ARPEGGIO: Automated Reproducible Polyploid EpiGenetic GuIdance workflOw

Stefan Milosavljevic[1,2], Tony Kuo[3], Samuele Decarli[4], Lucas Mohn[1], Jun Sese[5,6], Kentaro K. Shimizu[1,7], Rie Shimizu-Inatsugi[1] and Mark D. Robinson[2,8,*]

[1] Department of Evolutionary Biology and Environmental Studies, University of Zurich, Switzerland
[2] SIB Swiss Institute of Bioinformatics, University of Zurich, Switzerland
[3] Centre for Biodiversity Genomics, Guelph, Canada
[4] Department of Computer Science, ETH Zurich, Switzerland
[5] AIST Artificial Intelligence Research Center, Tokyo, Japan
[6] Humanome Lab Inc., Chuo-ku, Tokyo, Japan
[7] Kihara Institute for Biological Research, Yokohama City University, Yokohama, Japan
[8] Institute of Molecular Life Sciences, University of Zurich, Switzerland
* Corresponding author

Abstract (341 words, max 350)

Whole genome duplication (WGD) events are common in the evolutionary history of many living organisms. For decades, researchers have been trying to understand the genetic and epigenetic impact of WGD and its underlying molecular mechanisms. Particular attention was given to allopolyploid study systems, species resulting from an hybridization event accompanied by WGD. Investigating the mechanisms behind the survival of a newly formed allopolyploid highlighted the key role of DNA methylation. With the improvement of high-throughput methods, such as whole genome bisulfite sequencing (WGBS), an opportunity opened to further understand the role of DNA methylation at a larger scale and higher resolution. However, only a few studies have applied WGBS to allopolyploids, which might be due to lack of genomic resources combined with a burdensome data analysis process. To overcome these problems, we developed the Automated Reproducible Polyploid EpiGenetic GuIdance workflOw (ARPEGGIO): the first workflow for the analysis of epigenetic data in polyploids. This workflow analyzes WGBS data from allopolyploid species via the genome assemblies of the allopolyploid's parent species. ARPEGGIO utilizes an updated read classification algorithm (EAGLE-RC), to tackle the challenge of sequence similarity amongst parental genomes. ARPEGGIO offers automation, but more importantly, a complete set of analyses including spot checks starting from raw WGBS data: quality checks, trimming, alignment, methylation extraction, statistical analyses and downstream analyses. A full run of ARPEGGIO outputs a list of genes showing differential methylation. ARPEGGIO's design focuses on ease of use and reproducibility. ARPEGGIO was made simple to set up, run and interpret, and its implementation includes both package management and containerization. Here we discuss all the steps, challenges and implementation strategies; example datasets are provided to show how to use ARPEGGIO. In

addition, we also test EAGLE-RC with publicly available datasets given a ground truth, and we show that EAGLE-RC decreases the error rate by 3 to 4 times compared to standard approaches. The goal of ARPEGGIO is to promote, support and improve polyploid research with a reproducible and automated set of analyses in a convenient implementation.

Keywords (10, max 10)

snakemake, epigenetics, bisulfite-sequencing, polyploidy, allopolyploids, reproducibility, automation, workflow, dna-methylation, whole-genome-bisulfite-sequencing

Background

Polyploidy, also known as whole genome duplication (WGD), is a process leading to the formation of an organism with more than two sets of chromosomes. There are two types of polyploidy: autopolyploidy, the doubling of an entire genome in a single species, and allopolyploidy, the hybridization of two different species followed by whole genome duplication (1). Both of these processes influenced the evolutionary history of many living organisms such as nematodes, arthropods, chordates, fungi, oomycetes and plants (1–3). Of all these lineages, the most extensive research on polyploidy has been done on land plants (1–8), where about 35% of all species were estimated to be recent polyploids (7,8) and at least one ancient WGD was inferred in the ancestry of every lineage (3).

To understand the successful prevalence of WGD and the underlying mechanisms, particular attention was given to early stages of polyploidy in allopolyploids (4,9–11). Among several observed genomic and epigenomic changes (4,10,12), DNA methylation was shown to play an important role to ensure the survival of a newly formed allopolyploid (13–19). A well-studied example comes from Madlung and colleagues (13) in which they chemically treated synthetic *Arabidopsis suecica* allotetraploids to remove DNA methylation over the whole genome. With this treatment, they observed many phenotypic disorders such as abnormal branching or homeotic abnormalities in flowers, mostly leading to sterility. These abnormalities were not observed when treating the parent species or the natural allopolyploid, highlighting the importance of DNA methylation in the first generations after allopolyploidization. Follow-up studies focused on the epigenetic regulation in other resynthesized allopolyploid species with varying outcomes. In allopolyploid wheat, *Tragopogon*, *Spartina* and rice, DNA methylation changes indicated gene repression favoring one parental genome over the other (15–20). This was not the case in *Arabidopsis*, where similar DNA methylation and expression changes were observed on both parental genomes (21). In *Brassica*, both previously mentioned outcomes were reported (15,22), while in cotton no changes were found (23). All these studies proposed different mechanisms to clarify the role of methylation and its short and long term evolutionary impact, but the discussion remains open (4). One reason that might complicate the grounds of such discussion, is the variety of tools and methods used to analyze DNA methylation data. To better control discrepancies between findings caused by methodological differences, a standardized set of tools would be ideal.

Despite the potential significance of DNA methylation in allopolyploid evolution, many of the previously mentioned findings were limited by low-throughput methods. These methods, such as methylation-sensitive amplified length polymorphisms (MSAP), were unable to capture changes at a whole genome level (24). With advances in technology, new high-throughput methods such as whole genome bisulfite sequencing (WGBS) are able to obtain methylation information at individual nucleotides over the whole genome (25).
At the whole genome level, DNA methylation is separated into three different sequence contexts: CG, CHG and CHH (where H = A, T or C). Each context is

regulated by different families of enzymes and depending on the species, some contexts might be more important than others (26). For example, in mammals, methylation occurs mainly in CG context, while in plants it occurs in all three contexts (26).

Although WGBS is considered to be the gold standard in whole-genome DNA methylation studies (24,27), research on allopolyploid species using WGBS is limited, with most of the studies coming from crop study systems (28–30). On the one hand, these systems have excellent genomic resources to provide valuable insights, while on the other, it is unclear whether these insights can be extended to wild organisms in nature given their artificial selection (4).

In other polyploid study systems, two major challenges prevent the use of WGBS: limited genomic resources (i.e. genome assemblies) and a laborious data analysis process. The number of plant genome assemblies has been increasing exponentially in the last years (31), but polyploid genome assemblies are still an intensive, complex and expensive task (32,33), preventing the development of genetic and epigenetic studies using polyploids. For allopolyploids, this obstacle can be avoided by using the genome assemblies of the two (known) parent species (34), usually diploid.
Besides limited genomic resources, another challenge in WGBS comes from a laborious and complex data analysis process (35–37). In standard WGBS data analysis pipelines, complexities related to polyploids are often not taken into account. For example when mapping reads originating from an allopolyploid, high sequence similarity between parents can be challenging for read mapping algorithms (38,39) and the outcome can have strong bias, especially when the quality of the assemblies is asymmetric (40). To tackle this problem, several methods were developed to improve the categorization of allopolyploids' reads to the correct parental genome. HomeoRoq (41) and PolyDog (40) take into account alignment quality from both parental genomes to assign reads, while PolyCat (42) and EAGLE-RC (34) also use explicit genotype differences between parent genomes to classify reads. EAGLE-RC outperformed HomeoRoq in estimating homeolog expression with data from tetraploid *Arabidopsis* and hexaploid wheat (34). When comparing EAGLE-RC and PolyCat using *Gossypium* RNA-seq data, both tools outperformed other pipelines and had similar performance (43). Among all the tools, only PolyCat supports bisulfite-treated WGBS data, but only with available variant information (i.e. SNPs) between subgenomes, which represents an additional obstacle for most allopolyploid systems (44).

To promote and support allopolyploid DNA methylation research, we developed the Automated Reproducible Polyploid EpiGenetic GuIdance workflOw (ARPEGGIO). ARPEGGIO is a specialized workflow to process raw WGBS data utilizing the assemblies of the allopolyploid's parent species (hereafter referred to as progenitors) or independently phased subgenomes of an allopolyploid. ARPEGGIO includes all the steps from raw WGBS data to a list of genes showing differential methylation: conversion check, quality check, trimming, alignment, read classification, methylation extraction, statistical analysis and downstream analysis. More details about the prerequisites, setup, tools and outputs are discussed in the implementation section.

To handle sequence similarity between two genomes, ARPEGGIO exploits an updated version of EAGLE-RC that supports bisulfite-treated reads and does not require variant information between subgenomes. This version of EAGLE-RC was evaluated using three WGBS datasets, and showed better performance compared to a genome concatenation approach.

ARPEGGIO's implementation combines the Snakemake workflow management system (45) with the Conda package manager (46) and Singularity containers (47) to ensure both ease of use and reproducibility. For ease of use, a centralized configuration file controls all parameters related to ARPEGGIO and through Conda, all the tools required by the workflow are automatically installed.

Implementation

Design, concepts and challenges

ARPEGGIO's design had three main objectives, each dealing with different aspects and challenges of the workflow: allopolyploid support, ease of use and reproducibility. These aspects will be discussed at high-level here and more details about their implementation can be found in the following sections.

To support allopolyploids, ARPEGGIO first needed to allow for different experimental designs (i.e. sample comparisons). For allopolyploids without a genome assembly, but progenitor assemblies available, there are two possible comparisons: allopolyploid against progenitors or allopolyploid against allopolyploid (Fig. 1a,b). The former compares the two allopolyploid's subgenomes to the progenitors, while the latter compares directly the two subgenomes in different experimental conditions. An additional third comparison allows two groups of individuals from a species with an available (phased) genome assembly (Fig. 1c), regardless of the ploidy level. After choosing a comparison, the next allopolyploid-specific step is read classification.

To analyze allopolyploid data with progenitor assemblies, we run two separate workflows in parallel, one per progenitor (Fig. 2). The separation occurs at the alignment and deduplication step, where two separate alignments are performed for the same allopolyploid data, one for each progenitor. With each allopolyploid read being mapped twice, a read classification algorithm must choose one of the two progenitors; tor the classification, ARPEGGIO uses EAGLE-RC. In short, EAGLE-RC applies a probabilistic method that compares the two mappings for each read and classifies its progenitor origin or deems it ambiguous (equal probabilities for both progenitors sides). Two parameters were added to EAGLE-RC to deal with bisulfite data from allopolyploids. The first is called "no genotype information" (NGI) and allows EAGLE-RC to be used with no information about variants in the genome. This mode is especially useful to reduce prerequisites for using ARPEGGIO. The second parameter is called "bisulfite" (BS) and it causes bisulfite treatment to be taken into

account when a bisulfite-treated read is mapped to a genome. This parameter considers C-T as a match (forward strand), G-A as a match (reverse strand) or both.

Both experimental design and EAGLE-RC's inclusion had a major impact on ARPEGGIO's structure and implementation, but other important aspects were also taken into account. For example, allopolyploids can be found in different lineages such as plants and mammals, meaning that different approaches should be considered for conversion efficiency checks and the selection of methylation contexts.

Once the general design of ARPEGGIO was established, the next challenge was to make the workflow easy to set up, run and interpret. ARPEGGIO requires the users to install the Conda package management system (46), then Snakemake (45) via Conda and, optionally, Singularity (48). No other tools need to be installed as ARPEGGIO will take care of automatically installing what is needed. To prepare ARPEGGIO for a new dataset, input files have to be prepared and ARPEGGIO's settings have to be defined. Input files include raw data in FASTQ format and the progenitors' reference genome assemblies. To run downstream analyses, annotation files for both assemblies are also required. ARPEGGIO's settings are defined with a configuration file and a metadata file. The configuration file has different sections, each including parameters that define how ARPEGGIO will be run, while the metadata file contains information about samples such as filename, sequencing strategy, origin (allopolyploid or progenitor) and experimental condition (if present). A small dataset with its own configuration and metadata file are provided in ARPEGGIO's repository as an example. To run ARPEGGIO, only one command is needed and its main options are related to reproducibility (discussed below) and parallelization (i.e. multiple core usage). After ARPEGGIO is successfully run, the number of files in the output folder can be significant. For this reason, a map of the output is available in ARPEGGIO's user documentation: this map shows the general output structure with all the main folders and their contents. For each folder, there's a section describing the folder itself, sub-folders and all the files included in it.

Another key goal of ARPEGGIO was to ensure reproducibility. Considering the variety of tools and the amount of steps in the workflow, by letting users (or Conda) define the version of each tool, the outcome could be variable and lead to future reproducibility problems. To overcome this, we fixed all the versions of the tools and we combined ARPEGGIO with Conda and Singularity containers. The user can choose to use either only Conda or Conda and Singularity together. The main difference between the two modes lies on potential issues between the user's system and Conda. When these issues happen, Singularity offers a containerized run of Conda. Both these options can be specified with one or two parameters respectively when running ARPEGGIO. Aside from tool version differences, which we addressed above, the configuration file specifies all parameters that were used in a workflow run. Associating results to a specific set of parameters further aids reproducibility. The configuration file may also be shared to other researchers aiming to reanalyze a given dataset.

Workflow overview

ARPEGGIO includes eight processes: conversion check, quality checks, trimming, alignment and deduplication, read classification, methylation extraction, differential methylation analysis and downstream analyses (Fig. 2). These processes are divided into six steps, each represented by a black diamond in Fig. 2. Step 1 includes conversion check, a quality check specific to WGBS data, where reads are aligned to an unmethylated control genome (usually plastid genome for plants and lambda genome for others) to assess the efficiency of the bisulfite conversion; the lower the mapping rate, the better the conversion (27). This process is executed by Bismark (49). The conversion check is followed by quality checks and trimming (step 1 and 2), executed by FastQC (50) and Trim Galore (51), respectively. Both processes are common procedures to assess read quality and remove noise. Step 3 performs read alignment to a reference genome, followed by deduplication, which removes duplicated reads. Both of these are carried out by the Bismark suite (49). From this point of the workflow allopolyploid data is separated into two parallel workflows: one per progenitor side. These workflows intersect in the next, allopolyploid-specific read classification step (step 4), executed by the updated version of EAGLE-RC (34). Here, EAGLE-RC will classify allopolyploid reads after comparing the read alignment on each progenitor's side. After read classification (from step 5 on), the two workflows are independent, but execute the same steps. During methylation extraction via Bismark, methylation information is extracted for each cytosine from classified reads to produce a methylation count table. This table is used for differential methylation analyses (step 5), performed by the R/Bioconductor package `dmrseq` (52), to output a list of tested differentially methylated regions (DMRs). Finally, downstream analyses (step 6) consist of a series of R scripts for computing overlaps between statistically significant DMRs and annotated gene regions provided by the user (if available). More specifically, by default ARPEGGIO uses q-value < 0.05 to define a significant DMR. With this cutoff, ARPEGGIO looks for overlaps of at least 1 base pair between significant regions and gene regions based on the annotations. Before ARPEGGIO finishes a run, all reports (conversion check, quality checks, trimming, alignment, deduplication and methylation extraction) are combined into one interactive HTML report with MultiQC (53).

Each part in ARPEGGIO is optional and the user can specify which parts of the workflow to execute in the configuration file. It must be noted that skipping some parts will stop the workflow at a specific step (Fig. 2). Assuming that all prerequisites are met, ARPEGGIO goes from raw sequencing data to a list of genes showing differential methylation. Some useful intermediate outputs are also produced: an interactive HTML report merging all quality, alignment and methylation reports and an Rdata file with the output from the dmrseq analysis, which can be used to visualize DMRs or for other custom analyses.

Implementation details

ARPEGGIO is written in Snakemake, a Python based language for workflow development (45). With Snakemake, a workflow is broken down into a series of rules. One rule can be seen as one step in the workflow with a defined input and

output. Rules are related to each other based on their input and output files. Once all the rules are set, to run a Snakemake workflow, a target file (or multiple) needs to be requested. Snakemake will automatically build the workflow to obtain the target file based on the input/output relationships between rules (dependencies). If the relationships are successfully established, the workflow will be run. To illustrate these principles, an example with ARPEGGIO's rules is given in Additional File 1. This figure shows all the input/output relationships between rules when running ARPEGGIO with single-end data, comparing an allopolyploid to its progenitor species (default experimental design).

In addition to the core features of Snakemake, ARPEGGIO takes advantage of the integrated Conda package management system (46). Conda creates environments containing a specific set of software and users can switch between different environments depending on the software package(s) they need. An environment can be created in several ways. ARPEGGIO creates environments through YAML files, specifying all the packages to be included and the channels from which the packages are searched. The integration of Conda in Snakemake allows rules to be run within a specific environment and during the execution of a workflow, Snakemake takes care of switching between environments if different rules require different environments. From a user perspective, once Conda and Snakemake are installed, ARPEGGIO will take care of installing all the tools needed for the analyses, running them and switching automatically between environments when needed (Fig. 2).

Making the workflow specific for allopolyploids presented major challenges with both Snakemake and Conda. Snakemake rules in ARPEGGIO had to be structured to allow for any combination between sequencing strategies and experimental designs. This meant combining rules for six workflows in one: three experimental designs, each with two sequencing strategies. In addition, since EAGLE-RC could not be installed as a Conda package, a Conda environment with a specific set of rules was created to take care of downloading, extracting and installing EAGLE-RC.

In practice, any user can take advantage of all the Conda and Snakemake features discussed above with a central configuration file. Here, we will discuss the first three sections of this file, that consist of parameters concerning the workflow as a whole: general parameters, conditional rules and experimental designs. All the other sections in the configuration file are related to tool-specific parameters for each of the main steps in ARPEGGIO. More details about these parameters can be found in ARPEGGIO's user documentation. General parameters include the location of the output folder, the location of the metadata file and a parameter to define the sequencing strategy. Conditional rules are shown as black diamonds on Fig. 2. Those rules are set to "True" or "False" to define which parts of the workflow to run. Practically, only the initial steps of ARPEGGIO, quality check and trimming, can be skipped; otherwise, the workflow will stop for any other step that is set to "False". Finally, experimental designs are implemented via special modes. By default, ARPEGGIO compares a polyploid species against its two progenitor species (Fig. 1a). With the special mode "POLYPLOID_ONLY", ARPEGGIO compares a polyploid species from two different experimental conditions (Fig. 1b), while the mode

"DIPLOID_ONLY" compares a diploid species from two different conditions (or a polyploid species with an available phased assembly, Fig. 1c).

Results & Discussion

Performance of read classification

A simple and common way to analyze polyploid datasets is to concatenate the genome assemblies of the two progenitor species and let the aligner assign a mapping position. The position would define the origin of the read depending on which of the two subgenomes the read was mapped to. We define this approach as the "concatenated" approach.

The performance of EAGLE-RC was assessed using ARPEGGIO while shell scripts were used to evaluate the concatenated approach (see Availability of data and materials). In both cases, the same versions of tools as in ARPEGGIO were used.

For the evaluation, we used six datasets from three pairs of progenitor species that form an allopolyploid or a hybrid, and we compared EAGLE-RC's classification error to that of the concatenated approach in a similar fashion as (34). In short, each progenitor dataset was treated as an allopolyploid dataset, meaning that all the reads were assigned to a progenitor's side. With datasets coming from progenitors, the true origin of the reads was known, thus reads assigned to the wrong progenitor's side were used to calculate a classification error rate.

Two datasets were from *Mimulus guttatus* and *Mimulus luteus*, obtained from (54), with four technical replicates each. Those two species are the progenitors of the allopolyploid *Mimulus peregrinus*. Data from *Gossypium arboreum* and *Gossypium raimondii* was obtained from (29) and consisted of two technical replicates each. Those two species are the progenitors of the hybrid *Gossypium arboreum x raimondii*. The last datasets were produced in-house (Additional File 3) from *Arabidopsis halleri* and *Arabidopsis lyrata* with two biological replicates each. Those two species are the progenitors of the allopolyploid *Arabidopsis kamchatica* (55).

EAGLE-RC showed a lower error rate in all datasets compared to the concatenated approach (Table 1). The error rate was consistently between 3 to 4 times less with EAGLE-RC. When looking at absolute values, the improvement from read classification varied: from changes below 0.1% in *Gossypium* to almost 20% when using *Mimulus* data. These differences could be attributed to many factors, such as divergence between diploids, quality of genome assembly, and sequence data quality. From a qualitative point of view, *Mimulus* had lower quality assemblies compared to the other species, and this difference might also explain the higher error rates in both methods. Also, even though the *Gossypium* data was treated as allopolyploid, the large divergence between the two *Gossypium* species, particularly in terms of genome size, made the read classification task easier (Additional File 4).

Overall, EAGLE-RC showed a lower error rate with minimal loss of reads classified as ambiguous (Additional File 4). On the one hand, EAGLE-RC showed a lower error rate, while on the other, the absolute number of correctly assigned reads was lower in EAGLE-RC compared to the concatenated approach (Additional File 4). This happened because the reads classified as “ambiguous” reduced the amount of the correctly classified reads (both true negative and true positive reads). When focusing on the difference in true positive reads between EAGLE-RC and concatenation, values are negligible for both *Arabidopsis* and *Gossypium* datasets, representing <0.01% of uniquely mapped reads. In the case of *Mimulus*, the number of true positive reads is ~10% higher in the concatenated approach, but the error-rate is also 3 to 4 times higher compared to EAGLE-RC. Taken together, these results suggest that EAGLE-RC has a clear advantage when analyzing allopolyploid WGBS data, where higher accuracy in subgenome recognition is required.

In this evaluation, we have not examined in detail the effect of the genetic divergence between progenitor genomes and allopolyploid genomes. Divergence results from DNA mutations happening after polyploidization and leading to changes on both progenitor sides in the polyploid’s genome. The amount of differences is proportional to the number of generations, i.e. time, since polyploidization. As an example *M. peregrinus* is a 140-years old polyploid, and thus the changes in its genome might be very few. We speculate that ARPEGGIO should be tolerant for older allopolyploids, as both EAGLE-RC and HomeoRoq have shown good performance with both DNA and RNA-seq data of A. kamchatica, which is estimated to have originated around 20,000-250,000 years ago (41,56,57).

Example run with Mimulus data

To illustrate a full run of ARPEGGIO, we analyzed publicly available data coming from the natural allopolyploid *Mimulus peregrinus* and its progenitors *M. guttatus* and *M. luteus* (58).

First, we downloaded the raw WGBS data consisting of four technical replicates for each species, the genome assemblies of the progenitors with their annotation and a chloroplast genome to check conversion efficiency (details in Availability of data and materials). For WGBS data, genome assemblies and annotations we made sure that all files were formatted according to ARPEGGIO’s user guidelines.

Second, we created a metadata file specifying for each sample the sequencing strategy, single end, and the origin of the samples, i.e. *M. guttatus* samples were labeled “parent1”, *M. luteus* samples “parent2” and *M. peregrinus* samples “allopolyploid”.

With the input files ready, the configuration file was set up in two rounds. In the first round the general parameters were configured with the locations of output folder and metadata file, and data was specified as single end. By default, ARPEGGIO compares allopolyploid to progenitors (Fig 1a), meaning that no specific changes needed to be done to include the experimental design for this dataset. Then, all

conditional rules were set to false and ARPEGGIO was run to only perform quality checks. With this round we were able to get more details for the trimming step. In the second round, all the parameters were set for all the different steps in the workflow and all conditional rules were set to true to perform a full run of ARPEGGIO with eight cores. The configuration file, the MultiQC report and ARPEGGIO's output for the statistical and downstream analyses can be found in Availability of data and materials. The runtime of the full run on a Debian system, using eight CPU cores Intel(R) Xeon(R) CPU E5-4640 at 2.40GHz was approximately 24 hours. The average times for each step can be found in Additional File 2.

After comparing the methylation pattern of *M. peregrinus* to its progenitors, a total of 760 significant DMRs were found in the allopolyploid, most of them coming from the *M. luteus* side (Table 2). Downstream analyses found very few genes overlapping with these significant regions, suggesting that most of the methylation changes occur in intergenic rather than genic regions. For the *M. guttatus* side, 35 genes were found, mostly associated with changes in CG and CHG context, while for the *M. luteus* side only 2 genes were found in CG context. These genes represent a very small proportion of the total number of annotated genes in *M. guttatus*, almost 30'000, and *M. luteus*, almost 50'000. Taken all together, these results suggest almost no change in the global methylation pattern of genes in the natural allopolyploid compared to the two progenitors.

Our analyses use a different approach and different tools compared to (54), but Edger and colleagues also looked at changes in methylation pattern from progenitor to allopolyploid. The authors observed were similar methylation patterns within gene bodies, when comparing progenitors to natural allopolyploids. This is consistent with ARPEGGIO's downstream analyses showing few genes overlapping with DMRs. Additionally, further analyses in (54) showed that most of the methylation changes happened in transposable elements, another result in agreement with the number of intergenic DMRs found by ARPEGGIO.

User's experience and best practices

ARPEGGIO's user documentation, available through the GitHub Wiki, offers additional information for more and less experienced users. For less experienced users, the documentation offers a step by step guide of how to setup and run ARPEGGIO on a given dataset: data and system requirements, input files needed, configuration file instructions, commands to run the workflow and a map of the output structure. For experienced users, we tried to be as transparent as possible about ARPEGGIO's code and its architecture to make any customization of scripts and code easier.

As a whole, ARPEGGIO is meant to simplify reproducible data analysis, but best practices, such as data diagnostics and information sharing should be kept in mind. The complete ARPEGGIO pipeline should be run once data quality and potential sources of errors are assessed. To have more control over the analysis process, users also have the option to run ARPEGGIO steps one by one. By modifying the

configuration file to add further steps, the workflow will rerun only the parts that need to be updated. To ensure reproducibility when using ARPEGGIO, there are three specifications that need to be included with the datasets: the configuration file settings, the metadata file and the version of ARPEGGIO.

Software choice

Many alternative tools exist to perform some of ARPEGGIO's steps. For example, several aligners exist for short-read bisulfite sequencing data such as `bwa-meth` (59), `BSmap` (60), `BitMapperBS` (61), `SNAP` (62) and `gemBS` (63). The `Bismark` suite was selected because it included tools to perform alignment, deduplication and methylation extraction for any context all in one centralized package. Most if not all of the other aligners depend on external packages for downstream analyses of alignment files.
Similarly, many tools exist for DMRs discovery in whole-genome bisulfite sequencing data for all methylation contexts: `BSmooth` (64), `metilene` (65), `MOABS` (66), `BiSeq` (67), `MethylKit` (68) and others (69).
In the case of `dmrseq`, the tool was chosen because of its two step approach: first selecting candidate regions and then evaluating their statistical significance by taking into account both biological variability and spatial correlation. This approach offers important advantages such as limited loss of power and better FDR control, both critical aspects when detecting DMRs (70).
The selection of an appropriate alignment or statistical tool for WGBS data would require an independent benchmark of such tools. An ideal benchmark should evaluate tools on a variety of conditions and provide some guidelines about their suitability and use. Currently, no such benchmarks exist, and a thorough evaluation was out of the scope of this paper. ARPEGGIO provides a convenient implementation of the selected tools and its architecture allows future modifications as long as the input/output structure of the Snakemake rules is preserved.
This means that if any of the tools included in the workflow are shown to be underperforming compared to others, ARPEGGIO can be adapted accordingly.

Comparison to other workflows

To compare ARPEGGIO to other workflows, we selected key steps specifically related to WGBS data analysis (Table 3). The results included workflows able to work with raw bisulfite reads from WGBS and excluded highly specialized (i.e. alignment only or downstream only) and commercial workflows.

ARPEGGIO is the only workflow specifically targeted at polyploids, making it the main unique feature compared to other available workflows. Other features that were lacking in other workflows, but present in ARPEGGIO, were downstream analyses and reproducibility. Around half of the workflows investigated included downstream analyses (71–75). The lack of this feature might be due to downstream analyses being highly variable according to biological context, question, and aim of the research. With ARPEGGIO, the aim was to consolidate performant tools into a common approach that could be used as a start for further investigation; in our case

downstream analyses leading to a list of genes. Reproducibility was another main feature present in ARPEGGIO that was lacking in many workflows, but appeared to be more prevalent in more recent publications (71,75–77). Enhancing and promoting reproducibility is essential to ensure that discoveries stand the test of time (78). Other features were very similar across workflows. All workflows support diploid data, which is considered the same as polyploid data with an available polyploid phased assembly. When comparing the presence of quality check, alignment and statistical analyses, most workflows included them all together, but some didn't include either quality check (74–76) or statistical analyses (79). For methylation contexts, only two workflows focused on CpG context only (77,80), while all the other allowed analyses for all contexts (CpG, CHG and CHH).

One feature not implemented in ARPEGGIO, but present in other workflows, is visualization of DMRs. This step, similar to downstream analyses, is highly context dependent. The `dmrseq` package offers ways to visualize DMRs, but this was not included in ARPEGGIO. Instead, the workflow outputs an `Rdata` file with all information concerning DMRs that users can use in their custom analyses. It is important to stress that visualization is essential for high-throughput data analysis, and should happen at any step in the data analysis process.

It is important to note that Table 3 focuses only on features related to WGBS data analysis, the only data type supported by ARPEGGIO. Some of the workflows support additional data types and analyses: QuasR supports ChIP-seq, RNA-seq, smRNA-seq and allele-specific data analyses, RUBioSeq supports single-nucleotide and copy number variants (SNVs and CNVs) analyses and snakePipes supports simple DNA-mapping, ChIP-seq, ATAC-seq, HiC, RNA-seq and scRNA-seq data.

Overall, ARPEGGIO was the only workflow supporting polyploid data, and among all the different aspects considered, one of the few workflows including downstream analyses that explicitly set reproducibility as one of its main goals.

Conclusions

Research on DNA methylation in allopolyploids at a whole genome level seems to be favoring established allopolyploid species (i.e. crops). This can be partially attributed to two factors: 1) challenges in generating allopolyploid genome assemblies; and, 2) a laborious data analysis process. Here we presented ARPEGGIO: the first workflow for the analysis of allopolyploid WGBS data. ARPEGGIO includes a read classification algorithm, EAGLE-RC, to assign allopolyploid reads to the correct progenitor's side. EAGLE-RC showed better performance against a common concatenation for six different WGBS datasets. Read classification is part of a full set of analyses included in ARPEGGIO, going from raw sequencing data up to a list of genes showing differential methylation. The implementation of ARPEGGIO aimed at ease of use and reproducibility, both essential factors to have an accessible yet up-to-standard tool.

With ARPEGGIO, we provide a first step towards a future of standardized tools and workflows in polyploid research.

Availability and requirements

Project name: ARPEGGIO
Project home page: https://github.com/supermaxiste/ARPEGGIO
Operating system: Linux
Programming language: Python and R
Other requirements: Python 3, Conda, [Singularity]
License: GNU GPL v3.0

List of abbreviations

WGBS = Whole genome bisulfite sequencing
DMRs = Differentially methylated regions

Consent for publication

Not applicable

Availability of data and materials

Data from cotton taken from (29), available in the NCBI Nucleotide and Sequence Read Archive (SRA) under [SRA:SRP071640]. Data from *Mimulus* taken from (58), available in the NCBI Gene Expression Omnibus (GEO) under [GSE95799]. Data from *Arabidopsis* available in the DDBJ Sequence Read Archive (DRA) under [DRA009902].
The *Gossypium raimondii* v2.0 genome assembly (81) and *Mimulus guttatus* v2.0 (82) genome assembly and annotation were downloaded from Pythozome v12.1 (83). The *Gossypium arboreum* v2_a1 (84) genome assembly was downloaded from CottonGen (85). The *Mimulus luteus* (54) assembly and its annotation were downloaded from Dryad (58). The *Arabidopsis halleri* v2.2 genome assembly was taken from (86) and the *Arabidopsis lyrata* v2.2 genome assembly was taken from (57).
The scripts used for the evaluation of EAGLE-RC and genome concatenation together with the details about the *Mimulus* example run can be found on: https://github.com/supermaxiste/ARPEGGIO_paperAnalyses

Competing interests

The authors declare that they have no competing interests.

Funding

This work was supported by the University Research Priority Program (URPP) Evolution in Action of the University of Zurich.

Authors contributions

SM started the bisulfite treatment and library preparation protocol optimization, wrote most of the ARPEGGIO code and manuscript and tested EAGLE-RC. TK updated EAGLE-RC to a new version supporting bisulfite sequencing data and contributed to describe the model with its new features. TK, SD and MR tested ARPEGGIO on their own devices. SD helped with bug fixing in ARPEGGIO, optimized Conda support and implemented Singularity support. RSI managed, coordinated and optimized the bisulfite treatment protocol, library preparation and sequencing process of the *Arabidopsis* samples. LM executed the bisulfite treatment and library preparation. JS, MR, RSI and KS supervised the project and provided important insights at several stages of the project. All authors read and approved the final manuscript.

Acknowledgements

We thank A. Morishima and M. Wyler for all the support in the DNA extraction, bisulfite treatment and library preparation steps and the optimization of the protocol; the Functional Genomic Center Zurich and M. Hatakeyama for sequencing and data handling; the URPP Evolution in Action program for the opportunity to present ARPEGGIO; the Robinson and Shimizu group members for the feedback during different stages of the project, in particular R. Huang, K. Hembach, S. Orjuela and C. Soneson for the support with Snakemake and the workflow development process.

References

1. Van de Peer Y, Mizrachi E, Marchal K. The evolutionary significance of polyploidy. Nat Rev Genet [Internet]. 2017 Jul 15;18(7):411–24. Available from: http://www.nature.com/articles/nrg.2017.26
2. Blischak PD, Mabry ME, Conant GC, Pires JC. Integrating Networks, Phylogenomics, and Population Genomics for the Study of Polyploidy. Annu Rev Ecol Evol Syst [Internet]. 2018 Nov 2;49(1):253–78. Available from: https://www.annualreviews.org/doi/10.1146/annurev-ecolsys-121415-032302
3. One thousand plant transcriptomes and the phylogenomics of green plants. Nature [Internet]. 2019 Oct 23;574(7780):679–85. Available from: http://www.nature.com/articles/s41586-019-1693-2
4. Soltis DE, Visger CJ, Marchant DB, Soltis PS. Polyploidy: Pitfalls and paths to a paradigm. Am J Bot [Internet]. 2016 Jul;103(7):1146–66. Available from: http://doi.wiley.com/10.3732/ajb.1500501
5. Soltis PS, Soltis DE. Ancient WGD events as drivers of key innovations in angiosperms. Curr Opin Plant Biol [Internet]. 2016 Apr;30:159–65. Available from: https://linkinghub.elsevier.com/retrieve/pii/S1369526616300425
6. Clark JW, Donoghue PCJ. Whole-Genome Duplication and Plant Macroevolution. Trends Plant Sci [Internet]. 2018 Oct;23(10):933–45. Available from: https://linkinghub.elsevier.com/retrieve/pii/S1360138518301596
7. Wood TE, Takebayashi N, Barker MS, Mayrose I, Greenspoon PB, Rieseberg LH. The frequency of polyploid speciation in vascular plants. Proc Natl Acad Sci [Internet]. 2009 Aug 18;106(33):13875–9. Available from: http://www.pnas.org/cgi/doi/10.1073/pnas.0811575106
8. Mayrose I, Zhan SH, Rothfels CJ, Magnuson-Ford K, Barker MS, Rieseberg LH, et al. Recently Formed Polyploid Plants Diversify at Lower Rates. Science

(80-) [Internet]. 2011 Sep 2;333(6047):1257–1257. Available from: https://www.sciencemag.org/lookup/doi/10.1126/science.1207205

9. Soltis DE, Buggs RJA, Barbazuk WB, Chamala S, Chester M, Gallagher JP, et al. The Early Stages of Polyploidy: Rapid and Repeated Evolution in Tragopogon. In: Polyploidy and Genome Evolution [Internet]. Berlin, Heidelberg: Springer Berlin Heidelberg; 2012. p. 271–92. Available from: http://link.springer.com/10.1007/978-3-642-31442-1_14
10. Chen ZJ. Genetic and Epigenetic Mechanisms for Gene Expression and Phenotypic Variation in Plant Polyploids. Annu Rev Plant Biol [Internet]. 2007 Jun;58(1):377–406. Available from: http://www.annualreviews.org/doi/10.1146/annurev.arplant.58.032806.103835
11. Wendel JF, Lisch D, Hu G, Mason AS. The long and short of doubling down: polyploidy, epigenetics, and the temporal dynamics of genome fractionation. Curr Opin Genet Dev [Internet]. 2018 Apr;49:1–7. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0959437X17301557
12. Wendel JF. Genome evolution in polyploids. In: Plant Molecular Evolution [Internet]. Dordrecht: Springer Netherlands; 2000. p. 225–49. Available from: http://link.springer.com/10.1007/978-94-011-4221-2_12
13. Madlung A, Masuelli RW, Watson B, Reynolds SH, Davison J, Comai L. Remodeling of DNA Methylation and Phenotypic and Transcriptional Changes in Synthetic Arabidopsis Allotetraploids. Plant Physiol [Internet]. 2002 Jun 1;129(2):733–46. Available from: http://www.plantphysiol.org/lookup/doi/10.1104/pp.003095
14. Salmon A, Ainouche ML, Wendel JF. Genetic and epigenetic consequences of recent hybridization and polyploidy in Spartina (Poaceae). Mol Ecol. 2005;14(4):1163–75.
15. Xu Y, Zhong L, Wu X, Fang X, Wang J. Rapid alterations of gene expression and cytosine methylation in newly synthesized Brassica napus allopolyploids. Planta. 2009;229(3):471–83.
16. Shaked H, Kashkush K, Ozkan H, Feldman M, Levy AA. Sequence Elimination and Cytosine Methylation Are Rapid and Reproducible Responses of the Genome to Wide Hybridization and Allopolyploidy in Wheat. Plant Cell [Internet]. 2001 Aug;13(8):1749–59. Available from: http://www.plantcell.org/lookup/doi/10.1105/TPC.010083
17. Sehrish T, Symonds VV, Soltis DE, Soltis PS, Tate JA. Gene silencing via DNA methylation in naturally occurring Tragopogon miscellus (Asteraceae) allopolyploids. BMC Genomics. 2014;15(1):1–7.
18. RAN L, FANG T, RONG H, JIANG J, FANG Y, WANG Y. Analysis of cytosine methylation in early generations of resynthesized Brassica napus. J Integr Agric [Internet]. 2016 Jun;15(6):1228–38. Available from: https://linkinghub.elsevier.com/retrieve/pii/S2095311915612771
19. Bao Y, Xu Q. Extensive reprogramming of cytosine methylation in Oryza allotetraploids. Genes Genomics [Internet]. 2015 Jun 28;37(6):517–24. Available from: http://link.springer.com/10.1007/s13258-015-0279-0
20. Parisod C, Salmon A, Zerjal T, Tenaillon M, Grandbastien M-A, Ainouche M. Rapid structural and epigenetic reorganization near transposable elements in hybrid and allopolyploid genomes in Spartina. New Phytol [Internet]. 2009 Dec;184(4):1003–15. Available from: http://doi.wiley.com/10.1111/j.1469-

8137.2009.03029.x

21. Wang J, Tian L, Madlung A, Lee H-S, Chen M, Lee JJ, et al. Stochastic and Epigenetic Changes of Gene Expression in Arabidopsis Polyploids. Genetics [Internet]. 2004 Aug;167(4):1961–73. Available from: http://www.genetics.org/lookup/doi/10.1534/genetics.104.027896
22. Gaeta RT, Pires JC, Iniguez-Luy F, Leon E, Osborn TC. Genomic Changes in Resynthesized Brassica napus and Their Effect on Gene Expression and Phenotype. Plant Cell [Internet]. 2007 Nov;19(11):3403–17. Available from: http://www.plantcell.org/lookup/doi/10.1105/tpc.107.054346
23. Liu B, Brubaker CL, Mergeai G, Cronn RC, Wendel JF. Polyploid formation in cotton is not accompanied by rapid genomic changes. Genome [Internet]. 2001 Jun 1;44(3):321–30. Available from: http://www.nrcresearchpress.com/doi/10.1139/g01-011
24. Kurdyukov S, Bullock M. DNA Methylation Analysis: Choosing the Right Method. Biology (Basel) [Internet]. 2016 Jan 6;5(1):3. Available from: http://www.mdpi.com/2079-7737/5/1/3
25. Laird PW. Principles and challenges of genome-wide DNA methylation analysis. Nat Rev Genet [Internet]. 2010 Mar 2;11(3):191–203. Available from: http://www.nature.com/articles/nrg2732
26. Law JA, Jacobsen SE. Establishing, maintaining and modifying DNA methylation patterns in plants and animals. Nat Rev Genet [Internet]. 2010 Mar 9;11(3):204–20. Available from: http://www.nature.com/articles/nrg2719
27. Urich MA, Nery JR, Lister R, Schmitz RJ, Ecker JR. MethylC-seq library preparation for base-resolution whole-genome bisulfite sequencing. Nat Protoc [Internet]. 2015;10(3):475–83. Available from: http://dx.doi.org/10.1038/nprot.2014.114
28. Li N, Xu C, Zhang A, Lv R, Meng X, Lin X, et al. DNA methylation repatterning accompanying hybridization, whole genome doubling and homoeolog exchange in nascent segmental rice allotetraploids. New Phytol [Internet]. 2019 Jul 30;223(2):979–92. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1111/nph.15820
29. Song Q, Zhang T, Stelly DM, Chen ZJ. Epigenomic and functional analyses reveal roles of epialleles in the loss of photoperiod sensitivity during domestication of allotetraploid cottons. Genome Biol [Internet]. 2017 Dec 31;18(1):99. Available from: http://genomebiology.biomedcentral.com/articles/10.1186/s13059-017-1229-8
30. Bird KA, Niederhuth C, Ou S, Gehan M, Chris Pires J, Xiong Z, et al. Replaying the evolutionary tape to investigate subgenome dominance in allopolyploid Brassica napus bioRxiv [Internet]. 2019 Jan 1;814491. Available from: http://biorxiv.org/content/early/2019/10/22/814491.abstract
31. Kersey PJ. Plant genome sequences: past, present, future. Curr Opin Plant Biol [Internet]. 2019 Apr;48:1–8. Available from: https://linkinghub.elsevier.com/retrieve/pii/S1369526618300918
32. Claros MG, Bautista R, Guerrero-Fernández D, Benzerki H, Seoane P, Fernández-Pozo N. Why Assembling Plant Genome Sequences Is So Challenging. Biology (Basel) [Internet]. 2012 Sep 18;1(2):439–59. Available from: http://www.mdpi.com/2079-7737/1/2/439

33. Kyriakidou M, Tai HH, Anglin NL, Ellis D, Strömvik M V. Current Strategies of Polyploid Plant Genome Sequence Assembly. Front Plant Sci [Internet]. 2018 Nov 21;9. Available from: https://www.frontiersin.org/article/10.3389/fpls.2018.01660/full
34. Kuo TCY, Hatakeyama M, Tameshige T, Shimizu KK, Sese J. Homeolog expression quantification methods for allopolyploids. Brief Bioinform [Internet]. 2018 Dec 27; Available from: https://academic.oup.com/bib/advance-article/doi/10.1093/bib/bby121/5251019
35. Bock C. Analysing and interpreting DNA methylation data. Nat Rev Genet [Internet]. 2012 Oct 18;13(10):705–19. Available from: http://www.nature.com/articles/nrg3273
36. Yong W-S, Hsu F-M, Chen P-Y. Profiling genome-wide DNA methylation. Epigenetics Chromatin [Internet]. 2016;9(1):26. Available from: http://epigeneticsandchromatin.biomedcentral.com/articles/10.1186/s13072-016-0075-3
37. Wreczycka K, Gosdschan A, Yusuf D, Grüning B, Assenov Y, Akalin A. Strategies for analyzing bisulfite sequencing data. J Biotechnol [Internet]. 2017 Nov;261:105–15. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0168165617315936
38. Boatwright JL, McIntyre LM, Morse AM, Chen S, Yoo M-J, Koh J, et al. A Robust Methodology for Assessing Differential Homeolog Contributions to the Transcriptomes of Allopolyploids. Genetics [Internet]. 2018 Nov;210(3):883–94. Available from: http://www.genetics.org/lookup/doi/10.1534/genetics.118.301564
39. Gerard D, Ferrão LFV, Garcia AAF, Stephens M. Genotyping Polyploids from Messy Sequencing Data. Genetics [Internet]. 2018 Nov;210(3):789–807. Available from: http://www.genetics.org/lookup/doi/10.1534/genetics.118.301468
40. Page JT, Udall JA. Methods for mapping and categorization of DNA sequence reads from allopolyploid organisms. BMC Genet [Internet]. 2015;16(Suppl 2):S4. Available from: http://bmcgenet.biomedcentral.com/articles/10.1186/1471-2156-16-S2-S4
41. Akama S, Shimizu-Inatsugi R, Shimizu KK, Sese J. Genome-wide quantification of homeolog expression ratio revealed nonstochastic gene regulation in synthetic allopolyploid Arabidopsis. Nucleic Acids Res. 2014;42(6).
42. Page JT, Gingle AR, Udall JA. PolyCat: A Resource for Genome Categorization of Sequencing Reads From Allopolyploid Organisms. G3: Genes|Genomes|Genetics [Internet]. 2013 Mar;3(3):517–25. Available from: http://g3journal.org/lookup/doi/10.1534/g3.112.005298
43. Hu G, Grover CE, Arick MA, Liu M, Peterson DG, Wendel JF. Homoeologous gene expression and co-expression network analyses and evolutionary inference in allopolyploids. Brief Bioinform [Internet]. 2020 Mar 27; Available from: https://academic.oup.com/bib/advance-article/doi/10.1093/bib/bbaa035/5811916
44. GARVIN MR, SAITOH K, GHARRETT AJ. Application of single nucleotide polymorphisms to non-model species: a technical review. Mol Ecol Resour [Internet]. 2010 Nov;10(6):915–34. Available from:

http://doi.wiley.com/10.1111/j.1755-0998.2010.02891.x
45. Koster J, Rahmann S. Snakemake--a scalable bioinformatics workflow engine. Bioinformatics [Internet]. 2012 Oct 1;28(19):2520–2. Available from: https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/bts480
46. Anaconda. Anaconda Software Distribution [Internet]. 2014. Available from: https://anaconda.com
47. Merkel D. Docker: lightweight linux containers for consistent development and deployment. Linux J. 2014;2014(239):2.
48. Kurtzer GM, Sochat V, Bauer MW. Singularity: Scientific containers for mobility of compute. Gursoy A, editor. PLoS One [Internet]. 2017 May 11;12(5):e0177459. Available from: https://dx.plos.org/10.1371/journal.pone.0177459
49. Krueger F, Andrews SR. Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. Bioinformatics [Internet]. 2011 Jun 1;27(11):1571–2. Available from: https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btr167
50. Andrews S. FastQC: a quality control tool for high throughput sequence data [Internet]. 2010. Available from: https://www.bioinformatics.babraham.ac.uk/projects/fastqc/
51. Krueger F. Trim Galore [Internet]. 2012. Available from: http://www.bioinformatics.babraham.ac.uk/projects/trim_galore/
52. Korthauer K, Chakraborty S, Benjamini Y, Irizarry RA. Detection and accurate false discovery rate control of differentially methylated regions from whole genome bisulfite sequencing. Biostatistics [Internet]. 2019 Jul 1;20(3):367–83. Available from: https://academic.oup.com/biostatistics/article/20/3/367/4899074
53. Ewels P, Magnusson M, Lundin S, Käller M. MultiQC: summarize analysis results for multiple tools and samples in a single report. Bioinformatics [Internet]. 2016 Oct 1;32(19):3047–8. Available from: https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btw354
54. Edger PP, Smith R, McKain MR, Cooley AM, Vallejo-Marin M, Yuan Y, et al. Subgenome Dominance in an Interspecific Hybrid, Synthetic Allopolyploid, and a 140-Year-Old Naturally Established Neo-Allopolyploid Monkeyflower. Plant Cell [Internet]. 2017 Sep;29(9):2150–67. Available from: http://www.plantcell.org/lookup/doi/10.1105/tpc.17.00010
55. SHIMIZU-INATSUGI R, LIHOVÁ J, IWANAGA H, KUDOH H, MARHOLD K, SAVOLAINEN O, et al. The allopolyploid Arabidopsis kamchatica originated from multiple individuals of Arabidopsis lyrata and Arabidopsis halleri. Mol Ecol [Internet]. 2009 Oct;18(19):4024–48. Available from: http://doi.wiley.com/10.1111/j.1365-294X.2009.04329.x
56. Briskine R V., Paape T, Shimizu-Inatsugi R, Nishiyama T, Akama S, Sese J, et al. Genome assembly and annotation of Arabidopsis halleri , a model for heavy metal hyperaccumulation and evolutionary ecology. Mol Ecol Resour [Internet]. 2017 Sep;17(5):1025–36. Available from: http://doi.wiley.com/10.1111/1755-0998.12604
57. Paape T, Briskine R V., Halstead-Nussloch G, Lischer HEL, Shimizu-Inatsugi

R, Hatakeyama M, et al. Patterns of polymorphism and selection in the subgenomes of the allopolyploid Arabidopsis kamchatica. Nat Commun [Internet]. 2018 Dec 25;9(1):3909. Available from: http://www.nature.com/articles/s41467-018-06108-1

58. Edger PP, Smith RD, McKain MR, Cooley AM, Vallejo-Marin M, Yuan Y-W, et al. Data from: Subgenome dominance in an interspecific hybrid, synthetic allopolyploid, and a 140-year-old naturally established neo-allopolyploid monkeyflower [Internet]. Dryad; 2017. Available from: https://datadryad.org/stash/dataset/doi:10.5061/dryad.d4vr0
59. Pedersen BS, Eyring K, De S, Yang I V., Schwartz DA. Fast and accurate alignment of long bisulfite-seq reads. 2014 Jan 6; Available from: http://arxiv.org/abs/1401.1129
60. Xi Y, Li W. BSMAP: whole genome bisulfite sequence MAPping program. BMC Bioinformatics [Internet]. 2009 Dec 27;10(1):232. Available from: https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-10-232
61. Cheng H, Xu Y. BitMapperBS: a fast and accurate read aligner for whole-genome bisulfite sequencing. bioRxiv. 2019;
62. Zaharia M, Bolosky WJ, Curtis K, Fox A, Patterson D, Shenker S, et al. Faster and More Accurate Sequence Alignment with SNAP. 2011 Nov 23; Available from: http://arxiv.org/abs/1111.5572
63. Merkel A, Fernández-Callejo M, Casals E, Marco-Sola S, Schuyler R, Gut IG, et al. gemBS: high throughput processing for DNA methylation data from bisulfite sequencing. Hancock J, editor. Bioinformatics [Internet]. 2019 Mar 1;35(5):737–42. Available from: https://academic.oup.com/bioinformatics/article/35/5/737/5077236
64. Hansen KD, Langmead B, Irizarry RA. BSmooth: from whole genome bisulfite sequencing reads to differentially methylated regions. Genome Biol [Internet]. 2012;13(10):R83. Available from: http://genomebiology.biomedcentral.com/articles/10.1186/gb-2012-13-10-r83
65. Jühling F, Kretzmer H, Bernhart SH, Otto C, Stadler PF, Hoffmann S. metilene: fast and sensitive calling of differentially methylated regions from bisulfite sequencing data. Genome Res [Internet]. 2016 Feb;26(2):256–62. Available from: http://genome.cshlp.org/lookup/doi/10.1101/gr.196394.115
66. Sun D, Xi Y, Rodriguez B, Park H, Tong P, Meong M, et al. MOABS: model based analysis of bisulfite sequencing data. Genome Biol [Internet]. 2014;15(2):R38. Available from: http://genomebiology.biomedcentral.com/articles/10.1186/gb-2014-15-2-r38
67. Hebestreit K, Dugas M, Klein H-U. Detection of significantly differentially methylated regions in targeted bisulfite sequencing data. Bioinformatics [Internet]. 2013 Jul;29(13):1647–53. Available from: https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btt263
68. Akalin A, Kormaksson M, Li S, Garrett-Bakelman FE, Figueroa ME, Melnick A, et al. MethylKit: a comprehensive R package for the analysis of genome-wide DNA methylation profiles. Genome Biol [Internet]. 2012;13(10):R87. Available from: http://genomebiology.com/2012/13/10/R87
69. Shafi A, Mitrea C, Nguyen T, Draghici S. A survey of the approaches for

identifying differential methylation using bisulfite sequencing data. Brief Bioinform [Internet]. 2017;(January):1–17. Available from: https://academic.oup.com/bib/article/3064341/A

70. Robinson MD, Kahraman A, Law CW, Lindsay H, Nowicka M, Weber LM, et al. Statistical methods for detecting differentially methylated loci and regions. Front Genet [Internet]. 2014 Sep 16;5. Available from: http://journal.frontiersin.org/article/10.3389/fgene.2014.00324/abstract
71. Gaidatzis D, Lerch A, Hahne F, Stadler MB. QuasR: quantification and annotation of short reads in R. Bioinformatics [Internet]. 2015 Apr 1;31(7):1130–2. Available from: https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btu781
72. Song Q, Garvin T, Smith A, Qu J. The Smithlab DNA Methylation Data Analysis Pipeline (MethPipe) Methylome construction Mapping reads. 2014;1–19.
73. Liang F, Tang B, Wang Y, Wang J, Yu C, Chen X, et al. WBSA: Web Service for Bisulfite Sequencing Data Analysis. Pellegrini M, editor. PLoS One [Internet]. 2014 Jan 30;9(1):e86707. Available from: http://dx.plos.org/10.1371/journal.pone.0086707
74. Jiang P, Sun K, Lun FMF, Guo AM, Wang H, Chan KCA, et al. Methy-Pipe: An Integrated Bioinformatics Pipeline for Whole Genome Bisulfite Sequencing Data Analysis. Zhu D, editor. PLoS One [Internet]. 2014 Jun 19;9(6):e100360. Available from: http://dx.plos.org/10.1371/journal.pone.0100360
75. Lebrón R, Barturen G, Gómez-Martín C, Oliver JL, Hackenberg M. MethFlow^{VM}: a virtual machine for the integral analysis of bisulfite sequencing data. bioRxiv [Internet]. 2016 Jan 1;66795. Available from: http://biorxiv.org/content/early/2016/07/31/066795.abstract
76. Graña O, López-Fernández H, Fdez-Riverola F, González Pisano D, Glez-Peña D. Bicycle: a bioinformatics pipeline to analyze bisulfite sequencing data. Berger B, editor. Bioinformatics [Internet]. 2018 Apr 15;34(8):1414–5. Available from: https://academic.oup.com/bioinformatics/article/34/8/1414/4683461
77. Bhardwaj V, Heyne S, Sikora K, Rabbani L, Rauer M, Kilpert F, et al. snakePipes: facilitating flexible, scalable and integrative epigenomic analysis. Berger B, editor. Bioinformatics [Internet]. 2019 Nov 1;35(22):4757–9. Available from: https://academic.oup.com/bioinformatics/article/35/22/4757/5499080
78. Nekrutenko A, Taylor J. Next-generation sequencing data interpretation: enhancing reproducibility and accessibility. Nat Rev Genet [Internet]. 2012 Sep 17;13(9):667–72. Available from: http://www.nature.com/articles/nrg3305
79. Rubio-Camarillo M, Gomez-Lopez G, Fernandez JM, Valencia A, Pisano DG. RUbioSeq: a suite of parallelized pipelines to automate exome variation and bisulfite-seq analyses. Bioinformatics [Internet]. 2013 Jul 1;29(13):1687–9. Available from: https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btt203
80. Kumaki Y, Oda M, Okano M. QUMA: quantification tool for methylation analysis. Nucleic Acids Res [Internet]. 2008 May 19;36(Web Server):W170–5. Available from: https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkn294

81. Paterson AH, Wendel JF, Gundlach H, Guo H, Jenkins J, Jin D, et al. Repeated polyploidization of Gossypium genomes and the evolution of spinnable cotton fibres. Nature [Internet]. 2012 Dec 19;492(7429):423–7. Available from: http://www.nature.com/articles/nature11798
82. Hellsten U, Wright KM, Jenkins J, Shu S, Yuan Y, Wessler SR, et al. Fine-scale variation in meiotic recombination in Mimulus inferred from population shotgun sequencing. Proc Natl Acad Sci [Internet]. 2013 Nov 26;110(48):19478–82. Available from: http://www.pnas.org/cgi/doi/10.1073/pnas.1319032110
83. Goodstein DM, Shu S, Howson R, Neupane R, Hayes RD, Fazo J, et al. Phytozome: a comparative platform for green plant genomics. Nucleic Acids Res [Internet]. 2012 Jan;40(D1):D1178–86. Available from: https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkr944
84. Li F, Fan G, Wang K, Sun F, Yuan Y, Song G, et al. Genome sequence of the cultivated cotton Gossypium arboreum. Nat Genet [Internet]. 2014 Jun 18;46(6):567–72. Available from: http://www.nature.com/articles/ng.2987
85. Yu J, Jung S, Cheng C-H, Ficklin SP, Lee T, Zheng P, et al. CottonGen: a genomics, genetics and breeding database for cotton research. Nucleic Acids Res [Internet]. 2014 Jan;42(D1):D1229–36. Available from: https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkt1064
86. Briskine R V., Paape T, Shimizu-Inatsugi R, Nishiyama T, Akama S, Sese J, et al. Genome assembly and annotation of Arabidopsis halleri, a model for heavy metal hyperaccumulation and evolutionary ecology. Mol Ecol Resour. 2016;
87. International Organization for Standardization. Information processing — Documentation symbols and conventions for data, program and system flowcharts, program network charts and system resources charts [Internet]. 1985 [cited 2019 Dec 19]. p. 25. Available from: https://www.iso.org/standard/11955.html
88. Song Q, Decato B, Hong EE, Zhou M, Fang F, Qu J, et al. A Reference Methylome Database and Analysis Pipeline to Facilitate Integrative and Comparative Epigenomics. El-Maarri O, editor. PLoS One [Internet]. 2013 Dec 6;8(12):e81148. Available from: https://dx.plos.org/10.1371/journal.pone.0081148
89. Luu P-L, Gerovska D, Arrospide-Elgarresta M, Retegi-Carrión S, Schöler HR, Araúzo-Bravo MJ. P3BSseq: parallel processing pipeline software for automatic analysis of bisulfite sequencing data. Bioinformatics [Internet]. 2016 Oct 6;btw633. Available from: https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btw633

Figures, tables and additional files

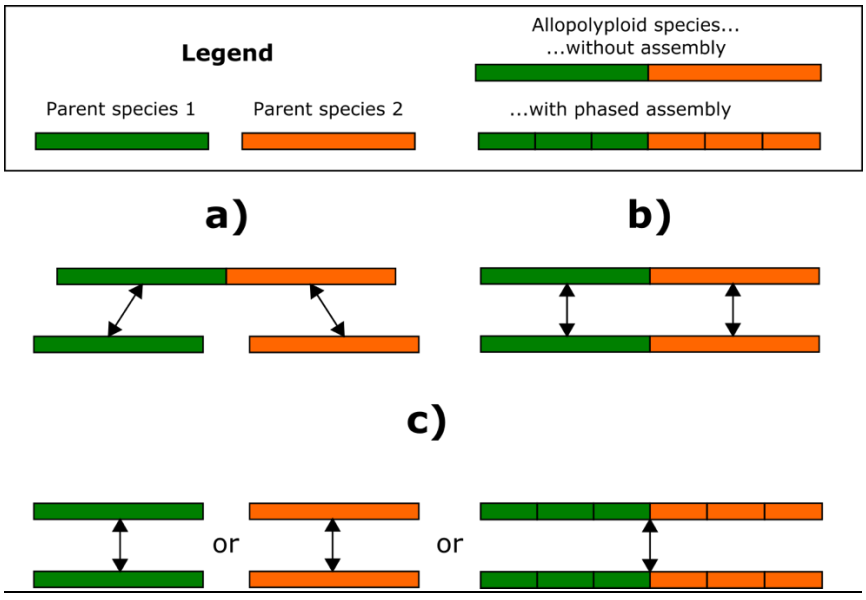

Figure 1: schematic view of the experimental designs supported by ARPEGGIO. *There are 3 possible comparisons: a) polyploid species without assembly against its progenitors, b) same polyploid without assembly in two different experimental conditions and c) diploid species or polyploid species with an available phased assembly in two different experimental conditions. All comparisons are about whole genome DNA methylation patterns.*

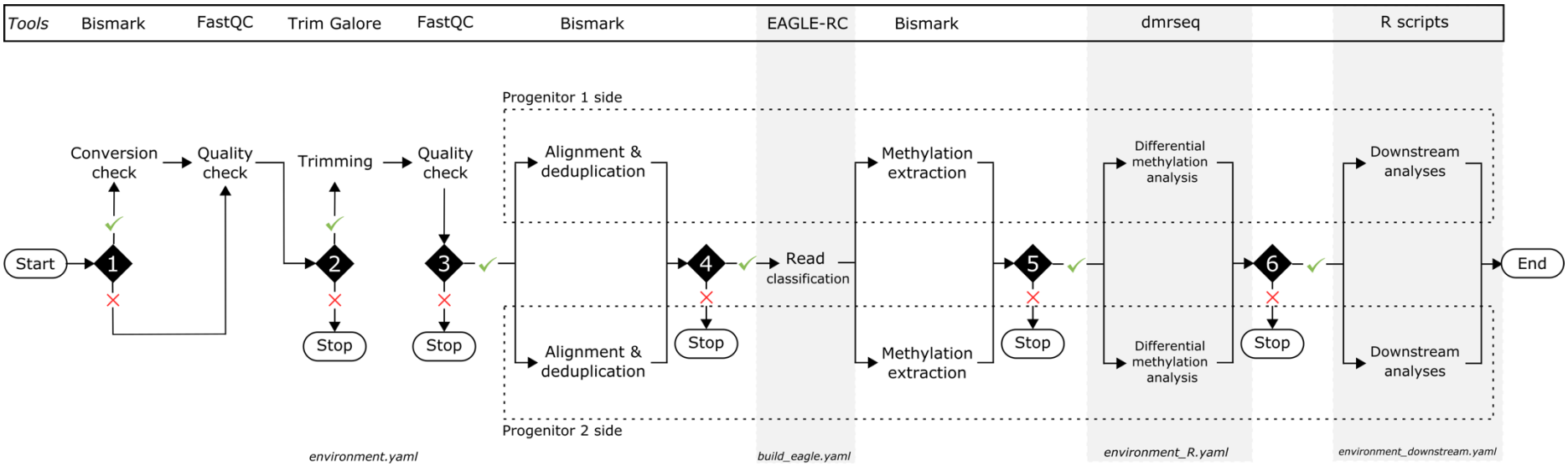

Figure 2: schematic overview of ARPEGGIO's structure. *All the shapes follow the flowchart standardized symbols* (87). *Ovals show the beginning and ends of the workflow. Diamonds represent conditional rules in ARPEGGIO's configuration file and those rules make ARPEGGIO more adaptable to the needs of the user. Each conditional rule can be set to "true" (tick) or "false" (cross). Besides the first conditional rule, all other rules stop the workflow at the given point when set to "false". The different grey backgrounds and the white background represent different Conda environments used by ARPEGGIO to carry out different steps of the analyses. In the scheme, the background of each step represents the environment that the step is part of. The bottom of each background shows the name of the file used to create the environment. At the top all the tools used by ARPEGGIO are shown and vertically aligned their corresponding step in the workflow. From the alignment and deduplication step, ARPEGGIO executes two workflows in parallel for each progenitor side, both highlighted by the dashed areas.*

Table 1: overview of the read assignment accuracy of EAGLE-RC against the concatenation method with real datasets. *The first column shows the species name behind the dataset, the second specifies if the replicates consisted of biological or technical replicates, the third and the fourth one show the error rate of the two approaches: concatenating genomes or read classification. The error rate was obtained by the number of reads assigned to the wrong genome divided by the total number of reads that were uniquely mapped and deduplicated.*

Datasets		**Average number of uniquely mapped reads**		**Average error rate**	
Species	**Type of replicate (#)**	**Concatenated genome**	**Read classification**	**Concatenated genome**	**Read classification**
Arabidopsis halleri	Biological (2)	17'258'758	18'311'330	3.98 %	1.16 %
Arabidopsis lyrata	Biological (2)	22'204'342	23'301'056	5.94 %	1.45 %
Mimulus guttatus	Technical (4)	1'420'116	1'288'800	26.78 %	7.52 %
Mimulus luteus	Technical (4)	3'889'458	3'760'614	9.80 %	2.29 %
Gossypium arboreum	Technical (2)	253'912'667	254'261'702	0.0044 %	0.0013 %
Gossipyium raimondii	Technical (2)	242'590'069	246'935'598	0.0039 %	0.0019 %

Table 2: summary of ARPEGGIO's downstream analyses on the dataset from Edger and colleagues. *The table is divided in two parts, one per progenitor. For each progenitor, the table shows the number of differentially methylated regions (DMRs) for each context, the number of genes overlapping with DMRs and the total number of genes found over all contexts.*

	Mimulus guttatus			*Mimulus luteus*		
Methylation context	CG	CHG	CHH	CG	CHG	CHH
DMRs	65	126	23	277	211	58
Total DMRs	214			546		
Genes overlapping DMRs	13	20	2	2	0	0
Total genes	35			2		

Table 3: comparison between ARPEGGIO and other available, non-commercial and general workflows able to work with raw WGBS data. *There were a total of 12 workflows found and different features were selected for this comparison. The language indicates the main language(s) used to program the workflow. Polyploid support refers to support analysis of data from a polyploid with no official genome assembly available. Diploid support refers to analysis of data from a diploid or a polyploid with an available official genome assembly. Quality check, alignment, statistical and downstream analyses are all different steps in the data analysis process with downstream analyses being defined as follow-up analyses on DMRs found by the statistical analyses. Methylation contexts are 3 in total: CpG, CHG and CHH and this feature is sometimes limited to CpG only. Visualization represents any script or function allowing the user to visualize the DMRs found by the statistical analyses. Reproducibility is difficult to quantify and in this table a tool was considered reproducible if the corresponding paper mentioned reproducibility as one of their goals.*

	ARPEGGIO	QUMA	MOABS	QuasR	MethPipe	bicycle	RUbioSeq	WBSA	P3BSseq	Methy-Pipe	MethFlow	snakePipes
Language	Python, R	HTML, Perl, Javascript	C++, Perl	R	C++	Java	Perl	Perl, R	Python	Perl, R	Python, Perl, Java	Python, R
Polyploid support	✓	X	X	X	X	X	X	X	X	X	X	X
Diploid support	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Quality check	✓	✓	✓	✓	✓	X	✓	✓	✓	X	X	✓
Alignment	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Statistical analyses	✓	✓	✓	✓	✓	✓	X	✓	✓	✓	✓	✓
Methylation context	All	CpG only	All	All	All	All	All	All	All	All	?	CpG only
Downstream analyses	✓	X	X	✓	✓	X	X	✓	X	✓	✓	X
Visualization	X	X	X	✓	✓	X	X	✓	X	✓	✓	X
Reproducibility	✓	-	-	✓	-	✓	-	-	-	-	✓	✓
Paper	-	(80)	(66)	(71)	(88)	(76)	(79)	(73)	(89)	(74)	(75)	(77)

Additional file 1:

- **Format: pdf**
- **Title of data:** Example of relationships between rules in ARPEGGIO
- **Description:** A graph showing the input/output relationships between different rules in ARPEGGIO in an example “default” run with single end reads.

Additional file 2:

- **Format: pdf**
- **Title of data:** Plot with average runtimes in ARPEGGIO with Mimulus data
- **Description:** A plot with the average runtime for each main step in the ARPEGGIO pipeline: conversion check, quality check, trimming, alignment, deduplication, read classification, methylation extraction and statistical analyses.

Additional file 3:

- **Format: pdf**
- **Title of data:** Plant material and WGBS library synthesis
- **Description:** Details about the plant conditions, sampling, DNA extraction, bisulfite treatment and sequencing strategies.

Additional file 4:

- **Format: pdf**
- **Title of data:** Read statistics about datasets used to compare EAGLE-RC against concatenation method
- **Description:** All the numbers related to the datasets used to compare EAGLE-RC to the concatenation method: total reads, uniquely mapped reads (and not), duplicated reads, correct, ambiguous and wrongly classified reads and error rate.