Title

AFLAP: assembly-free linkage analysis pipeline using k-mers from genome sequencing data

Permalink
https://escholarship.org/uc/item/6zd4z6hc

Journal
Genome Biology, 22(1)

ISSN
1474-760X

Authors
Fletcher, Kyle
Zhang, Lin
Gil, Juliana
et al.

Publication Date
2021-12-01

DOI
10.1186/s13059-021-02326-x

Peer reviewed

Genome Biology

METHOD Open Access

AFLAP: assembly-free linkage analysis pipeline using k-mers from genome sequencing data

Kyle Fletcher[1], Lin Zhang[1], Juliana Gil[1], Rongkui Han[1], Keri Cavanaugh[1] and Richard Michelmore[1,2*]

* Correspondence: rwmichelmore@ucdavis.edu
[1]The Genome Center, University of California, Davis, USA
[2]Departments of Plant Sciences, Molecular & Cellular Biology, Medical Microbiology & Immunology, University of California, Davis, USA

Abstract

Our assembly-free linkage analysis pipeline (AFLAP) identifies segregating markers as k-mers in the raw reads without using a reference genome assembly for calling variants and provides genotype tables for the construction of unbiased, high-density genetic maps without a genome assembly. AFLAP is validated and contrasted to a conventional workflow using simulated data. AFLAP is applied to whole genome sequencing and genotype-by-sequencing data of F1, F2, and recombinant inbred populations of two different plant species, producing genetic maps that are concordant with genome assemblies. The AFLAP-based genetic map for *Bremia lactucae* enables the production of a chromosome-scale genome assembly.

Keywords: Genetic map, *K*-mer, *Bremia*, Oomycete, *Arabidopsis*, Lettuce, Genotype-by-sequencing, GBS

Background

The complexity of contemporary maps has increased drastically since linkage, the tendency for co-segregation of two or more genetic loci during meiosis relative to their proximity on a chromosome, was introduced as a concept at the start of the twentieth century [1, 2]. The first genetic map was based on six sex-linked phenotypes in *Drosophila melanogaster* [3]. Now, due to technical advances, particularly in DNA sequencing, it is common to construct genetic maps with thousands of markers that greatly exceed the number of genetic bins observed in segregating progeny. Typically, markers are derived from aligning sequencing reads to a reference assembly and calling polymorphisms. However, reference assemblies are not available for all species and may not be available for all genotypes, even in well-studied species. In addition, a single reference assembly of a species does not capture all variation within a species. Therefore, some variants segregating in the gametes of a specific cross may not be accessible when a genome assembly is used to identify polymorphisms, particularly if neither parent was the reference genotype. Therefore, we developed a pipeline for generating ultra-high-density genetic maps that does not require a genome assembly and can be applied to any species regardless of genomic resources.

Bremia lactucae is an outbreeding oomycete that causes the economically important downy mildew disease of lettuce. Multiple linkage studies have achieved increasing marker density over time. The first genetic map for *B. lactucae*, reporting 13 linkage groups, was generated using 53 restriction fragment length polymorphisms (RFLPs) and nine phenotypic loci spanning 230 cM [4]. A second map was based on 83 RFLPs, 347 amplified fragment length polymorphisms (AFLPs), and seven phenotypic loci [5]. Genomic investigations of *B. lactucae* revealed genetic characteristics that complicated map construction. Although *B. lactucae* was shown to be a diploid species, many isolates were determined to be heterokaryotic with genetically distinct, diploid nuclei sharing the same coenocytic cytoplasm [6]. Single nucleotide polymorphism (SNP) analysis of the F_1 population previously used for map construction revealed two groups of half-siblings, indicating that one of the parents was contributing two sets of gametes. In addition, analysis of multiple isolates revealed high levels of heterozygosity (> 1%) and the reference assembly consisted of over 70% long terminal repeat retrotransposons [6]. These genomic features may reduce the accuracy of short-read mapping and SNP calling [7], which are critical for genetic map construction with next-generation sequencing data. Therefore, at the beginning of this study, the genetic architecture of *B. lactucae* remained unresolved.

Our assembly-free linkage analysis pipeline (AFLAP) enables the construction of genetic maps without mapping or SNP calling against a reference genome assembly. Instead, reads are reduced to *k*-mers ($k = 31$, onward referred to as 31-mer) and surveyed to identify those that segregate uniquely in the gametes of each parent. AFLAP was benchmarked against a conventional linkage analysis pipeline. Simulations were used to investigate the impacts of varying genome size, heterozygosity, and sequencing depth on run times and results produced by AFLAP. Running AFLAP on an F_2 population of *Arabidopsis thaliana* land races Colombia (Col) x Landsberg (Ler) whole genome sequenced to low coverage generated the five expected linkage groups. Testing AFLAP on reduced representation genotype-by-sequencing (GBS) data of a recombinant inbred line population of a *Lactuca serriola* x *L. sativa* interspecific-cross generated the expected nine linkage groups for both parents. AFLAP was then used to construct an ultra-dense genetic map for *B. lactucae*, using F_1 isolates that had been sequenced to over 5x by whole genome sequencing (WGS).

Results

A novel assembly free-linkage analysis pipeline (AFLAP [8, 9]) was designed to rapidly construct genetic maps without requiring the alignment of reads to and subsequent variant calling against a reference genome assembly (Fig. 1). Briefly, AFLAP generated 31-mers from the parental read sets. Single-copy 31-mers were identified by analyzing peaks contained within JELLYFISH count files. Single-copy 31-mers common to both parents were discarded; 31-mers unique to each parent were assembled creating variants unique to either parent. One 31-mer was extracted from each fragment of 61 bp and larger to be used as markers. These markers were then surveyed in 31-mers present in progeny individuals. Genotypes were scored as present or absent, resulting in a genotype table, which was then inspected to obtain segregation statistics of the

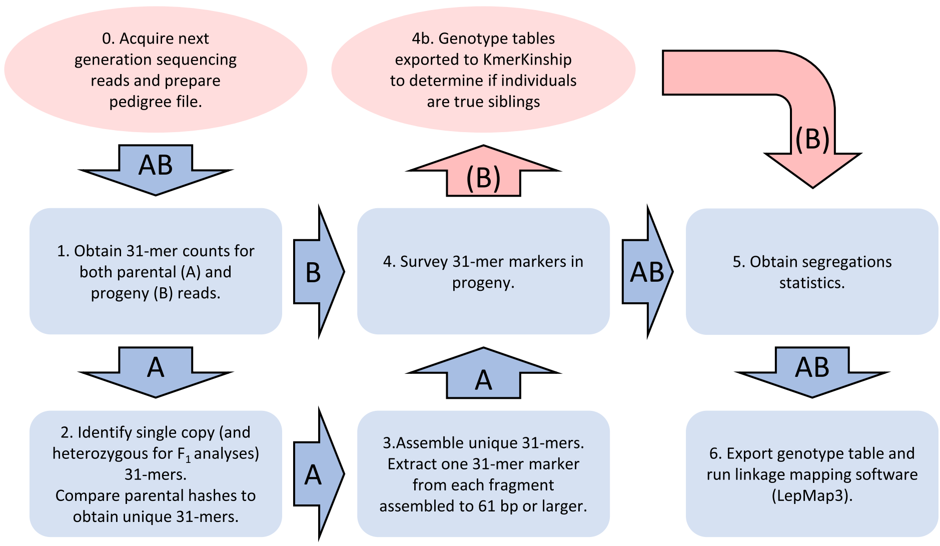
0. Acquire next generation sequencing reads and prepare pedigree file.
4b. Genotype tables exported to KmerKinship to determine if individuals are true siblings
(B)
AB
(B)
1. Obtain 31-mer counts for both parental (A) and progeny (B) reads.
B
4. Survey 31-mer markers in progeny.
AB
5. Obtain segregations statistics.
A
A
AB
2. Identify single copy (and heterozygous for F1 analyses) 31-mers. Compare parental hashes to obtain unique 31-mers.
A
3.Assemble unique 31-mers. Extract one 31-mer marker from each fragment assembled to 61 bp or larger.
6. Export genotype table and run linkage mapping software (LepMap3).

Fig. 1 Assembly-free linkage analysis pipeline (AFLAP). Blue rectangles summarize core steps of the pipeline and numbers correspond to script numbering found at https://github.com/kfletcher88/AFLAP [8]. Arrows indicate where one step outputs the input of the next step. Arrows labeled A indicate the path taken in genotyping and identifying markers from parental individuals. Arrows labeled B indicate the path taken to genotype progeny individuals. Red ovals and arrows are supplementary to AFLAP and not required to complete the pipeline. An example pedigree file is provided for the *A. thaliana* data (Additional file 2: Table S1)

markers. Markers were filtered for segregation distortion and finally exported to LepMap3 [10] for linkage analysis. The pipeline is described in detail in the "Materials and methods" section.

Simulations

To benchmark AFLAP, we simulated a test-cross population of 100 F_1 individuals and compared the results from AFLAP to those from a conventional pipeline. The 119 Mb, five-chromosome *Arabidopsis thaliana* genome assembly was used as a template to simulate one parent to be 0.2% heterozygous with ~ 118,000 SNPs and ~ 1200 indels. One hundred F_1 progeny were simulated by introducing one or two crossovers per chromosome (see the "Materials and methods" section). Running AFLAP on synthetic reads derived from the simulated F_1 progeny resulted in a 699 cM genetic map containing five linkage groups. The average Kendall rank coefficient (τ) per linkage group was 0.986, demonstrating that the results were colinear with the reference assembly (Table 1, Additional file 1). This dataset was compared to a conventional read mapping, variant calling, and linkage analysis pipeline, with the same linkage analysis software. The conventional pipeline produced similar results, calculating 701 cM genetic maps containing five linkage groups that correlated with the reference assembly ($\tau = 0.992$; Table 1, Additional file 1). Significantly, AFLAP was approximately three times as fast as the conventional pipeline and could be further accelerated by downsampling markers. The average τ across linkage groups showed that downsampling markers did not alter the correlation of the genetic map with the reference assembly nor did the map length or number of linkage groups change (Table 1). The longest step of AFLAP

Table 1 Comparison of AFLAP with a conventional linkage analysis pipeline. Maps aligned to assemblies in Additional file 1

Method	**Assembly size**	**Chromosomes**	**Segregating markers detected.**	**Markers used in linkage analysis.**	**Minimum LOD score**	**Linkage groups calculated**	**Map length estimated**	**Percent of markers placed in linkage groups (%)**	**Collinearity of genetic map with genome assembly (average Kendall rank coefficient; τ)**	**Run time (wall clock)**
AFLAP	119 Mb	5	186,523	186,523	7	5	699 cM	99.8	0.986	59 h 32 min
AFLAP	119 Mb	5	186,523	10,000	7	5	696 cM	99.8	0.986	6 h 13 min
Conventional	119 Mb	5	225,797	225,797	20	5	701 cM	99.6	0.992	173 34 min
Conventional	119 Mb	5	226,087	10,000	7	5	698 cM	99.8	0.992	107 h 5 min

is *k*-mer counting performed by JELLYFISH; however, the pipeline is written in such a way that previously calculated JELLYFISH results can be used if the *k*-mer size is not changed. This saves a large amount of time should the user wish to permute their analysis through downsampling, or the addition/exclusion of individuals from the linkage analysis. Therefore, AFLAP can produce accurate genotype tables for linkage analysis much faster without using a whole-genome assembly, even if one is available, and results are comparable to conventional pipelines.

Performance of AFLAP was investigated by simulating different biological and experimental inputs into the pipeline. The most significant factor impacting AFLAP was the sequencing depth of the progeny (Table 2, Additional file 1). When progenies were simulated to have adequate sequence coverage (≥ 5x), the linkage groups were highly colinear with the reference assembly with ≥ 98% of markers placed and an average $\tau >$ 0.98 across linkage groups. When the simulated sequencing depth was reduced to 3x, only 86.7% of the markers were placed in linkage groups and marker order was less correlated with the reference assembly ($\tau = 0.826$; Table 2, Additional file 1). The sequencing coverage of the parents had less of an effect on the final map quality, with maps being colinear ($\tau > 0.98$); however, the number of markers reported reduced as the coverage dropped (Table 2). Therefore, AFLAP can use low coverage (10x) parental sequencing data to produce accurate genetic maps, albeit with potentially lower information content. The effect of heterozygosity was tested by varying the synthetic heterozygosity (from 0.001 to 2%) of the mapped parent. Genetic maps produced were colinear with the original assembly ($\tau > 0.98$), suggesting that heterozygosity had little effect on the calling of markers (Table 2, Additional file 1). Through downsampling markers to 10,000, AFLAP was able to construct genetic maps from all simulations with the 119 Mb, five-chromosome genome in under 9 h under the test conditions (Table 2). The impact of different genome sizes and chromosome numbers was tested by simulating crosses using the genomes of *Vitis riparia* (19 chromosomes, ~ 500 Mb reference assembly) and *Atriplex hortensis* (nine chromosomes, ~ 900 Mb reference assembly), synthesizing the mapped parent to be 0.2% heterozygous and sequencing coverage to be 50x for the parents, 10x for the progeny making it directly comparable with previous simulations. As expected, the number of reported markers increased with genome size. As with the smaller 119 Mb genome, the genetic maps produced with downsampled markers were colinear with genome assembly from which they were derived ($\tau > 0.98$; Table 2, Additional file 1); however, more markers were required after downsampling. Fifty thousand markers produced a synthetic genetic map colinear with the 19-chromosome, 500 Mb reference; simulations with 10,000 and 25,000 markers resulted in at least one error. For the largest genome (900 Mb), 25,000 markers were able to reconstitute the nine chromosomes (Table 2, Additional file 1), suggesting that the number of chromosomes, not the genome size, is the driver for requiring more markers in these simulations. The time required to run AFLAP increased with genome size, although it was still faster than the conventional pipeline on the smallest simulated cross (Tables 1 and 2). These simulations demonstrated that AFLAP can accommodate genomes of differing sizes, chromosome numbers, and heterozygosity provided adequate sequencing depth of the progeny is available.

Finally, AFLAP was tested on 100 synthetic F_2 individuals. For this simulation, the five-chromosome, 119 Mb genomes of both parents were 100% homozygous, varying

Table 2 Simulation of AFLAP testing different genome and sequencing coverage variations. Maps aligned to assemblies in Additional file 1

Assembly size	Chromosomes	Cross structure	Simulated progeny sequencing depth	Simulated parental sequencing depth	Simulated heterozygosity	Segregating markers detected by AFLAP	Markers used in linkage analysis	Minimum LOD score	Linkage groups calculated	Percent of markers placed in linkage groups	Collinearity of genetic map with genome assembly (average Kendall rank coefficient; τ)	Run time (wall clock)
i) Altering progeny sequencing depth.												
119 Mb	5	F_1 (ABxCC)	**20**	50	0.2%	186,522	10,000	7	5	100%	0.993	8 h 44 min
119 Mb	5	F_1 (ABxCC)	**10**	50	0.2%	186,523	10,000	7	5	99.8%	0.986	6 h 13 min
119 Mb	5	F_1 (ABxCC)	**7**	50	0.2%	186,523	10,000	7	5	> 99.9%	0.988	5 h 18 min
119 Mb	5	F_1 (ABxCC)	**5**	50	0.2%	186,525	10,000	7	5	98%	0.983	4 h 22 min
119 Mb	5	F_1 (ABxCC)	**3**	50	0.2%	186,525	10,000	3	6	86.7%	0.826	4 h 17 min
ii) Altering parental sequencing depth.												
119 Mb	5	F_1 (ABxCC)	10	**50**	0.2%	186,523	10,000	7	5	99.8%	0.986	6 h 13 min
119 Mb	5	F_1 (ABxCC)	10	**40**	0.2%	189,329	10,000	7	5	99.8%	0.985	6 h 5 min
119 Mb	5	F_1 (ABxCC)	10	**30**	0.2%	179,218	10,000	7	5	99.8%	0.985	6 h 20 min
119 Mb	5	F_1 (ABxCC)	10	**20**	0.2%	127,628	10,000	7	5	99.8%	0.985	6 h 0 min
119 Mb	5	F_1 (ABxCC)	10	**10**	0.2%	112,655	10,000	7	5	99.8%	0.986	5 h 40 min
iii) Altering parental heterozygosity.												
119 Mb	5	F_1 (ABxCC)	10	50	**0.01%**	9996	9996	7	5	99.8%	0.985	6 h 49 min
119 Mb	5	F_1 (ABxCC)	10	50	**0.1%**	97,446	10,000	7	5	99.8%	0.984	7 h 3 min
119 Mb	5	F_1 (ABxCC)	10	50	**0.2%**	186,523	10,000	7	5	99.8%	0.986	6 h 13 min
119 Mb	5	F_1 (ABxCC)	10	50	**0.5%**	401,020	10,000	7	5	99.8%	0.983	7 h 6 min
119 Mb	5	F_1 (ABxCC)	10	50	**1%**	736,293	10,000	7	5	99.6%	0.984	6 h 40 min
119 Mb	5	F_1 (ABxCC)	10	50	**2%**	1,148,814	10,000	7	5	99.2%	0.984	7 h 12 min
iv) Altering genome size and chromosome number.												
119 Mb	**5**	F_1 (ABxCC)	10	50	0.2%	186,523	10,000	7	5	99.8%	0.986	6 h 13 min
500 Mb	**19**	F_1 (ABxCC)	10	50	0.2%	490,397	50,000	7	19	99.2%	0.984	12 h 23 min
900 Mb	**9**	F_1 (ABxCC)	10	50	0.2%	801,712	25,000	7	9	99.3%	0.982	22 h 0 min
v) Altering cross type												
119 Mb	5	**F_1 (ABxCC)**	10	50	0.2%	186,523	10,000	7	5	99.8%	0.986	6 h 13 min
119 Mb	5	**F_2 (AAxCC)**	10	50	0%	94,428 104,106	10,000 10,000	7 7	5 5	100% > 99.9%	0.978 0.977	7 h 49 min

from one another by ~ 118,000 SNPs and ~ 1200 indels. Whole genome sequencing coverage was simulated to 50x for both parents and 10x for the F_2 progeny. The resultant five linkage groups in the genetic maps of each parent were colinear with the reference assembly ($\tau > 0.97$; Table 2, Additional file 1). Therefore, AFLAP can be applied to F_1 and F_2 populations and could be extended to other types of populations such as recombinant inbred lines.

AFLAP analysis of an F_2 population of *Arabidopsis thaliana* using WGS data

AFLAP was then validated on real sequencing data using an F_2 population generated from *A. thaliana* Col x Ler that had previously been sequenced to low coverage (1x to 8x; Additional file 2: Table S1) and analyzed genetically [11–13]. Based on the distributions of 31-mers from both parents, the boundaries for classification as a single-copy 31-mer were defined as 32 to 105x for *A. thaliana* Col and 41 to 177x for *A. thaliana* Ler (Fig. 2a), totaling 109,276,920 and 114,129,149 single-copy 31-mers, respectively. The unique number of single-copy 31-mers was 21,704,129 for Col and 27,891,980 for Ler.

Assembly of unique, single-copy Col 31-mers resulted in 499,936 fragments ranging from 25 to 6324 bp. Of these, 162,206 fragments were 61 bp and had a SNP in the middle at their 31st base; 156,106 fragments were larger than 61 bp, representing more complex variants; and 181,624 fragments were smaller than 61 bp and contained repetitive or low complexity sequences that were difficult to assemble and were therefore not used in downstream analyses because they represented potentially unreliable variants (Fig. 2b). Complex variants identified from Col had high percent identity and query coverage when aligned to the Col assembly, supporting that they were accurately assembled. When aligned to the Ler assembly, the percent identity and query coverage was much lower, supporting that the fragments were unique to Col (Additional file 3: Figure S1). 31-mers were extracted from the 318,312 fragments ≥ 61 bp, 285,492 (89.9%) of which were confirmed to be within the single-copy limits of Col and absent from Ler. On average, these 31-mer markers were scored as present in 55% of F_2 progeny (Fig. 2c).

Assembly of unique, single-copy Ler 31-mers resulted in 519,493 fragments ranging from 25 bp to 26,666 bp. Of these, 161,955 fragments equaled 61 bp (SNPs), 161,174 fragments were larger than 61 bp (complex variants), and 196,364 fragments were < 61 bp (Fig. 2b). 31-mers were extracted from the 323,129 fragments ≥ 61 bp, 321,026 (99.3%) of which were confirmed to be within the single-copy limits of Ler and absent from Col. On average, these 31-mer markers were scored as present in 51% of F_2 progeny (Fig. 2c). Additionally, 5329 markers, nearly all of which were derived from markers > 61 bp, were not detected in any F_2 plants (Fig. 2c). Only 13 of these markers could be aligned to the reference genome assembly, consistent with these markers being derived from contaminant reads in the SRA dataset for Ler.

Of the 285,492 markers derived from Col, 50.9% were ordered into 315 genetic bins across five linkage groups using a logarithm of the odds (LOD) score ≥ 7. The total map length was 379 cM and linkage groups ranged from 64 to 86 cM (Fig. 3a). Of the genetically placed markers, 144,395 markers (99.4%) were aligned to the *A. thaliana* reference assembly. Over 99.99% of the markers were concordant with the physical map (Fig. 3b). Of the 321,026 markers derived from Ler, 68.0% were ordered into 366 genetic bins across six linkage groups (LOD ≥ 7). The number of markers assigned to

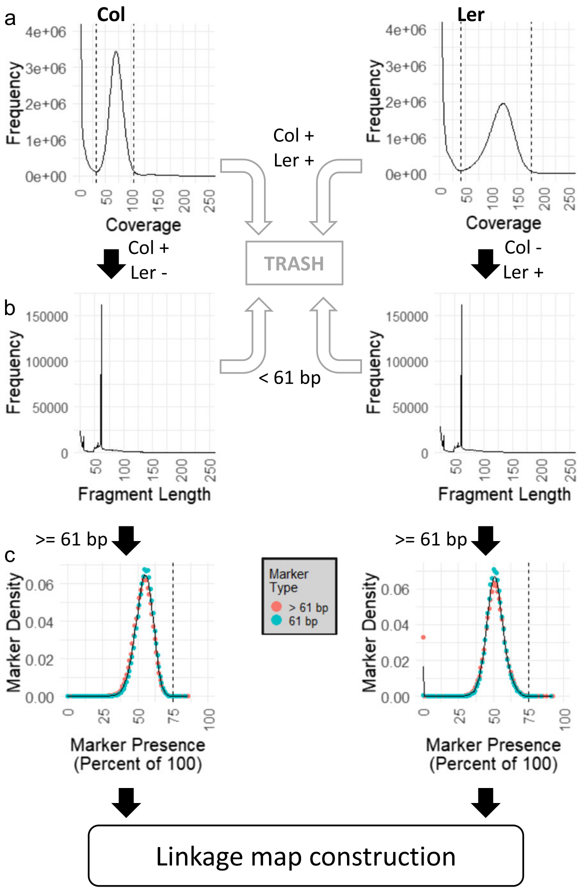

Fig. 2 Intermediate data produced by AFLAP when analyzing an *A. thaliana* F_2 population. Analysis of parental accession Colombia (Col) is plotted on the left and Landsberg (Ler) on the right. **a** JELLYFISH histograms were plotted to determine the lower and upper boundaries of single-copy 31-mers (indicated by dotted lines). 31-mers found in both parents were discarded (gray arrows) from further analysis. 31-mers unique to either accession were assembled for each parent (black arrows). **b** The distribution of the assembled fragments from 31-mers showing that most of these fragments are 61 bp (SNPs). Assembled fragments < 61 bp were discarded (gray arrows). A representative 31-mer marker was extracted from each fragment ≥ 61 bp, verified against the parents, and used for downstream genotyping (black arrows). **b** Segregation of 31-mer markers in the F_2 population. Markers extracted from fragments > 61 bp segregate at the same frequency as markers derived from fragments exactly = 61 bp, although > 3% of Ler-derived markers from fragments > 61 bp were not observed in the F_2 population and most were found to be contaminants. The modal segregation of Colombia markers was 55%, for Landsberg it was 51%. The black vertical line indicated the expected marker presence of 75% in the F2 (Aa x Aa)

each genetic bin ranged from 1 to 5857. The total map length was 421 cM and linkage groups ranged from 53 to 98 cM (Fig. 3c). Of the genetically placed markers, 138,434 (63.4%) were placed unambiguously on the *A. thaliana* reference assembly, with 98.8% concordance between the genetic map and the physical assembly. Two linkage groups mapped to Chromosome 3, each covering a chromosome arm (Fig. 3d).

AFLAP analysis of a RIL population of *Lactuca* spp. using GBS data

AFLAP was run on 235 F6 RILs generated by crossing Armenian999 (*L. serriola*) and PI251246 (*L. sativa*), which had previously been sequenced using GBS with 249.8 Mb generated for each line [14]. It was not possible to obtain the single-copy boundaries

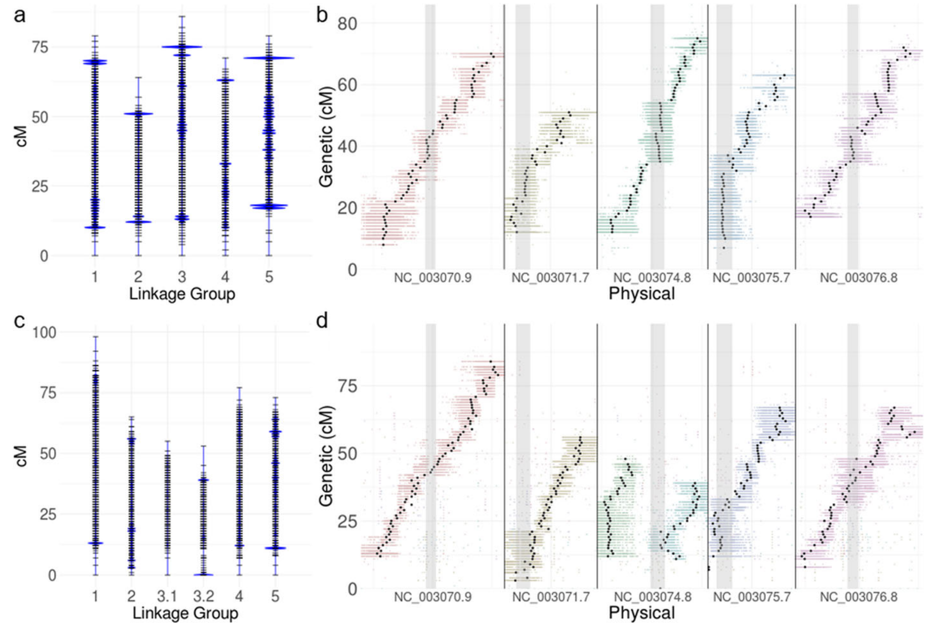

Fig. 3 Genetic map of *A. thaliana* generated by AFLAP. **a** 145,330 markers unique to *A. thaliana* accession Col ordered into five linkage groups. Each horizontal black bar indicates a genetic bin with > 1 marker. The sizes of the blue marks indicate the number of markers in each bin relative to every other genetic bin. Genetic bins with the highest marker densities were near the ends of linkage groups. **b** Alignment of the genetic map constructed from markers unique to Col against the genome assembly of Col-0. All 31-mers assigned to each genetic bin were aligned to the assembly and plotted. Points are colored by linkage group assignment resulting in a bar depicting the noise for placement of each genetic bin. The average position of each genetic bin is plotted in black, demonstrating that the linkage groups were nearly colinear with the physical assembly. Gray columns indicate centromeric positions. Over 99.99% of the markers were placed on the correct chromosomes and were concordant with the assembly. **c** 115,810 markers unique to *A. thaliana* accession Ler ordered into six linkage groups. Sizes of blue marks indicate the number of markers per genetic bin. **d** Alignment of the genetic map constructed from markers unique to Ler displayed as for Col (B). Arms of Chromosome 3 were generated as independent linkage groups. 98.8% of the markers were placed on the correct chromosomes and were nearly concordant with the assembly. In both **b** and **d**, the inexact placement of markers into genetic bins demonstrated by the range of the colored bars is probably due to missing data because of the low coverage of the sequencing data

from GBS data as had been calculated for *A. thaliana*. Instead, limits were set to ≥ 20x and ≤ 45x read depths so markers would not be derived from under- or over-represented sequences. After filtering Armenian999 against PI251246 31-mers, 496,010 unique 31-mers remained. These were assembled into 3681 markers; 631 of which were 61 bp and 3050 were > 61 bp. The reverse, filtering PI251246 31-mers against Armenian999 identified 650,616 unique 31-mers belonging to PI251246 that were assembled into 5264 markers; 915 of which were 61 bp and 4349 were > 61 bp. The assembly step also produced many potential markers < 61 bp for both parents, which were not used in the subsequent analysis. Markers ≥ 61 bp segregated approximately 1:1, as expected, indicating that the markers derived from GBS reads by AFLAP were robust (Fig. 4a, b). Linkage analysis with Armenian999-derived markers produced a nine-linkage group, 1730 cM genetic map, containing 3241 markers (88% of the total identified) placed in 1656 genetic bins (Fig. 4c). Linkage analysis with PI251246-derived markers also produced a nine-linkage group genetic map of 1705 cM, containing 4947 markers (94% of the total identified) placed in 2037 genetic bins (Fig. 4d). Both parental maps were aligned to the *L. sativa* cultivar (cv.) Salinas genome assembly [15] to determine if they

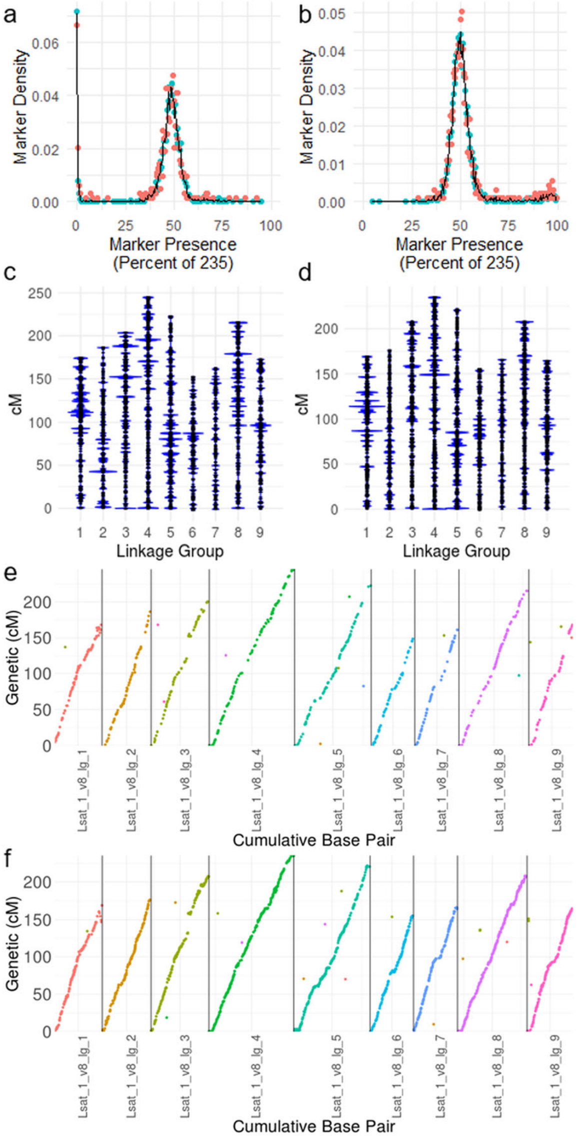

Fig. 4 Genetic maps of *Lactuca* spp. produced by AFLAP. Parental and 235 F_6 recombinant inbred lines of an interspecific *L. serriola* x *L. sativa* cross were analyzed using genotyping-by-sequencing. The segregation of markers derived from **a** Armenian999 (*L. serriola*) and **b** PI251246 (*L. sativa*) peaked at approximately 0.5, as expected for an F_6 RIL population. **c** The female genetic map was 1730 cM and contained 3241 Armenian999-derived markers in 1656 genetic bins across nine linkage groups. **d** The male genetic map was 1705 cM and contained 4947 PI251246-derived markers placed in 2037 genetic bins across nine linkage groups. Unique alignments of **e** Armenian999-derived markers and f PI251246-derived markers along the nine chromosomes of *L. sativa* genome demonstrated collinearity with the genome. Because these RILs were predominantly homozygous, a compound map could also be calculated from a combined genotype table (Additional file 3: Figure S2)

were colinear. Unique alignments were found for 747 Armenian999-derived markers, across 602 genetic bins in the map (Fig. 4e). For PI251246-derived markers, unique alignments for 2235 markers placed in 1290 genetic bins were identified (Fig. 4f). Both maps were colinear with the genome assembly. As the RILs were largely homozygous, a compound map could be calculated by combining the genotype calls for PI251246-derived and Armenian999-derived markers. The compound map was 1711 cM across

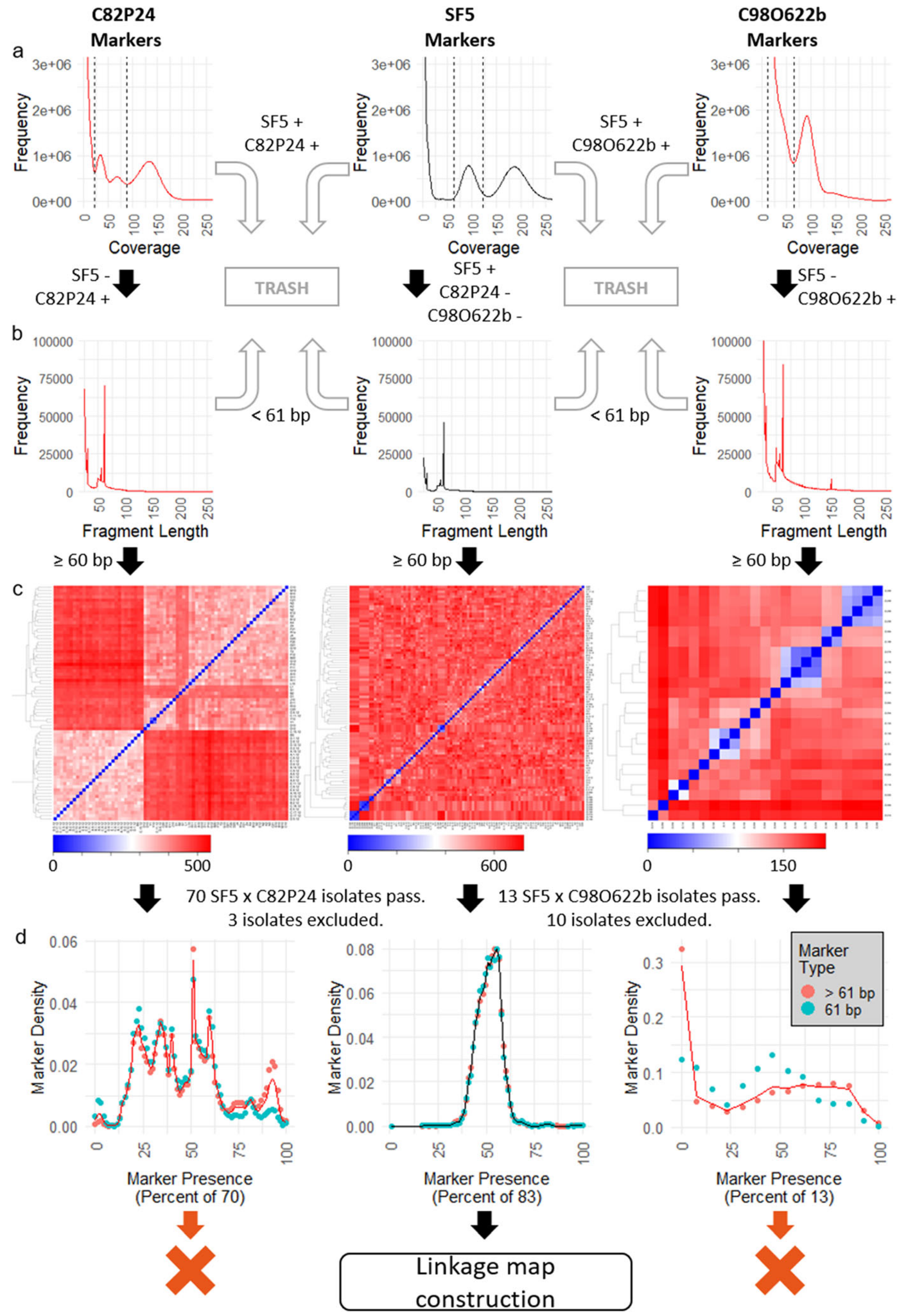

Fig. 5 (See legend on next page.)

(See figure on previous page.)
Fig. 5 Intermediate data produced by AFLAP when analyzing two F_1 populations of *B. lactucae*. Two populations generated by crossing SF5 (center) to C82P24 (left) or C98622b (right). Black arrows indicate the path of 31-mers derived from each parental isolate. Ultimately, only those derived from SF5 were used for linkage analysis. **a** JELLYFISH histograms were generated to determine the boundaries of single-copy, heterozygous 31-mers, (dashed lines). 31-mers common to both parents of each cross were discarded (gray arrows). Unique 31-mers to each parental isolate were assembled (vertical arrows). **b** The modal assembled fragment size was 61 bp (SNPs), except for C98O622b. Fragments < 61 bp were discarded (gray arrows). One representative 31-mer marker was extracted from each fragment ≥ 61 bp, verified against the parents, and used for downstream genotyping (vertical arrows). **c** Kinship heatmaps for parent specific, pseudo-test cross markers from each parent. Left: 73 progeny isolates from SF5 x C82P24 tested with C82P24-derived markers ≥ 61 bp. Center: 96 progeny isolates from both crosses tested with SF5-derived markers ≥ 61 bp. Right: 23 progeny isolates from SF5 x C98O622b tested with C98O622b-derived markers equal to 61 bp. One isolate was selected when clusters of isolates (colored blue) were formed using markers of both parents. The additional pattern observed for C82P24 (left) is due to heterokaryosis (see 6). **d** Segregation of 31-mer pseudo-test cross markers in the F_1. For C82P24 (left) and SF5 (center), both marker types ≥ 61 bp segregate at the same frequency. The discordance observed for markers derived from C98O622b may be due to contamination (see **a**). Distortion is observed for C82P24 because this isolate is heterokaryotic (see **c** and 6). SF5 is homokaryotic so all gametes are derived from a single nucleus. Therefore, only SF5-derived markers were suitable for linkage map construction

nine linkage groups, containing 8191 markers in 2497 genetic bins. The compound map was also colinear with the genome assembly, with unique alignments found for 2984 markers across 1556 genetic bins (Additional file 3: Figure S2). Therefore, AFLAP can be used to analyze RIL populations and can effectively genotype individuals using GBS data.

AFLAP analysis of an F_1 population of *Bremia lactucae* using WGS data

AFLAP was then used to genetically analyze the obligately biotrophic oomycete *Bremia lactucae,* for which there was only a partial genetic map and incomplete genome assembly. Eighty-three F_1 progeny isolates that had been generated by crossing *B. lactucae* isolate SF5 to either isolate C82P24 or isolate C98O622b [5, 6] were whole genome sequenced to greater than 5x coverage. Based on the distributions of 31-mers from isolate SF5 of *B. lactucae* (Fig. 5a), the boundaries for classification as a single-copy, heterozygous 31-mer were 63x to 123x, identifying 27,691,779 potentially useful 31-mers. When compared to the 31-mer compositions of the other parental isolates, C82P24 and C98O622b, 591,159 informative 31-mers were found to be unique to SF5.

Assembly of heterozygous 31-mers unique to SF5 resulted in fragments ranging from 25 bp to 2686 bp. Of these, 45,849 fragments equaled 61 bp representing a SNP in the middle at their 31st base; 59,712 fragments were larger than 61 bp, representative of more complex variants; 132,323 fragments were smaller than 61 bp, representing potentially unreliable variants including repetitive or low complexity sequences that were difficult to assemble (Fig. 5b). The set of 31-mer markers unique to SF5 was extracted from the 105,561 fragments equal to or greater than 61 bp, 103,246 (97.8%) of which were confirmed to be within the heterozygous boundaries of isolate SF5 and absent from the two other parental isolates.

The same process was repeated for the two heterokaryotic parental isolates, C82P24 and C98O622b, to analyze kinship. For C82P24, boundaries of 22x to 88x were identified after visual inspection of the 31-mer distribution (Fig. 5a), totaling 40,209,579 31-mers. Compared to the SF5 hash, 19,179,719 were unique to C82P24. Assembly of the unique, heterozygous 31-mers resulted in fragments ranging from 25 bp to 20,022 bp. Of these, 69,868 fragments were 61 bp (SNPs), 123,926 fragments were > 61 bp (complex variants), and 335,149 fragments were < 61 bp (unreliable variants; Fig. 5b). Of the 193,794 markers ≥ 61 bp, 190,753 (98.4%) were confirmed to be absent in SF5 and heterozygous in C82P24. For C98O622b, manual inspection of the 31-mer distribution curve indicated that it was not possible to differentiate the lower limits of the heterozygous 31-mer peaks from contaminant sequences (Fig. 5a) resulting from sequencing xenic cultures of the biotrophic *B. lactucae*. Therefore, 142,669,686 31-mers with a lower limit of 12x and an upper limit of 65x were selected as the heterozygous component. When compared to the SF5 hash, 117,091,383 31-mers were found to be unique to C98O622b, which assembled into fragments ranging from 25 bp to 59,784 bp. Of these, 84,435 fragments were 61 bp, 512,948 fragments were > 61 bp, and 836,933 fragments were < 61 bp (Fig. 5b). Because of the high count and inability to resolve 31-mers heterozygous to C98O622b from those belonging to contaminant organisms, only the representative markers of C98O622b SNPs (equal to 61 bp) were used for further analysis. Of these 84,435 markers, 72,496 (85.9%) were absent in SF5 and heterozygous in C98O622b.

Kinship was analyzed by clustering unique and heterozygous markers to identify near-identical progeny isolates that shared highly similar genotypes. From 73 sexual progeny generated by crossing SF5 by C82P24, six were consistently identified as duplicates with low Euclidean distances between them when considering markers derived from both parents (Fig. 5c). One of each pair was omitted from downstream analysis (Additional file 4: Table S2). Fifteen of the 23 progeny generated by crossing SF5 by C98O622b clustered into one of five replicate groups, each with high within group similarity, based on both SF5 and C98O622b markers (Fig. 5c). Consequently, ten isolates were excluded, and five representative isolates were used for downstream analysis (Additional file 4: Table S2). This resulted in 83 isolates available for further analysis (Additional file 3: Figure S3).

Additional structure is visible in the kinship analysis of heterokaryotic isolates C82P24 and C98O622b. For C82P24, two large sub-populations of progeny, consisting of 28 progeny (bottom left) and 42 progeny isolates (excluding duplicates, top right), can be identified (Fig. 5c); this reflects the heterokaryotic nature of C82P24, where two nuclei contribute independent sets of gametes to the progeny [6]. The same pattern can be seen in C98O622b (Fig. 5c), which is also a heterokaryon, although only two isolates make up the first sub population (bottom left) and 11 isolates the second (excluding replicates, top right). These patterns are not observed in the SF5 markers (Fig. 5c) because it is a homokaryon and therefore only contributes one set of gametes. Finally, for C82P24, four isolates in the second heterokaryon group appear to cluster with one another at a greater Euclidean distance from the rest of the progeny (Fig. 5c). The reason for this clustering is unclear because clustering of these isolates was not observed with SF5 markers (Fig. 5c); therefore, these isolates were retained for downstream analysis.

Presence of markers in progeny isolates was used to filter for segregation distortion. On average, SF5 31-mer markers were scored as present in 55% of the 83 F_1 progeny. A total of 5845 markers (5.6%) present in ≤ 33 or ≥ 52 isolates were filtered out due to segregation distortion (Fig. 5d). Therefore, 97,401 markers were used for linkage analysis. Subsequent analysis that increased the cutoff values from 0.4–0.6 to 0.2–0.8 did not greatly increase the number of markers nor did it result in more sequence being captured in the linkage map; this reflects the selection of the boundaries as indicated in Fig. 5d.

Markers originating from SF5 were ordered into 1337 genetic bins across 19 linkage groups, placing 98.8% (96,226) of the markers. The total map length was 1769 cM and linkage groups ranged from 52 to 148 cM (Fig. 6a). Of these markers, 61.4% (59,087) aligned unambiguously to the previously reported *B. lactucae* assembly, 96.9% (57,247) of which aligned to scaffolds larger than 1 Mb. This accounted for 96.6% of the 115.9 Mb genome assembly [6]. Long stretches of genetic markers were colinear with this assembly; however, 19 of the 21 scaffolds larger than 1 Mb contained sequences from different linkage groups and therefore appeared to be chimeric (Fig. 6b). Chimeras may have resulted from false joins due to the highly repetitive architecture of the *B. lactucae* genome [6]. The chimeric scaffolds were broken using the linkage data. Reorienting and re-scaffolding resulted in 97 Mb organized into 19 scaffolds, each encompassing a single linkage group (Fig. 6c). The remaining 18.6 Mb compromising 200 scaffolds did not have genetic markers aligned and so remained unplaced. This included six scaffolds over 1 Mb. Repeat-masking revealed that 70% of the genetically placed contig sequence and 73% of the unplaced contig sequence was repetitive. A higher percentage of the unplaced large scaffolds were covered by C82P24 and C98O622b-derived markers than SF5-derived markers (Additional file 3: Figure S4). Of the 9767 annotated genes, 8349 were placed into the 19 large scaffolds and 1418 were on genetically unassigned scaffolds; 261 out of a total 280 candidate effector genes were located across all linkage groups, except linkage group 17 (Fig. 6c). The dark diagonal obtained when analyzing Hi-C contact frequency demonstrated that the genetically orientated assembly was consistent with the Hi-C data (Fig. 6d). Attempts to refine the assembly using Hi-C reads and scaffolding software did not improve the assembly. Synteny of the genetically oriented assembly of *B. lactucae* with *Phytophthora sojae* showed that the gene order between the two oomycete assemblies was highly conserved (Fig. 7). Therefore, scaffolding using the AFLAP genetic map was able to correct errors in the genome assembly of *B. lactucae* and produce linkage-group-scale scaffolds that are coherent with Hi-C data and largely syntenic with a distantly related oomycete.

Discussion

We developed the assembly-free linkage analysis pipeline, AFLAP, to generate ultra-dense genetic maps based on single-copy *k*-mers without reference to a genome assembly. This approach to linkage analysis does not require reads to be mapped and variants called against a reference assembly for marker identification. Instead, variants are identified using assembled *k*-mers, of fixed length *k*, unique and single copy to each parent. Assembled fragments equal to $k + k$ - 1 are considered equivalent to SNPs, with the variant position present at the center of the fragment. Fragments larger than $k + k$ *- 1* are likely to be complex multi-nucleotide variants or insertions relative to the other haplotype. Therefore, AFLAP uses markers generated from fragments larger or equal to $k + k$ *- 1*. These fragments are reduced to a representative marker, containing the variant, equal in length to *k* so that (a) all markers have the same size and (b) a constant

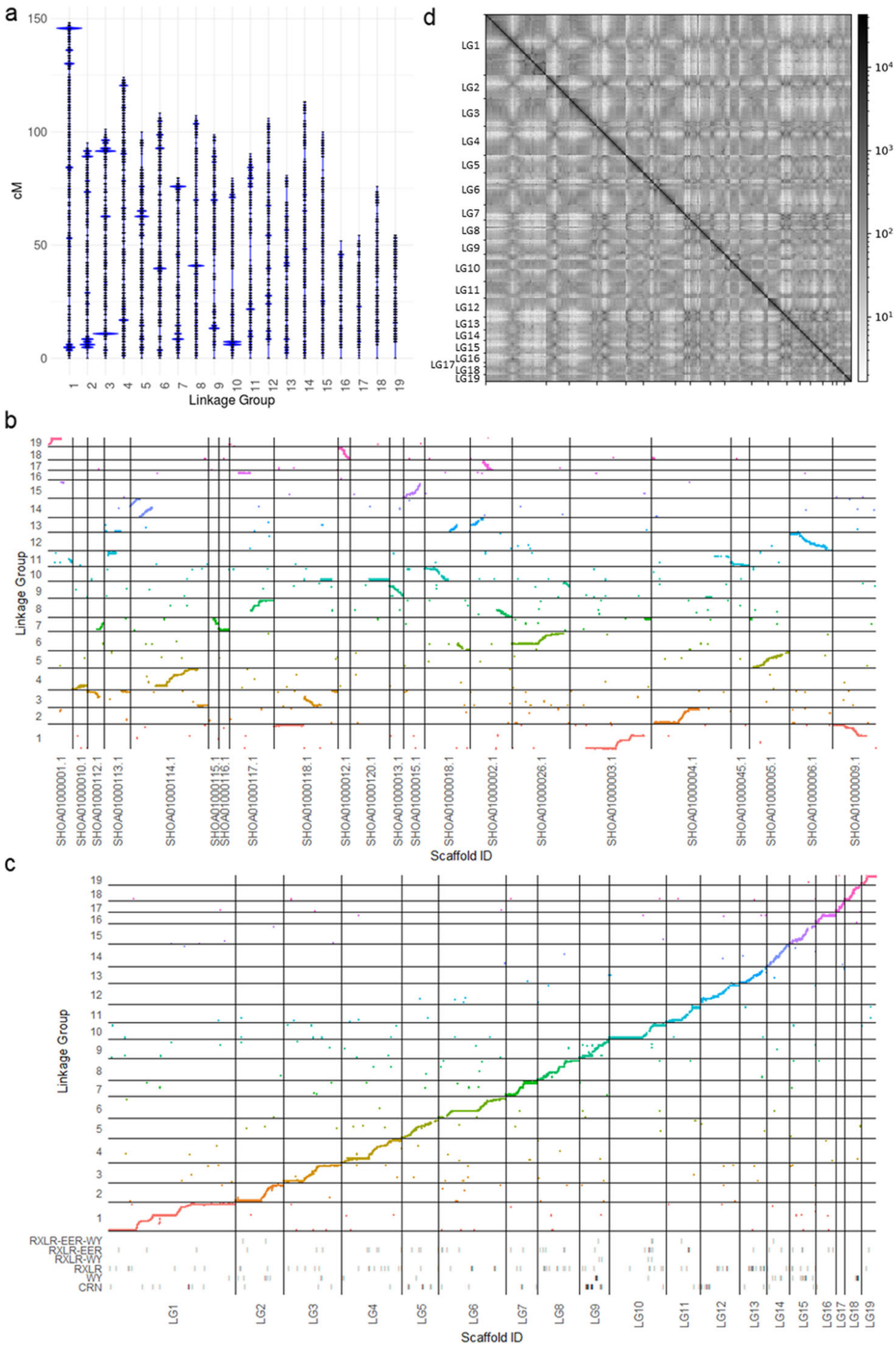

Fig. 6 (See legend on next page.)

(See figure on previous page.)
Fig. 6 *Bremia lactucae* AFLAP results. **a** Markers unique to SF5 ordered in 19 linkage groups. Each horizontal bar indicates a genetic bin with > 1 marker. The sizes of the blue marks indicate the relative number of markers in each genetic bin. Genetic bins with large numbers of markers were observed at the end of some, but not all linkage groups. **b** Alignment of the genetic map on the 21 scaffolds larger than 1 Mb in the draft reference assembly (6). Multiple scaffolds spanned more than one linkage group, indicative of mis-assembly. Linkage groups included multiple scaffolds providing genetic support for reorienting and rejoining the scaffolds. **c** Scatter plot demonstrating collinearity between the revised genome assembly and the genetic map after fragmenting, reorienting, and re-scaffolding of the assembly. Coordinates of annotated genes that encode six categories of candidate effector proteins are plotted on tracks below the scatter plot. **d** Contact frequency obtained by aligning paired Hi-C reads to the genetically reoriented assembly. The strong dark diagonal indicates that the assembly is consistent with Hi-C data. Crosses off the diagonal indicate intra-chromosomal contacts, possibly involving centromeres

marker size can be surveyed in the progeny downstream. AFLAP enables the rapid construction of a genotype table for subsequent linkage analysis.

We tested AFLAP using 100 F_2 individuals of *A. thaliana* sequenced to low coverage. The genetic architecture of *A. thaliana* has been studied in detail; over 2000 F_2 individuals, generated by crossing Colombia (Col) x Landsberg (Ler), have been sequenced to low coverage [11–13]. The 100 individuals with the largest number of reads from this population were selected to create a test population of similar size to the total progeny of the two *B. lactucae* experimental populations. The *A. thaliana* markers were expected to segregate in a 1:2:1 ratio; however, this was not the case. The modal percentage of progeny markers detected was 55% for Col markers and 51% for Ler markers (Fig. 2c). Missing data can therefore be estimated as between 26% and 32%. Despite this, the size of the genetic map is very similar to that reported previously [16, 17]. In addition, the average physical positions of genetic bins were highly concordant with the

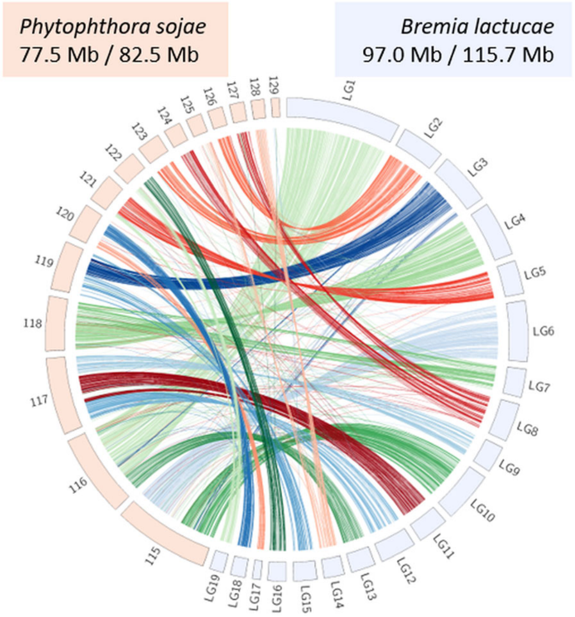

Fig. 7 Synteny between *Bremia lactucae* and *Phytophthora sojae*. Single-copy orthologs of *B. lactucae* and *P. sojae* were used to link genetically revised scaffolds of *B. lactucae* (light blue) with scaffolds of *P. sojae* larger than 1 Mb (pink). *P. sojae* scaffolds are labeled with the three-digit suffix of their NCBI accession (i.e., 115 is NW_009258115.1). Links are colored based on their assignment on *B. lactucae* scaffolds. Gene order is highly conserved between the two assemblies providing strong support for the quality of both

genome assembly with 99% of the markers assigned to the correct chromosome (Fig. 3b, d). The noise in the precise placement of the genetic bins and the low percentage of genetically placed markers (50.9% for Col, 63.4% for Ler-derived markers) was likely caused by missing data due to low sequence coverage, as shown with the simulated data (Table 2, Additional file 1). Despite the imperfect input data, AFLAP was able to produce a good genetic map using markers from each parent, concordant with the chromosome-scale genome assembly.

AFLAP was then applied to F_1 progeny isolates of *B. lactucae* generated by crossing isolate SF5 with either C82P24 or C98O622b, both of which are heterokaryotic; therefore, SF5 was effectively crossed to four different nuclei [6]. Progeny isolates were whole genome sequenced to at least 5x coverage to provide reliable identification of unique 31-mers. Heterokaryosis was reflected by half-sib clusters of isolates in the progeny (Fig. 5c). In addition to heterokaryosis, clustering of 31-mer markers also demonstrated that some isolates were genetically more similar to other isolates than expected, allowing potential duplicate individuals to be removed. Therefore, 83 isolates were genotyped with SF5-specific markers and used for linkage analysis (Additional file 3: Figure S3). The small population sizes for individual nuclei from the heterokaryotic parents meant that maps of C82P24 and C98O622b could not be constructed.

The genetic map of isolate SF5 of *B. lactucae* produced by AFLAP placed 98.8% of the SF5-specific markers into 19 linkage groups (Fig. 6a). The genetic map was highly concordant with large portions of the published genome assembly (Fig. 6b; 6). Discordance between the genetic map and the assembly was used to identify mis-assemblies; linkage data was then used to guide binning, orienting, and scaffolding, resulting in a much-improved genome assembly with 19 linkage-group-scale scaffolds (Fig. 6c). The more accurate placement of genetic bins on the assembly and higher percentage of mapped makers when compared to *A. thaliana* (Fig. 3b, d) is likely due to the higher coverage in the *B. lactucae* dataset. The size of the genetic map produced for *B. lactucae* is similar to that reported previously [4]. Therefore, with adequate sequencing depth, AFLAP was able to rapidly produce a genetic map of a non-model organism with a highly repeated genome [6]; the high marker density enabled genetically guided fragmentation and re-scaffolding of the genome assembly.

Not all scaffolds were placed on the linkage map. The marker sparse regions totaling 18.6 Mb of the current assembly were only marginally more repetitive than the genetically oriented sequence. Pseudo-test cross markers derived from isolates C82P24 and C98O622b did align to the large unplaced scaffold in the *B. lactucae* assembly (Additional file 3: Figure S4); therefore, it is possible that the unplaced regions over 1 Mb are homozygous in isolate SF5. In the current study, not enough progeny isolates were obtained from any of the nuclei of the heterokaryotic isolates C82P24 or C98O622b for genetic analysis. Therefore, additional genetic analysis of other isolates will further refine the *B. lactucae* genome assembly. Genotyping more progeny isolates and generating a consensus map will determine if *B. lactucae* has fewer than 19 chromosomes. Aligning the assemblies of *B. lactucae* and *P. sojae* allowed potential joins to be inferred based on synteny (e.g., Fig. 7; scaffold 117 of *P. sojae* suggests that linkage groups 9, 11, and 12 of *B. lactucae* might belong to a single chromosome). Alternatively, enhanced genetic resolution may demonstrate that the genomes of these distant relatives have undergone large-scale structural variation since divergence from their common ancestor. It is possible that applying AFLAP to *P. sojae* could further refine the *P. sojae* assembly, investigating syntenic joins suggested by the new assembly of *B. lactucae* (e.g., Fig. 7; linkage group 2 of *B. lactucae* joins scaffolds 123 and 127 of *P. sojae*).

Markers derived from fragments under 61 bp were investigated by rerunning AFLAP including markers derived from fragments equaling 60 bp. Many fragments smaller than $k + k - 1$ are probably derived from low complexity, repetitive, or hard to assemble sequences and are therefore uninformative. Some fragments equal to 60 bp will contain instances of deletions at a locus and therefore will be informative (Additional file 3: Figure S5). Rerunning AFLAP including markers derived from 60 bp fragments only added 4070 markers to the 96,226 markers used to construct the *B. lactucae* map (Fig. 6) and did not alter the ordering of markers (Additional file 3: Figure S6). Given that the very large number of markers far exceeded the number of crossovers, using markers derived from smaller fragments was unnecessary to generate accurate genetic maps. Indeed, AFLAP can generate robust genetic maps using only markers derived from 61 bp fragments. Depending on the genetics of the organism under study, there may be advantages to including markers derived from smaller fragments or only using markers derived from 61 bp fragments.

AFLAP has several technical benefits over other strategies for linkage analysis. It is not subject to biases that may be introduced by a reference assembly due to reads from reference alleles mapping more readily to an assembly than reads from alternative alleles [18] or associated SNP calling errors. In addition, AFLAP enables access to all single-copy portions of the genomes, some of which may not be present in the reference assembly. This may be particularly important when the parents of the mapping population are distantly related to the reference genotype. AFLAP makes it possible to genotype multi-nucleotide polymorphisms and indels in addition to SNPs; such variants are often inaccessible in conventional mapping approaches [19, 20]. The frequency of markers derived from fragments > 61 bp was ~ 50% for the *A. thaliana* F_2 maps, ~ 56% for the *B. lactucae* F_1 map, and > 80% for the *Lactuca* interspecific map (GBS data). Therefore, AFLAP removes bias in marker calling and increases access to variants and genetic markers, resulting in high-density maps. GBS/RADseq can be used to obtain high coverage, reduced representation sequencing data, and using ustacks, linkage analysis may be performed without an assembly [21]; however, library preparation for GBS/RADseq may introduce allele bias caused by restriction site distribution [22], a lack of robust genotype calls, and much lower marker density. The utility of AFLAP was demonstrated on a lettuce RIL population genotyped using GBS [14]. AFLAP generated nine linkage groups for each parent colinear with the 2.4 Gb genome assembly (Fig. 4) [15]. A nine-linkage group, 1711 cM compound map (Additional file 3: Figure S2) was concordant with a 1883 cM genetic map previously obtained via a conventional read alignment and variant calling workflow [14]. Therefore, AFLAP can efficiently generate accurate genotype tables for linkage analysis from GBS data. AFLAP also allows facile addition of data from new progeny individuals from the same or different populations that have a common parent to increase the genetic resolution of the map. Because each isolate is genotyped independently, adding new isolates generated from the same parents is equivalent to appending additional columns to the genotype table. Adding data from a new cross, but sharing one parent, can be achieved by filtering the 31-mer marker set against the new parent and removing markers from the genotype table that are no longer unique to the common parent. When analyzing the interspecific *Lactuca* spp. RILs (Additional file 3: Figure S2), the compound map was generated by concatenating the genotype table containing *L. serriola*-derived markers to the end of the genotype table containing *L. sativa*-derived markers (i.e., genotypes did not require recalculation). Therefore, AFLAP can incorporate new data easily, enabling rapid maturation of genetic maps.

AFLAP enables the construction of accurate genotype tables resulting in high-quality genetic maps for any organism using a segregating population sequenced to adequate depth. Analyses using simulated and real data demonstrated that the sequence depth obtained on progeny affects the accuracy of marker placement in the genetic map. Even low coverage sequencing (3x) is adequate to assign a marker to a correct linkage group with approximate placement. Simulations demonstrated that 5x WGS coverage was adequate for highly accurate marker placement (Table 2). The accuracy of the genetic placement of markers increased as progeny sequencing depth increased. The desired sequencing depth will therefore vary depending on the aims of the project. For validation of a chromosome scale assembly, low coverage may be adequate. For genetic orientation of a fragmented assembly, at least 5x coverage in the progeny is required. In simulated data, more markers were required to accurately place markers in genomes containing more chromosomes. AFLAP may be applicable to many datasets already generated or being generated. Also, WGS data generated for AFLAP can be easily repurposed for use in numerous other projects. AFLAP was validated with short-read data but could also be applied to high accuracy long-read data. Reads containing multiple errors would reduce the quality of genotyping and may impede the accuracy of AFLAP. Workflows, such as AFLAP, that use unbiased WGS as input will become increasingly desirable as the costs of library generation and sequencing continue to decrease. This may be critical to validating genome assemblies of non-model species generated in projects such as the Earth BioGenome Project [23].

Conclusions

AFLAP is a novel *k*-mer based approach to linkage analysis able to produce a high-density genetic map, without a prerequisite reference genome assembly. In addition, AFLAP can analyze complex variants and genomic regions that may not be accessible to variant calling approaches that use a reference genome assembly. AFLAP was benchmarked using multiple simulations, varying the sequencing depth of progeny and parents, parental heterozygosity, genome size, and chromosome number and contrasted to a conventional read alignment, variant calling workflow. AFLAP was validated using low coverage *A. thaliana* F_2 WGS data and was able to construct linkage groups that were coherent to the reference assembly when aligned; however, the use of low coverage data introduced significant noise. The utility of AFLAP when using GBS data was demonstrated by analyzing a RIL population of a *Lactuca* interspecific cross. AFLAP was then deployed to analyze 83 F_1 isolates of the non-model oomycete *B. lactucae* that had been whole genome sequenced to ≥ 5x. The genetic maps produced were unambiguously aligned to the reference assembly and resulted in significant improvements of the assembly. AFLAP can therefore be used to generate saturated genetic maps and to improve draft genome assemblies of non-model organisms provided a mapping population with adequate sequencing coverage is available.

Materials and methods

Whole genome sequencing

Whole genome sequencing of *A. thaliana* accessions Colombia (Col) and Landsberg (Ler) were downloaded from NCBI Short Read Archive (SRA) accessions SRR5882797 and SRR3166543 [24], respectively. Low coverage WGS reads of F_2 individuals generated from Col x Ler were obtained from previous studies [11–13]; 100 individuals with the largest gzipped files were selected for analysis (Additional file 2: Table S1). Reads for parents and RILs of a previously analyzed *L. serriola* Armenian999 x *L. sativa* PI251246 interspecific-cross were downloaded from NCBI BioProjects PRJNA642889, PRJNA510128, and PRJNA478460 [14].

Two *B. lactucae* crosses were analyzed with parental isolate SF5 in common. For both crosses, WGS of *B. lactucae* parental isolates SF5, C82P24, and C98O622b have been described previously (6; NCBI BioProject PRJNA387613). Thirty-seven previously reported F_1 progeny (6; NCBI BioProject PRJNA387454) and an additional 36 F_1 progeny from the same cross, made earlier [5], were added. Extracted DNA that had previously been used for RFLP linkage analysis [5] was used to construct ~ 350 bp libraries using the Lucigen (Middleton, WI, USA) NxSeq HT dual indexing kit, per the manufacturer's instruction, and 150 bp paired end reads were generated on an Illumina NovaSeq 6000 lane. For the second cross, oospores were obtained by co-inoculating isolates SF5 and C98O622b onto cv. Cobham Green. Oospores matured for several weeks in decaying plant tissue prior to maceration. Isolates were generated by growing cv. Cobham Green in a dilute oospore suspension, which was titrated via serial dilution so that on average a single seedling would be infected per culture box. DNA was extracted by vortexing sporangia for two minutes in a microcentrifuge tube with approximately 200 μL of Rainex-treated beads and 0.5 mL of 2× extraction buffer (100 mM Tris-HCl pH 8.0, 1.4 M NaCl, 20 mM EDTA, 2% [wt/vol] cetyltrimethylammonium bromide, and B-mercaptoethanol at 20 μL/mL), and then transferred to a fresh 2 mL tube. Material was treated with RNase (20 μL/mL; 65 °C for 30 min). An equal volume of 1:1 phenol/chloroform was added, mixed, and centrifuged at maximum speed (8000 rpm; 15 min). The aqueous phase was retained and further washed twice with equal volumes of 24:1 chloroform/isoamyl alcohol, obtaining the aqueous phase each time by centrifuging at maximum speed for 15 min. The resulting aqueous phase was mixed with 0.7 volumes of isopropanol and DNA was precipitated at − 20 °C for 1 h. DNA was pelleted by centrifuging at maximum speed for 30 min, washing with 70% ethanol, drying, and suspending in 10 mM Tris-HCl. Quantity and quality of DNA was determined by spectrometry, as well as estimated by TAE gel electrophoresis. Single index libraries of 23 F_1 isolates were generated by sonicating DNA to ~ 220 bp (Covaris, Woburn, MA, USA), cleaning and concentrating (1 part DNA: 1.2 parts AMPure, Beckman Coulter, Pasadena, CA USA), end repairing (End Repair Module #E6050L, New England Biolabs, Ipswich, MA, USA), cleaning (1 part DNA: 1.2 parts AMPure), A-base ligating (Klenow, Enzymatics/Qiagen, Hilden, Germany), and adapter ligating (T4 DNA Ligase #L603-HC-L, Enzymatics/Qiagen). Final cleanup and size selection was performed using 1-part DNA: 0.8 parts AMPure. Paired end, 150 bp reads were generated by sequencing the libraries with an Illumina HiSeq 4000. Reads new to this study have been deposited under BioProjects PRJNA387454 and PRJNA634525.

Assembly-free linkage analysis pipeline

Figure 1 provides an overview of the pipeline. For *A. thaliana*, *B. lactucae*, and *Lactuca* spp., 31-mer hashes were produced independently for the read sets of each parent and each progeny individual using JELLYFISH [25] sub-command count, parameters *-m31* -C -s *10G*. Different *k*-mer lengths are enabled, though longer *k*-mers will increase the complexity and running time of the pipeline, while shorter *k*-mers will reduce the ability to capture closely linked variants as a single marker. The parental (F_0) hashes were then inspected to identify single-copy 31-mers using the JELLYFISH sub-command histo to produce histograms of parental hashes, which were manually inspected to determine lower and upper bounds for filtering. For *A. thaliana*, where both parental accessions are highly homozygous and the population analyzed was an F_2, all single-copy 31-mers were retained. For *Lactuca* spp., where both parental lines were sequenced by reduced representation GBS, no single-copy 31-mers could be recovered, so user-specified limits of ≥ 20 to ≤ 45x were supplied to avoid sampling markers from under- or over-represented sequence. For *B. lactucae*, where both parental isolates are highly heterozygous and the population was an F_1, only the heterozygous 31-mers from either parent were retained. FASTA files were obtained for each parent using the JELLYFISH sub-command dump, parameters *-L [LowerLimit] -U [UpperLimit]*. For *A. thaliana* and *B. lactucae*, single-copy 31-mers were then queried against the opposite parental hash using JELLYFISH sub-command query and filtered for zero counts. The resulting 31-mers were single-copy and homozygous (*A. thaliana*) or heterozygous (*B. lactucae*) in one parent and absent in the alternate parent. For *Lactuca* spp., retained 31-mers were also queried against the opposite parental hash using JELLYFISH sub-command query and filtered for zero counts, so the resulting 31-mers were unique to either parent, although may not be single copy.

To reduce redundancy, 31-mers for each parent were assembled using ABYSS v2.2.2 [26], with the parameters *-k25 -c0 -e0* [27]. Assembling *k*-mers directly means that at least one variant position in the fragment produced is known and that downstream genotyping can rapidly occur using the same *k*-mer size. Assembled fragments equal to or greater than 61 bp were extracted and a single representative 31-mer equal to coordinates 10 to 41 was selected for each fragment, though any 31-mer would have sufficed. Fragments equal to 61 bp were considered single nucleotide variants and therefore were used as markers for the subsequent genetic analysis. Fragments larger than 61 bp were considered complex, multi-nucleotide variants or insertions relative to the alternate haplotype and were also used. These were validated by aligning the 1000 longest *A. thaliana* Col markers to the genome assemblies of both Col (GCF_000001735.4) and Ler (GCA_001651475.1) with BLASTn [28]. Fragments smaller than 61 bp were likely to contain repetitive or low complexity sequences that were not easily assembled and therefore not considered suitable for use as markers. Fragments equal to 60 bp would also contain deletions, relative to the alternate haplotype. The representative 31-mers were verified against the parental hashes to ensure they were (a) within the boundaries set for single-copy markers and (b) absent in the second parent. This set of markers was then scored against every progeny hash to obtain progeny genotypes. For the low coverage WGS (1x to 8x) *A. thaliana* and GBS *Lactuca* inter-cross lines, the presence of the marker was established by a single observation in the raw reads. For the high coverage (5x to >50x) *B. lactucae*, two or more observations were required for the marker to be scored, filtering out the majority of errors. It was not necessary to alter the thresholds for individuals sequenced to higher depths. Errors unique to one library will not segregate as expected so will be filtered out in later steps (Fig. 1). The scores were then collated into a genotype table, where 0 = marker not observed and 1 = marker observed.

To filter for siblings of *B. lactucae* with high identity, individuals were clustered by genotype. The Euclidean distance between progeny was calculated for the marker scores using the *dist* function of R [29] package *proxy* [30]. These were clustered with the *hclust* function, from which a dendrogram was calculated (*as.dendrogram* function) and plotted with the *heatmap.2* function from the package gplots [31]. Progeny with high identity to one another were reduced to a single representative isolate.

The genotype table was then converted to be compatible with Lep-MAP3 [10]. For *A. thaliana* (F_2 population) and *Lactuca* spp. (RIL population), the F_0 are coded as grandparents, and markers unique to each parent are coded AA in the accession that it was identified in and CC in the alternate parent, for which it was not identified. Inferred, genetically homogeneous F_1 parents were inserted with every marker coded AC. For F_2 progeny for which the genotype table was constructed, marker presence was coded AA and marker absence was coded CC. Genotype tables were constructed for markers sourced independently from either parent as well as a combined table from both. For *B. lactucae* (F_1 population), the identification of a marker was coded AC and marker absence was coded CC in both parents and progeny.

Lep-MAP3 subcommand SeperateChromosomes2 was run on the genotype table to assign markers to linkage groups using the following parameters: *lodLimit = 7* for *A. thaliana*, *lodLimit = 20* for *L. sativa*, and *lodLimit = 3* for *B. lactucae*. Lep-MAP3 subcommand OrderMarkers2 was subsequently used to order markers within each linkage group using a Morgan mapping function with 20 iterations. The native output file of LepMap3 OrderMarkers2 did not retain the original marker names; instead, these are derived from the genotype table using Linux join. A shell script is provided in the GitHub repository [8] and will generate the output described here. Linkage groups produced were visualized using the R [29] packages dplyr [32], ggplot2 [33], and ungeviz [34].

To validate the genetic maps produced by AFLAP, the linkage groups were aligned against corresponding reference assemblies (*A. thaliana*; GCF_000001735.4, *B. lactucae*; GCA_004359215.1, *L. sativa* GCA_002870075.2) by mapping 31-mers to the assembly with bwa aln [35], converting to a sorted BAM file with SAMtools v1.9.1 subcommand sort [36], and filtering for uniquely mapped 31-mers with a maximum one edit distance. Genetic coordinates were visualized across the physical assembly as scatter plots in R using ggplot2 [33], colored by linkage group. For *A. thaliana*, the average physical position of each genetic bin was calculated and plotted due to the low coverage of the F_2 individuals.

For the simulations, parents and progeny were simulated for multiple organisms representing different genome sizes. The genome assemblies of *A. thaliana* (119 Mb, five chromosomes, NCBI: GCF_000001735.4), *Vitis riparia* (500 Mb, 19 chromosomes, NCBI: GCA_004353265.1), and *Atriplex hortensis* (900 Mb, nine chromosomes, CoGe Genome ID: 56906) were used to simulate F_1 crosses where markers only segregate from one haplotype (ABxCC). For all assemblies, SNPs and indels were introduced to produce a synthetic parent (AB) containing 0.2% heterozygosity using mutate.sh [37]. The other parent was 100% homozygous and represented by the reference assembly. The impact of varying heterozygosity was tested at levels approximate to 0.01%, 0.1%, 0.2%, 0.5%, 1%, and 2% using the smallest simulated genome. Progeny were generated by randomly assigning an inherited haplotype and either one or two cross-overs along each chromosome (https://github.com/kfletcher88/CrossSimulator). Parental sequencing depth was simulated to 50x and progeny sequencing depth to 10x whole genome coverage using randomreads.sh [37], generating 150 bp paired-end reads, with default error rates and insert lengths. Simulations were also run varying the parental sequencing depth to 10x, 20x, 30x, and 40x and the progeny sequencing depth to 3x, 5x, 7x, and 20x. F_2 crosses were simulated using the smallest genome assembly simulating both parents to be 100% homozygous, varying from one another by 0.1%, and synthesizing reads for the parents at 50x and the progeny at 10x. All AFLAP simulations were run using a subsampled marker set using the shell script AFLAP.sh available at https://github.com/kfletcher88/AFLAP/archive/v0.03.tar.gz.

To compare AFLAP with a contemporary SNP based pipeline, the simulated paired-end reads, described above, for the smallest genome assembly were mapped back to the reference assembly using BWA mem (v0.7.17) [38] and SNPs were called using FreeBayes (v1.3.1) [39]. The VCF file was recoded using VCFtools (v0.1.16) and converted into a LepMap3 compatible file using the template R-code available at https://github.com/rkbhan/GeneticsTools.git. LepMap3 was run using the entire marker set and a downsampled subset for comparison with AFLAP. All simulated data was run in series using 12 threads writing to a scratch drive enabling time comparisons. The same cross-over coordinates were used for these simulations, so that the map length calculated by AFLAP and the conventional pipelines were comparable. A minimum LOD score of seven was used for AFLAP runs with both the full marker set and the downsampled marker set. For the conventional run, a minimum LOD score of seven was used for the downsampled marker set. For the full marker set, a minimum LOD score of 20 was required to resolve the five linkage groups; lower minimum LOD scores resulted in some linkage groups being erroneously joined to one another. Correlation of each synthetic genetic map with the original genome assembly, from which the synthetic data was derived, was inferred by plotting the genetic coordinates by physical coordinates and calculating the Kendall rank coefficient [40]. All computations were done on the UC Davis Genome Center computing cluster.

Curation of the reference genome assembly of *B. lactucae*

Linkage data was used to identify and break chimeric scaffolds outside of gene boundaries followed by linkage-guided reorientation and rejoining of genetically congruent scaffolds. To identify chimeric scaffolds, continuous runs of genetically placed markers were quantified along each scaffold. Spurious results, where a run of less than 10 markers was identified on a scaffold, were disregarded. Scaffolds were broken on the final marker of a run of 10 or more markers, or beyond the boundaries of any genes that were identified to contain the marker. A gene was retained in the linkage group when there was at least one marker providing evidence for the segregation of that gene with the linkage group. The average genetic position of each scaffold was determined based on the genetic position of markers mapped to it. Scaffolds upon which recombination could be detected were oriented based on the average physical position of the genetic bins mapped to each scaffold. Oriented scaffolds were joined with a string of 100 Ns. Unresolved marker sparse or void regions were excised from the scaffold and retained as unlinked scaffolds within the assembly file. Genetic markers were remapped onto the genetically oriented assembly for validation. Hi-C reads previously generated from the same isolate (6; BioProject PRJNA387613) were aligned to the genetically oriented assembly and contact frequencies visualized with Hi-C explorer v2.2 [41] to validate the assembly. Repeats were masked with RepeatMasker v4.0.9 [42] and a previously defined repeat library [6]. Genic annotations were lifted over to the new assembly using Liftoff v1.3.0 [43]. Single-copy orthologs previously identified [6] between *B. lactucae* and *Phytophthora sojae* (GCF_000149755.1) were used to plot synteny with Circos v0.69-8 [44].

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13059-021-02326-x.

Additional file 1. Plots of simulated genetic maps aligned back to their respective assemblies, in support of Table 1 and Table 2.

Additional file 2: Table S1. A pedigree file which could be used to run AFLAP, includes accession numbers for *A. thaliana* WGS reads used.

Additional file 3: Supplementary figures. **Figure S1.** Alignment of *Arabidopsis thaliana* land race Columbia markers to *A. thaliana* reference assemblies. **Figure S2.** Compound and parental maps of *Lactuca* spp. generated with AFLAP. **Figure S3.** Corrected *B. lactucae* kinship clustering heatmap. **Figure S4.** Percent of bases covered by *B. lactucae* pseudo-test cross markers derived from each parent. **Figure S5.** Associating variants with different fragment sizes. **Figure S6.** Addition of markers derived from 60 bp fragments to the SF5 genetic map.

Additional file 4: Table S2. High identity *B. lactucae* progeny isolates identified in Fig. 5c.

Additional file 5. Review history.

Acknowledgements

We thank B. Rowan (UC Davis) for discussion of results produced by AFLAP on the *A. thaliana* data, W. J. Palmer (UC Davis) for discussions on linkage analysis software to integrate into the pipeline, H. Xu (UC Davis) for managing raw data submissions, and E. Georgian (UC Davis) for editorial services. The sequencing was carried out by the DNA Technologies and Expression Analysis Cores at the UC Davis Genome Center, supported by NIH Shared Instrumentation Grant 1S10OD010786-01. The bioinformatic analysis was carried out using the UC Davis LSSC0 High Performance Computing cluster maintained by the UC Davis Bioinformatics Core.

Review history

The review history is available as Additional file 5.

Peer review information

Anahita Bishop was the primary editor of this article and managed its editorial process and peer review in collaboration with the rest of the editorial team.

Authors' contributions

KF conceptualized the project, performed the bioinformatic analysis, and drafted the manuscript. LZ and JG produced isolates of *B. lactucae* and prepared libraries for sequencing. RH conducted the GBS analysis and contributed code to the project. KC prepared libraries of additional isolates. RM supervised the project, contributed to data analysis and all drafts. All authors contributed to writing and editing the paper and have approved the final submission.

Funding

The work was supported by The Novozymes Inc. Endowed Chair in Genomics to RWM.

Availability of data and materials

All *Bremia* data reported in this paper are available from NCBI BioProjects PRJNA387613, PRJNA387454, PRJNA387192, and PRJNA634525 [6, 45, 46]. The reoriented genome assembly and annotations are available under NCBI GenBank accession GCA_004359215.2. Previously published *A. thaliana* reads are available at NCBI short read archive SRR3166543 and SRR5882797 [24] and EBI ArrayExpress E-MTAB-4657, E-MTAB-5476, and E-MTAB-8165 [11–13]. Previously published *Lactuca* reads are available in NCBI BioProjects PRJNA642889, PRJNA510128, and PRJNA478460 [14]. Scripts to run AFLAP are available at https://github.com/kfletcher88/AFLAP [8] and archived at https://doi.org/10.5281/zenodo.4552613 [9], both freely distributed under an MIT license.

Declarations

Ethics approval and consent to participate

N/A

Consent for publication

N/A

Competing interests

The authors declare that there are no competing interests.

Received: 10 September 2020 Accepted: 25 March 2021
Published online: 21 April 2021

References

1. Bateson W, Saunders E, Punnett R. Experimental studies in the physiology of heredity. Reports to the Evolution Committee. Proc R Soc B. 1906;77:236–8.
2. Sutton WS. The chromosomes in heredity. Biol Bull. 1903;4(5):231–50. https://doi.org/10.2307/1535741.
3. Sturtevant AH. The linear arrangement of six sex-linked factors in *Drosophila*, as shown by their mode of association. J Exp Zool. 1913;14(1):43–59. https://doi.org/10.1002/jez.1400140104.
4. Hulbert SH, Ilott TW, Legg EJ, Lincoln SE, Lander ES, Michelmore RW. Genetic analysis of the fungus, *Bremia lactucae*, using restriction fragment length polymorphisms. Genetics. 1988;120(4):947–58. https://doi.org/10.1093/genetics/120.4.947.
5. Sicard D, Legg E, Brown S, Babu NK, Ochoa O, Sudarshana P, et al. A genetic map of the lettuce downy mildew pathogen, *Bremia lactucae*, constructed from molecular markers and avirulence genes. Fungal Genet Biol. 2003;39(1):16–30. https://doi.org/10.1016/S1087-1845(03)00005-7.
6. Fletcher K, Gil J, Bertier LD, Kenefick A, Wood KJ, Zhang L, et al. Genomic signatures of heterokaryosis in the oomycete pathogen *Bremia lactucae*. Nat Commun. 2019;10(1):2645. https://doi.org/10.1038/s41467-019-10550-0.
7. Treangen TJ, Salzberg SL. Repetitive DNA and next-generation sequencing: computational challenges and solutions. Nat Rev Genet. 2011;13(1):36–46. https://doi.org/10.1038/nrg3117.
8. Fletcher K. Assembly Free Linkage Analysis Pipeline. GitHub. 2021; https://github.com/kfletcher88/AFLAP.
9. Fletcher K. Assembly Free Linkage Analysis Pipeline. Zenodo. 2021; https://zenodo.org/record/4552613#.YFjuR69KiUk.
10. Rastas P. Lep-MAP3: robust linkage mapping even for low-coverage whole genome sequencing data. Bioinformatics. 2017;33(23):3726–32. https://doi.org/10.1093/bioinformatics/btx494.
11. Choi K, Reinhard C, Serra H, Ziolkowski PA, Underwood CJ, Zhao X, et al. Recombination rate heterogeneity within *Arabidopsis* disease resistance genes. PLoS Genet. 2016;12(7):e1006179. https://doi.org/10.1371/journal.pgen.1006179.
12. Rowan BA, Heavens D, Feuerborn TR, Tock AJ, Henderson IR, Weigel D. An ultra high-density *Arabidopsis thaliana* crossover map that refines the influences of structural variation and epigenetic features. Genetics. 2019;213(3):771–87. https://doi.org/10.1534/genetics.119.302406.
13. Underwood CJ, Choi K, Lambing C, Zhao X, Serra H, Borges F, et al. Epigenetic activation of meiotic recombination near *Arabidopsis thaliana* centromeres via loss of H3K9me2 and non-CG DNA methylation. Genome Res. 2018;28(4):519–31. https://doi.org/10.1101/gr.227116.117.
14. Han R, Wong AJY, Tang Z, Truco MJ, Lavelle DO, Kozik A, et al. Drone phenotyping and machine learning enable discovery of loci regulating daily floral opening in lettuce. J Exp Bot. 2021;72(8):2979–94. https://doi.org/10.1093/jxb/erab081.
15. Reyes-Chin-Wo S, Wang Z, Yang X, Kozik A, Arikit S, Song C, et al. Genome assembly with in vitro proximity ligation data and whole-genome triplication in lettuce. Nat Commun. 2017;8(1). https://doi.org/10.1038/ncomms14953.
16. Giraut L, Falque M, Drouaud J, Pereira L, Martin OC, Mézard C. Genome-wide crossover distribution in *Arabidopsis thaliana* meiosis reveals sex-specific patterns along chromosomes. PLoS Genet. 2011;7(11):e1002354. https://doi.org/10.1371/journal.pgen.1002354.
17. Kuittinen H, de Haan AA, Vogl C, Oikarinen S, Leppälä J, Koch M, et al. Comparing the linkage maps of the close relatives *Arabidopsis lyrata* and *A. thaliana*. Genetics. 2004;168(3):1575–84. https://doi.org/10.1534/genetics.103.022343.
18. Ballouz S, Dobin A, Gillis JA. Is it time to change the reference genome? Genome Biol. 2019;20(1):159. https://doi.org/10.1186/s13059-019-1774-4.
19. Kaplanis J, Akawi N, Gallone G, McRae JF, Prigmore E, Wright CF, et al. Exome-wide assessment of the functional impact and pathogenicity of multinucleotide mutations. Genome Res. 2019;29(7):1047–56. https://doi.org/10.1101/gr.239756.118.
20. Sandmann S, de Graaf AO, Karimi M, van der Reijden BA, Hellström-Lindberg E, Jansen JH, et al. Evaluating variant calling tools for non-matched next-generation sequencing data. Sci Rep. 2017;7(1):43169. https://doi.org/10.1038/srep43169.
21. Hohenlohe PA, Bassham S, Etter PD, Stiffler N, Johnson EA, Cresko WA. Population genomics of parallel adaptation in threespine stickleback using sequenced RAD tags. PLoS Genet. 2010;6(2):e1000862. https://doi.org/10.1371/journal.pgen.1000862.
22. Andrews KR, Good JM, Miller MR, Luikart G, Hohenlohe PA. Harnessing the power of RADseq for ecological and evolutionary genomics. Nat Rev Genet. 2016;17(2):81–92. https://doi.org/10.1038/nrg.2015.28.
23. Lewin HA, Robinson GE, Kress WJ, Baker WJ, Coddington J, Crandall KA, et al. Earth BioGenome Project: sequencing life for the future of life. Proc Natl Acad Sci. 2018;115(17):4325–33. https://doi.org/10.1073/pnas.1720115115.
24. Zapata L, Ding J, Willing EM, Hartwig B, Bezdan D, Jiao WB, et al. Chromosome-level assembly of *Arabidopsis thaliana* Ler reveals the extent of translocation and inversion polymorphisms. Proc Natl Acad Sci. 2016;113(28):E4052–60. https://doi.org/10.1073/pnas.1607532113.

25. Marçais G, Kingsford C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. Bioinformatics. 2011;27(6):764–70. https://doi.org/10.1093/bioinformatics/btr011.
26. Jackman SD, Vandervalk BP, Mohamadi H, Chu J, Yeo S, Hammond SA, et al. ABySS 2.0: resource-efficient assembly of large genomes using a Bloom filter. Genome Res. 2017;27(5):768–77. https://doi.org/10.1101/gr.214346.116.
27. Rahman A, Hallgrímsdóttir I, Eisen M, Pachter L. Association mapping from sequencing reads using k-mers. eLife. 2018;7:e32920. https://doi.org/10.7554/eLife.32920.
28. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990;215(3):403–10. https://doi.org/10.1016/S0022-2836(05)80360-2.
29. R Development Core Team. R: a language and environment for statistical computing. Vienna: R Foundation for Statistical Computing; 2012.
30. Meyer D, Buchta C. proxy: distance and similarity measures; 2019.
31. Warnes G, Bolker B, Bonebakker L, Gentleman R, Huber W, Liaw A, et al. gplots: various R programming tools for plotting data. 2020.
32. Wickham H, Francois R, Henry L, Müller K. dplyr: a grammar of data manipulation; 2019.
33. Wickham H. ggplot2: elegant graphics for data analysis. 2nd ed. New York City: Springer International Publishing; 2016. https://genomebiology.biomedcentral.com/articles/10.1186/s13059-021-02293-3.
34. Wilke CO. ungeviz: tools for visualizing uncertainty with ggplot2; 2020.
35. Li H, Durbin R. Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics. 2009;25(14):1754–60. https://doi.org/10.1093/bioinformatics/btp324.
36. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The sequence alignment/map (SAM) format and SAMtools. Bioinformatics. 2009;25(16):2078–9. https://doi.org/10.1093/bioinformatics/btp352.
37. Bushnell B. BBMap short read aligner. Berkeley: University of California; 2016. https://sourceforge.net/projects/bbmap/.
38. Li H et al. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. 2013;arXiv:1303.3997v2. https://genomebiology.biomedcentral.com/articles/10.1186/s13059-020-02217-7#Bib1.
39. Garrison E, Marth G. Haplotype-based variant detection from short-read sequencing. arXiv. 2012:12073907.
40. Kendall MG. A new measure of rank correlation. Biometrika. 1938;30(1/2):81–93. https://doi.org/10.1093/biomet/30.1-2.81.
41. Ramírez F, Bhardwaj V, Arrigoni L, Lam KC, Grüning BA, Villaveces J, et al. High-resolution TADs reveal DNA sequences underlying genome organization in flies. Nat Commun. 2018;9(1):189. https://doi.org/10.1038/s41467-017-02525-w.
42. Smit A, Hubley R, Green P. RepeatMasker open-4.0; 2013.
43. Shumate A, Salzberg SL. Liftoff: an accurate gene annotation mapping tool. Bioinformatics. 2020:btaa1016. https://doi.org/10.1093/bioinformatics/btaa1016.
44. Krzywinski MI, Schein JE, Birol I, Connors J, Gascoyne R, Horsman D, et al. Circos: an information aesthetic for comparative genomics. Genome Res. 2009;19(9):1639–45. https://doi.org/10.1101/gr.092759.109.
45. Fletcher K. AFLAP: assembly-free linkage analysis pipeline using k-mers from genome sequencing data. Gene Expression Omnibus https://www.ncbi.nlm.nih.gov/bioproject/PRJNA634525 (2021).
46. Fletcher K. AFLAP: assembly-free linkage analysis pipeline using k-mers from genome sequencing data. Gene Expression Omnibus. 2021. https://www.ncbi.nlm.nih.gov/bioproject/PRJNA387454.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.