Syst. Biol. 62(3):424–438, 2013

DOI:10.1093/sysbio/syt009
Advance Access publication February 15, 2013

Phylogenetic Signal Variation in the Genomes of *Medicago* (Fabaceae)

[1]*Department of Plant Biology, University of Minnesota, Saint Paul MN 55108;* [2]*Department of Computer Science, University of Minnesota, Saint Paul, MN 55455;* [3]*National Center for Genome Resources, Santa Fe, NM 87505, USA;* [4]*Institute of Evolutionary Biology and Environmental Studies, 8057 Zurich, Switzerland;* [5]*Department of Applied Sciences and Mathematics, Arizona State University Polytechnic, Mesa, AZ 85212; and* [6]*Department of Plant Pathology, University of Minnesota, Saint Paul, MN 55108, USA*
**Correspondence to be sent to: Department of Plant Biology, University of Minnesota, Saint Paul MN 55108, USA; Email: ptiffin@umn.edu.*

Received 27 July 2012; reviews returned 23 October 2012; accepted 4 February 2013
Associate Editor: Mark Fishbein

Abstract.—Genome-scale data offer the opportunity to clarify phylogenetic relationships that are difficult to resolve with few loci, but they can also identify genomic regions with evolutionary history distinct from that of the species history. We collected whole-genome sequence data from 29 taxa in the legume genus *Medicago*, then aligned these sequences to the *Medicago truncatula* reference genome to confidently identify 87 596 variable homologous sites. We used this data set to estimate phylogenetic relationships among *Medicago* species, to investigate the number of sites needed to provide robust phylogenetic estimates and to identify specific genomic regions supporting topologies in conflict with the genome-wide phylogeny. Our full genomic data set resolves relationships within the genus that were previously intractable. Subsampling the data reveals considerable variation in phylogenetic signal and power in smaller subsets of the data. Even when sampling 5000 sites, no random sample of the data supports a topology identical to that of the genome-wide phylogeny. Phylogenetic relationships estimated from 500-site sliding windows revealed genome regions supporting several alternative species relationships among recently diverged taxa, consistent with the expected effects of deep coalescence or introgression in the recent history of *Medicago*. [*Medicago*; phylogenomics; whole-genome resequencing.]

Genome-scale phylogenetic (i.e., phylogenomic) data offer large numbers of informative sites, dramatically reducing stochastic sampling error (Brinkmann et al. 2005; Gatesy et al. 2007). To date, phylogenomic data in the form of sequences of many individual loci, fully sequenced genomes, or expressed sequence tag libraries have been primarily applied to clarify deep relationships among primates (Siepel 2009), birds (Hackett et al. 2008; Harshman et al. 2008), fish (Steinke et al. 2006), placental mammals (Murphy et al. 2004; Nishihara et al. 2006; Decker et al. 2009), fungi (Marcet-Houben and Gabaldón 2009), green plants (Sanderson et al. 2010; Burleigh et al. 2011), the Metazoa (Philippe et al. 2009), and the Eukarya (Burki et al. 2007). The focus on these deep relationships reflects, at least in part, priorities in genome sequencing projects, which have captured an increasingly broad sample of the tree of life. However, advances in sequencing technology now allow collection of genomic data from many (sometimes hundreds) of individuals from a single species (Gibbs et al. 2003; Kim et al. 2007; Gore et al. 2009; Emerson et al. 2010; Branca et al. 2011) or multiple closely related species (Decker et al. 2009). These data offer an opportunity to clarify taxonomic relationships in groups that have previously proven challenging, but they also present unknown territory for current phylogenetic methods, which are only beginning to grapple with the characteristics of genomic data sets (Brinkmann et al. 2005; Philippe et al. 2005; Jeffroy et al. 2006; Siepel 2009).

The challenges that have faced most phylogenomic studies—such as identifying deeply separated orthologs, long-branch attraction, and the appropriate fit of substitution models (Brinkmann et al. 2005; Gatesy et al. 2007; Marcet-Houben and Gabaldón 2009)—are less likely to create error when inferring relationships among closely related taxa. On the other hand, resolving relationships within a single genus must contend with phylogenetic signal that varies across the genome due to incomplete lineage sorting and hybridization or horizontal gene transfer between closely related taxa (Tajima 1983; Pamilo and Nei 1988; Maddison and Knowles 2006; Kubatko and Degnan 2007; Degnan and Rosenberg 2009; Knowles 2009)—and may still be complicated by uncertain orthology relationships if gene duplication events occur at a rate similar to nucleotide substitution. For multilocus data sets, these issues have received considerable empirical attention (Jennings 2005; Carstens and Knowles 2007; Cranston et al. 2009; Degnan and Rosenberg 2009), yet few of these studies have applied truly genome-scale data for more than a handful of taxa. One recent example infers phylogenetic relationships among 10 closely related species of *Oryza* using a full data set with an alignment of 2.45 million nucleotides representing 9481 protein-coding genes obtained from BAC-end sequence data (Cranston et al. 2010). Analyses of the entire data set confirmed previously identified relationships among species, although there was considerable incongruence among gene trees.

Here, we present a phylogenomic analyses of more than 87 000 polymorphic single-nucleotide sites identified by whole-genome resequencing of 29 taxa (representing 26 named species and infra-specific entities) in the legume genus *Medicago*, and alignment to the *Medicago truncatula* reference genome (Young et al. 2011). This genus includes economically important

forage crops such as alfalfa (*Medicago sativa*) in addition to *M. truncatula*, a model for legume genomics (Branca et al. 2011; Young et al. 2011) and legume-rhizobia symbiosis (Barker et al. 1990; Cook et al. 1997) and the focus of the current *Medicago* HapMap Project (www.medicagohapmap.org). *Medicago* is estimated to have diverged from its sister taxon, the genus *Trigonella*, ~15.9 million years ago (Lavin et al. 2005), and previous attempts to resolve its phylogeny have been complicated by gene tree conflict, which has been proposed to reflect historical and ongoing hybridization across the genus (Maureira-Butler et al. 2008; Steele et al. 2010). This gene tree conflict within the species tree of *Medicago* is more typical of the issues faced by the emerging body of species-level, genome-wide phylogenetic analyses than most previous phylogenomic studies.

The properties of our data set make it similar to those collected by other high-throughput sequencing pipelines, particularly those that identify variable sites from reduced-representation libraries, such as restriction-site associated DNA (RAD-tag) methods (Baird et al. 2008; Ekblom and Galindo 2010; Catchen et al. 2011), and which have already been used in phylogenetic studies (Decker et al. 2009; Emerson et al. 2010). The mean distance between adjacent sites in our data set (3357 bp) is approximately the distance over which linkage disequilibrium decays to half its observed maximum in a range-wide sample of *M. truncatula* (mean $r^2 < 0.25$ between single-nucleotide polymorphisms [SNPs] separated by 3000 bp; Branca et al. 2011). Considering that we sample across a much deeper phylogenetic scale, and that we sample a number of self-incompatible species, most of the nuclear sites in our data set probably represent independent loci. Modern systematics has at its disposal multiple methods for estimating species trees from freely recombining loci (Liu et al. 2008; Bryant and Bouckaert 2009; Kubatko et al. 2009), but our data set comprises tens of thousands of SNPs, far beyond what is computationally tractable for current coalescent methods of species-tree estimation (see the Discussion section). We therefore performed our primary phylogenetic estimate on the concatenated nuclear data set—an approach that has been applied widely in phylogenomic studies (Philippe et al. 2005; Decker et al. 2009; Wiens et al. 2010). We then assessed variation in phylogenetic signal within the complete data set (i.e., the potential for different regions of the genome to support gene tree topologies in conflict with the topology supported by the genome as a whole) by estimating trees from subsets of the data, either as random samples of varying size or as "sliding windows" across the genome.

We examine (i) the resolution of species relationships obtained with the genome-wide data set; (ii) the effect of sample size (i.e., the number of sites sampled from the data set) on our ability to recover the phylogeny obtained with the entire data set; (iii) whether phylogenetic relationships inferred from our genome-wide data set conflict with hypotheses generated from previous multilocus studies; and (iv) whether particular regions of the genome strongly support conflicting phylogenetic signal, as may be expected from deep coalescence, hybridization, or introgression over the history of *Medicago*.

Methods

We collected whole-genome sequence data from each of 29 *Medicago* accessions. Twenty-four of these are generally recognized as separate species (Small 2011). We also included the 2 infra-specific taxa *M. truncatula* var. *tricycla* and the alfalfa subspecies *M. sativa* ssp. *caerulea* for their interest in the ongoing HapMap Project, and to sample more recent relationships within the genus. Finally, we included 3 accessions that had been identified as *M. truncatula* (hereafter referred to by their HapMap accession codes HM017, HM018, and HM022), but which were recently found, in phylogenetic analysis based on whole-chloroplast sequence, to be more deeply diverged from a range-wide sample of *M. truncatula* accessions than they are from one another (Branca et al. 2011), and which exhibit some morphological differences from *M. truncatula* (Kelly Steele, unpublished data). We include these accessions in our analyses to see whether this placement is corroborated with nuclear data, and in analyses including data from more deeply diverged taxa. Our complete data set represents about one-third of the maximum number of species recognized in *Medicago* (estimated at ~87; Maureira-Butler et al. 2008; Steele et al. 2010), but samples all major clades within the genus. All of the sampled taxa are diploid with the exception of *Medicago arborea* ($2n = 32$) and *Medicago cancellata* ($2n = 48$). Six species are $2n = 14$ (*Medicago constricta*, *Medicago murex*, *Medicago polymorpha*, *Medicago praecox*, *Medicago rigidula*, and *Medicago rigiduloides*) while the remainder are $2n = 16$. Our sequencing, alignment, and filtering methods should robustly identify homologous loci despite this variation in ploidy; but synteny among the species in our data set is not known in any detail, and we are only able to refer to genomic locations in terms of the *M. truncatula* reference genome (Mt 3.5 assembly; Young et al. 2011).

Our sequence data collection, alignment, and variant site identification methods follow Branca et al. (2011), but we describe them briefly here. We extracted DNA for Illumina library construction using a modified cTAB procedure, and used Illumina paired-end sequencing with 90 bp reads to sequence DNA from each of the 29 accessions to an average aligned coverage of 10.2× and an average of 26.3% uniquely aligned coverage (Table 1). Genomic paired-end Illumina sequencing libraries were prepared for sequencing by synthesis according to standard methods (Bentley et al. 2008). Insert sizes (not including the adapters) ranged from ~200 to 450 nt. We sequenced libraries using GAII or GAIIx Illumina sequencing instruments, which yield paired 90mer reads. We used default settings of the Illumina image analysis pipeline for image analysis, base-calling, and read filtering. Further filtering was

TABLE 1. Taxa sampled for the present study, with description of the genomic data collected from each

Species[a]	MHP accession number	USDA/GRIN population	Collection location	Depth of total coverage[b]	Percent uniquely aligned[c]	Per-SNP coverage[d]	Percent missing data	Proportion of sites differing from HM101	
								Nuclear	Chloroplast
Medicago truncatula var. *truncatula*	HM101	A17_Varma	Australia	47.7	44.5	—	0.47	—	—
Medicago sp.	HM017	A10	Tunisia	35.0	40.7	93.5±2.3	1.08	0.117	0.359
Medicago sp.	HM018	A20	Tunisia	33.3	31.9	61.2±1.3	1.49	0.114	0.302
Medicago sp.	HM022	TN3.23	Tunisia	26.3	46.6	85.5±2.2	0.88	0.118	0.364
M. arabica	HM318	PI495200	France	2.5	19.8	9.9±0.2	13.77	0.474	0.583
M. arborea	HM319	PI504540	Greece, Aegean island	2.8	23.5	16.0±0.4	18.86	0.466	0.585
M. cancellata	HM320	PI440491	Russia, Stavropol	2.6	20.7	9.9±0.3	24.56	0.530	0.539
M. constricta	HM321	PI534177	Bulgaria	2.0	26.2	8.1±0.2	13.88	0.497	0.502
M. coronata	HM322	PI498790	Greece	2.2	11.9	9.0±0.3	45.99	0.70	0.641
M. doliata	HM323	PI495278	Lebanon	3.0	42.7	12.5±0.4	3.98	0.184	0.336
M. italica	HM324	PI385014	Tunisia	2.4	50.1	7.2±0.2	8.14	0.141	0.256
M. laciniata	HM325	PI498902	Morocco	2.6	15.0	8.7±0.2	24.30	0.510	0.512
M. littoralis	HM030	F11013-27	Algeria	36.1	37.2	123.9±3.3	0.86	0.117	0.376
M. lupulina	HM326	PI251834	Italy	2.5	18.2	10.1±0.3	13.16	0.498	0.633
M. minima	HM327	PI499080	Turkey	3.4	20.6	13.7±0.4	6.64	0.459	0.523
M. murex	HM328	PI495379	France, Corsica	2.5	23.6	9.9±0.3	10.33	0.425	0.564
M. noeana	HM329	PI495414	Turkey, Icel	1.8	10.6	5.4±0.3	51.13	0.661	0.507
M. polymorpha	HM330	PI566877	Italy, Sicily	2.6	21.7	13.3±0.6	13.82	0.454	0.624
M. praecox	HM337	PI495434	Corsica	3.4	17.7	13.6±0.4	18.64	0.446	0.504
M. prostrate	HM339	PI577445	Abruzzi,	2.4	19.1	11.8±0.3	25.56	0.520	0.572
M. rigidula	HM331	PI495517	Greece	2.7	26.6	12.5±0.4	6.56	0.461	0.519
M. rididuloides	HM332	PI534250	Turkey, Ankara	2.6	27.1	12.5±0.3	0.47	0.457	0.500
M. rotate	HM333	PI495581	Turkey, Icel	2.1	17.6	10.4±0.3	7.46	0.518	0.633
M. ruthenica	HM335	PI568100	China, Nei Mongol	2.7	10.8	7.7±0.2	26.30	0.584	0.590
M. sativa ssp. *Sativa*	HM102	CADL	USA	28.6	13.1	52.5±1.3	36.27	0.403	0.516
M. sativa ssp. *Caerulea*	HM336	PI314275	Uzbekistan	3.4	19.2	12.1±0.3	5.09	0.491	0.558
M. soleirolii	HM338	PI537240	Algeria	3.3	31.6	12.4±0.3	19.39	0.194	0.447
M. truncatula var. *tricycle*	HM029	R108-C3	Algeria	29.9	29.7	64.1±1.7	5.60	0.128	0.362
M. turbinate	HM334	PI535555	Tunisia	3.0	43.3	12.4±0.3	2.24	0.183	0.406

[a]Taxonomy follows USDA listings. [b]Depth of unique coverage aligned to the *M. truncatula* reference genome (Young et al. 2011). [c]Percent of all reads aligned to a single location in the *M. truncatula* reference genome. [d]Mean depth of aligned coverage at each SNP in the alignment, ± 95% CI.

done to remove adapter and PhiX contamination based on blast alignment (pairs with $\geqslant$14 nt aligned at $\geqslant$98% were removed) and reads with quality (Q-score) <10. All Illumina sequence data have been deposited in the NCBI Short Read Archive (accession SRP001874). Coverage data (Table 1) and called variant sites are available at www.medicagohapmap.org.

All sequence reads that passed the initial quality control filter were aligned to the *M. truncatula* reference genome (Mt 3.5 assembly; Young et al. 2011) using the Genomic Short-read Nucleotide Alignment Program (GSNAP; [Wu and Nacu 2010]) following protocols similar to those previously used for aligning multiple accessions of *M. truncatula* (Branca et al. 2011): First, only reads with only 8 mismatched bases out of every 90 bp of sequence aligning to a region in the reference genome (i.e., at least 91% identical to the aligned genome location), and which aligned to fewer than 5 locations were included in the alignment output file. This mismatch threshold was determined by Branca et al. (2011) based on an empirical assessment of error in the alignment and variant-calling pipeline using Sanger sequencing of 100 randomly selected genome regions, and it maximizes true variant calls while minimizing false positives. Then, a variant site (i.e., a site differing from the reference genome) was called only if 70% of the reads covering the site called the variant nucleotide and at least 2 of the reads covering the variant aligned only to a single region in the reference genome. Requiring a 91% match to the reference genome means that we will have excluded any highly divergent regions, and the $>70\%$ requirement means that we excluded most heterozygous sites. Our filtering for number of aligned locations and the 91% match threshold cannot entirely ensure against the possibility of sequencing paralogous regions in species with differing ploidy or other differences in gene copy number relative to the *M. truncatula* reference genome, and these may be a source of phylogenetic error in our data.

From the aligned reads, we identified all sites for which no more than 5 of the 29 taxa were missing data. The analyses we present were conducted on sites for which at least 2 accessions carried the less-frequent variant, a minor allele frequency (MAF) threshold that provides an additional safeguard against incorrectly called variants and is necessary to ensure that all characters are informative within the ingroup. A total of 87 596 nuclear sites met these criteria, with between

TABLE 2. Properties of the variable sites identified on the 8 chromosomes and the chloroplast in the *M. truncatula* genome

Sites mapped to	Variable sites	Percent coding[a]
Chloroplast	432	25.23
Chromosome 1	10 292	92.59
Chromosome 2	10 339	96.50
Chromosome 3	13 462	96.69
Chromosome 4	13 295	95.31
Chromosome 5	15 467	96.44
Chromosome 6	2705	92.50
Chromosome 7	12 420	96.46
Chromosome 8	9616	97.38
Total for nuclear data set	87 596	95.86

[a]Percent of sites annotated as coding in the *M. truncatula* reference genome (Young et al. 2011).

2700 and 15 000 sites mapping to each chromosome in the *M. truncatula* reference genome (Table 2) and another 432 sites mapping to the chloroplast genome (Table 2). Alignments of the polymorphic sites from each chromosome and the chloroplast are available online in the Dryad data repository (DOI: 10.5061/dryad.vp634306).

Phylogenetic Estimation and Comparison

We performed our primary phylogenetic estimation using the Bayesian phylogenetic methods implemented in MrBayes v. 3.2.1 (Huelsenbeck and Ronquist 2001; Ronquist and Huelsenbeck 2003). Because our data set contains only variable sites, most models of DNA sequence evolution will overestimate rates of evolution, and thereby branch lengths (Lewis 2001; Felsenstein 2004). We therefore performed this and all other phylogenetic estimation using the Mkv model developed by Lewis (2001), which provides a likelihood framework for estimating phylogenies from data sets containing only variable characters, and which is implemented in MrBayes by setting the parameters *datatype* to "standard" and *coding* to "variable." We recoded the SNP alignments in such a way that each site had 4 potential states (i.e., the 4 nucleotide bases), and the rate of evolution varied across sites with rates drawn from a gamma distribution. We performed the analysis as a Metropolis-coupled Markov-chain Monte Carlo (MCMCMC) with 4 chains for each of 2 independent runs and determined that the Markov chain had converged when the standard deviation of split frequencies across runs dropped < 0.01; we set MrBayes to discard the first 25% of sampled states as burn-in, after confirming that the chains had sampled from a stationary distribution by inspecting the time-series plot of the post-burn-in parameter states for pilot runs. We ran MrBayes in parallel on a cluster of SGI Altix (x86_64) computing nodes maintained by the Minnesota Supercomputing Institute (MSI; msi.umn.edu), using 8 processor cores and 16 GB of memory.

In some cases, the exceptionally large size of phylogenomic data sets have led researchers to analyze them using simple parsimony (e.g., Decker et al. 2009), which is both computationally efficient and suitable for matrices consisting solely of variable characters. We therefore also used parsimony-based methods to estimate a phylogeny from our complete nuclear data set. We identified the maximum parsimony tree for the concatenated nuclear matrix and estimated confidence in the clades it identified using 500 bootstrap replicates, run using the *dnapars* and *seqboot* components of the Phylip package (Felsenstein 1989, 2005). For all parsimony tree searches, we used the default settings in Phylip, using ordinary parsimony and retaining 10 000 trees while searching.

Following phylogenetic estimation from the complete data set, we examined variation in phylogenetic signal across the genome by performing Bayesian phylogenetic estimation on various subsets of the data. In all cases, we ran the analysis on a concatenated data set in MrBayes, using Lewis's (2001) Mkv model as above, as MCMCMC with 4 chains for each of 2 independent runs. For all analyses, we set MrBayes to search for 10^6 iterations, or until the average standard deviation of split frequencies between independent runs dropped < 0.01, whichever was shorter. In almost every case, the analysis achieved convergence well before completing 10^6 iterations. We discarded the first 25% of each run as burn-in (the default setting for MrBayes) after confirming, from inspection of the time-series plots of parameter states in initial runs, that this was sufficient to ensure stationarity. We ran these analyses in parallel on a cluster of Dell PowerEdge Rack 900 computing nodes with Intel Xeon 2.67 GHz processers maintained by MSI; using either 8 or 16 computing cores (depending on the size of the data set under analysis) and 16 GB of memory.

To identify variation in phylogenetic signal within the nuclear data set, we estimated phylogenies based on (i) the concatenated SNPs from each of the 8 chromosomes and the chloroplast; (ii) replicate subsamples of varying size from the complete nuclear data set; (iii) for each window in a "sliding window" analysis run across the aligned genome; and (iv) selected genomic regions revealed by the sliding window analysis as supporting unusually divergent phylogenies. (Each of these analyses is described in more detail below.) We also used neighbor-net network analysis implemented in SplitsTree v. 4.11.3 (Huson and Bryant 2006), to visualize the phylogenetic signal present in our genome-wide data set. The network visualization provides an illustration of alternative groupings that may not be obvious from a traditional phylogeny.

In all phylogenetic reconstructions, we rooted the estimated trees using *Medicago ruthenica* as an outgroup, because this species was strongly resolved as the most deeply diverged taxon in our sample by the most recent multilocus phylogenetic estimate for the genus by Steele et al. (2010), who used a more distantly related outgroup. We created the subsample alignments and analyzed trees output by MrBayes using the analysis packages *ape* and *geiger* for the statistical programming language R (Paradis et al. 2004; R Core Team 2012).

Effect of Genomic Sample Size on Phylogeny Estimation

To consider how the estimated topology changed with the number of sites used for analyses, we drew 100 random samples, without replacement, of 500, 1000, 2000, or 5000 sites from the nuclear genomic alignment. We then estimated phylogenies from each random sample using the MrBayes procedure described above, and compared the consensus topologies with the genome-wide phylogenetic estimate. We examine several possible measures of phylogenetic conflict: the topological distance metric, Dt, of Penny and Hendy (1985); the number of nodes in the majority rule consensus topology of a subsample that conflict with the genome-wide topology; and the number of such conflicting nodes with strong ($\geqslant$0.95) posterior support. Because none of these captured both the degree of conflict (i.e., number of nodes in conflict with the genome-wide tree) and the strength of posterior support for conflicting nodes, we also quantified conflict with the genome-wide tree as the sum of posterior probability for all conflicting nodes in the consensus topology divided by the total number of nodes resolved in the consensus topology, which we term "weighted conflict." This value scales between zero, when there are no resolved nodes in conflict with the genome-wide tree, and one, when all resolved nodes are in conflict and have posterior probability = 1.0. It therefore reflects the degree to which a given subsample supports a topology that robustly conflicts with the genome-wide estimate. This weighted conflict score is related to Dt, but it captures different aspects of conflict, as will be seen below.

Phylogenetic Signal and Genomic Location

To examine whether there are discrete genomic regions that harbor a phylogenetic signal in conflict with the genome-wide signal, we performed a "sliding window" analysis, drawing "windows" of 500 consecutive SNPs from the concatenated alignment. (Chromosome position of the sites is based on mapped position in the Mt3.5 assembly of the *M. truncatula* reference genome.) The windows "slid" in increments of 250 SNPs; thus, 2 consecutive windows would overlap by 250 SNPs. We estimated a phylogeny from the concatenated sites in each window as the Bayesian posterior consensus from the MrBayes analysis conducted as described above. We then examined the weighted topological conflict (as well as the total number of resolved nodes in conflict, the total number of strongly supported conflicting nodes, and Penny and Hendy's Dt) between the tree estimated from each window and the genome-wide estimate. To identify genome regions with greater than expected conflict, we compared the conflict scores of the sliding window trees to conflict scores for consensus topologies estimated from 100 replicate samples of 500 sites drawn from across the genome or to conflict scores for 100 replicate samples of 500 sites drawn from each chromosome. We considered regions where sliding windows had greater weighted conflict scores more than 97.5% of these randomly sampled 500 site data sets to be in significantly greater conflict with the genome-wide topology than expected by sampling effects alone, and focused our closer examination of specific phylogenetic conflicts on these regions.

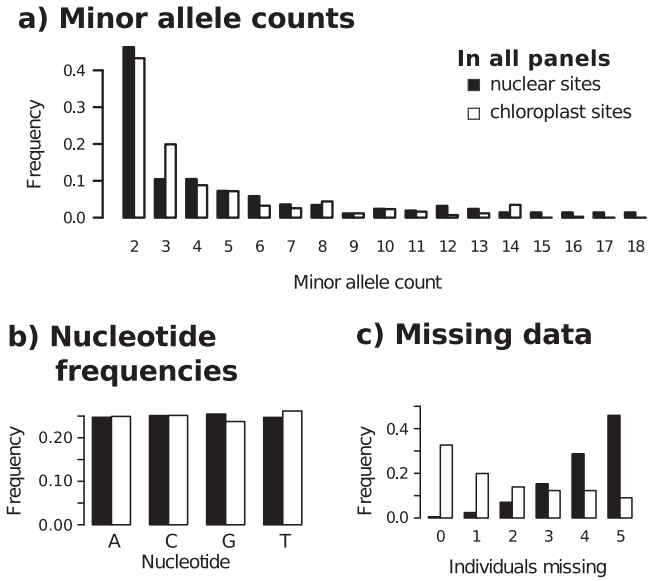

FIGURE 1. Properties of our phylogenomic data set. In all panels, filled bars represent results from the nuclear genome, and empty bars, results from the chloroplast. a) Distribution of minor allele counts. b) Frequency of each nucleotide in the nuclear and chloroplast data sets. c) Distribution of missing data (number of individuals with missing data); sites with more than 5 individuals missing data were excluded from the data set.

RESULTS

After resequencing, alignment to the *M. truncatula* reference genome, and quality filtering, our data set consists of 87 596 nuclear sites and 432 chloroplast sites that are polymorphic within our taxonomic sample, and for which data are missing for no more than 5 taxa. Over 90% of the nuclear sites are located in regions identified as protein-coding (Table 2), which probably reflects greater divergence of intergenic regions. The distribution of observed MAFs is similar in the nuclear and chloroplast data (Fig. 1a), as is the base composition (Fig. 1b), but the distribution of missing data is markedly different (Fig. 1c). For the nuclear data, fewer than 0.5% of the 87 596 polymorphic sites are sampled from all of the 29 accessions, and almost half of the nuclear sites have data missing for 5 of 29 accessions, the maximum allowed by our filters (Fig. 1c). By contrast, 33% of the chloroplast sites have data from all 29 accessions, and only 8% are missing data from 5 accessions. This may reflect greater conservation of sites in the chloroplast genome than the nuclear genome, or greater coverage of the chloroplast genome. Mean depth of aligned coverage at the called SNPs varied across accessions (Table 1); the mean across taxa was 25.7×, but the range of coverage depths ran from a minimum of 5.4× in *Medicago*

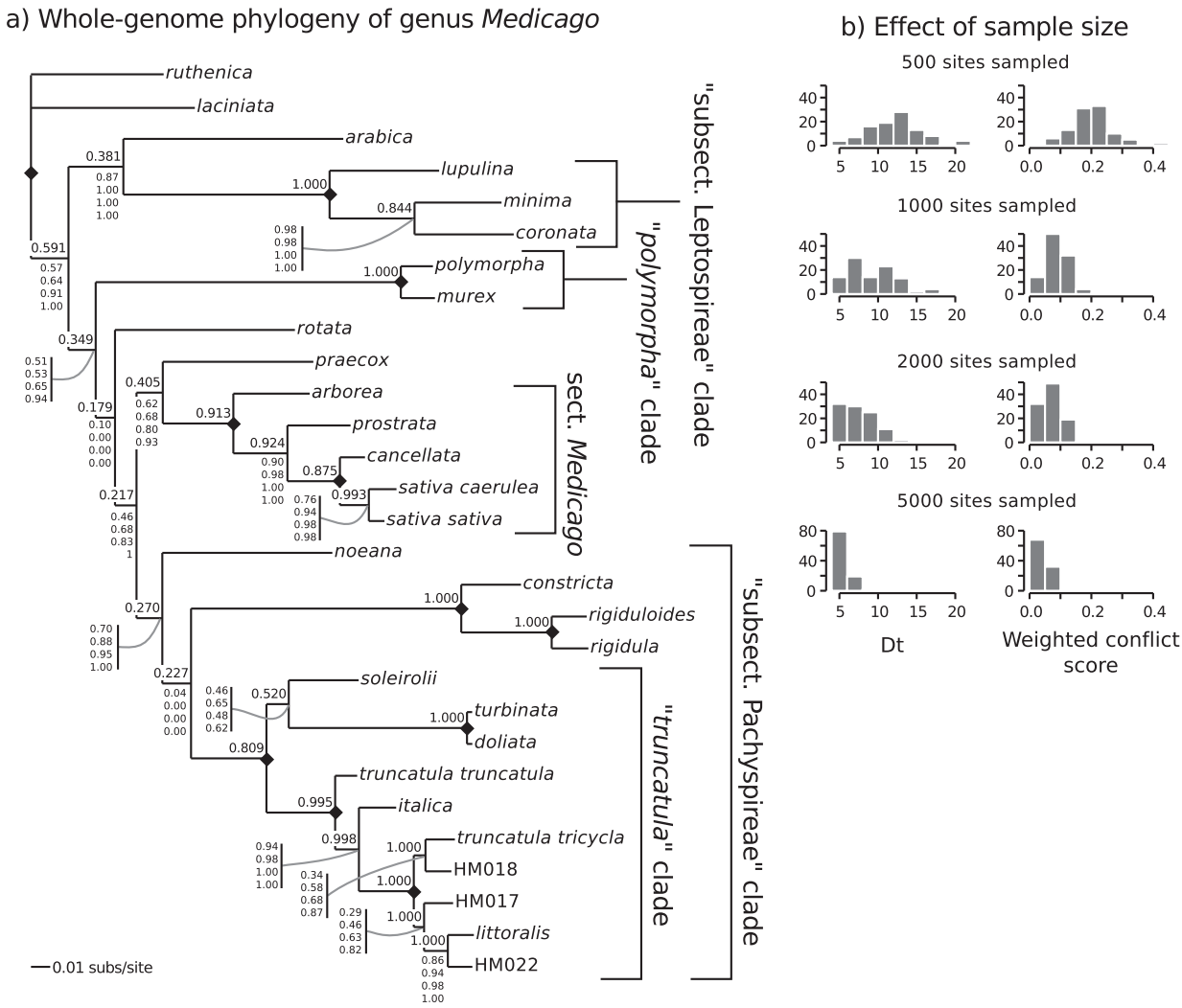

FIGURE 2. Phylogenetic estimates from the nuclear genome. a) Bayesian consensus tree estimated from a concatenated matrix of 87 596 sites on all 8 chromosomes, with proportion of posterior support given above each node (nodes with support = 1.0 have no label). Clade names are modified from Steele et al. (2010); the scale bar at the lower left gives length in expected substitutions per site. Numbers below each node give the proportion of support for that node from the posterior consensus of MrBayes estimation from 100 replicate subsamples of (from highest to lowest position) 500, 1000, 2000, or 5000 sites; nodes marked with a filled diamond were recovered by 100% of subsamples at all sizes. b) Distribution of topological conflict with the genome-wide estimate (left, Penny and Hendy's Dt; right, our weighted conflict score) in MrBayes consensus trees estimated from random samples of increasing size drawn randomly from across the genome.

noeana to a maximum of 123.9× in *Medicago littoralis*. As might be expected, there was a strong negative correlation between an accession's phylogenetic distance from HM101, *M. truncatula* var. *truncatula* (as estimated on the genome-wide Bayesian consensus, Fig. 2a) and depth of coverage aligning to the *M. truncatula* genome, which is sequenced from accession HM101 (Pearson's product-moment correlation = −0.588, $P < 0.001$).

Sites in this final filtered data set are spaced across the genome, with mean distance between adjacent SNPs > 3000 bp. However, the distribution of between-site distances is strongly right-skewed, with the median distance between adjacent markers equal to 27 bp, while the maximum distance between adjacent sites is 657 044 bp; the mean between-site distance, 3357 bp, is in the 88th percentile of between-site distances for the whole data set. Branca et al. (2011) found that linkage between sites diminished to 50% of its maximum at a distance of 3000 bases in a sample of *M. truncatula*, which is highly selfing; since our data set samples across a much deeper phylogenetic scale and includes multiple predominantly outcrossing species, this suggests that the SNP markers in our data set represent thousands of loci that assort independently at the genus-wide scale.

Many phylogenomic studies have resolved phylogenetic relationships with exceptionally high confidence, which may be deceptive given the large data sets involved (Jeffroy et al. 2006). However, our MrBayes analysis of the complete concatenated nuclear data set returned a topology with quite low confidence for many relationships, particularly in deep internal nodes (Fig. 2a). Visualization of splits in the nuclear

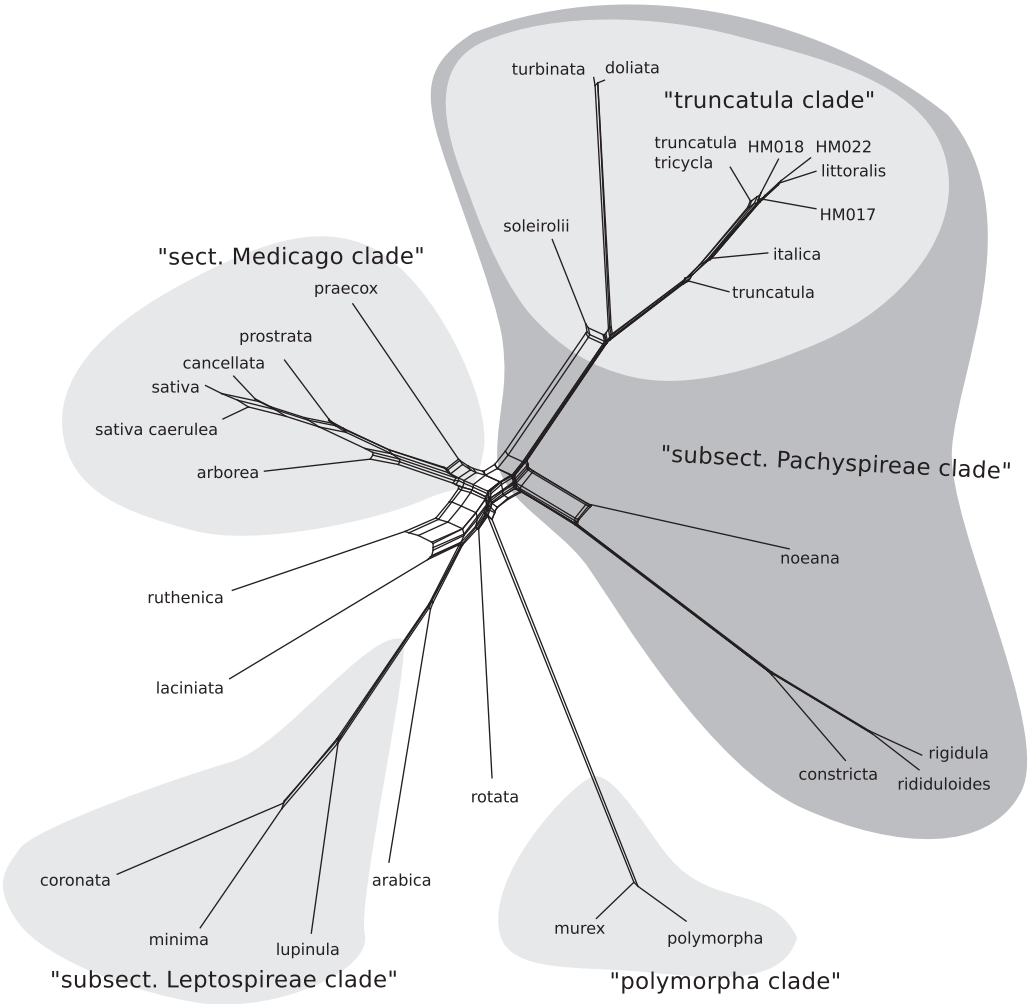

FIGURE 3. Neighbor-net network generated from the complete nuclear data set.

genomic data set as a neighbor-net network revealed support for the major groupings found by the Bayesian analysis of the concatenated data set, but also a number of alternate relationships not evident from the MrBayes analysis (Fig. 3). In contrast, the parsimony bootstrap analysis found exceptionally strong confidence in the MP topology estimated from the complete concatenated nuclear data set, with only 2 relationships having $<100\%$ bootstrap support: the placement of *Medicago soleirolii* as sister to *Medicago doliata*+*Medicago turbinata* (97%), and the placement of *Medicago rotata* as sister to *M. polymorpha*+*M. murex* (99%). The MP tree differed from the Bayesian consensus in the placement of only 2 taxa, *M. rotata* and *M. noeana* (Supplementary Fig. S01; Dryad DOI 10.5061/dryad.vp634306). The Bayesian consensus weakly favored a placement of *M. rotata* as sister to the entire "subsect. Pachyspirareae clade," while the MP topology placed it as sister to *M. polymorpha*+*M. murex*; the Bayesian consensus placed *M. noeana* as the deepest diverging member of the "truncatula clade," while the MP topology has it as sister to a subclade consisting of *M. constricta*, *M. rigidula*, and *M. rigiduloides*.

Effect of Genomic Sample Size on Phylogeny Estimation

Random samples of the nuclear data never supported topologies identical to that of the full data set. Trees estimated from 100 randomly sampled sets of 500 SNPs had weighted conflict scores from 0.085 to 0.408, with mean $=0.205$; and Dt scores between 6 and 22 (Fig. 2b). This means that trees estimated from these subsamples never supported the same topology as that of the genome-wide tree, but at the same time gave strong support to few alternative phylogenetic relationships. Larger subsamples had lower mean conflict scores (Fig. 2b), but even among subsamples of 5000 sites there were none that supported a topology identical to that from the full nuclear data set—none had Dt <4. Conflict between phylogenies estimated from the random samples and from the full nuclear data set was not localized to a single region of the tree, but conflict was generally associated with shorter branches, and larger samples were more likely to agree with the genome-wide estimates of recent relationships (Fig. 2a). However, samples as small as 500 sites were consistently able to resolve a few (mostly more recent) relationships,

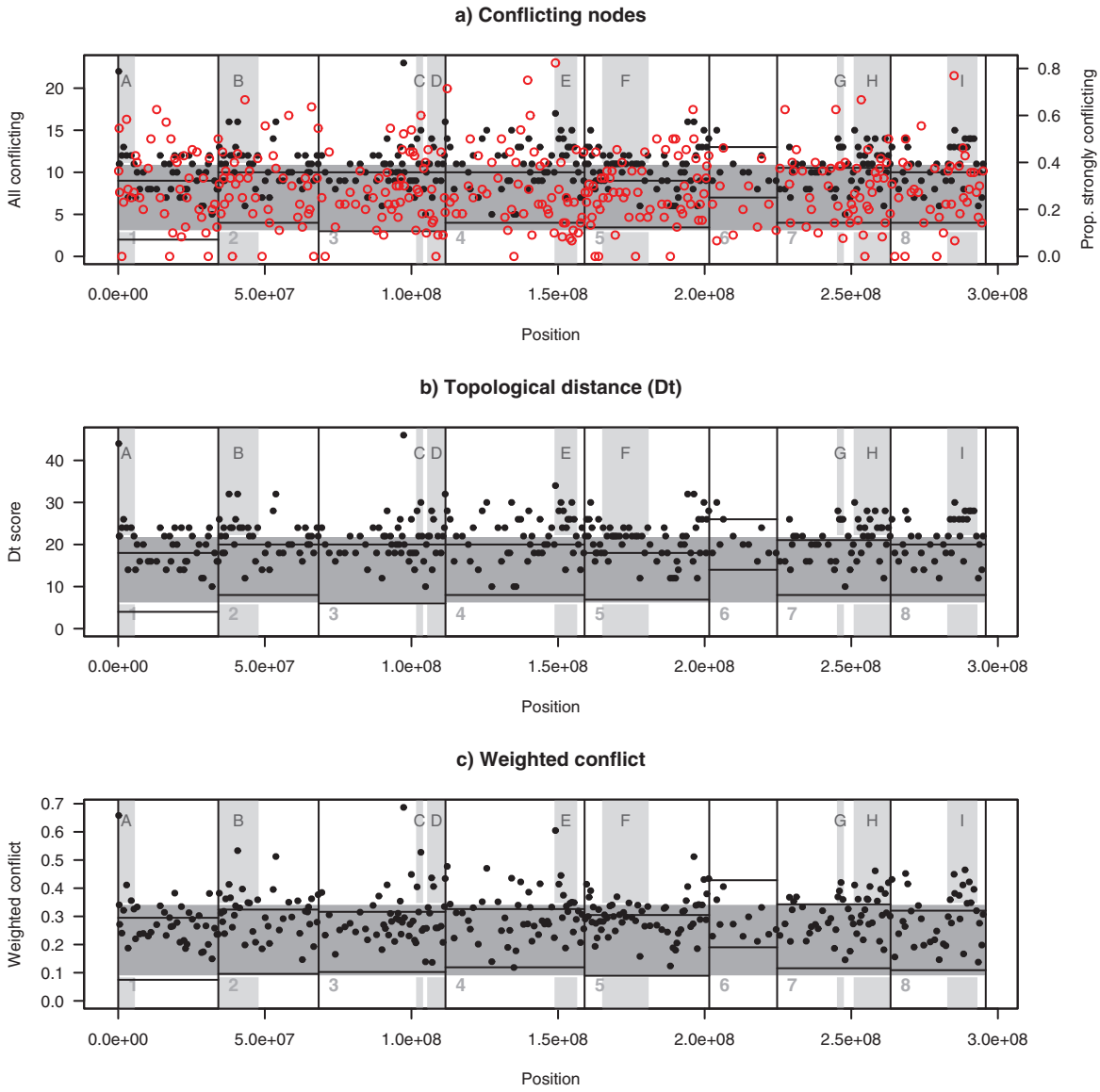

FIGURE 4. Genome position and phylogenetic signal. Conflict between the genome-wide phylogenetic estimate (Fig. 2a) and the Bayesian consensus phylogeny estimated from sliding windows of 500 consecutive variable sites, as measured by a) the number of nodes resolved in the Bayesian consensus that conflict with the genome-wide tree (black filled circles, total number of conflicting nodes; open circles, proportion of conflicting nodes with posterior support $\geqslant$0.95); b) topological distance, the Dt statistic of Penny and Hendy (1985); and c) weighted conflict, the sum of posterior support for all conflicting nodes in the consensus topology, divided by the number of nodes resolved in the Bayesian consensus. In all panels, chromosomes are separated by vertical lines; dark gray regions indicate the range of the conflict measure seen for 95% of random samples of 500 sites drawn from across the genome, and black horizontal lines mark the same for random samples of 500 sites from each chromosome. Pale gray regions highlight areas where 4 or more consecutive windows have weighted conflict scores > 95% random samples from across the genome (phylogenetic estimates from these regions are given in Fig. 5).

suggesting widespread support for these relationships in the genomic data set. Interestingly, there was a significant positive relationship between the posterior support for a given clade in the MrBayes analysis of the genome-wide data set and the frequency with which subsampled data sets recovered that clade, regardless of the sample size (correlation for 500-SNP samples = 0.646, $P < 0.001$; for 1000-SNP samples = 0.641, $P < 0.001$; for 2000-SNP samples = 0.610, $P < 0.001$; for 5000-SNP samples = 0.534, $P = 0.005$).

Phylogenetic Signal and Genomic Location

Across the nuclear genome, 332 of 349 sliding windows produced consensus topologies with strong posterior support (posterior probability $\geqslant$0.95) for one or more nodes that conflicted with the genome-wide tree ("strongly conflicting," Fig. 4a). One hundred eighty-one windows supported consensus topologies with unusually high Dt, which measures all topological conflict with the genome-wide tree ("topological distance," Fig. 4b). Finally, 89 windows supported

consensus topologies with unusually high weighted conflict scores—the proportion of resolved nodes in conflict with the genome-wide tree, weighted by the posterior support for each ("weighted conflict," Fig. 4c). Consistent with the distribution of conflict in trees estimated from random genome-wide samples of 500 sites (Fig. 2b), no sliding window supported a consensus topology identical to the genome-wide tree. This asymmetry in conflict—many regions show greater-than-expected conflict, but few show *lower*-than-expected conflict—may be a consequence of the boundaries of the possible range of topological conflict. That is, there are many possible ways to be different from the genome-wide tree, but only one way to be identical to it. Measures of topological conflict in trees estimated from the sliding windows did not increase with the proportion of missing data in each window; in fact, the windows with the highest Dt and weighted conflict scores were also those with the lowest proportion of missing data (data not shown).

Phylogenetic conflict with the genome-wide estimate also differs among chromosomes, with all windows on Chromosome 6 showing significantly higher weighted conflict scores more than 500-site samples drawn from across the genome (one-sided *t*-test $P = 8.5 \times 10^{-8}$). Chromosome 6 is the smallest chromosome in the *M. truncatula* genome, containing large stretches of repetitive elements, and regions coding for proteins in the large and evolutionarily dynamic nucleotide binding site leucine-rich repeat (NBS-LRR) gene family (Ameline-Torregrosa et al. 2008). Inference of homology relationships in NBS-LRR genes is complicated by frequent duplication and deletion events, and genes in this family may also harbor unresolved ancestral polymorphism as a result of balancing selection arising from their role in immune defense (Tiffin and Moeller 2006). In fact, BAC coverage of Chromosome 6 is lower than for any other chromosome (Young et al. 2011), and fewer SNPs on Chromosome 6 passed our quality filtering than any other chromosome (Table 2). The phylogeny reconstructed from sites on Chromosome 6 conflicts with the genome-wide tree in several nodes, and many of these conflicts have strong posterior support (Supplementary Fig. S02; Dryad DOI: 10.5061/dryad.vp634306).

Out of 332 sliding-window regions supporting at least one node in conflict with the genome-wide topology with $\geqslant$0.95 posterior probability, 267 supported alternative (relative to the genome-wide tree) relationships within the "*truncatula* clade," which includes *Medicago italica*, *M. littoralis*, *M. truncatula* var. *tricycla*, and the 3 unidentified accessions HM017, HM018, and HM022 (Supplementary Fig. S03a; Dryad DOI: 10.5061/dryad.vp634306). Rearrangements among the rest of the "subsect. Pachyspireae clade" were also quite common, with 151 windows giving strong support for relationships among these taxa that conflicted with the genome-wide estimate (Supplementary Fig. S03b). Alternative relationships among the representatives of section *Medicago* included in our analysis were strongly supported by 102 windows (Supplementary Fig. S03c), and 37 windows strongly supported rearrangements within subsection Leptospireae (Supplementary Fig. S03d). Finally, 260 windows strongly supported placement of *Medicago laciniata* in a position other than that found by the whole-genome estimate (Supplementary Fig. S03e), which placed this species as the most deeply diverged member of the ingroup. The placement of *M. laciniata* was also ambiguous in our neighbor-net analysis (Fig. 3).

We further examined 9 genomic regions where a run of consecutive windows had weighted conflict scores $>$97.5% of 500-site windows drawn randomly from across the genome (light gray regions in Fig. 4), by estimating phylogenies from the sites in each region using the same MrBayes procedure applied to each individual window (Fig. 5). The phylogenies estimated from these "highly conflicting" regions recapitulate what we found in examining recurrent conflicts across the individual window regions and highlight conflicting reconstructions of relationships that are revealed as ambiguous in our network analysis (Fig. 3). Many strongly differ with the genome-wide tree in recent relationships, but many also include rearrangements of the most early diverging clades. Six of the highly conflicting regions strongly support (i.e., give posterior probability $\geqslant$0.95 for) rearrangements among *M. italica*, *M. littoralis*, *M. truncatula* var. *tricycla*, and the 3 unidentified accessions HM017, HM018, and HM022; 4 strongly support a rearrangement of earlier relationships among the "*truncatula* clade" species (Fig. 5). One region on Chromosome 3 (Region D, Figs. 4 and 5) gives strong support for a sister relationship between *M. sativa* ssp. *sativa* and *M. cancellata*. Many of the highly conflicting regions also give strong posterior support for rearrangements at deeper levels in the tree, particularly in the relative placement of *M. polymorpha*+*M. murex*, whose position is weakly supported even in the genome-wide estimate (Figs. 2a and 5), which is ambiguously placed in the network analysis (Fig. 3), and which was given alternative placements by many random subsamples of the data set (Fig. 2a).

Discussion

Advances in sequencing and information technology are making data sets such as the one presented here—consisting of variable sites collected for many close evolutionary relatives—practical to collect for many groups in which phylogenetic relationships remain poorly resolved (Catchen et al. 2011; Davey et al. 2011; Etter et al. 2011). Our data set for the genus *Medicago* therefore provides an opportunity to gain insight into the phylogenetic utility of genome-wide variant data that can be collected using next-generation sequencing, reduced-representation libraries (Davey et al. 2011; Etter et al. 2011), or genome-wide SNP-chip genotyping (Decker et al. 2009). Using relatively shallow sequence coverage (Table 1; a mean of 25.7$\times$ coverage at all SNPs

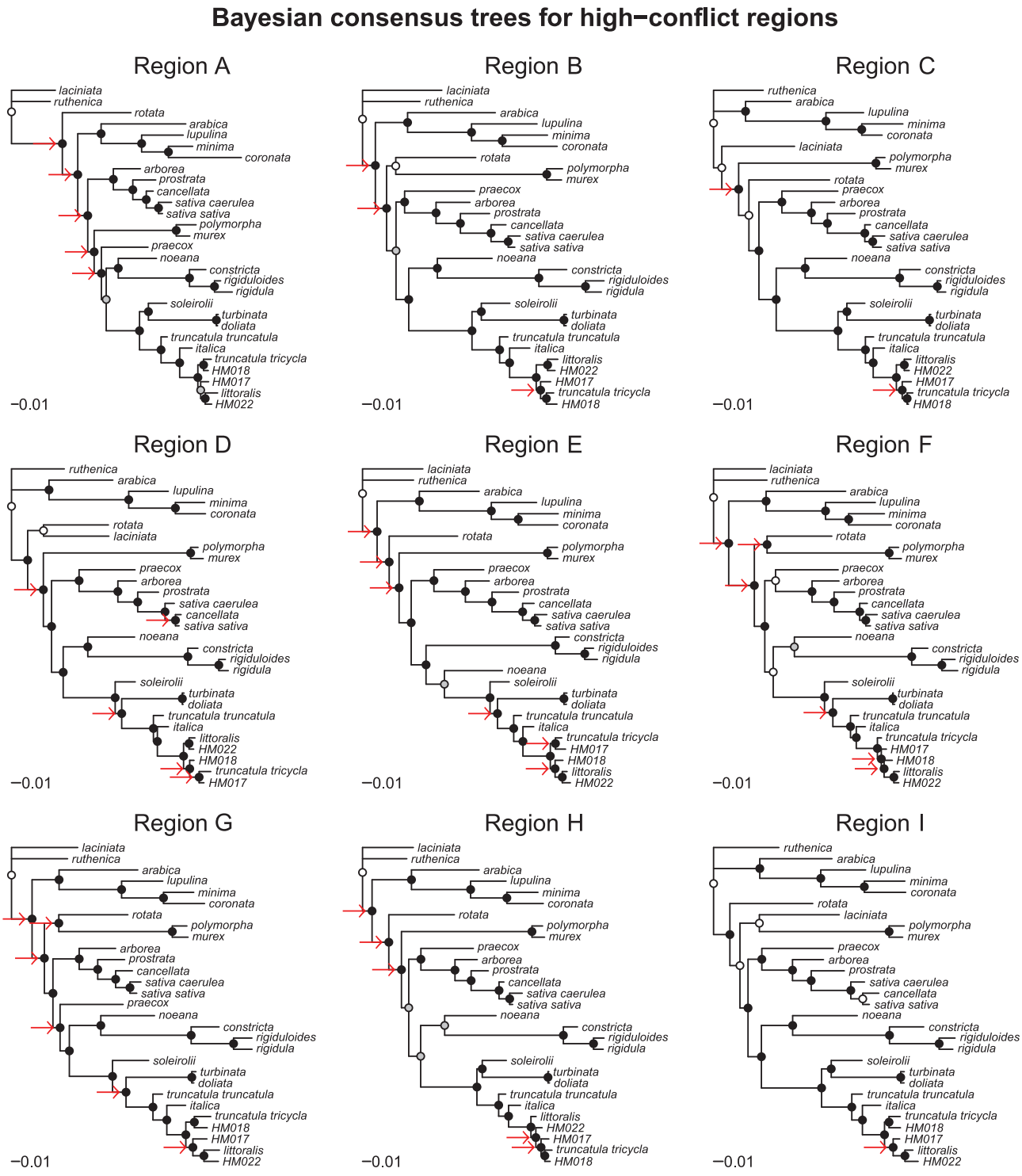

FIGURE 5. Phylogenetic estimates from genome regions with unusually high topological weighted conflict scores. Each tree is the Bayesian posterior consensus tree estimated from the concatenated sites between the given chromosome positions (light gray regions in Fig. 3), with posterior support indicated at the nodes: black filled circles, $\geqslant 0.95$; gray, between 0.95 and 0.75; and white, < 0.75. Nodes in conflict with the genome-wide estimate that have $\geqslant 0.95$ posterior support are highlighted with arrows. Scale bars indicate branch length in expected substitutions per site.

across all taxa, with 9 of 29 taxa having a coverage depth of $<10\times$), we identified 87 596 variable nuclear sites. In our genome-wide analysis, these data resolve relationships among several *Medicago* species that were ambiguous in previous multilocus studies (Maureira-Butler et al. 2008; Steele et al. 2010).

We assembled our phylogenomic data set by resequencing the genomes of the sampled taxa, then aligning this sequence data to the *M. truncatula* reference genome and identifying variable sites, an approach used for identification of SNPs within species (Branca et al. 2011; Davey et al. 2011; Etter et al. 2011). This approach allows us to rapidly identify informative, homologous sites across the genomes of the sampled taxa, in contrast to earlier phylogenomic studies that have required extensive effort to identify homologous loci in deeply diverged, independently assembled genomes (Marcet-Houben and Gabaldón 2009). The data set thus produced is overwhelmingly composed of SNPs mapping to annotated protein-coding regions (Table 2). The overrepresentation of coding regions is not surprising given that intergenic regions evolve rapidly in plant species, and our alignment criteria excluded reads that were not $>91\%$ identical to the *M. truncatula* reference genome. This overrepresentation of coding regions is also likely to affect other methods of genomic data collection, particularly those approaches used to sample transcriptomes, and may often be necessary in order to identify homologous sites with confidence.

Estimated Phylogenetic Relationships Within Medicago

We sampled only about a third of the estimated species diversity of genus *Medicago* (up to 87 species; Maureira-Butler et al. 2008; Steele et al. 2010; Small 2011), and it is difficult to predict how the addition of more taxa could alter the topology we recover. Nevertheless, the consensus topology from our analysis (Fig. 2a) is largely consistent with major section and subsection groupings identified from previous multilocus analyses (Maureira-Butler et al. 2008; Steele et al. 2010) while also resolving relationships that were not well-resolved in previous studies. For example, our data provide more confident resolution of relationships among recently diverged species and support a sister-group relationship between section *Medicago* and subsection *Pachyspireae* (Fig. 2a). Some of the strongly supported species-level relationships we resolved are in conflict with results from earlier studies based on sequence data on from 2 or 3 loci, totaling 3000–4000 bp—of which ~20% were variable sites (Maureira-Butler et al. 2008; Steele et al. 2010). These previous studies are thus comparable to the number of informative sites to the 500-site subsets—which produced a wide range of topologies, all conflicting with the topology produced using the full data set (Fig. 2b).

With genomic data, we are able to resolve relationships within the section *Medicago* clade (Fig. 2a), which contains the economically important *M. sativa* ssp. *sativa* (alfalfa). Our data provide strong support for *M.s. sativa* as sister to *M.s. caerulea*; and for *M. sativa*, *M. cancellata*, and *M. prostrata* being more closely related to one another than any are to *M. arborea*. Robust resolution of the species relationships within section *Medicago* did not necessarily require a large sample of the data set (Fig. 2a), though 3 of the genome regions that showed greatest conflict with the genome-wide topology strongly supported a different placement of *M. cancellata* (Fig. 5). This may reflect the suspected allopolyploid origins of *M. cancellata* (Small 2011), which is thought to have arisen through polyploid hybridization between *Medicago rupestris*, which is not included in our data set, and some *M. sativa* subspecies.

Our data also provide good resolution within the *Pachyspireae* clade, which contains the model legume *M. truncatula* var. *truncatula*. Based on our nuclear data set, *M. turbinata* and *M. doliata* are sister species, and, with *M. soleirolii*, they form a sister group to the rest of the "*truncatula* clade." Our data also find that *M. truncatula* var. *tricycla*, *M. littoralis*, and the 3 accessions formerly identified as *M. truncatula* var. *truncatula* (HM017, HM018, and HM022; Branca et al. 2011) together form a sister clade to *M. italica* rather than to *M. t. truncatula*. This relationship is recovered with strong posterior support in the full genomic data set (Fig. 2a), though not in many smaller genome regions (Fig. 5 and Supplementary Fig. S02). Placement of the 3 unidentified accessions in a clade with *M. littoralis* corroborates earlier genetic analyses (Branca et al. 2011) and preliminary morphological examination (Kelly Steele, unpublished data), which indicate that their identification as *M. t. truncatula* is incorrect. The placement of *M. t. tricycla* in that clade is worthy of note in light of the most recent comprehensive treatment of *Medicago* by Small (2011), which does not recognize *M. t. tricycla* as a taxon, but considers that most, if not all, accessions identified as *M. t. tricycla* are actually entities with considerable genetic material from *M. littoralis*. Our finding that an overwhelming majority of sliding window regions with strong support for one or more node in conflict with the genome-wide tree support rearrangements among the group of *M. italica*, *M. littoralis*, *M. t. tricycla*, and the unknown accessions (Supplementary Fig. S03) supports the hypothesis that introgression (recent or ongoing) and/or deep coalescence in these species are responsible for difficulty resolving their relationships; but rigorously determining the relative contributions of these processes would require wider sampling within the taxa concerned.

Variation in Phylogenetic Signal across the Nuclear Genome

The practice of concatenating thousands of probably unlinked sites into a single alignment for "supermatrix" analysis (Philippe et al. 2011) elides variation in the genealogical history of different chromosomal regions—that is, gene trees that conflict with the species tree (Nichols 2001; Lee and Edwards 2008). This practice is

nevertheless common for analysis of genome-scale data sets, due to the computational challenges of analyzing hundreds or thousands of independently assorting loci (e.g., Decker et al. 2009; Emerson et al. 2010; discussion in Philippe et al. 2011).

Indeed, we found that any analysis of our data which treated each SNP as an independent locus is computationally impractical. A promising method for estimating a species tree from an alignment of unlinked SNP markers has recently been published, implemented in the program SNAPP (Bryant et al. 2012; http://www.maths.otago.ac.nz/software/snapp last accessed April 2, 2012). We investigated the possibility of using this method for our data with guidance from the lead author (D. Bryant, personal communication to J.B.Y.). However, we found that SNAPP was not practical for our data. Analyzing a random sample of 1000 SNPs, and using computational resources comparable to those we employed for the MrBayes analysis, SNAPP's Markov chain sampler progressed at a rate that would have required 190 days to sample 1 million states; analyzing the complete data set, it would have required almost 55 years. Although the SNAPP framework shows promise as a rigorous method for phylogenetic analysis of unlinked SNP markers, it seems clear to us that improvements in both software and hardware performance are required before it can be routinely applied for genome-scale data.

As an alternative method, we considered the possibility of a method somewhere on the spectrum between the computationally challenging treatment of tens of thousands of SNPs as independent loci, and the biologically unrealistic method of concatenating thousands of independent loci for a single gene-tree estimate. Such a middle ground could be found if we could identify a defensible way to divide the data set into a computationally tractable number of smaller alignments, which might then be treated as separate "loci" in a coalescent species-tree estimation. For example, we found that we were able to complete an analysis in BEST, a coalescent method built on top of MrBayes (Liu et al. 2008) by treating the concatenated SNPs mapping to each of the 8 chromosomes of the Mt 3.5 genome as 8 pseudo-loci. This analysis achieved convergence within 96 h, running on 16 cores in the same computing cluster we used for our other analyses, and it supported an almost identical topology to our Bayesian estimate from the concatenated data set but with substantially higher posterior support. However, this approach violates the same assumptions our analysis of the complete concatenated alignment does, since it is demonstrably not the case that the 8 chromosomes of *M. truncatula* represent non-recombining loci. Similar schemes based on smaller subdivisions of the data might come closer to biological reality, but we can expect that the computational power required to analyze our entire data set under such an approach would increase accordingly.

Given these limitations, we believe that our alternative approach, scanning the genome for regions supporting topologies in substantial conflict with the phylogeny estimated from the complete data set, has some promise for illustrating variation in phylogenetic signal across the genome, and to assess the confidence we should have in relationships reconstructed by our MrBayes analysis of the concatenated data set. Previous multilocus estimates of the *Medicago* phylogeny have identified substantial conflict between loci (Maureira-Butler et al. 2008; Steele et al. 2010), and hybridization and incomplete lineage sorting have been proposed as potential explanations for this conflict. If these phenomena have been important in the evolution of genus *Medicago*, we would expect discrete genomic regions to strongly support distinct and conflicting evolutionary histories. Our sliding window analysis finds that specific genome regions strongly support alternative relationships in conflict with the genome-wide topology (Figs. 4 and 5); but it identifies few large, discrete regions of conflicting signal.

In spite of uncertainty arising from the effects of sample size (Fig. 2b), our analyses do provide indications of alternative gene histories within the genome-wide data set. Our neighbor-net analysis finds support for alternative groupings that are not represented in the Bayesian consensus tree (Figs. 2a and 3), and each of the highly conflicting regions we examine provides strong support for multiple relationships in conflict with the genome-wide phylogenetic estimate (Fig. 5 and Supplementary Fig. S03). In some cases, these strongly supported conflicts recapitulate the alternative groupings revealed in our neighbor-net analysis (Fig. 3) and found in earlier phylogenetic studies of *Medicago* (Maureira-Butler et al. 2008; Steele et al. 2010), and morphological studies (Small 2011). The strong correlations between Bayesian posterior support for each clade (Fig. 2a, Results), and the frequency with which the clades were recovered in estimates from smaller subsamples of the data, are consistent with the hypothesis that poor posterior support reflects conflict among many independent gene trees over these relationships.

First, there are the rearrangements among the most recently diverged taxa in our sample, which are first indicated by the ambiguity of these relationships in the network analysis (Fig. 3) and strongly supported by windows on every chromosome (Supplementary Fig. S03a and b). This result recapitulates the earlier finding of Branca et al. (2011) that data from the chloroplast genome place the HM017, HM018, and HM022 accessions as more closely related to *M. t. tricycla* and *M. littoralis* than to *M. t. truncatula*. It is also consistent with hybridization or shared ancestral polymorphism among these taxa, corroborating Small's (2011) proposal that *M. t. tricycla* is nested within *M. littoralis*; although sampling multiple accessions from each of these taxa is necessary to conclusively test that hypothesis.

Similarly, *M. soleirolii* is nested within what we call the "*truncatula* clade" in the genome-wide estimate (Fig. 2a), but is placed as basal to the rest of this clade by several of the "highly conflicting" genome regions (Fig. 4).

This is consistent with the neighbor-net analysis, which finds support for 2 major alternative rootings of the "*truncatula* clade" group (Fig. 3). It also echoes past multilocus studies, which have found disagreement in the placement of *M. soleirolii* among independent loci. Maureira-Butler et al. (2008) found that phylogenetic estimates from 2 nuclear genes placed *M. soleirolii* in 2 different polytomies within subsection Pachyspireae; Steele et al. (2010) found *M. soleirolii* strongly supported as sister to *M. turbinata* based on sequence data from chloroplast loci, but sister to *M. truncatula*+*M. littoralis* based on nuclear sequence data.

Within section *Medicago*, the relationships among *Medicago sativa* ssp. *sativa*, *M. sativa* ssp. *caerulea*, and *M. cancellata* are rearranged in trees estimated from 3 highly conflicting regions (Fig. 5), both times by the placement of *M. cancellata* as sister to *M. sativa*. Nuclear gene trees estimated by Maureira-Butler et al. (2008) placed *M. sativa* ssp. *sativa* and *M.s.*ssp. *caerulea* as members of the same monophyletic group (though this study did not sample *M. cancellata*); Steele et al. (2010) found support for relationships compatible with our topology in a tree estimated from a nuclear locus, but not in an estimate based on a chloroplast locus. *Medicago cancellata* is thought to be allopolyploid, with contributions from *M. rupestris* (which we do not sample) and an unknown subspecies of *M. sativa* (Small 2011)—this may contribute to uncertainty in the placement of *M. cancellata* relative to *M. sativa*.

Conclusions

We have used shallow resequencing and alignment to a reference genome to identify 87 596 variable, homologous sites across the genomes of 29 accessions sampled from the genus *Medicago*. Not surprisingly, these data support a phylogeny with high posterior support, resolving many species relationships that were ambiguous in previous, smaller multilocus studies. Resolving these relationships, however, required a large proportion of our data set. Samples of 500 sites showed considerable inconsistencies as a set, and never resolved the same phylogeny as the complete data set; even random samples of 5000 sites often supported phylogenies that were not identical to the genome-wide reconstruction. A neighbor-net network analysis of the complete nuclear data set reveals support for alternative topologies that is not evident from our coalescent species-tree estimate. Finally, our examination of genome regions supporting phylogenetic estimates that strongly conflict with the genome-wide phylogeny finds, in some cases, strong support for alternative relationships that recapitulates both the results of the network analysis conflicts found among loci in previous phylogenetic studies of *Medicago*.

We hope our data provide useful insight into the size of phylogenomic data sets that will be needed for the resolution of recently diverged species. We also believe that the approaches we use here—subsampling to visualize variation in phylogenetic signal within the data set, and examining the relationship between genomic position and phylogenetic signal—will be useful in assessing the robustness of species relationships estimated from phylogenomic data sets in the future.

Supplementary Material

Data files and supplementary figures can be found in the Dryad data repository at http://datadryad.org, DOI: 10.5061/dryad.vp634306.

Funding

The *Medicago* HapMap Project was funded by the US National Science Foundation (PGRP-0820005) and by the Noble Foundation.

Acknowledgments

The authors thank Roxanne Denny for extensive assistance with data collection, and Keith Barker for invaluable guidance on analysis. Germplasm for the *Medicago* HapMap Project was provided by the Institut National de la Recherche Agronomique in Montpellier, France; and by the Noble Foundation. Computation resources for sequence alignment, data processing, and phylogenetic analysis were conducted on systems provided by the Minnesota Supercomputing Institute.

References

Ameline-Torregrosa C., Wang B.-B., O'Bleness M.S., Deshpande S., Zhu H., Roe B., Young N.D., Cannon S.B. 2008. Identification and characterization of nucleotide-binding site-leucine-rich repeat genes in the model plant *Medicago truncatula*. Plant Physiol. 146:5–21.

Baird N.A., Etter P.D., Atwood T.S., Currey M.C., Shiver A.L., Lewis Z.A., Selker E.U., Cresko W.A., Johnson E.A. 2008. Rapid SNP discovery and genetic mapping using sequenced RAD markers. PLoS One 3:e3376.

Barker D., Bianchi S., Blondon F. 1990. *Medicago truncatula*, a model plant for studying the molecular genetics of the rhizobium-legume symbiosis. Plant Mol. Biol. Rep. 8:40–49.

Bentley D.R., Balasubramanian S., Swerdlow H.P., Smith G.P., Milton J., Brown C.G., Hall K.P., Evers D.J., Barnes C.L., Bignell H.R., Boutell J.M., Bryant J., Carter R.J., Keira Cheetham R., Cox A.J., Ellis D.J., Flatbush M.R., Gormley N.A., Humphray S.J., Irving L.J., Karbelashvili M.S., Kirk S.M., Li H., Liu X., Maisinger K.S., Murray L.J., Obradovic B., Ost T., Parkinson M.L., Pratt M.R., Rasolonjatovo I.M., Reed M.T., Rigatti R., Rodighiero C., Ross M.T., Sabot A., Sankar S.V., Scally A., Schroth G.P., Smith M.E., Smith V.P., Spiridou A., Torrance P.E., Tzonev S.S., Vermaas E.H., Walter K., Wu X., Zhang L., Alam M.D., Anastasi C., Aniebo I.C., Bailey D.M., Bancarz I.R., Banerjee S., Barbour S.G., Baybayan P.A., Benoit V.A., Benson K.F., Bevis C., Black P.J., Boodhun A., Brennan J.S., Bridgham J.A., Brown R.C., Brown A.A., Buermann D.H., Bundu A.A., Burrows J.C., Carter N.P., Castillo N., Chiara E. Catenazzi M., Chang S., Neil Cooley R., Crake N.R., Dada O.O., Diakoumakos K.D., Dominguez-Fernandez B., Earnshaw D.J., Egbujor U.C., Elmore D.W., Etchin S.S., Ewan M.R., Fedurco M., Fraser L.J., Fuentes Fajardo K.V., Scott Furey W., George D., Gietzen K.J., Goddard C.P., Golda G.S., Granieri P.A., Green D.E., Gustafson D.L., Hansen N.F., Harnish K., Haudenschild

C.D., Heyer N.I., Hims M.M., Ho J.T., Horgan A.M., Hoschler K., Hurwitz S., Ivanov D.V., Johnson M.Q., James T., Huw Jones T.A., Kang G.D., Kerelska T.H., Kersey A.D., Khrebtukova I., Kindwall A.P., Kingsbury Z., Kokko-Gonzales P.I., Kumar A., Laurent M.A., Lawley C.T., Lee S.E., Lee X., Liao A.K., Loch J.A., Lok M., Luo S., Mammen R.M., Martin J.W., McCauley P.G., McNitt P., Mehta P., Moon K.W., Mullens J.W., Newington T., Ning Z., Ling Ng B., Novo S.M., O'Neill M.J., Osborne M.A., Osnowski A., Ostadan O., Paraschos L.L., Pickering L., Pike A.C., Pike A.C., Chris Pinkard D., Pliskin D.P., Podhasky J., Quijano V.J., Raczy C., Rae V.H., Rawlings S.R., Chiva Rodriguez A., Roe P.M., Rogers J., Rogert Bacigalupo M.C., Romanov N., Romieu A., Roth R.K., Rourke N.J., Ruediger S.T., Rusman E., Sanches-Kuiper R.M., Schenker M.R., Seoane J.M., Shaw R.J., Shiver M.K., Short S.W., Sizto N.L., Sluis J.P., Smith M.A., Ernest Sohna Sohna J., Spence E.J., Stevens K., Sutton N., Szajkowski L., Tregidgo C.L., Turcatti G., Vandevondele S., Verhovsky Y., Virk S.M., Wakelin S., Walcott G.C., Wang J., Worsley G.J., Yan J., Yau L., Zuerlein M., Rogers J., Mullikin J.C., Hurles M.E., McCooke N.J., West J.S., Oaks F.L., Lundberg P.L., Klenerman D., Durbin R., Smith A.J. 2008. Accurate whole human genome sequencing using reversible terminator chemistry. Nature 456:53–9.

Branca A., Paape T.D., Zhou P., Briskine R., Farmer A.D., Mudge J., Bharti A.K., Woodward J.E., May G.D., Gentzbittel L., Ben C., Denny R., Sadowsky M.J., Ronfort J., Bataillon T., Young N.D., Tiffin P. 2011. Whole-genome nucleotide diversity, recombination, and linkage disequilibrium in the model legume *Medicago truncatula*. Proc. Natl Acad. Sci. U S A 108:E864–E870.

Brinkmann H., Van der Giezen M., Zhou Y., Poncelin de Raucourt G., Philippe H. 2005. An empirical assessment of long-branch attraction artefacts in deep eukaryotic phylogenomics. Syst. Biol. 54:743–757.

Bryant D., Bouckaert R. 2009. Inferring species trees directly from biallelic genetic markers: bypassing gene trees in a full coalescent analysis. Mol. Biol. Evol. 29:1917–1732.

Bryant D., Bouckaert R., Felsenstein J., Rosenberg N.A., RoyChoudhury A. 2012. Inferring species trees directly from biallelic genetic markers: bypassing gene trees in a full coalescent analysis. Mol. Biol. Evol. 29:1917–1932.

Burki F., Shalchian-Tabrizi K., Minge M., Skjaeveland A., Nikolaev S.I., Jakobsen K.S., Pawlowski J. 2007. Phylogenomics reshuffles the eukaryotic supergroups. PLoS One 2:e790.

Burleigh J.G., Bansal M.S., Eulenstein O., Hartmann S., Wehe A., Vision T.J. 2011. Genome-scale phylogenetics: inferring the plant tree of life from 18,896 gene trees. Syst. Biol. 60:117–125.

Carstens B., Knowles L. 2007. Estimating species phylogeny from gene-tree probabilities despite incomplete lineage sorting: an example from *Melanoplus* grasshoppers. Syst. Biol. 56:400–411.

Catchen J.M., Amores A., Hohenlohe P., Cresko W., Postlethwait J.H. 2011. Stacks: building and genotyping loci de novo from short-read sequences. G3 1:171.

Cook D., VandenBosch K., de Bruijn F., Huguet T. 1997. Model legumes get the nod. Plant Cell 9:275–282.

Cranston K.A., Hurwitz B., Ware D., Stein L., Wing R.A. 2009. Species trees from highly incongruent gene trees in rice. Syst. Biol. 58: 489–500.

Cranston K., Hurwitz B., Sanderson M. 2010. Phylogenomic analysis of BAC-end sequence libraries in *Oryza* (Poaceae). Syst. Bot. 35:512–523.

Davey J.W., Hohenlohe P.A., Etter P.D., Boone J.Q., Catchen J.M., Blaxter M.L. 2011. Genome-wide genetic marker discovery and genotyping using next-generation sequencing. Nat. Rev. Genet. 12:499–510.

Decker J.E., Pires J.C., Conant G.C., McKay S.D., Heaton M.P., Chen K., Cooper A., Vilkki J., Seabury C.M., Caetano A.R., Johnson G.S., Brenneman R.A., Hanotte O., Eggert L.S., Wiener P., Kim J.-J., Kim K.S., Sonstegard T.S., Van Tassell C.P., Neibergs H.L., McEwan J.C., Brauning R., Coutinho L.L., Babar M.E., Wilson G.A., McClure M.C., Rolf M.M., Kim J., Schnabel R.D., Taylor J.F. 2009. Resolving the evolution of extant and extinct ruminants with high-throughput phylogenomics. Proc. Natl Acad. Sci. U S A 106:18644–18649.

Degnan J.H., Rosenberg N.A. 2009. Gene tree discordance, phylogenetic inference and the multispecies coalescent. Trends Ecol. Evol. 24:332–340.

Ekblom R., Galindo J. 2010. Applications of next generation sequencing in molecular ecology of non-model organisms. Heredity 107:1–15.

Emerson K.J., Merz C.R., Catchen J.M., Hohenlohe P.A., Cresko W.A., Bradshaw W.E., Holzapfel C.M. 2010. Resolving postglacial phylogeography using high-throughput sequencing. Proc. Natl Acad. Sci. U S A 107:16196–16200.

Etter P., Bassham S., Hohenlohe P.A., Johnson E.A. Cresko W.A. 2011. SNP discovery and genotyping for evolutionary genetics using RAD sequencing. Methods Mol. Biol. 772: 157–178.

Felsenstein J. 1989. PHYLIP—Phylogeny Inference Package (Version 3.2). Cladistics 5:164–166.

Felsenstein J. 2004. Inferring phylogenies. Sunderland (MA): Sinauer Associates.

Felsenstein J. 2005. PHYLIP (Phylogeny Inference Package). Available from: URL http://evolution.genetics.washington.edu/phylip.html (last accessed December 7, 2011).

Gatesy J., DeSalle R., Wahlberg N. 2007. How many genes should a systematist sample? Conflicting insights from a phylogenomic matrix characterized by replicated incongruence. Syst. Biol. 56: 355–363.

Gibbs R., Belmont J., Hardenbol P., Willis T. 2003. The international HapMap project. Nature 426:789–796.

Gore M.A., Chia J.-M., Elshire R.J., Sun Q., Ersoz E.S., Hurwitz B.L., Peiffer J.A., McMullen M.D., Grills G.S., Ross-Ibarra J., Ware D.H., Buckler E.S. 2009. A first-generation haplotype map of maize. Science 326:1115–11117.

Hackett S., Kimball R., Reddy S. 2008. A phylogenomic study of birds reveals their evolutionary history. Science 320:1763–1768.

Harshman J., Braun E.L., Braun M.J., Huddleston C.J., Bowie R.C.K., Chojnowski J.L., Hackett S.J., Han K.-L., Kimball R.T., Marks B.D., Miglia K.J., Moore W.S., Reddy S., Sheldon F.H., Steadman D.W., Steppan S.J., Witt C.C., Yuri T. 2008. Phylogenomic evidence for multiple losses of flight in ratite birds. Proc. Natl Acad. Sci. U S A 105:13462–13467.

Huelsenbeck J., Ronquist F. 2001. MrBayes: Bayesian inference of phylogenetic trees. Bioinformatics 17:754–755.

Huson D.H., Bryant D. 2006. Application of phylogenetic networks in evolutionary studies. Mol. Biol. Evol. 23:254–267.

Jeffroy O., Brinkmann H., Delsuc F. Philippe H. 2006. Phylogenomics: the beginning of incongruence? Trends Genet. 22:225–231.

Jennings W. 2005. Speciational history of Australian grass finches (Poephila) inferred from thirty gene trees. Evolution 59: 2033–2047.

Kim S., Plagnol V., Hu T.T., Toomajian C., Clark R.M., Ossowski S., Ecker J.R., Weigel D., Nordborg M. 2007. Recombination and linkage disequilibrium in *Arabidopsis thaliana*. Nat. Genet. 39: 1151–1155.

Knowles L.L. 2009. Estimating species trees: methods of phylogenetic analysis when there is incongruence across genes. Syst. Biol. 58: 463–467.

Kubatko L.S., Carstens B.C., Knowles L.L. 2009. STEM: species tree estimation using maximum likelihood for gene trees under coalescence. Bioinformatics 25:971–973.

Kubatko L.S., Degnan J.H. 2007. Inconsistency of phylogenetic estimates from concatenated data under coalescence. Syst. Biol. 56:17–24.

Lavin M., Herendeen P.S., Wojciechowski M.F. 2005. Evolutionary rates analysis of Leguminosae implicates a rapid diversification of lineages during the Tertiary. Syst. Biol. 54:575–594.

Lee J.Y., Edwards S.V. 2008. Divergence across Australia's Carpenterian barrier: statistical phylogeography of the red-backed fairy wren (*Malurus melanocephalus*). Evolution 62:3117–3134.

Lewis P. 2001. A likelihood approach to estimating phylogeny from discrete morphological character data. Syst. Biol. 50:913–925.

Liu L., Pearl D.K., Brumfield R.T., Edwards S.V. 2008. Estimating species trees using multiple-allele DNA sequence data. Evolution 62:2080–2091.

Maddison W., Knowles L. 2006. Inferring phylogeny despite incomplete lineage sorting. Syst. Biol. 55:21–30.

Marcet-Houben M., Gabaldón T. 2009. The tree versus the forest: the fungal tree of life and the topological diversity within the yeast phylome. PLoS One 4:e4357.

Maureira-Butler I.J., Pfeil B.E., Muangprom A., Osborn T.C., Doyle J.J. 2008. The reticulate history of *Medicago* (Fabaceae). Syst. Biol. 57: 466–482.

Murphy W.J., Pevzner P.A., O'Brien S.J. 2004. Mammalian phylogenomics comes of age. Trends Genet. 20:631–639.
Nichols R. 2001. Gene trees and species trees are not the same. Trends Ecol. Evol. 16:358–364.
Nishihara H., Hasegawa M., Okada N. 2006. Pegasoferae, an unexpected mammalian clade revealed by tracking ancient retroposon insertions. Proc. Natl Acad. Sci. U S A 103:9929–9934.
Pamilo P., Nei M. 1988. Relationships between gene trees and species trees. Mol. Biol. Evol. 5:568–583.
Paradis E., Claude J., Strimmer K. 2004. APE: analysis of phylogenetics and evolution in R language. Bioinformatics 20: 289–290.
Penny D., Hendy M. 1985. The Use of Tree Comparison Metrics. Syst. Zool. 34: 75–82.
Philippe H., Brinkmann H., Lavrov D.V., Littlewood D.T.J., Manuel M., Wörheide G., Baurain D. 2011. Resolving difficult phylogenetic questions: why more sequences are not enough. PLoS Biol. 9:e1000602.
Philippe H., Delsuc F., Brinkmann H., Lartillot N. 2005. Phylogenomics. Annu. Rev. Ecol. Evol. Syst. 36:541–562.
Philippe H., Derelle R., Lopez P., Pick K., Borchiellini C., Boury-Esnault N., Vacelet J., Renard E., Houliston E., Quéinnec E., Da Silva C., Wincker P., Le Guyader H., Leys S., Jackson D.J., Schreiber F., Erpenbeck D., Morgenstern B., Wörheide G., Manuel M. 2009. Phylogenomics revives traditional views on deep animal relationships. Curr. Biol. 19:706–712.
R Core Team 2012. R: a language and environment for statistical computing. Available from: URL www.r-project.org.
Ronquist F., Huelsenbeck J. 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics 19: 1572–1574.
Sanderson M.J., McMahon M.M., Steel M. 2010. Phylogenomics with incomplete taxon coverage: the limits to inference. BMC Evol. Biol. 10:155.
Siepel A. 2009. Phylogenomics of primates and their ancestral populations. Genome Res. 19:1929–1941.
Small E. 2011. Alfalfa and relatives: evolution and classification of *Medicago*. Ottawa: NRC Research Press.
Steele K.P., Ickert-Bond S.M., Zarre S., Wojciechowski M.F. 2010. Phylogeny and character evolution in *Medicago* (Leguminosae): evidence from analyses of plastid trnK/matK and nuclear GA3ox1 sequences. Am. J. Bot. 97:1142–1155.
Steinke D., Salzburger W., Meyer A. 2006. Novel relationships among ten fish model species revealed based on phylogenomic analysis using ESTs. J. Mol. Evol. 62:772–784.
Tajima F. 1983. Evolutionary relationship of DNA sequences in finite populations. Genetics 105:437–460.
Tiffin P., Moeller D.A. 2006. Molecular evolution of plant immune system genes. Trends Genet. 22:662–670.
Wiens J.J., Kuczynski C.A., Townsend T., Reeder T.W., Mulcahy D.G., Sites J.W. 2010. Combining phylogenomics and fossils in higher-level squamate reptile phylogeny: molecular data change the placement of fossil taxa. Syst. Biol. 59:674–688.
Wu T.D., Nacu S. 2010. Fast and SNP-tolerant detection of complex variants and splicing in short reads. Bioinformatics 26:873–881.
Young N.D., Debellé F., Oldroyd G.E.D., Geurts R., Cannon S.B., Udvardi M.K., Benedito V.A., Mayer K.F., Gouzy J., Schoof H., Van de Peer Y., Proost S., Cook D.R., Meyers B.C., Spannagl M., Cheung F., De Mita S., Krishnakumar V., Gundlach H., Zhou S., Mudge J., Bharti A.K., Murray J.D., Naoumkina M.A., Rosen B., Silverstein K.A., Tang H., Rombauts S., Zhao P.X., Zhou P., Barbe V., Bardou P., Bechner M., Bellec A., Berger A., Bergès H., Bidwell S., Bisseling T., Choisne N., Couloux A., Denny R., Deshpande S., Dai X., Doyle J.J., Dudez A.M., Farmer A.D., Fouteau S., Franken C., Gibelin C., Gish J., Goldstein S., González A.J., Green P.J., Hallab A., Hartog M., Hua A., Humphray S.J., Jeong D.H., Jing Y., Jöcker A., Kenton S.M., Kim D.J., Klee K., Lai H., Lang C., Lin S., Macmil S.L., Magdelenat G., Matthews L., McCorrison J., Monaghan E.L., Mun J.H., Najar F.Z., Nicholson C., Noirot C., O'Bleness M., Paule C.R., Poulain J., Prion F., Qin B., Qu C., Retzel E.F., Riddle C., Sallet E., Samain S., Samson N., Sanders I., Saurat O., Scarpelli C., Schiex T., Segurens B., Severin A.J., Sherrier D.J., Shi R., Sims S., Singer S.R., Sinharoy S., Sterck L., Viollet A., Wang B.B., Wang K., Wang M., Wang X., Warfsmann J., Weissenbach J., White D.D., White J.D., Wiley G.B., Wincker P., Xing Y., Yang L., Yao Z., Ying F., Zhai J., Zhou L., Zuber A., Dénarié J., Dixon R.A., May G.D., Schwartz D.C., Rogers J., Quétier F., Town C.D., Roe B.A. 2011. The Medicago genome provides insight into the evolution of rhizobial symbioses. Nature 480: 520–524.