *S. cerevisiae* reference genome (version: R64-2-1_20150113) retrieved from *Saccharomyces* Genome Database (SGD) with the mitochondrial genome excluded. Based on the same reference genome, the genome of P2 was further generated by simuG (Yue and Liti 2019) (GitHub commit version: 212ea1f) in a two-pass manner to randomly introduce 60,000 SNVs and 6,000 Indels (options: -refseq SGDref.genome.fa -snp_count 60000 -titv_ratio 2.0 -indel_count 6000 -seed 20190518 -prefix yeast_60kSNP_6kINDEL) as well as six CNVs (options: -refseq yeast_60kSNP_6kINDEL.simseq.genome.fa -cnv_count 6 -cnv_gain_loss_ratio 1 -duplication_tandem_dispersed_ratio 1 -cnv_max_copy_number 4 -centromere_gff SGDref.centromere.gff). The centromere annotation of the *S. cerevisiae* reference genome (distributed with the reference genome) was used for the second pass to prevent the simulated CNVs from surpassing centromeres.

For the simulated genome of P1 and P2, 150-bp paired-end Illumina reads were further simulated by ART (Huang et al. 2012) (version: MountRainier-2016-06-05; options: “-p -l 150 --qprof1 HiSeq2500L150 -qprof2 HiSeq2500L150 -f <depth> -m 500 -s 10 -na -rs 20210210”). Here we simulated a wide range of sequencing depths (10X, 20X, ..., 100X) for exploring the influence of sequencing depth on parental marker identification. The simulated parental genome and reads for P1 and P2 were fed into RecombineX for parental marker identification. The identified markers following both reference-based and parent-based protocols were compared with the initially simulated SNVs between P1 and P2.

Recombinant tetrad simulation for gamete genotyping and recombination profiling

The aforementioned P1 genome and the consensus markers that RecombineX identified based on 100X parental reads were used as the inputs for tetrad genome simulation. A total of 90 COs and 65 GCs were simulated with their size distribution parameters determined based on real yeast tetrads: mean Type 1 GC size = 2250 bp, standard deviation of Type 1 GC size = 2200, mean Type 2 GC size

= 2500 bp, standard deviation of Type 2 GC size = 2000 bp, min GC size = 100 bp, max GC size = 5000 bp (Mancera et al. 2008). Random seed was set to 20210210 for this simulation. The introduced recombination events and the resulting genotypes were used as ground truth sets for downstream comparison. For each simulated gamete genome, paired-end Illumina reads were further generated by ART with varied depths (1X, 2X, 4X, 8X, 16X, 32X, 64X). The simulated gamete reads were processed with RecombineX in both reference-based and parent-based modes. The generated gamete genotyping and recombination profiling results were compared with the ground truth generated during our simulation. The impacts of different Q_{net} cutoffs (e.g., 10, 20, 30, ..., 100) were thoroughly explored during this process.

Real tetrad-based analyses

Two real tetrad sequencing datasets retrieved from previous studies (Callender et al. 2016; Liu et al. 2018) were used to run the full workflow of RecombineX. The first dataset includes five tetrads derived from the budding yeast *S. cerevisiae* cross S288C x SK1, for which both reference-based and parent-based analyses were performed. The SGD yeast reference genome (version: R64-2-1_20150113) was used for the reference-based analysis, while our previously generated long-read-based genome assemblies for S288C and SK1 (Yue et al. 2017) were used for the parent-based analysis. The second dataset includes five tetrads derived from the green alga *C. reinhardtii* cross: CC408 x CC2936 (Table S8 and S9), for which only reference-based analysis was performed. For this analysis, we retrieved the alga *C. reinhardtii* (v5.5) reference genome from Ensembl Plants (<https://plants.ensembl.org>). For both yeast and green alga datasets, a Q_{net} cutoff of 50 was used and no adjacent recombination event merging was applied. The RecombineX profiled recombination events were compared with the events reported in the original studies, which were retrieved via the following links respectively:

yeast tetrads: <http://dx.doi.org/10.5061/dryad.g6s2k>

green alga tetrads: <https://figshare.com/s/a95156f0ed5272b9109e>

In our comparison, two events were considered “match” only if they completely agreed with each other in event types, genomic locations, and involved gametes. For events that were identified by RecombineX or the original studies alone, we further examined the gamete read alignment in the IGV browser (Robinson et al. 2011) (version: 2.8.13) to understand the specific causes of such disagreement.

Oxford Nanopore sequencing of yeast gametes AND1702-8:a and AND1702-12:a

To take a closer look at the structural rearrangements identified by RecombineX for yeast gamete AND1702-8:a and AND1702-12:a, we retrieved the corresponding strains stocked in Dr. Alain Nicolas’s lab at Institut Curie (Paris, France). Upon receiving the yeast cells, we grew them in 10 - 15

ml YPD (2% peptone, 1% yeast extract, 2% glucose) at 30 °C for overnight (220 rpm). A total number of cells less than 7×10^9 were used for DNA extraction. High molecular weight (HMW) DNA was extracted by QIAGEN® Genomic-tip 100/g according to the “QIAGEN Genomic DNA handbook” for Yeast. DNA quantity and length were controlled by the Qubit dsDNA HS Assay. Library preparation and ONT sequencing were performed based on the protocol of “1D Native barcoding genomic DNA with EXP-NBD104 and SQK-LSK108” obtained from Oxford Nanopore Technologies Community. The FLO-MIN106 MinION flow cell was used for sequencing.

Genome assembly, annotation, and comparison for yeast gametes AND1702-8:a and AND1702-12:a

The nanopore reads were processed with our previously developed LRSDAY pipeline (Yue and Liti 2018) (version: 1.6.0) for *de novo* genome assembly and comprehensive feature annotation. The internal protocols employed by LRSDAY are briefly described as follows. The raw nanopore-sequencing fast5 reads are processed with Guppy (version: 3.2.4) for basecalling and demultiplexing. The resulting fastq reads are further trimmed with Porechop (version: 0.2.4; options: --discard_middle) and filtered with Filtlong (version: 0.2.0; options: --min_length 1000 --mean_q_weight 10 --target_bases 750000000). The filtered reads are assembled with Canu (version: 1.8; options: -s genomeSize=12.5m -nanopore-raw). The raw Canu-assembly is further polished with both nanopore (sequenced in this study) and Illumina reads (retrieved from the original study). Three successive rounds of long-read-based polishing are performed by Racon (Vaser et al. 2017) (version: 1.4.7) and Medaka (<https://github.com/nanoporetech/medaka>) (version: 0.8.1 ; options: -m r941_flip235). Another three successive rounds of short-read-based polishing are performed by Pilon (Walker et al. 2014) (version: 1.23; --fix snps,indels). The polished assembly is further processed with Ragout (Kolmogorov et al. 2018) (version: 2.2) and circulator (Hunt et al. 2015) (version: 1.5.5; option: fixstart --genes_fa ATP6.cds.fa --min_id 90) for reference-based scaffolding and mitochondrial assembly improvement. The resulting final nuclear and mitochondrial genome assemblies are further annotated by Maker3 (Holt and Yandell 2011) (version: 3.00.0-beta) and Mfannot (<https://github.com/BFL-lab/Mfannot>) (version: 1.35) respectively, with additional reference-based gene orthology identification performed by Proteinortho (Lechner et al. 2011) (version: 5.16b). Other important genomic features such as centromere, tRNA, Ty transposable elements, core-X elements, Y' elements were also annotated by dedicated modules implemented in LRSDAY. The fully assembled and annotated genomes of the gametes AND1702-8:a and AND1702-12:a obtained in this way were further compared with the native genome assembly of their crossing parents (S288C and SK1) by Mummer3 (Kurtz et al. 2004), BLAST+ (Camacho et al. 2009) (version: 2.2.31+; options: -blastn) and Easyfig (Sullivan et al. 2011) (version: 2.2.3; options: -i 98 -min_length 1000 -filter -f1 T -f2 1000) regarding both sequence similarity and annotated genomic features. The Spo11 DSB hotspot annotation used in such comparison was retrieved from the literature (Pan et al. 2011).

Acknowledgements

We thank Dr. Johan Hallin (Gothenburg University, Gothenburg, Sweden) and Ms. Jessica Chevallier (CRCM, Marseille, France) for beta testing and valuable discussions. We thank Dr. Alain Nicolas (Institut Curie, Paris, France) and Ms. Sophie Loeillet (Institut Curie, Paris, France) for locating and sharing the original gamete samples of AND1702-8:a and AND1702-12:a.

References

- Anderson CM, Chen SY, Dimon MT, Oke A, DeRisi JL, Fung JC. 2011. ReCombine: A Suite of Programs for Detection and Analysis of Meiotic Recombination in Whole-Genome Datasets. *PLOS ONE* **6**: e25509.
- Boeva V, Zinovyev A, Bleakley K, Vert J-P, Janoueix-Lerosey I, Delattre O, Barillot E. 2011. Control-free calling of copy number alterations in deep-sequencing data using GC-content normalization. *Bioinformatics* **27**: 268–269.
- Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**: 2114–2120.
- Brion C, Legrand S, Peter J, Caradec C, Pflieger D, Hou J, Friedrich A, Llorente B, Schacherer J. 2017. Variation of the meiotic recombination landscape and properties over a broad evolutionary distance in yeasts. *PLoS Genetics* **13**: e1006917.
- Broman KW, Wu H, Sen S, Churchill GA. 2003. R/qtl: QTL mapping in experimental crosses. *Bioinformatics* **19**: 889–890.
- Callender TL, Laureau R, Wan L, Chen X, Sandhu R, Laljee S, Zhou S, Suhandynata RT, Prugar E, Gaines WA, et al. 2016. Mek1 Down Regulates Rad51 Activity during Yeast Meiosis by Phosphorylation of Hed1. *PLOS Genetics* **12**: e1006226.
- Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. 2009. BLAST+: architecture and applications. *BMC Bioinformatics* **10**: 421.
- Coop G, Wen X, Ober C, Pritchard JK, Przeworski M. 2008. High-Resolution Mapping of Crossovers Reveals Extensive Variation in Fine-Scale Recombination Patterns Among Humans. *Science* **319**: 1395–1398.
- Garrison E, Kronenberg ZN, Dawson ET, Pedersen BS, Prins P. 2021. *Vcflib and tools for processing the VCF variant call format*. <https://www.biorxiv.org/content/10.1101/2021.05.21.445151v1> (Accessed November 20, 2021).

- Garrison E, Marth G. 2012. Haplotype-based variant detection from short-read sequencing. *arXiv:12073907 [q-bio]*. <http://arxiv.org/abs/1207.3907> (Accessed October 30, 2020).
- Hinch AG, Zhang G, Becker PW, Moralli D, Hinch R, Davies B, Bowden R, Donnelly P. 2019. Factors influencing meiotic recombination revealed by whole-genome sequencing of single sperm. *Science* **363**: eaau8861.
- Holt C, Yandell M. 2011. MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. *BMC Bioinformatics* **12**: 491.
- Hou Y, Fan W, Yan L, Li R, Lian Y, Huang J, Li J, Xu L, Tang F, Xie XS, et al. 2013. Genome Analyses of Single Human Oocytes. *Cell* **155**: 1492–1506.
- Huang W, Li L, Myers JR, Marth GT. 2012. ART: a next-generation sequencing read simulator. *Bioinformatics* **28**: 593–594.
- Hunt M, Silva ND, Otto TD, Parkhill J, Keane JA, Harris SR. 2015. Circlator: automated circularization of genome assemblies using long sequencing reads. *Genome Biology* **16**: 294.
- Kirkness EF, Grindberg RV, Yee-Greenbaum J, Marshall CR, Scherer SW, Lasken RS, Venter JC. 2013. Sequencing of isolated sperm cells for direct haplotyping of a human genome. *Genome Res* **23**: 826–832.
- Kolmogorov M, Armstrong J, Raney BJ, Streeter I, Dunn M, Yang F, Odom D, Flicek P, Keane TM, Thybert D, et al. 2018. Chromosome assembly of large and complex genomes using multiple references. *Genome Res*. <https://genome.cshlp.org/content/early/2018/10/19/gr.236273.118> (Accessed November 21, 2021).
- Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C, Salzberg SL. 2004. Versatile and open software for comparing large genomes. *Genome Biology* **5**: R12.
- Lechner M, Findeiß S, Steiner L, Marz M, Stadler PF, Prohaska SJ. 2011. Proteinortho: Detection of (Co-)orthologs in large-scale analysis. *BMC Bioinformatics* **12**: 124.
- Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**: 1754–1760.
- Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**: 2078–2079.
- Li X, Li L, Yan J. 2015. Dissecting meiotic recombination based on tetrad analysis by single-microspore sequencing in maize. *Nature Communications* **6**: 6648.
- Liu H, Huang J, Sun X, Li J, Hu Y, Yu L, Liti G, Tian D, Hurst LD, Yang S. 2018. Tetrad analysis in plants and fungi finds large differences in gene conversion rates but no GC bias. *Nature Ecology & Evolution* **2**: 164–173.

- Liu H, Maclean CJ, Zhang J. 2019. Evolution of the Yeast Recombination Landscape. *Molecular Biology and Evolution* **36**: 412–422.
- Lu P, Han X, Qi J, Yang J, Wijeratne AJ, Li T, Ma H. 2012. Analysis of Arabidopsis genome-wide variations before and after meiosis and meiotic recombination by resequencing Landsberg erecta and all four products of a single meiosis. *Genome Res* **22**: 508–518.
- Lukaszewicz A, Lange J, Keeney S, Jasin M. 2021. De novo deletions and duplications at recombination hotspots in mouse germlines. *Cell* **184**: 5970–5984.e18.
- Mancera E, Bourgon R, Brozzi A, Huber W, Steinmetz LM. 2008. High-resolution mapping of meiotic crossovers and non-crossovers in yeast. *Nature* **454**: 479–485.
- Marco-Sola S, Sammeth M, Guigó R, Ribeca P. 2012. The GEM mapper: fast, accurate and versatile alignment by filtration. *Nat Methods* **9**: 1185–1188.
- Marsolier-Kergoat M-C, Khan MM, Schott J, Zhu X, Llorente B. 2018. Mechanistic View and Genetic Control of DNA Recombination during Meiosis. *Molecular Cell* **70**: 9–20.e6.
- Martini E, Borde V, Legendre M, Audic S, Regnault B, Soubigou G, Dujon B, Llorente B. 2011. Genome-Wide Analysis of Heteroduplex DNA in Mismatch Repair–Deficient Yeast Cells Reveals Novel Properties of Meiotic Recombination Pathways. *PLOS Genetics* **7**: e1002305.
- McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, et al. 2010. The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* **20**: 1297–1303.
- Morgulis A, Gertz EM, Schäffer AA, Agarwala R. 2006. WindowMasker: window-based masker for sequenced genomes. *Bioinformatics* **22**: 134–141.
- Murakami H, Keeney S. 2008. Regulating the formation of DNA double-strand breaks in meiosis. *Genes Dev* **22**: 286–292.
- Pan J, Sasaki M, Kniewel R, Murakami H, Blitzblau HG, Tischfield SE, Zhu X, Neale MJ, Jasin M, Socci ND, et al. 2011. A Hierarchical Combination of Factors Shapes the Genome-wide Topography of Yeast Meiotic Recombination Initiation. *Cell* **144**: 719–731.
- Peter J, De Chiara M, Friedrich A, Yue J-X, Pflieger D, Bergström A, Sigwalt A, Barre B, Freel K, Llored A, et al. 2018. Genome evolution across 1,011 *Saccharomyces cerevisiae* isolates. *Nature* **556**: 339–344.
- Ptak SE, Hinds DA, Koehler K, Nickel B, Patil N, Ballinger DG, Przeworski M, Frazer KA, Pääbo S. 2005. Fine-scale recombination patterns differ between chimpanzees and humans. *Nat Genet* **37**: 429–434.

- Qi J, Chen Y, Copenhaver GP, Ma H. 2014. Detection of genomic variations and DNA polymorphisms and impact on analysis of meiotic recombination and genetic mapping. *PNAS* **111**: 10007–10012.
- Qi J, Wijeratne AJ, Tomsho LP, Hu Y, Schuster SC, Ma H. 2009. Characterization of meiotic crossovers and gene conversion by whole-genome sequencing in *Saccharomyces cerevisiae*. *BMC Genomics* **10**: 475.
- Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**: 841–842.
- Rasmussen MD, Hubisz MJ, Gronau I, Siepel A. 2014. Genome-Wide Inference of Ancestral Recombination Graphs. *PLOS Genetics* **10**: e1004342.
- Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, Mesirov JP. 2011. Integrative genomics viewer. *Nat Biotechnol* **29**: 24–26.
- Sakhanenko NA, Cromie GA, Dudley AM, Galas DJ. 2019. Computational Inference Software for Tetrad Assembly from Randomly Arrayed Yeast Colonies. *G3 Genes|Genomes|Genetics* **9**: 2071–2088.
- Spence JP, Song YS. 2019. Inference and analysis of population-specific fine-scale recombination maps across 26 diverse human populations. *Science Advances* **5**: eaaw9206.
- Sullivan MJ, Petty NK, Beatson SA. 2011. Easyfig: a genome comparison visualizer. *Bioinformatics* **27**: 1009–1010.
- Tan A, Abecasis GR, Kang HM. 2015. Unified representation of genetic variants. *Bioinformatics* **31**: 2202–2204.
- Turner DJ, Miretti M, Rajan D, Fiegler H, Carter NP, Blayney ML, Beck S, Hurles ME. 2008. Germline rates of de novo meiotic deletions and duplications causing several genomic disorders. *Nat Genet* **40**: 90–95.
- Vaser R, Sović I, Nagarajan N, Šikić M. 2017. Fast and accurate de novo genome assembly from long uncorrected reads. *Genome Res* **27**: 737–746.
- Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, Cuomo CA, Zeng Q, Wortman J, Young SK, et al. 2014. Pilon: An Integrated Tool for Comprehensive Microbial Variant Detection and Genome Assembly Improvement. *PLOS ONE* **9**: e112963.
- Wang J, Fan HC, Behr B, Quake SR. 2012. Genome-wide Single-Cell Analysis of Recombination Activity and De Novo Mutation Rates in Human Sperm. *Cell* **150**: 402–412.
- Wijnker E, Velikkakam James G, Ding J, Becker F, Klasen JR, Rawat V, Rowan BA, de Jong DF, de Snoo CB, Zapata L, et al. 2013. The genomic landscape of meiotic crossovers and gene conversions in *Arabidopsis thaliana* ed. G. McVean. *eLife* **2**: e01426.

- Yeadon PJ, Bowring FJ, Catcheside DEA. 2016. Meiotic Recombination in *Neurospora crassa* Proceeds by Two Pathways with Extensive Holliday Junction Migration. *PLOS ONE* **11**: e0147815.
- Yue J-X, Li J, Aigrain L, Hallin J, Persson K, Oliver K, Bergström A, Coupland P, Warringer J, Lagomarsino MC, et al. 2017. Contrasting evolutionary genome dynamics between domesticated and wild yeasts. *Nature Genetics* **49**: 913–924.
- Yue J-X, Liti G. 2018. Long-read sequencing data analysis for yeasts. *Nature Protocols* **13**: 1213–1231.
- Yue J-X, Liti G. 2019. simuG: a general-purpose genome simulator. *Bioinformatics* **35**: 4442–4444.

Supporting information

Supporting information captions

Figure S1. Overview of the RecombineX directory system. The pre-shipped top-level directories and individual files of RecombineX are denoted with solid lines. Additional directories and files to be generated during the installation of RecombineX are denoted with dashed lines.

Figure S2. Overview of the RecombineX parental marker identification algorithms. Two marker identification modes are supported: the reference-based mode (colored in yellow) and the parent-based mode (colored in orange).

Figure S3. Overview of the RecombineX gamete genotyping algorithms. Two genotyping modes are supported: the reference-based mode (colored in yellow) and the parent-based mode (colored in orange).

Figure S4. Overview of the RecombineX recombination event classification scheme. The definition and example of different CO and GC types are shown in panel a and panel b respectively. This classification scheme is designed based on the original ReCombine recombination event classification scheme (Anderson et al. PLoS One, 2011) with additional modifications.

Figure S5. Overview of the RecombineX recombination event identification algorithm. This algorithm is designed based on the original ReCombine algorithm (Anderson et al. PLoS One, 2011) with additional modifications.

Figure S6. Manual examination of unmatched recombination events from the yeast S288C-SK1 tetrads in IGV. The read alignments of parent and gamete reads are visualized in IGV with event-defining SNP markers shown in colors.

Figure S7. Manual examination of unmatched recombination events from the green alga CC408- CC2936 tetrads in IGV. The read alignments of parent and gamete reads are visualized in IGV with event-defining SNP markers shown in colors.

Table S7. Summary of parental marker identification for real cross examples with RecombineX.

Table S2. Yeast and green alga parental genome sequencing datasets employed in this study.

Table S3. RecombineX's performance in parental marker identification performance based on simulated data.

Table S4. RecombineX's gamete genotyping performance based on simulated tetrads with all four gametes.

Table S5. RecombineX's gamete genotyping performance based on simulated tetrads with only three viable gametes.

Table S6. RecombineX's recombination event profiling performance based on the simulated tetrad.

Table S7. Summary of parental marker identification for real cross examples with RecombineX.

Table S8. RecombineX's recombination event profiling performance based on real yeast tetrads.

Table S9. RecombineX's recombination event profiling performance based on real green alga tetrads.

Availability of data and materials

The RecombineX software is freely distributed under MIT license at GitHub (<https://github.com/yjx1217/RecombineX>). The nanopore reads as well as the corresponding genome assemblies of the gametes AND1702-8:a and AND1702-12:a have been deposited to the SRA database under the accession number of PRJNA698424. In addition, another copy of the genome assembly of AND1702-8:a and AND1702-12:a together with the corresponding genome annotations have been uploaded to RecombineX's GitHub depository under the data subfolder (<https://github.com/yjx1217/RecombineX/tree/master/data>).

Competing interests

The authors declare that they have no competing interests.

Funding

This work is supported by National Natural Science Foundation of China (32070592 to J.-X. Y. and 32000395 to J. Li.), Guangdong Basic and Applied Basic Research Foundation (2019A1515110762 to J.-X. Y.), Guangdong Pearl River Talents Program (2019QN01Y183 to J.-X. Y.), Microsoft Azure Research Award (CRM:074871 to J.-X. Y.), ANR (ANR-15-IDEX-01 and ANR-20-CE13-0010 to G. L. ; ANR-18-CE12-0013 to B. L.), Fondation pour la Recherche Médicale (EQU202003010413 to G. L.), and Guangzhou Municipal Science and Technology Bureau (202102020938 to J. L.), respectively.

Authors' contributions

JXY and GL designed the study. JXY developed the associated software. JL performed the nanopore sequencing of the yeast gametes AND1702-8:a and AND1702-12:a. JXY and JL performed the data analysis. All authors discussed the results and contributed to the final manuscript. All authors read and approved the final manuscript.