Chapter 12

Allele Identification in Assembled Genomic Sequence Datasets

Katrina M. Dlugosch and Aurélie Bonin

Abstract

Allelic variation within species provides fundamental insights into the evolution and ecology of organisms, and information about this variation is becoming increasingly available in sequence datasets of multiple and/or outbred individuals. Unfortunately, identifying true allelic variants poses a number of challenges, given the presence of both sequencing errors and alleles from other closely related loci. We outline the key considerations involved in this process, including assessing the accuracy of allele resolution in sequence assembly, clustering of alleles within and among individuals, and identifying clusters that are most likely to correspond to true allelic variants of a single locus. Our focus is particularly on the case where alleles must be identified without a fully resolved reference genome, and where sequence depth information cannot be used to infer the putative number of loci sharing a sequence, such as in transcriptome or post-assembly datasets. Throughout, we provide information about publicly available tools to aid allele identification in such cases.

Key words: Allelic variation, Paralogs, Gene duplication, Maximum likelihood clustering, Single-linkage clustering, AllelePipe, Granularity, Transcriptome data, Next-generation sequencing

1. Introduction

Surveys of intra-specific molecular genetic variation now form the core of evolutionary studies of individual organisms. The frequencies of mutations segregating within and across populations can reveal the history of migration, gene flow, demography, recombination, and natural selection in a species, as well as the genetic basis of its phenotypes (1–4). Since the early allozyme studies in the late 1960s, it has been clear that such genetic diversity is pervasive in living organisms (5); yet, only now are we getting a clear picture of the extent and nature of this diversity through a few model species. For example, the latest data from the 1000-genome project identified

François Pompanon and Aurélie Bonin (eds.), *Data Production and Analysis in Population Genomics: Methods and Protocols*, Methods in Molecular Biology, vol. 888, DOI 10.1007/978-1-61779-870-2_12, © Springer Science+Business Media New York 2012

15 million single nucleotide polymorphisms (SNPs), one million short insertion–deletion (indel) mutations, and 20,000 structural variants in the human genome, most of which were previously unknown (6). Similarly, a recent whole-genome resequencing effort revealed >800,000 unique SNPs and ~80,000 unique 1- to 3-bp indels in two divergent strains of the plant model *Arabidopsis thaliana*, relative to its reference genome (7).

At the gene or haplotype level, these individual polymorphisms combine to generate distinct allelic forms. New alleles are continually created with each new mutation, and these rise and fall in frequency in response to both drift and selection. In some cases, selection appears to have favored the retention and proliferation of large numbers of segregating alleles via negative frequency-dependent selection (e.g., those involved in pathogen recognition and plant self-incompatibility systems (8, 9)). For example, no less than 241 different alleles have been identified so far for the gene determining the ABO blood group in humans (10). Importantly, variation in the nature of drift and selection experienced by individual loci has led to striking variation in the number of mutations that separate a given pair of alleles: observations of intra-specific allelic divergence within model eukaryotes range over two orders of magnitude *within* species, from <0.1% to >10%, for synonymous site divergence in coding regions (11–14). Variation in the level of divergence among alleles poses a challenge for our analyses of genomic sequence surveys, which are increasingly available. If we hope to identify alleles that are segregating at the same locus, how similar should we expect their sequences to be? How do we distinguish these from sequences belonging to other loci?

Ideally, sequences representing alleles from different loci would at least show a consistently higher level of divergence from one another than exists among alleles of the same locus. Unfortunately, insights from whole genome sequencing indicate that this criterion will be violated frequently as a result of ongoing gene duplication (15–19). For inbred or haploid genotypes of most eukaryotes studied to date, analyses of synonymous site divergence among genes reveal a characteristic frequency distribution sensu (18), wherein the genome includes many highly similar paralogous loci and fewer and fewer paralogs at higher levels of divergence (Fig. 1). These patterns are consistent with a high rate of both gene duplication and loss (17), and indeed duplication rates have been estimated to meet or exceed rates of SNP mutations per generation for many loci (20). This frequent formation of paralogs means that the genome is populated with loci that are separated by levels of divergence (among one another) that span the divergence among their own alleles (e.g., (21)).

There are some options for working around the problem of alleles that cannot be disentangled among paralogous loci. Prior to widespread genomic studies of nonmodel and natural populations,

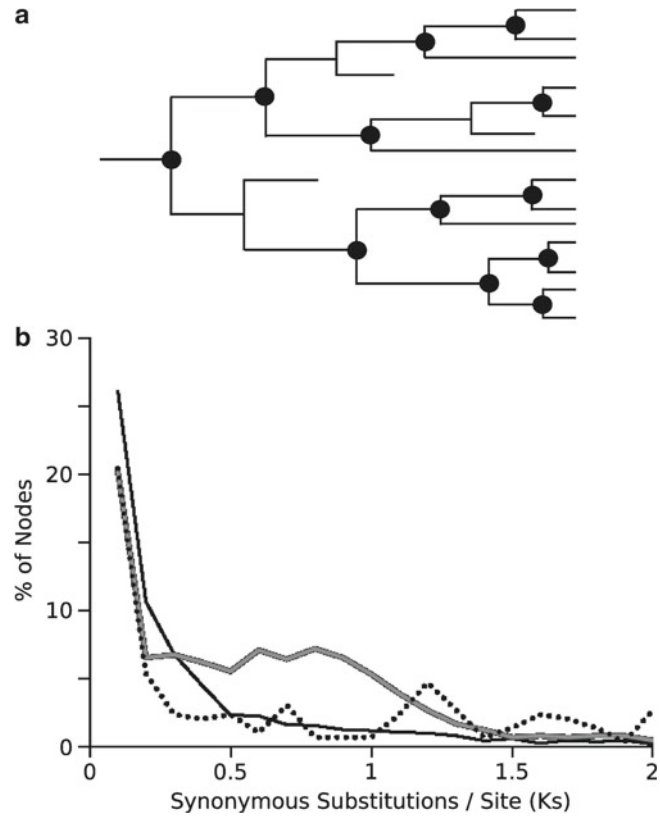

Fig. 1. (**a**) Gene tree showing gene duplication and loss in the history of a single genome. *Dots* indicate nodes that can be reconstructed from extant sequences. (**b**) Frequency distribution of synonymous site divergence at gene duplication nodes for human (*solid black line*, NCBI build 36.43), insect *Drosophila melanogaster* (*dotted line*, Berkeley *Drosophila* Genome Project build 4.3.4) and plant *Arabidopsis thaliana* (*gray line*, TAIR build 9). All species show a characteristic distribution with many recent duplication events; the older peaks in *A. thaliana* reveal ancient genome duplication (42). The most recent duplication events (those below 0.1 Ks divergence) are similar in divergence as are alleles, and these dominate at 20–30% of duplications in these genomes.

allelic variation was simply avoided through the use of inbred or haploid tissue. Once a full genome reference library was in hand, subsequent resequencing efforts identified (and continue to identify) alleles by mapping sequences onto these reference genomes (reviewed in (22, 23)). If a sequence cannot be uniquely mapped to one location on the reference, as expected due to problems with closely related paralogs, it is typically discarded from any further analysis of diversity. Note that variation in paralog number among individuals and incomplete assembly of the reference genome will both introduce error into this mapping process.

For nonmodel organisms without a reference genome, the situation becomes considerably more complicated because sequences must be clustered somehow into groups of putative alleles.

Existing methods for grouping similar sequences within and among species typically rely on the use of arbitrary sequence similarity thresholds or distributions for allowable allelic divergence (e.g., refs. 24–27 and references therein; 28). For surveys of genomic DNA, read depth information (the number of reads that align to a particular position) can provide a valuable indicator of potential problems with paralogs, where sudden increases in read depth, relative to the average, signal that multiple loci might have clustered together as one. This approach is presented in the chapter dealing with RAD tag assembly and analysis (see the chapter by Hohenlohe et al., this volume). For surveys of expressed transcripts (i.e., EST/transcriptome/cDNA sequencing, and RNA-seq), read depth information is unrelated to the frequency of occurrence in the genome and cannot be used to identify problematic clusters.

In this chapter, we examine the steps involved in identifying and clustering alleles in genomic data for which read depth is not informative (transcriptome or already assembled datasets), and a reference genome is not available. This situation is becoming increasingly common in transcriptome surveys of nonmodel organisms and in comparative genomic analyses using published assemblies. We leverage the information afforded by genomic data for *multiple individuals* within a species, when available. We also describe our own publicly available software AllelePipe, a pipeline to aid in moving data through analyses of allelic variation (http://EvoPipes.net/AllelePipe.html).

2. General Procedure to Identify Allelic Clusters

2.1. Sequence Assembly

Sequence assembly is a nontrivial task, making published assemblies a valuable resource for further analyses. To identify alleles in either previously assembled data or new assemblies, it is critical to consider the parameter decisions that have affected the reconstruction of alleles and paralogs during assembly. Fundamentally, sequence assembly is a process of merging reads that are highly similar. Typically, neither de novo nor reference-based assemblies require exact matches before merging reads. A certain amount of sequence divergence is allowed in successful matches because all sequencing methods are prone to error. A high depth of coverage (many reads aligned together at the same position) allows subsequent bioinformatic error estimation and correction via majority-rule, maximum likelihood, or Bayesian methods (e.g., (29–33)) The sequence divergence (or similarity) cut-offs for merging reads are set by the user, and their stringency will necessarily impact the degree to which highly similar alleles/paralogs are seen as error and are merged into the same contigs. A high depth of coverage will also help to avoid problems with allele/paralog merging

because most assembly programs detect strong support for multiple versions of a sequence, and separate these into different contigs—although this solution can be problematic if errors occur multiple times at high coverage positions (29). In some cases, authors of the assembly may intentionally adjust settings to try and collapse allelic variation in a sample, in order to obtain a single consensus genomic sequence from outbred individuals or multiple strains. It is also important to note that recently diverged alleles are by their nature separated by a very low density of mutations, which are less likely to be detected by short-read sequences and assembly programs.

There are a variety of ways to assess the dataset quality once the assembly has been completed. The number and length of contigs, the percentage of reads assembling, and the recovery of gene families known from other organisms are all common metrics of the completeness of a genomic or transcriptomic survey (e.g., (34)). None of these metrics explicitly examines the resolution of alleles and paralogs unless these are already identified by mapping to a reference genome. We advocate two approaches for datasets without a reference. First, known highly conserved single copy loci should be represented by roughly the expected number of alleles per individual for the ploidy and heterozygosity of a given species. A number of single copy gene datasets are available for different groups of organisms, including a recently developed list of highly conserved orthologs across all of eukaryotes (35) available at http://compgenomics.ucdavis.edu/compositae_reference.php. These can be searched against an assembly with discontiguous MegaBLAST or tBLASTx ((36, 37); available at http://blast.ncbi.nlm.nih.gov/Blast.cgi), and the number of matches examined closely.

Second, most genomic and transcriptomic datasets should yield the expected distribution of divergence events for paralogs and alleles in gene families (Fig. 1), where there is clear peak at low divergence. These distributions can be created using the DupPipe (38) at Evopipes.net (http://evopipes.net/dup_pipe.html). Over-assembled data will produce truncated curves (Fig. 2, dotted line), where close paralogs and alleles have been merged together and these recent divergence events are not seen. Under-assembled data, where many near-identical copies have failed to assemble will produce extreme front peaks in the distribution (Fig. 2, dashed line). For example, a pattern of under-assembly is common in output from the assembly software MIRA (39), which is otherwise outstanding for its resolution of highly similar copies, but has a tendency to produce many duplicates in areas of high coverage. The pipeline iAssembler (e.g., (40); available at http://bioinfo.bti.cornell.edu/tool/iAssembler/) has been created to combat this problem by iteratively assembling datasets with both MIRA and CAP3 (41), the latter being the standard assembly tool of the Sanger

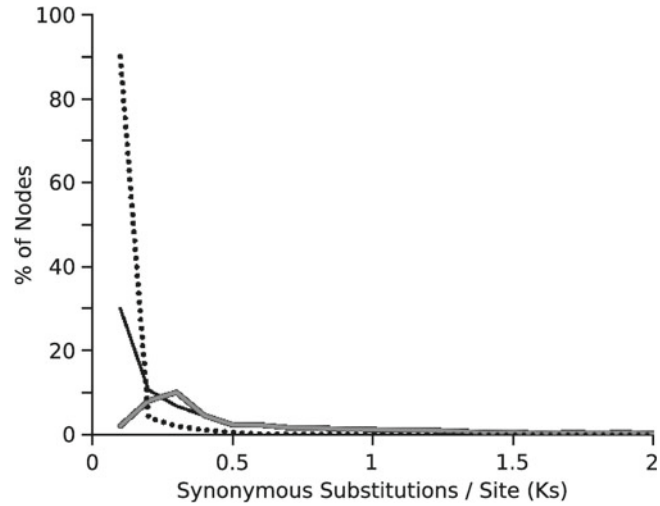

Fig. 2. Example frequency distributions of synonymous site divergence since gene duplication events within a genome for properly assembled (*solid black line*), over-assembled (*gray line*), and under-assembled sequences (*dotted line*). Over-assembled sequences will lack recent duplication events while under-assembled data will be strongly dominated by apparent close duplicates.

sequencing era. Note that additional peaks may appear in these distributions due to past genome duplication events (e.g., (38, 42)), but these will not result in the over- and under-assembly patterns described here.

2.2. Sequence Clustering

The first step in identifying allelic variation across a genomic dataset is to group similar contigs within and/or among individuals into clusters that might represent individual loci. Deciding what is meant by "similar" is a fundamental challenge. As noted above, alleles can vary widely in their level of divergence. Wang and colleagues (27) found that minimum similarity thresholds on the order of 90% might be appropriate for clustering transcripts of the same gene, in transcriptome data from *A. thaliana*. The most detailed information about the typical variation among alleles in a genome under study can come from the data themselves, by assessing the divergence of putative alleles identified between conspecific *individuals* in the dataset. Highly similar sequence matches among conspecifics will either be identical (same allele) or divergent because they represent different alleles segregating in the species, with error introduced by close paralogs. The median similarity of reciprocal best matches between contigs from two individuals should give an idea of the typical similarity expected between any two sequences that are part of the same locus (Fig. 3), and can be used as a starting minimum similarity for clustering contigs in a dataset. A list of reciprocal best hits can be created using the RBH pipeline at EvoPipes.net (http://evopipes.net/rbhpipe.html).

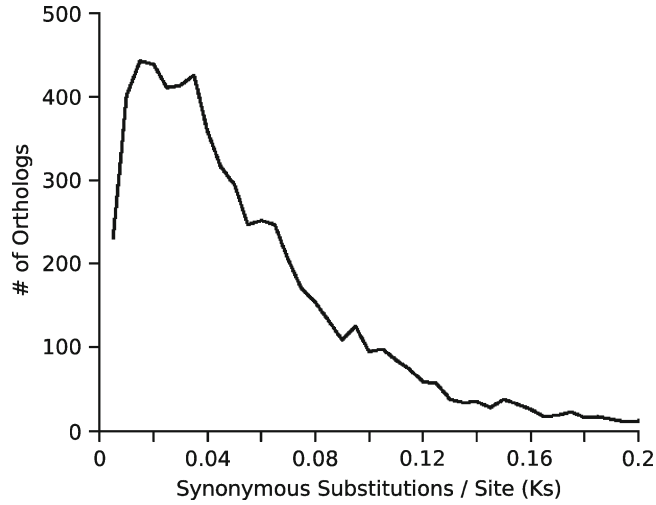

Fig. 3. Frequency distribution of synonymous site divergence of putative alleles of the same locus between individuals, based upon reciprocal best matches between transcriptomes of two *Centaurea solstitialis* plants (see Note 1).

Sequence similarity across the dataset can be found by searching for alignments of all sequences against all sequences (across one or many individuals). Many programs are available for this purpose, such as MegaBLAST (37) in the downloadable form of the BLAST package (available at http://blast.ncbi.nlm.nih.gov/Blast.cgi), and more recently developed high throughput programs, such as SSAHA2 and SMALT, from the Sanger Institute ((43); both available at http://www.sanger.ac.uk/resources/software/). Below, we describe a pipeline that we have developed to execute searching and subsequent clustering steps. In general, two key criteria are important when conducting these types of searches and filtering for desired matches. First, alignments should be continuous and include the entire region of overlap. Small indels may be acceptable, but large gaps are not expected among alleles within a species, with rare exception of intron variation in genomic DNA (44). Complete alignment throughout overlaps is a criterion also used by assembly software, but not by commonly used local alignment searches—such as BLAST—so alignments may have to be verified by custom scripts when these tools are used. Second, alignment lengths need to be sufficiently long to quantify the degree of similarity. For example, 95% "neutral" site similarity implies that an *average* of only five SNPs will separate alleles across 100 bp of alignment in noncoding regions. For coding regions, the number of codons and synonymous sites are a fraction of the total sequence length, and minimum overlaps on the order of ~300 bp (100 codons) may be prudent for properly assessing sequence similarity.

Once similarity between pairs of sequences is known, these must be aggregated into clusters. The simplest form of aggregation

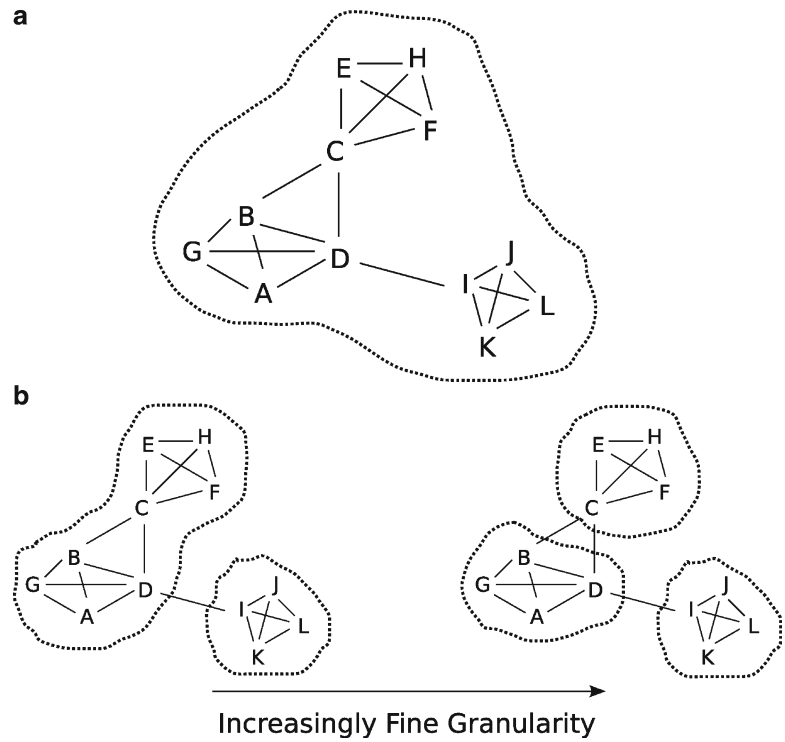

Fig. 4. Diagram of clusters (*dashed outlines*) for alleles (*letters*) connected by sequence similarity (*lines*) using either (**a**) single linkage clustering, or (**b**) graphical clustering with increasingly fine granularities.

is "single-linkage" clustering, where any sequences that share sufficient similarity are placed in the same cluster (Fig. 4a). This is highly inclusive, and can lead to large clusters where increasingly divergent sequences are added because they share sufficient similarity with at least one existing member of a cluster. To avoid the potential problem of over-aggregation of sequences by single-linkage clustering, a maximum likelihood approach to grouping clusters of highly similar sequences has been developed as the Markov Cluster Algorithm provided in the software MCL (http://micans.org/mcl/), and implemented for protein families as OrthoMCL (45). This approach considers all sequences in a network (graph) analysis, where sections of the network with many strong links (here sequence similarities) are identified as clusters. It is possible to control the scale of the clustering—whether groups are large or more fine-scaled—using the "Inflation" (-I) parameter to control granularity (Fig. 4b). Altering this parameter results in inevitable trade-offs between the appropriate separation of clusters that truly represent distinct loci and the maintenance of clusters of divergent alleles of the same locus (and associated Type I and Type II clustering errors, sensu (27)). These trade-offs can be observed by examining known single-copy loci (e.g., (35)), and identifying the

numbers of clusters that represent what should be a single locus (see Note 1). This is the single most important parameter for MCL clustering, and should be tuned for each dataset when used.

We point out that there are several sources of errors that can affect the accuracy of clustering, beyond the inevitable aggregation of paralogous loci into the same cluster: (a) large families of genes with a nearly continuous range of divergence among members will form clusters of sequences that cannot easily be separately and that will not form a single multiple alignment (hereafter "superclusters"); (b) highly divergent alleles (e.g., alleles maintained by frequency-dependent selection (8, 9)) are unlikely to cluster together under most clustering scenarios; (c) alternate splicing products in transcriptome data will generate separate clusters that are not true loci (e.g., (46)); and (d) sequencing chimeras, which are polymerase errors, will either produce erroneous loci or cause merging of clusters (47).

2.3. Haplotype (Allele) Calling

With clusters (putative loci) in hand, SNPs and indels can be identified among sequences in the clusters, allowing individual alleles to be defined. The most efficient current approach for identifying sequence variants is to create a consensus sequence for the cluster, and then to map the member sequences against this reference. This reference-guided approach necessarily means that we have created a consensus genomic library for the dataset, which can then conveniently be used to identify alleles in additional individuals as new data are collected. Sequence mapping to a reference is possible with a wide variety of current programs (e.g., (30, 31, 33, 39, 43, 48)). As in the case of sequence similarity parameters, mapping parameters must take into account thresholds for minimum alignment length and maximum sequence divergence, where the latter should be less stringent than for initial clustering, particularly if single-linkage clustering has been used.

At this stage, several types of potential error can be removed when defining alleles. Even the best assemblies can fail to merge some identical contigs, and these redundant sequences can be collapsed into single alleles. Unique SNPs or indels that are only observed once in a large set of individuals (possible sequencing errors) can be ignored if desired. Variants that may be associated with errors unique to a given sequencing platform may also be ignored; for example, erroneous indels may be common near mononucleotide repeats in 454 sequences (49), making these markers less reliable even if observed across multiple individuals. Quality and/or coverage information at polymorphic positions may also be used to help validate SNP and indel validity.

2.4. Single Locus Cluster Sorting

Even an optimal clustering strategy (i.e., best possible minimum sequence similarity requirements and clustering granularity settings) will still produce some allele clusters that circumscribe multiple loci.

Here, information from multiple individuals can offer critical insights. At the minimum, clusters can be filtered for those with no more than the expected number of alleles per individual (two for diploids). This strategy of course becomes more accurate with larger numbers of individuals and higher levels of heterozygosity, such that multilocus clusters will often reveal more than the expected number of alleles. With enough individuals, patterns of association between multiple SNPs in each allele offer the opportunity to infer mutually exclusive sets of recombining alleles (50). Currently, we are not aware of any software that automates this analysis, and it is a ripe area for future development. For the special case where many individuals are from the same population and data coverage is expected to be good for each individual, departure from HWE could be used to infer multilocus clusters (see the chapter by Hohenlohe et al., this volume). Note that particular care must be taken here because many common biological processes can cause departures from HWE, including selection (51), which is often the target of genomic sequencing projects.

2.5. The AllelePipe Software

We provide a bioinformatic pipeline called AllelePipe (Dlugosch et al. in prep; avail at http://EvoPipes.net/AllelePipe.html) to aid in clustering putative alleles across one or more individuals (see also Note 1). Briefly, our pipeline takes in assembled sequence contigs and passes them through the following steps:

1. Similarity is assessed among all sequences using SSAHA2 (43) according to user-defined minimum similarity and alignment length thresholds (Defaults: 95% similarity over 100 bp).
2. Alignment throughout the region of overlap is verified.
3. Sequences are clustered by either single-linkage clustering or MCL as desired, with the option of restarting the clustering with alternative methods/granularities.
4. Multiple alignments are created for sequences within each cluster and their consensus sequence generated, using CAP3 (41). A single consensus genomic reference fasta file is generated for the whole dataset which can be used again in other analyses.
5. Optionally, putatively chimeric clusters are removed, assuming that these are clusters where only one sequence bridges an internal region of the multiple alignment. This step is only appropriate for datasets with many individuals and good coverage of loci, where many sequences should be aligning across the length of each locus
6. SNPs (and optionally indels) are identified using SSAHAsnp (43) against the reference sequence for the same or different sets of individuals, as desired (the program can be restarted from this step for additional analyses).

7. Clusters are sorted as being single or multilocus, based upon user settings for the maximum number of alleles allowed per individual.

3. Summary

The availability of genomic polymorphism data within and among individuals of the same species is one of the most exciting outcomes of recent advances in the ease and affordability of genome-scale sequencing. Tapping that treasure trove of information is another matter, and will continue to challenge bioinformatic analyses until a time when all individuals under study are sequenced completely to their physically accurate full chromosomes. In the meantime, our best strategy is to use all available information to filter clusters of sequences for those most likely to represent true alleles of individual loci. We have proposed a few simple steps along this path, and there is substantial opportunity for further advancement, particularly using inference of recombination events to cluster segregating alleles.

4. Note

1. Single-linkage clustering, where sequences that share sufficient similarity with one member of a cluster are placed in this cluster (Fig. 4a), is the simplest and most inclusive way of clustering closely related sequences. Nevertheless, this approach can lead to problems for some clusters, when increasingly dissimilar sequences are aggregated inappropriately into one group. The current best alternative to single-linkage clustering sequences—without a reference genome for guidance—is maximum likelihood clustering, as implemented in the software MCL (http://micans.org/mcl/). In MCL, clusters are created among the sequences with the strongest links (by any measure of similarity set by the user) and the "Inflation" (-I) parameter controls whether groups are separated at a large or fine scale, known as "granularity" (Fig. 4b). This is the most important parameter for MCL clustering, and should be tuned for each dataset when used.

   We demonstrate this using a pair of transcriptome assemblies for two individuals of the thistle *Centaurea solstitialis*. One library includes 23,267 unigenes (19.3 total Mbp) generated by Sanger sequencing, from a publicly available completed assembly of an individual from North America (38). The second was

Table 1
The number of clusters inferred to be single-locus (no more than two variants per individual), multilocus, monomorphic (no allelic variation), and super clusters (those which will not align), using single linkage clustering and several granularities of MCL clustering of two *Centaurea solstitialis* individuals

		MCL clustering granularities (option -I)			
	Single linkage clustering	1.4	2	4	6
Single-locus	4,208	4,205	4,258	4,423	4,455
Multilocus	2,161	2,171	2,167	2,115	2,099
Monomorphic	288	289	295	319	332
Super cluster	23	21	14	7	5

obtained from publicly available 454 Life Sciences sequence data of an individual from South America (doi: 10.5061/dryad.cm7td/4). We cleaned these data with SnoWhite (http://evopipes.net), and assembled them using MIRA (39) and CAP3 (41) to yield 43,503 unigenes (32.3 Mbp). Putative alleles between these two individuals were most often separated by 2–4% *synonymous* divergence (total divergence will be lower typically) (Fig. 3).

Sequences were clustered within and between individuals using the AllelePipe software, with a minimum similarity of 95% (i.e., maximum 5% total divergence) and minimum alignment length of 300 bp. We created clusters with single linkage clustering and with MCL granularities (parameter -I) of 1.4, 2, 4, and 6 (Table 1). Most clusters were inferred to be single-locus under all clustering strategies, but a subset of loci were more variable. As granularities became finer, there were fewer cases of inferred multilocus clusters and clusters that would not previously form a single alignment (superclusters), but also more cases of monomorphic (invariant) clusters, which might indicate splitting apart of alleles of the same locus. The potential problem of locus splitting is demonstrated by alignment of cluster consensus sequences with known eukaryotic single copy loci ("Ultra-Conserved Orthologs" or UCOs) (35). A total of 310 of the 357 UCOs matched consensus sequences of clusters in our dataset (based on tblastx comparisons, with maximum e-value of 0.1 and minimum 30 protein residues). The majority of these loci matched a single cluster under single-linkage

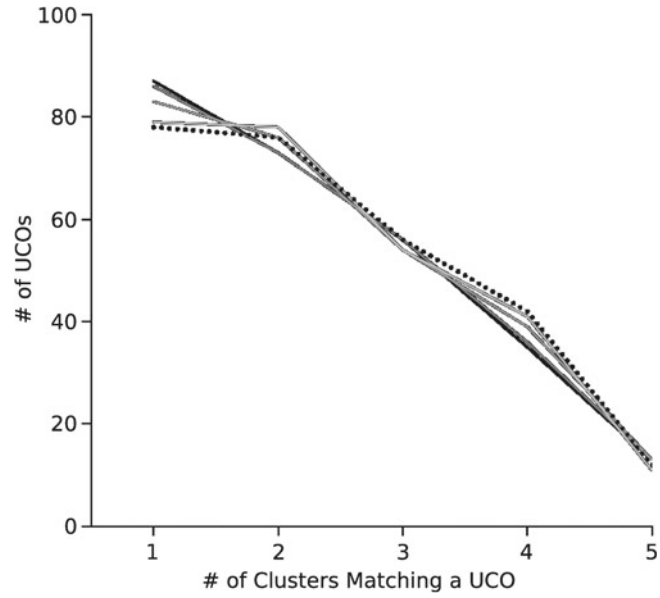

Fig. 5. Histograms of the number of clusters matching conserved single copy eukaryotic loci (UCOs) for single linkage clustering (*solid black line*) and MCL clustering with increasingly fine granularities (granularities 1.4, 2, and 4 shown with increasingly *light gray lines*; finest granularity of 6 shown with *dotted line*), of sequences from two *Centaurea solstitialis* individuals.

clustering, but increasingly fine MCL clustering granularities began to diminish the number of single matches and increase the number of UCOs matching 2–4 clusters each (Fig. 5). Matches to 2–4 clusters are consistent with the undesirable partitioning of single-locus clusters into individual alleles. We conclude that, while single linkage clustering results in some losses of useful data due to super clusters and multilocus clusters, this aggressive clustering strategy does retain most valid clusters of alleles while avoiding problems of locus-splitting. Likelihood-based clustering may be most appropriate for further partitioning of clusters that appear to be multilocus.

Acknowledgments

We thank MS Barker, LH Rieseberg, I Mayrose, and SP Otto for insightful discussions on this topic. We also thank Z Lai and LH Rieseberg for making available multi-individual genomic datasets that prompted our interests in this area.

References

1. Avise JC (2004) Molecular markers, natural history, and evolution. Sinauer Associates, Sunderland
2. Lynch M, Walsh B (1998) Genetics and analysis of quantitative traits. Sinauer Associates, Sunderland
3. Wakeley J (2008) Coalescent theory: an introduction. Roberts & Company, Greenwood Village
4. McCarthy MI, Abecasis GR, Cardon LR et al (2008) Genome-wide association studies for complex traits: consensus, uncertainty and challenges. Nat Rev Genet 9:356–369
5. Zuckerkandl E, Pauling L (1965) Evolutionary divergence and convergence in proteins. In: Bryson V, Vogel H (eds) Evolving genes and proteins. Academic, New York
6. Altshuler DL, Durbin RM, Abecasis GR et al (2010) A map of human genome variation from population-scale sequencing. Nature 467:1061–1073
7. Ossowski S, Schneeberger K, Clark RM et al (2008) Sequencing of natural strains of *Arabidopsis* thaliana with short reads. Genome Res 18:2024–2033
8. Charlesworth D, Vekemans X, Castric V, Glemin S (2005) Plant self-incompatibility systems: a molecular evolutionary perspective. New Phytol 168:61–69
9. Hulbert SH, Webb CA, Smith SM, Sun Q (2001) Resistance gene complexes: evolution and utilization. Annu Rev Phytopathol 39: 285–312
10. Patnaik SK, Blumenfeld OO (2011) Patterns of human genetic variation inferred from comparative analysis of allelic mutations in blood group antigen genes. Hum Mutat 32:263–271
11. Bergelson J, Kreitman M, Stahl EA, Tian D (2001) Evolutionary dynamics of plant R-genes. Science 292:2281–2285
12. Lawlor DA, Ward FE, Ennis PD et al (1988) HLA-A and B polymorphisms predate the divergence of humans and chimpanzees. Nature 335:268–271
13. Li WH, Sadler LA (1991) Low nucleotide diversity in man. Genetics 129:513–523
14. Moriyama EN, Powell JR (1996) Intraspecific nuclear DNA variation in *Drosophila*. Mol Biol Evol 13:261–277
15. Demuth JP, De Bie T, Stajich JE et al (2006) The evolution of mammalian gene families. PLoS One 1:e85
16. Hahn MW, De Bie T, Stajich JE et al (2005) Estimating the tempo and mode of gene family evolution from comparative genomic data. Genome Res 15:1153–1160
17. Hahn MW, Han MV, Han S-G (2007) Gene family evolution across 12 *Drosophila* genomes. PLoS Genet 3:e197
18. Lynch M, Conery JS (2000) The evolutionary fate and consequences of duplicate genes. Science 290:1151–1155
19. Sebat J, Lakshmi B, Troge J et al (2004) Large-scale copy number polymorphism in the human genome. Science 305:525–528
20. Lynch M (2007) The origins of genome architecture. Sinauer Associates, Sunderland
21. Fredman D, White SJ, Potter S et al (2004) Complex SNP-related sequence variation in segmental genome duplications. Nat Genet 36:861–866
22. Bentley DR (2006) Whole-genome re-sequencing. Curr Opin Genet Dev 16:545–552
23. Charlesworth B (2010) Molecular population genomics: a short history. Genet Res 92: 397–411
24. Li L, Stoeckert CJ, Roos DS (2003) OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res 13: 2178–2189
25. Nagaraj SH, Gasser RB, Ranganathan S (2007) A hitchhiker's guide to expressed sequence tag (EST) analysis. Brief Bioinform 8:6–21
26. Tang J, Vosman B, Voorrips RE et al (2006) QualitySNP: a pipeline for detecting single nucleotide polymorphisms and insertions/deletions in EST data from diploid and polyploid species. BMC Bioinformatics 7:438
27. Wang J-PZ, Lindsay BG, Leebens-Mack J et al (2004) EST clustering error evaluation and correction. Bioinformatics 20:2973–2984
28. Hazelhurst S, Hide W, Lipták Z et al (2008) An overview of the wcd EST clustering tool. Bioinformatics 24:1542–1546
29. Lynch M (2009) Estimation of allele frequencies from high-coverage genome-sequencing projects. Genetics 182:295–301
30. Malhis N, Jones SJM (2010) High quality SNP calling using Illumina data at shallow coverage. Bioinformatics 26:1029–1035
31. Li H, Ruan J, Durbin R (2008) Mapping short DNA sequencing reads and calling variants using mapping quality scores. Genome Res 18:1851–1858
32. Zerbino DR, Birney E (2008) Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res 18:821–829

33. Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics 25:1754–1760
34. Gibbons JG, Janson EM, Hittinger CT et al (2009) Benchmarking next-generation transcriptome sequencing for functional and evolutionary genomics. Mol Biol Evol 26:2731–2744
35. Kozik A, Matvienko M, Michelmore RW (2010) Effects of filtering, trimming, sampling and k-mer value on de novo assembly of Illumina GA reads. In: Plant and Animal Genomes XVIII Conference, San Diego
36. Altschul SF, Gish W, Miller W et al (1990) Basic local alignment search tool. J Mol Biol 215: 403–410
37. Zhang Z, Schwartz S, Wagner L, Miller W (2000) A greedy algorithm for aligning DNA sequences. J Comput Biol 7:203–214
38. Barker MS, Kane NC, Matvienko M et al (2008) Multiple paleopolyploidizations during the evolution of the Compositae reveal parallel patterns of duplicate gene retention after millions of years. Mol Biol Evol 25: 2445–2455
39. Chevreux B, Pfisterer T, Suhai S (2000) Automatic assembly and editing of genomic sequences. In: Suhai S (ed) Genomics and proteomics: functional and computational aspects. Kluwer Academic/Plenum Publishers, New York
40. Guo S, Zheng Y, Joung JG et al (2010) Transcriptome sequencing and comparative analysis of cucumber flowers with different sex types. BMC Genomics 11:384
41. Huang X, Madan A (1999) CAP3: a DNA sequence assembly program. Genome Res 9:868–877
42. Barker MS, Vogel H, Schranz ME (2009) Paleopolyploidy in the brassicales: analyses of the cleome transcriptome elucidate the history of genome duplications in *Arabidopsis* and other brassicales. Genome Biol Evol 1:391–399
43. Ning Z, Cox AJ, Mullikin JC (2001) SSAHA: a fast search method for large DNA databases. Genome Res 11:1725–1729
44. Omilian AR, Scofield DG, Lynch M (2008) Intron presence-absence polymorphisms in Daphnia. Mol Biol Evol 25:2129–2139
45. Enright AJ, Van Dongen S, Ouzounis CA (2002) An efficient algorithm for large-scale detection of protein families. Nucleic Acids Res 30:1575–1584
46. Gupta S, Zink D, Korn B et al (2004) Genome wide identification and classification of alternative splicing based on EST data. Bioinformatics 20:2579–2585
47. Bragg LM, Stone G (2009) k-link EST clustering: evaluating error introduced by chimeric sequences under different degrees of linkage. Bioinformatics 25:2302–2308
48. Li R, Yu C, Li Y et al (2009) SOAP2: an improved ultrafast tool for short read alignment. Bioinformatics 25:1966–1967
49. Margulies M, Egholm M, Altman WE et al (2005) Genome sequencing in microfabricated high-density picolitre reactors. Nature 437: 376–380
50. Griffin PC, Robin C, Hoffmann AA (2011) A next-generation sequencing method for overcoming the multiple gene copy problem in polyploid phylogenetics, applied to *Poa* grasses. BMC Biol 9:19
51. Hartl DL, Clark AG (2006) Principles of population genetics, 4th edn. Sinauer Associates, Sunderland
52. Lai Z, Kane N, Kozik A et al (2012) Genomics of compositae weeds: EST libraries, microarrays, and evidence of introgression. American Journal of Botany 99:209–218