BaitFisher: A Software Package for Multispecies Target DNA Enrichment Probe Design

Christoph Mayer,[†,1] Manuela Sann,[†,1,2] Alexander Donath,[1] Martin Meixner,[3] Lars Podsiadlowski,[4] Ralph S. Peters,[5] Malte Petersen,[1] Karen Meusemann,[1,6] Karsten Liere,[3] Johann-Wolfgang Wägele,[7] Bernhard Misof,[1] Christoph Bleidorn,[8,9,10] Michael Ohl,[*,2] and Oliver Niehuis[*,1]

[1]Center for Molecular Biodiversity Research, Zoological Research Museum Alexander Koenig, Bonn, Germany

[2]Museum für Naturkunde, Leibniz Institute for Evolution and Biodiversity Science, Berlin, Germany

[3]Services in Molecular Biology GmbH, Rüdersdorf, Germany

[4]University of Bonn, Institute of Evolutionary Biology and Ecology, Bonn, Germany

[5]Department Arthropoda, Zoological Research Museum Alexander Koenig, Bonn, Germany

[6]Australian National Insect Collection, CSIRO National Research Collections Australia, Acton, Canberra, ACT, Australia

[7]Zoological Research Museum Alexander Koenig, Bonn, Germany

[8]Molecular Evolution and Systematics of Animals, Institute for Biology, University of Leipzig, Leipzig, Germany

[9]German Centre for Integrative Biodiversity Research (iDiv) Halle-Jena-Leipzig, Leipzig, Germany

[10]Museo Nacional de Ciencias Naturales, Spanish National Research Council (CSIC), Madrid, Spain

[†]These authors contributed equally to this work.

***Corresponding author:** E-mail: o.niehuis@zfmk.de; michael.ohl@mfn-berlin.de.

Associate editor: Michael Rosenberg

Abstract

Target DNA enrichment combined with high-throughput sequencing technologies is a powerful approach to probing a large number of loci in genomes of interest. However, software algorithms that explicitly consider nucleotide sequence information of target loci in multiple reference species for optimizing design of target enrichment baits to be applicable across a wide range of species have not been developed. Here we present an algorithm that infers target DNA enrichment baits from multiple nucleotide sequence alignments. By applying clustering methods and the combinatorial 1-center sequence optimization to bait design, we are able to minimize the total number of baits required to efficiently probe target loci in multiple species. Consequently, more loci can be probed across species with a given number of baits. Using transcript sequences of 24 apoid wasps (Hymenoptera: Crabronidae, Sphecidae) from the 1KITE project and the gene models of *Nasonia vitripennis*, we inferred 57,650, 120-bp-long baits for capturing 378 coding sequence sections of 282 genes in apoid wasps. Illumina reduced-representation library sequencing confirmed successful enrichment of the target DNA when applying these baits to DNA of various apoid wasps. The designed baits furthermore enriched a major fraction of the target DNA in distantly related Hymenoptera, such as Formicidae and Chalcidoidea, highlighting the baits' broad taxonomic applicability. The availability of baits with broad taxonomic applicability is of major interest in numerous disciplines, ranging from phylogenetics to biodiversity monitoring. We implemented our new approach in a software package, called BaitFisher, which is open source and freely available at https://github.com/cmayer/BaitFisher-package.git.

Key words: **hybrid enrichment, comparative genomics, phylogenetics, phylogenomics, Hymenoptera.**

Article

Introduction

Target DNA enrichment combined with high-throughput sequencing technology is a highly promising approach to studying and characterizing a large number of loci in genomes, at reasonable costs. Target DNA enrichment comprises various molecular techniques that augment target DNA in a given next-generation sequencing (NGS) library by means of oligonucleotide probes (hereafter also synonymously referred to as baits), either in solution (Faircloth et al. 2012; Lemmon et al. 2012) or on an array (Albert et al. 2007; Hodges et al. 2007, 2009; Liu et al. 2016). The nucleotide sequences of these baits are selected for high nucleotide sequence similarity to target DNA sequence sections of interest. The baits can then be hybridized to the target sequence sections in a DNA sample, which allows enriching these sequence sections. This technique has been named differently depending on which target regions are enriched (e.g., exome or gene capture when exons/coding DNA sequences are enriched [Ng et al. 2009; Cosart et al. 2011; Fisher et al. 2011; Li et al. 2013]; anchored hybrid enrichment when the flanking region of [ultra] conserved regions are of interest [Bejerano et al. 2004; Crawford et al. 2012; Faircloth et al. 2012, 2014; Lemmon et al. 2012; Bragg et al. 2015; Hawkins et al. 2015; Vinner et al. 2015];

hyRAD when specific RAD segments are enriched [Suchan et al. 2016]). Various laboratory protocols for enriching target loci have been developed (Bashiardes et al. 2005; Blumenstiel et al. 2010; Meyer and Kircher 2010; Bodi et al. 2013; Peñalba et al. 2014). Furthermore, molecular procedures have been described, which allow the capture of more dissimilar target loci with a given set of baits and extend the reach of the method considerably (Li et al. 2013; Paijmans et al. 2016). However, because target locus enrichment efficacy decreases with increasing bait-to-target DNA sequence distance (Bragg et al. 2015; Hawkins et al. 2015; Paijmans et al. 2016; present study), design of bait sets to be applied across a range of distantly related species can still pose a challenge. For example, a given bait can exhibit a high nucleotide sequence similarity to the target DNA of species in one ingroup lineage and consequently effectively enrich the target DNA in species of this lineage. But if the same bait differs significantly from the target DNA in species of another ingroup lineage, it will not enrich it to the same extent (or at all) in the species of the second lineage. In such a situation, one might want to design more than one bait to cope with the significant ingroup target locus sequence divergence and thereby improving the odds that the target locus is evenly enriched across all ingroup species.

No software algorithm is available so far that allows formally optimizing the number of baits for enriching target loci across a diverse group of species by dynamically adjusting the number of baits to the known taxonomic ingroup target locus divergence. Ideally, baits are developed by exploiting target locus nucleotide sequence information from multiple reference species that representatively capture ingroup nucleotide sequence divergence of all target loci. Baits should then be designed in a way that 1) for every target locus and reference species there is a bait that differs in less than a user-defined nucleotide sequence similarity threshold value from the target DNA in the reference species and 2) the total number of baits that fulfil criterion 1) is minimized. There is a growing need for such an approach, because the costs for bait sets scale with the number of different baits in such a set and comparative (phylo-)genomic studies, in which target genes are sampled across a wide range of species, are frequently conducted (Bejerano et al. 2004; Faircloth et al. 2012, 2014; Lemmon et al. 2012; Li et al. 2013; McCormack, Harvey, et al. 2013; McCormack, Hird, et al. 2013; Bragg et al. 2015; Hawkins et al. 2015; Hugall et al. 2015; Prum et al. 2015; Vinner et al. 2015).

Here we present a novel approach for the design of hybridization baits to be applied to DNA of a range of species. It infers baits by exploiting user-provided nucleotide sequence information of target loci in a representative set of species. It optimizes the total number and the nucleotide sequences of baits so that for each target locus in a reference species there is exactly one designed bait that differs in less than the user-defined bait-to-target nucleotide sequence similarity threshold value from the target locus. It furthermore allows the user to specify any intended tiling design and to thus compensate edge effects that may arise from shifts of, for example, exon–intron boundaries and other local but substantial changes in the target DNA (Bi et al. 2012). We implemented this approach in a software package called BaitFisher, which comprises two programs. The first one, BaitFisher, provides all possible bait designs suitable for enriching a given target locus (e.g., a gene or the exon of a specific gene). The output from BaitFisher can be passed to the second program, BaitFilter, for selecting a specific bait set. BaitFilter enables choosing the optimal start position for a given tiling design in a given locus based on a user-specified optimality criterion (i.e., minimizing the number of baits required to enrich the target locus or maximizing the number of nucleotide sequences which baits were inferred from). BaitFilter is also able to assess whether or not baits are likely to bind to multiple genomic regions and whether baits are likely to bind to contiguous genomic DNA (e.g., in case the applied gene model of a reference species was not correct when extracting CDS regions from user-provided transcript sequences). Both procedures require a user-provided reference species genome assembly.

We empirically tested baits inferred with BaitFisher in a pilot project for a study on the phylogeny of apoid wasps and bees, exploiting the comprehensive transcriptome libraries of insects compiled in the international 1KITE project (www.1kite.org). We present the result from this pilot project, which demonstrates that the BaitFisher-inferred baits were able to efficiently enrich a major fraction of the target genes in the taxonomic target group. Our results furthermore provide insights into what bait-to-target distance threshold value to choose when designing baits with the BaitFisher software in order to ensure that target loci are consistently and effectively enriched across species when applying molecular procedures similar to those used by us.

New Approaches

Software Implementation

BaitFisher designs baits for target DNA enrichment on the basis of multiple nucleotide sequence alignments that contain contiguous template DNA. Suitable templates are 1) genomic DNA sequences (gDNA) for designing baits that are meant to enrich gDNA, 2) transcript-complementary DNA (cDNA) sequences for designing baits that are meant to enrich cDNA, or 3) cDNA sequences for designing baits that are meant to enrich gDNA (fig. 1). When using the latter as templates, the software requires nucleotide sequence alignments that additionally include the cDNA sequence of a user-defined reference species with an annotated genome. By providing the genome assembly and corresponding gene feature format (GFF) annotation file of the reference species to BaitFisher, the program is able to split aligned cDNA sequences into genome-feature-specific sections, such as exons or coding sequence (CDS) regions (fig. 1, step 1). Specifically, the program uses the gene ID of the reference species in the multiple cDNA alignment to fetch nucleotide sequences from each corresponding feature (e.g., exons or CDS sections) in the assembled genome by using the corresponding coordinates in the GFF file. Each retrieved genomic feature sequence is subsequently aligned to the cDNA sequence of the reference species with the Needleman–

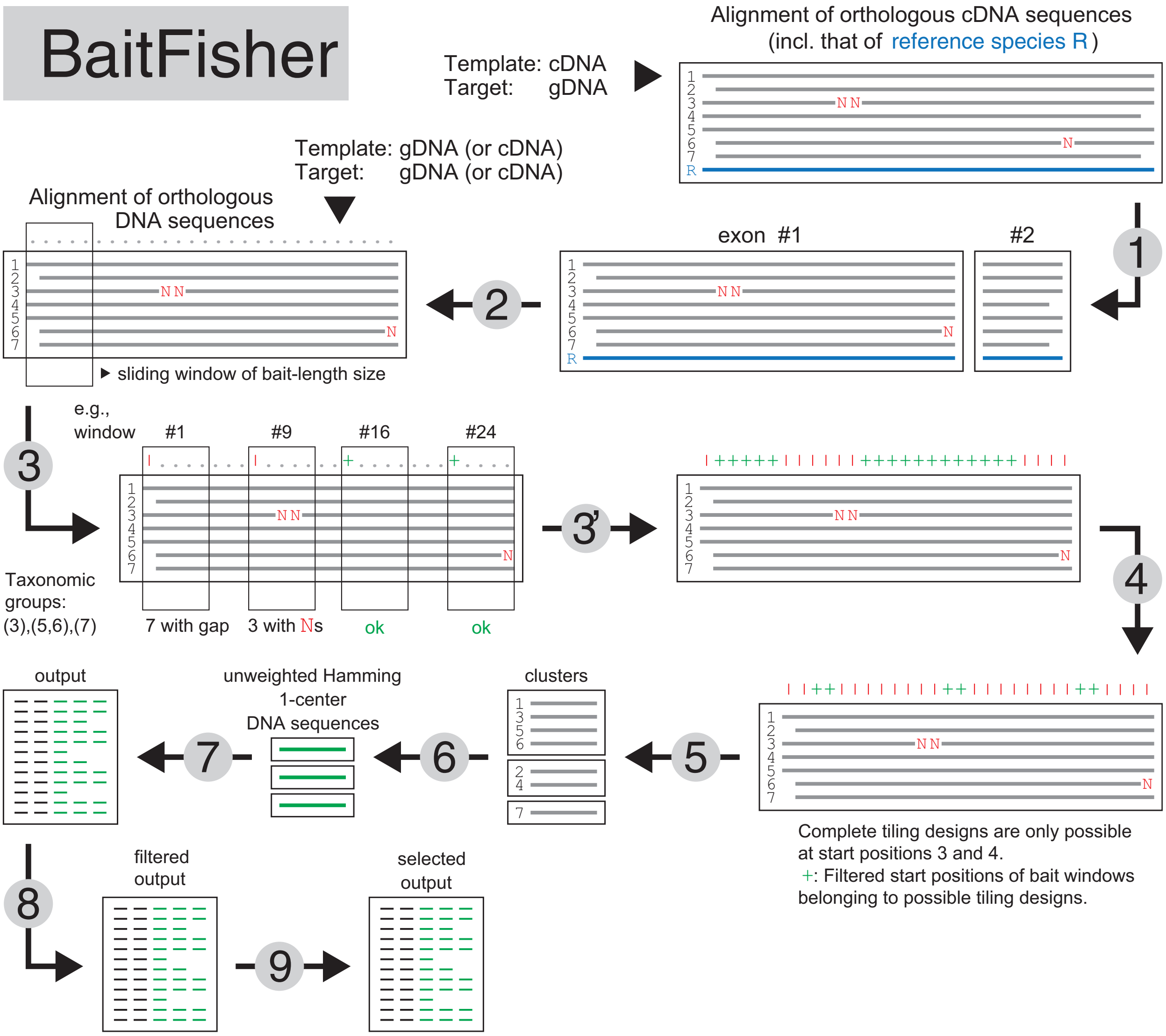
BaitFisher
Template: cDNA
Target: gDNA
Alignment of orthologous cDNA sequences (incl. that of reference species R)
1 2 3 4 5 6 7 R
N N
N
1
exon #1
#2
2
Template: gDNA (or cDNA)
Target: gDNA (or cDNA)
Alignment of orthologous DNA sequences
sliding window of bait-length size
3
e.g., window
#1
#9
#16
#24
Taxonomic groups: (3),(5,6),(7)
7 with gap
3 with Ns
ok
ok
3'
4
Complete tiling designs are only possible at start positions 3 and 4.
+: Filtered start positions of bait windows belonging to possible tiling designs.
5
clusters
6
unweighted Hamming 1-center DNA sequences
7
output
8
filtered output
9
selected output

Fig. 1. Procedure for designing target DNA enrichment probes (= baits) as implemented in BaitFisher. MSAs that directly serve as templates for bait design are used as input. Alternatively, the user can provide MSAs of cDNA that are afterwards split into individual exons/CDS region based on the gene models of a user-defined reference species (R). To enable the latter approach, the MSAs must include the cDNA of a reference species with a sequenced and annotated genome. The cDNA sequence plus the genome assembly and official gene set of the reference species (all user-provided) allow identifying exon boundaries in the MSAs and split of the latter according to these boundaries (step 1). The cDNA sequence of the reference species can be optionally discarded (step 2). BaitFisher next identifies with a sliding window of bait-length size (5 bp in the illustrated example) start positions in the MSA that are deemed suitable for bait design (+) (step 3). Suitable start positions are those with windows in which user-defined taxonomic groups are represented by at least one gap- and ambiguity code-free nucleotide sequence. In the example shown, after having removed all sequences with gaps and/or ambiguity codes from a given window, it must still include the nucleotide sequences of taxon 3 and taxon 7 and the nucleotide sequence of taxon 5 or taxon 6. The nucleotide sequences of all remaining taxa (1, 2, and 4) are considered during bait design if they are gap- and ambiguity code-free, but their presence is not mandatory. After all windows have been analyzed (step 3'), BaitFisher filters positively evaluated start positions for those compatible with a user-defined tiling design (step 4). In the example given, the tiling design requires three consecutive baits of 5-bp length with a new bait every 10 bp. From the gap- and ambiguity code-free nucleotide sequences of each retained positively evaluated window, BaitFisher clusters sequences according to a user-defined degree of nucleotide sequence similarity (step 5). It then calculates the 1-center nucleotide sequence (= bait) of each cluster (step 6). Finally, information about all inferred baits is summarized (step 7). The inferred baits can be optionally searched with the BaitFilter helper program against the genome assembly of a user-selected reference species to identify and remove potentially non target-binding baits (step 8). BaitFilter program furthermore allows selecting one optimal set of tiled baits per exon/CDS region or gene, based on a user-selected optimality criterion (step 9).

Wunsch algorithm (Needleman and Wunsch 1970). The region in the multiple sequence alignment (MSA), to which the genomic feature sequence is aligned, is subsequently extracted and stored in a separate file. BaitFisher allows the user to decide whether the nucleotide sequence of the reference species is subsequently also considered for designing baits (fig. 1, step 2).

BaitFisher identifies regions in MSAs of contiguous potential target DNA that are deemed suitable for bait design (fig. 1, steps 3 and 4). Specifically, for every MSA window of the user-defined bait length, BaitFisher first discards sequences with gaps and/or IUPAC ambiguity codes. BaitFisher then evaluates whether all user-defined taxonomic groups (fig. 1, step 3) are represented in a given MSA window. Only if they are present does BaitFisher mark this window as suitable for bait design. Finally, the software filters for those start positions that are compatible with the user-defined tiling design (e.g., three baits with a new bait every 10 bp; fig. 1, step 4).

In order to minimize the number of baits required to efficiently enrich all nucleotide sequences that are part of the MSA in a given window, BaitFisher infers baits in two steps: The software first calculates the uncorrected (= Hamming; *p*) distances between all sequences in a given window of bait-length size and clusters those sequences that differ by less than a user-defined maximum distance (e.g., 0.06) from each other (fig. 1, step 5). BaitFisher then infers from the nucleotide sequences of each cluster an artificial 1-center sequence (fig. 1, step 6). This 1-center sequence represents an artificial sequence that exhibits the smallest maximum distance to all nucleotide sequences in this cluster (Li et al. 2002). In case of multiple equivalent solutions, the software randomly picks a sequence from the pool of equivalent 1-center sequences. A detailed description of the 1-center problem and the algorithm used to compute the 1-center sequence is given in supplementary file S1, Supplementary Material online.

After having calculated all 1-center nucleotide sequences, BaitFisher provides the user a tab-delimited text file that contains the essential information about each possible bait region (fig. 1, step 7). A bait region is a sequence segment in the MSA that 1) hosts a complete tiling design and 2) fully contains the nucleotide sequences of all user-defined mandatory taxonomic groups (see BaitFisher manual for more information). The file lists for each possible start position in a given target DNA region (i.e., user-provided MSA, such as of a gene, or excised feature) an optimal set of baits compatible with the user-defined tiling design. A procedure to automatically select an optimal start position in a given target DNA region is described below.

If one or multiple baits of a given bait region exhibit a high sequence similarity to two or more regions in the user-provided reference genome assembly, it is likely that the inferred baits would also enrich nontarget DNA. Hence, the user might want to exclude such bait regions in favor of others that are more target specific. We therefore developed a helper program called BaitFilter, which is part of the BaitFisher software package. BaitFilter allows the user to identify and discard baits that are likely to bind to multiple regions, as judged from the baits' nucleotide sequence similarity to regions in a user-provided reference genome assembly (fig. 1, step 8). The sequence similarity search of baits against the reference genome assembly is accomplished using BLAST+ (Camacho et al. 2008). If a given bait shows no significant similarity to any region in the reference genome assembly, the bait and the corresponding bait region are retained. We hereby acknowledge the fact that the nucleotide sequence of the reference genome might not have been part of the sequence cluster from which the bait was inferred from. Using a similar approach, BaitFilter allows the user also to identify and remove baits that would likely not properly bind to the target DNA (e.g., because gene models used to splice MSAs consisting of cDNA were not correct), as judged from searching baits against the assembled genome of a reference species. Finally, BaitFilter enables selecting an optimal start position for a bait set in a given target DNA region (fig. 1, step 9). The user can apply one of the two optimality criteria for selecting the optimal start position: 1) Minimizing the number of baits required to enrich a given locus, which usually means placing the bait region (i.e, the genomic region spanned by all baits tiled across this region) in the most conserved segment of a given locus; or 2) maximizing the number of sequences that were considered when inferring baits, which results in selecting the bait region, in which the smallest number of nucleotide sequences is missing or contain gaps or ambiguous nucleotides.

The BaitFisher software package is written in the programing language C++. It is open source and freely available at https://github.com/cmayer/BaitFisher-package.git.

Empirical Evaluation of the Bait Enrichment Capabilities

To assess the capability of baits designed by BaitFisher for enriching target DNA, we inferred a set of 57,650 baits for studying target DNA of apoid wasps (Hymenoptera: Crabonidae, Sphecidae) with the SureSelect Target Enrichment System offered by Agilent Technologies, Inc. (Santa Clara, CA). Specifically, we exploited transcriptomes of adults of 24 apoid wasp species sequenced in the international 1KITE project, and listed in supplementary file S2, Supplementary Material online. By querying the OrthoDB 5 database (Waterhouse et al. 2010), we identified a set of 5,561 genes that likely are single copy in apoid wasps, judged from the genes' presence in a representative set of six Hymenoptera with well-sequenced genomes (i.e., *Acromyrmex echinatior*, *Apis mellifera*, *Camponotus floridanus*, *Harpegnathos saltator*, *Linepithema humile*, *Nasonia vitripennis*; Weinstock et al. 2006; Bonasio et al. 2010; Werren et al. 2010; Nygaard et al. 2011; Smith et al. 2011) (supplementary file S2, Supplementary Material online). We next searched for transcripts that are orthologous to these 5,561 genes in the apoid wasp transcriptomes. For this purpose, we made use of HaMStRad (Misof et al. 2014), a modified version of HaMStR 8 (Ebersberger et al. 2009), following the procedure described by Misof et al. (2014). We used the software Orthograph 0.5.6, which became more recently available (Petersen et al. 2015), to later assign assembled contigs from enriched and sequenced next-generation DNA sequencing libraries to target loci. Orthograph and HaMStRad both rely on the best reciprocal genome/transcriptome-wide hit (BRH) criterion to infer gene-transcript orthology. We only considered transcripts, for which the BRH criterion was fulfilled for each of the six (see above) reference taxa in the reciprocal searches. We used the amino acid sequence output to align orthologous transcripts on the translational level with MAFFT 7.017 (L-INS-i iterative refinement method; Katoh and Standley 2013) and inferred the corresponding nucleotide sequence alignment with the program Pal2Nal (Suyama et al. 2006) using a version modified as described by Misof et al. (2014).

We split the aligned cDNA sequences of apoid wasp into individual CDS regions using custom scripts which, in optimized form, are now integrated in BaitFisher. These scripts made use of the genome assembly and official gene set version 1.2 of the jewel wasp *N. vitripennis* (Werren et al. 2010) available from the Hymenoptera Genome Database (Muñoz-Torres et al. 2011). We specified a bait length of 120 bp and a tiling design of seven baits spanning a 240-bp window with a new bait every 20 bp. Furthermore, each bait region had to contain the cDNA sequence of at least one representative from each sampled taxonomic subfamily and tribe.

After all possible sets of baits had been inferred with the aid of BaitFisher and the above specified search parameters, we evaluated each bait for its potential to bind to nontarget regions with the BaitFilter program. For the present data set, we used the genomes of *Ap. mellifera* (assembly 4.0; Weinstock et al. 2006), *H. saltator* (assembly 3.3; Bonasio et al. 2010), and *N. vitripennis* (assembly 1.0; Werren et al. 2010) as references (supplementary file S2, Supplementary Material online). We discarded all bait regions that contained a bait that showed a significant match to ≥ 2 different loci in any of the reference genomes. To be more precise, the first BLASTN hit had to have an *E* value $<10^{-8}$ and the second BLASTN hit had to have an *E* value $<10^{-5}$ for the bait region to be considered to bind unspecifically. Finally, we removed baits of 131 CDS regions to lower the total number of baits to 57,650, the maximum number of baits to be included with the SureSelect Target Enrichment System at the time we ordered (July 31, 2013).

Results

Inference of Baits for Studying Target Genes in Apoid Wasps

We found orthologous transcripts to 5,555 selected single-copy target genes in 24 apoid wasp transcript libraries (with 2,767–4,406, average 4,033, genes per species). However, we discarded 256 of the resulting MSAs due to a missing *N. vitripennis* nucleotide sequence, which resulted in 5,299 target genes. Using the gene models of the *N. vitripennis* official gene set 1.2 as a basis for identifying CDS regions in the 5,299 MSAs suggested 10,854 CDS regions as suitable for bait design. Requiring the presence of at least one representative species per taxonomic subfamily and tribe (17 taxonomic groups in total) in each MSA resulted in 631 CDS regions in 424 single-copy genes as promising for bait design. When comparing the orthologous nucleotide sequences of the species included in the 631 CDS region MSAs, we found the maximum sequence distances to range between 6.7% and 68% when analyzing all possible 120-bp-long nucleotide sequence windows (i.e., the length of baits that we intended to design). Specifying a sequence similarity threshold of 6% for clustering the sequences of each given 120-bp-long sequence window, we inferred 12,177,558 promising baits likely to capture a total of 631 CDS regions. Searching the 12,177,558 bait sequences, referring to 79,174 bait regions against the genomes of the *Ap. mellifera*, *N. vitripennis*, and *H. saltator* (supplementary file S2, Supplementary Material online) indicated competing nontarget binding sites for baits in 23,910 bait regions. We deemed the remaining 55,264 bait regions suitable for capturing 509 CDS regions in 356 genes. Using BaitFilter to choose for each CDS region the bait region that requires the smallest number of baits resulted in 77,119 baits, which are required to enrich the 509 CDS regions under the requested tiling design and the cluster threshold parameter. However, given the maximum number of 57,650 baits that the SureSelect Target Enrichment System by Agilent Technologies, Inc. allowed to be designed on a single glass slide, we removed baits for enriching 131 CDS regions, thereby losing the ability to enrich 74 target genes. At this point, we were able to order 57,650 nonredundant baits to empirically test their capability to capture 378 CDS regions in 282 genes in various in- and outgroup species (supplementary file S3, Supplementary Material online). Due to later optimization of the BaitFisher code for extracting individual CDS regions in the MSAs (see Empirical Evaluation of the Bait Enrichment Capabilities), a small fraction (1.6%) of the ordered baits is not suggested by the current version of the software any more, because some baits do not full-length map to the target gDNA. Our subsequent empirical testing of the ordered baits (see Target DNA Enrichment Success) thus provides conservative estimates.

Computational Performance of BaitFisher and BaitFilter

We evaluated the computational performance of BaitFisher and BaitFilter on a 2.66 GHz Linux desktop computer with 36 GB of RAM using the above design of baits for enriching single-copy genes (see Inference of Baits for Studying Target Genes in Apoid Wasps). For this purpose, BaitFisher was provided the 5,299 MSAs (consisting on average of 19 sequences) specified in Inference of Baits for Studying Target Genes in Apoid Wasps and was run with the parameters outlined in Empirical Evaluation of the Bait Enrichment Capabilities. When applying various distance threshold values (0.06–0.30) for clustering of the nucleotide sequences, the total number of baits required to enrich the target loci significantly decreased when increasing the nucleotide sequence clustering threshold (supplementary file S1, Supplementary Material online). At the same time, the run-time for computing baits only slightly increased when increasing the nucleotide sequence clustering threshold.

To assess the impact of the number of nucleotide sequences within a given MSA on BaitFisher's memory consumption and computation time, we analyzed MSAs with an arbitrary set of 1) 500, 2) 1,500, and 3) 2,500 nonredundant and publicly available nucleotide sequences of the barcoding gene cytochrome c oxidase I (COI) using the same hardware as specified above. The required computation times were 6.2, 28.6h, and 289 h, respectively. The observed increase in run-time is roughly in line with the expected run time for a hierarchical clustering algorithm, which scales with the order of $O(N^2)$, where N is the number of nucleotide sequences. We also found the memory consumption to scale roughly with

O(N^2): 18, 180, and 396 MB, respectively. The software implementation limit for the maximum number of sequences in a MSA for BaitFisher to be able to handle is 32,767. The practical limit for the maximum number of sequences in a MSA is determined by the available computation time: Analyzing a MSA with 3,000 nucleotide sequences required BaitFisher about 16 days on the described hardware.

Applying BaitFilter on the output files from the apoid wasp data set with different cluster threshold values (see above and supplementary file S1, Supplementary Material online) showed that even for large output files of up to 1 GB in size, BaitFilter extracts the requested information in less than 2 min (supplementary file S1, Supplementary Material online). Users are thus unlikely to experience any practical limitations when filtering BaitFisher output files.

When BaitFilter was invoked for removing bait sets that contain baits with nontarget binding sites in a reference genome, the filtering took several hours, as BaitFisher relies on the BLAST+ software for searching baits against the reference genome. The time required for this step is thus primarily determined by the number of baits and the size of the reference genome through which it searches (supplementary file S1, Supplementary Material online).

Target DNA Enrichment Success

We collected between 1.38 and 2.25 M (deeply sequenced Illumina DNA sequencing test libraries; four species in total) and between 0.35 and 0.97 M (shallowly sequenced Illumina DNA sequencing libraries; nine species) quality-trimmed raw reads per species. These reads assembled into 4,508–19,100 (deeply sequenced libraries) and 1,884–13,035 (shallowly sequenced libraries) contigs with lengths between 414 and 803 bp (table 1).

When searching the 13 obtained assemblies with Orthograph for the 378 target CDS regions in 282 target genes, we identified 203–303 target CDS regions (average: 263; median: 275) and 26–279 (average: 253; median: 262) target genes (table 1). The fewest target CDS regions and target genes were identified in the *Crabro peltarius* sample, which had been stored in Vitzthum's solution for 21 years. We found no striking difference in the target DNA recovery between samples of ingroup species preserved in pure ethanol (274 and 262 target genes; the first value found for a deeply sequenced sample, the second for a shallowly sequenced sample) and those preserved in approximately 70% ethanol (251–276 target genes; values refer to both deeply and shallowly sequenced samples) (table 1).

The base-coverage depth of contigs that referred to target genes, C_t, was on average 38–94 in species with deeply sequenced libraries and 3–51 in species with shallowly sequenced libraries (table 1). The base-coverage depth of contigs that contained nontarget DNA, C_n, was on average 0.15 of that of contigs with target DNA, suggesting a relative enrichment coefficient of 6.8 (table 1). When comparing C_t with the base-coverage depth that one would expect to find in the assembled contigs if DNA fragments of the genome of the investigated species were randomly sequenced (C_g), we found C_t to be on average 71.1 times higher than C_g (table 1).

Table 1. Sequencing Depth and Target Gene Recovery Statistics Obtained from Tests of BaitFisher-Inferred Baits on Genomic DNA of Various Apoid Wasps and Selected Outgroup Species.

Species	Number of Clean Raw Reads	Number of Assembled Contigs	Number of Assembled and Cleaned Contigs	Number of Sequenced Target Genes	Length of Contigs Referring to Target DNA (bp) (min–median–max)	C_t	C_n	$C_t \times C_n^{-1}$	$C_t \times C_g^{-1}$	C_g	Sample Tissue Preservation History
Dynatus burmeisteri	1,379,306	7,407	7,261	270	328–588–1,717	54	7	7.7		NA	7 years in 70% ethanol
Isodontia mexicana[a]	2,084,280	17,567	17,374	274	323–792–3,202	94	6	15.7		NA	12 years in 70% ethanol
Stangeella cyaniventris	1,723,712	4,508	4,495	251	308–553–2,674	38	18	2.1		NA	7 years in 70% ethanol
Stictia heros	2,248,824	19,101	19,091	279	180–584–1,719	73	5	14.6		NA	<1 year in 96% ethanol
Ampulex compressa	596,982	7,926	7,731	253	329–821–2,659	31	4	7.8	64.8	0.08	<1 year in 96% ethanol
Apis mellifera	488,786	2,190	2,120	188	319–441–1,070	51	6	8.5	98.1	0.22	<1 year in 96% ethanol
Crabro peltarius[a]	707,220	7,264	7,020	26	334–472–1,092	3	2	1.5		NA	21 years in Vitzthum's solution
Clypeadon sculleni	833,770	13,035	12,876	271	333–803–1,967	29	19	1.5		NA	2 years in 96% ethanol
Dinetus pictus[a]	350,062	1,935	1,702	262	319–445–101	28	5	5.6		NA	9 years in 70% ethanol
Eremnophila melanaria	381,072	1,884	1,658	276	327–614–1,686	23	5	4.6		NA	7 years in 70% ethanol
Harpegnathos saltator	548,090	5,154	4,904	236	321–698–2,313	24	4	6.0	57.8	0.07	<1 year in 96% ethanol
Nasonia vitripennis	964,536	9,933	9,742	203	333–581–1,971	41	6	6.8	63.6	0.13	<1 year in 96% ethanol
Sphex funerarius[a]	394,392	3,495	3.203	278	383–757–2,456	28	5	5.6		NA	10 years in 70% ethanol

NOTE.—The libraries of the first four species were deeply sequenced, those of the following nine species were shallowly sequenced. C_t, average base-coverage depth of contigs referring to target genes; C_n, average base-coverage depth of contigs referring to nontarget genes; C_g, average base-coverage depth of contigs expected if DNA fragments of the genome of the investigated species were randomly sequenced (only given for species whose genome size was known); NA, not applicable.
[a]Target DNA of the species was known and exploited during bait design.

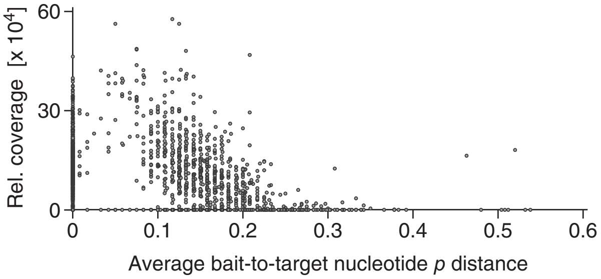

Fig. 2. Correlation between average Hamming (p) distances of baits designed by BaitFisher for enriching a given locus to the respective locus' actual nucleotide sequence and the relative base-coverage depth (normalized by dividing the base-coverage depth by the total amount of sequenced nucleotides) by which the locus was sequenced after applying the designed baits for enriching the target DNA. Shown are the results from analyzing *Apis mellifera*, *Dinetus pictus*, *Harpegnathos saltator*, *Isodontia mexicana*, *Nasonia vitripennis*, and *Sphex funerarius*.

To assess the impact of the bait-to-target sequence similarity on the enrichment efficiency, we plotted the base-coverage depth of contigs referring to target genes, C_t, normalized by dividing it by the total number of sequenced nucleotides, against the average bait-to-target sequence similarity in species with known target DNA (fig. 2). We found a strong negative correlation between bait-to-target sequence similarity and the relative base-coverage depth of the target DNA (fig. 2).

Discussion

The ability to selectively study the nucleotide sequences of hundreds or thousands of loci of interest in the genomes of different species can be considered as one of the most significant steps forward in targeted genomic data acquisition, relevant to many research disciplines (Bejerano et al. 2004; Hodges et al. 2007; Ng et al. 2009; Crawford et al. 2012; Brandley et al. 2015; Jones and Good 2016). Although the inference of oligonucleotides that serve as baits for enriching target DNA in a single species, whose genome has been sequenced, is well established, the design of baits to enrich target DNA in a wider range of species still remains a challenge. Researchers have applied different strategies to capture target loci across species. Li et al. (2013), for example, designed baits to capture target genes by using baits designed from analyzing the genome of a single species. By tuning the wet laboratory procedures (e.g., hybridization temperature profile; see below), the authors were able to extend the reach of the method considerably despite the potentially substantial bait-to-target distances associated with the applied bait design strategy. This approach is reasonable if no additional nucleotide sequence information of taxonomic ingroup species is available. Other authors considered nucleotide sequence information from in- and outgroup species in search for (ultra) conserved nucleotide sequence sections that can serve as anchors to capture and study (typically more variable) flanking nucleotide sequences across species (Crawford 2012; Faircloth et al. 2012; Lemmon et al. 2012; McCormack, Harvey, et al. 2013; McCormack, Hird, et al. 2013). Given the high conservation of the target nucleotide sequence, this approach also allows using the nucleotide sequence of a single species for bait design. Unfortunately, the approach cannot easily be used to study loci that are spatially distant from conserved sequence sections.

To capture and study variable exonic sequences in an entire class of marine invertebrates (Ophiuroidea), Hugall et al. (2015) recently suggested and applied an intriguing approach by exploiting transcriptomes. They inferred a phylogenetic tree from transcriptomes, which had been sampled in species across the class Ophiuroidea. The tree was then used to infer the ancestral nucleotide sequences of single-copy target genes in subordinated clades within Ophiuroidea. Sections of the inferred ancestral nucleotide sequences subsequently served as baits to capture the corresponding loci in other species of these clades. The number and size of the clades, from which one ancestral nucleotide sequence per locus was inferred, was chosen in a manner that the majority (>80%) of the resulting baits for capturing the target loci in species of a given clade did not differ in more than 12% from the known transcript sequences of species in this clade. The clustering of species and the inference of ancestral nucleotide sequences served two purposes: 1) To reduce redundancy in the taxonomic sampling by clustering species that share a high nucleotide sequence similarity and 2) to traceably select a single representative nucleotide sequence per locus and clade from which baits are designed.

Our approach to optimize the number and the efficacy of baits required to capture target loci across species relies on a strategy comparable with the one proposed by Hugall et al. (2015): It exploits user-provided nucleotide sequence information of target loci in different species for designing baits and automatically reduces (taxonomic) redundancy by clustering nucleotide sequences that differ in less than a user-defined threshold value from each other. In contrast to the approach applied by Hugall et al. (2015), our approach performs the clustering of nucleotide sequences, from which baits are inferred, for each nucleotide sequence window of bait-length size separately. By not clustering the reference species' nucleotide sequences by the species' phylogenetic relationships, but by clustering them according to the sequences' distances separately in each sequence section of bait-length size, we are able to reduce redundancy and bait-to-target distances even further (as compared with the approach applied by Hugall et al. 2015). We subsequently infer one artificial bait sequence per sequence window of bait-length size and group of clustered nucleotide sequences to reduce the bait-to-target distance across species, while Hugall et al. (2015) inferred an ancestral sequence for this purpose. The 1-center sequence guaranties that the bait-to-target sequence distance (as judged from the baits' sequence distance to the corresponding clustered nucleotide sequences; see supplemental file S1, Supplementary Material online) is indeed minimized. An ancestral sequence, a randomly picked ingroup sequence, or a consensus sequence, in contrast, do not guarantee to minimize the maximum bait-to-target sequence distance. For example, the nucleotide sequence of a locus can be highly derived in some of the sampled species in a clade. Using the presumed ancestral sequence of this locus as bait would thus possibly result in the bait's nucleotide sequence being more similar to that of species with more plesiomorphic sequences than to those of species with more derived sequences.

The availability of the nucleotide sequences of a representative set of ingroup species is an important prerequisite when designing baits that are meant to effectively enrich target DNA across species of the ingroup. Our empirical evaluation of 120-bp-long baits to enrich target DNA with known nucleotide sequence similarity using the molecular procedure outlined in Taxon Sampling and Molecular Procedures and in which we applied constant hybridization and posthybridization washing temperatures suggests that the bait-to-target DNA sequence distance should not exceed 15–20% for the enrichment to be efficient (fig. 2). Our results are in line with those reported by Bragg et al. (2015), Hawkins et al. (2015), and Paijmans et al. (2016). Li et al. (2013) reported the capture of nucleotide sequences exhibiting a bait-to-target distance of up to 39%. The libraries that we enriched contained target loci that differed in up to 52% from the corresponding baits, but there is a clear negative correlation between enrichment efficacy and bait-to-target nucleotide sequence dissimilarity (fig. 2). Thus, although it may appear that our experiments resulted in successful enrichment of target loci differing in up to 52% from the nucleotide sequence of the applied capture baits, we interpret these distant target nucleotide sequences as outliers (fig. 2). This is because any enriched library still contains nontarget nucleotide sequences with both low and high read coverage. Based on the currently available data and the applied wet laboratory protocol (Taxon Sampling and Molecular Procedures; see also discussion further below), we suggest using a cluster threshold of not more than 30% when designing baits with a length of 120 bp. Although this value may appear at first glance conservative, given that the nucleotide sequence of a bait would generally not differ in more than 15% from any nucleotide sequence in a given cluster, this value acknowledges that the sequences of some target species may have historically undergone accelerated evolutionary change and that a higher enrichment efficacy requires less deep sequencing of the enriched library. However, future experiments should investigate the relationship between bait-to-target DNA sequence similarity and enrichment efficacy as a function of the length of the baits. We decided to design baits with a length of 120 bp due to promising results in studies that employed baits of this length for in-solution target capture (Faircloth et al. 2012; Lemmon et. al. 2012); however, other investigators successfully applied baits with a length of 60–90 bp on capture microarrays (Hodges et al. 2009; Mamanova et al. 2010; Hancock-Hanser et al. 2013).

Depending on the research question, target locus specificity of baits can be important. BaitFilter allows evaluating the probability of baits to bind to nontarget DNA by searching all inferred baits against a user-provided genome assembly. Search of bait sequences against an ingroup genome assembly may also prove valuable for evaluating the enrichment success of target loci. BaitFilter therefore also allows the user to assess whether at least one bait, of a given stack of baits (see BaitFisher manual for details) that is meant to bind at a specific position of a target locus across species, indeed exhibits a high nucleotide sequence similarity to a unique locus in a user-provided reference genome assembly. This feature is useful when baits are designed for enriching specific genomic features, such as individual CDSs. If the identification of these genomic features in the nucleotide sequences of the ingroup species relied on gene models in an outgroup species, chances are higher that these features are not applicable to ingroup species. We use sequence nucleotide similarity as a proxy to assess the propensity of baits to bind to target and off-target nucleotide sequence stretches, but acknowledge that the hybridization of oligonucleotides to DNA is determined by thermodynamic properties, such as the number of hydrogen bonds. Consideration of these properties when searching tens or hundreds of thousands of baits against a reference genome of up to several giga base pairs in size is computationally challenging and would result in a reduction of BaitFilter's computational performance. Given the tight correlation between DNA hybridization energy and nucleotide sequence similarity (Wallace et al. 1979) and the fact that baits designed by BaitFisher are meant to enrich loci in species, whose nucleotide sequence is expected to be different from that of the reference species, consideration of thermodynamic properties is expected to result only in a marginal improvement of the predictive power. We therefore deliberately refrained from considering hybridization properties in the current version of the BaitFilter software. However, a promising approach to cope with this shortcoming could be combining nucleotide sequence similarity search-guided identification of reference genome candidate regions, to which baits could potentially bind, and an in-depth analysis of the thermal stability of bait-target DNA duplexes in these candidate regions. BaitFisher currently does not consider the baits' propensities for folding and dimerization either. Although DNA binding and folding energy calculations are often considered by polymerase chain reaction (PCR) oligonucleotide primer design software (Mann et al. 2009), the large size (in bp) and the disproportionately large number of disparate oligonucleotides typically employed in target DNA enrichment exacerbate explicit contemplation of these properties in the latter context.

Our empirical evaluation of baits inferred with the aid of BaitFisher and BaitFilter on DNA of apoid wasps showed that the baits worked very well. In fact, the overall enrichment coefficient (C_t/C_g) achieved by using the baits proved to be in the magnitude of 58- to 98-fold when comparing the base-coverage depths of target loci (C_t) with the base-coverage depths expected if no enrichment had taken place (C_g) (table 1). The recovery rate of the target DNA from samples that had been long-term stored in approximately 70% ethanol was also very high (table 1) and opens a wide range of new areas for the application of target DNA enrichment. We see, for example, a particular profit of target DNA enrichment in the field of museomics and biodiversity monitoring. In the former, investigators seek to recover the DNA from unique and often old samples stored in museum collections (Guschanski et al. 2013). The classical procedure of PCR-amplifying target loci and subsequent sequencing of the obtained amplicons using Sanger sequencing technology often cannot be applied, because the target DNA is too degraded (Hofreiter et al. 2015; Paijmans et al. 2016). Our recovery rate of target DNA from samples that had been stored for up to 12 years in approximately 70% ethanol, which resulted in a substantial degradation of the samples' DNA, was very high and is extremely promising (table 1). These results have been obtained by applying the molecular procedures outlined in Taxon Sampling and Molecular Procedures. The procedures involved constant hybridization and posthybridization temperature profiles and only a single round of target locus capture. Li et al. (2013) and Paijmans et al. (2016) assessed alternative temperature profiles in the hybridization step and in posthybridization steps and suggest modifications of the wet laboratory protocols that allow extending the reach of the method. Li et al. (2013) also suggested conducting a second round of target locus capture to further increase the enrichment success. We refer the reader to these two excellent articles when planning their wet laboratory procedures.

Materials and Methods

Taxon Sampling and Molecular Procedures

DNA Extraction and Library Preparation for Next-Generation DNA Sequencing

We tested the enrichment capacity of the 57,650 inferred baits on DNA extracts of nine ingroup species (i.e., apoid wasps, excluding cockroach wasp) and four outgroup taxa (including cockroach wasp; supplementary file S4, Supplementary Material online). Specifically, we selected four species of crabronid wasps and five species of sphecid wasps. The cDNA sequences of four of these species (i.e., *C. peltarius*, *Dinetus pictus*, *Isodontia mexicana*, *Sphex funerarius*) had also been used for bait design. Hence, these four species served as a positive control (i.e., the nucleotide sequences of a specific fraction of the designed baits were known to differ in less than 6% from that of the target genomic DNA of the four species). As outgroup taxa, we chose the cockroach wasp *Ampulex compressa* (Ampulicidae), the honeybee (*Ap. mellifera*), an ant (*H. saltator*), and a parasitoid wasp (*N. vitripennis*). The genomes of the latter three are sequenced (Weinstock et al. 2006; Bonasio et al. 2010; Werren et al. 2010) and enabled us to estimate the degree of target DNA enrichment in more distantly related taxa (as compared with ingroup species). The enriched DNA of these three taxa plus that of the four ingroup species serving as controls were also considered when exploring the correlation between bait-to-target DNA distance and target locus base-coverage depth. The DNA quality differed across the analyzed samples: Although we extracted some DNA from tissues that were short-term stored in absolute ethanol (i.e., the four outgroup species plus *Clypeadon sculleni* and *Stictia heros*), other DNAs were extracted from tissues that had been stored over much longer time (9–21 years) in either approximately 70% ethanol (i.e., *Di. pictus*, *Dynatus burmeisteri*, *Eremnophila melanaria*, *I. mexicana*, *Sph. funerarius*, *Stangeella cyaniventris*) or in Vitzthum's solution (80 g of 75% ethanol, 16 g glycerol, 4 g acetic acid glacial; Öttingen 1938) (i.e., *C. peltarius*).

The genomic DNA of all investigated species was extracted either with the Qiagen DNeasy Blood & Tissue Kit (Qiagen GmbH, Hilden, Germany) or by applying the CTAB DNA extraction protocol by Rogers and Bendich (1985) in combination with a DNA purification step using AMPure XP beads (Beckman Coulter GmbH, Krefeld, Germany). All extracted DNAs were dissolved in 100 µl of nuclease-free water and next-generation DNA sequencing libraries were prepared from the extracted DNA following the protocol given in supplementary file S5, Supplementary Material online.

Target DNA Enrichment and Illumina MiSeq Paired-End DNA Sequencing

We followed the SureSelect Target Enrichment System Kit protocol by Agilent Technologies, Inc. for Illumina Multiplexed Sequencing, published in 2013 (pp. 60–70), to capture target DNA fragments from the amplified NGS libraries using a pool of 57,650 baits that was designed by BaitFisher and synthesized by Agilent Technologies, Inc. when ordering the SureSelect Target Enrichment System Kit. Hybridization of the baits to the target DNA was allowed for 18 h at 65 °C in a GeneAmp PCR System 2700 thermocycler (Applied Biosystems, Inc., Waltham, USA). Posthybridization PCR amplification of the target-enriched libraries was also conducted with a GeneAmp PCR System 2700 thermocycler using the PCR Primer Cocktail and PCR Mastermix as described in supplementary file S5, Supplementary Material online, for NGS library PCR amplification. No additional indexing was done because we had already ligated indices to the DNA fragments of the NGS libraries during library preparation (supplementary file S5, Supplementary Material online). We applied the PCR protocol (consisting of 12 cycles) recommended by Agilent Technologies, Inc. for capturing >1.5 Mbp of DNA in all posthybridization PCRs. All amplified enriched libraries were purified with Agencourt AMPure XP beads, and the quality and quantity of the purified DNA fragments assessed with a Fragment Analyzer (Advanced Analytical Technologies GmbH, Heidelberg, Germany) and a Qubit 2.0 Fluorometer (Thermo Fisher Scientific Inc., Waltham, USA). In those cases in which the total yield of DNA in the range 200–800 bp proved to be too low (i.e., DNA concentration <1.5 ng/µl) for Illumina paired-end sequencing, we repeated the posthybridization PCR amplification as described above. Illumina MiSeq paired-end DNA sequencing of the enriched next-generation DNA sequencing libraries followed the protocol given in supplementary file S5, Supplementary Material online. In the first four samples sequenced (i.e., *Dy. burmeisteri*, *I. mexicana*, *Sta. cyaniventris*, *Sti. heros*), we collected 1.4–2.3 Mbp per species. Given the high base-pair coverage depth in the target genes achieved from assembling these data, we subsequently lowered the amount of raw data collected per species to 0. 35–0.96 Mbp.

De Novo Assembly of Reads

The quality of all obtained NGS raw reads was checked with FastQC 0.11.2 (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/). Adaptor and poor-quality regions were clipped with Trimmomatic 0.32 (Bolger et al. 2014; seed mismatches: 2, palindrome clip threshold: 30, simple clip threshold: 10, minimum quality required to keep a leading base: 3, minimum quality required to keep a trailing base: 3, sliding window size: 4, required average quality in window: 15, minimum length of reads to be kept: 25). The filtered paired-end reads were then assembled with the IDBA-UD de novo assembler 1.1.1 (Peng et al. 2012). The assembler is optimized for assembling contigs sequenced to a very uneven base-coverage depth. We recompiled IDBA-UD after applying slight changes in the source code, as suggested by the software developers, so that the assembler was able to handle reads of up to 320 bp in length. The iterative assembly process started with a *k*-mer size of 20 bp. The *k*-mer size was increased in steps of 5 bp during each iteration, until a *k*-mer size of 120 bp was reached.

Identification and Removal of Possible Contaminant Contigs

We discovered in context of the 1KITE project that single index-tagged libraries pooled on the same Illumina lane often exhibit a small percentage of cross contamination. To cope with this problem in the present investigation, we searched the contigs of those reduced-representation libraries that were sequenced on the same Illumina lane against each other with the BLASTN search engine of BLAST+ 2.2.29 (Camacho et al. 2008). In those instances, in which we identified nucleotide sequences that shared over a length of $\geq$200 bp a similarity of $\geq$98% with each other, we proceeded as follows: 1) If the relative read-coverage depths of the two contigs in question differed more than 2-fold, we removed the contig with the lower relative read-coverage depth from the corresponding assembly; and 2) if the relative read-coverage depth of the two contigs in question were sequenced to roughly the same depth (less than 2-fold difference), we conservatively removed both of the contigs from the corresponding assembly. If multiple highly similar contigs were found (because we searched the contigs of all assemblies in question simultaneously against each other; see above), we retained only the contig with the best coverage, given that its coverage was more than 2-fold higher than the coverage of the second-best matching contig. We defined as “read-coverage depth” of a given contig, the number of reads (as provided in IDBA-UD output) of this contig divided by the total amount of nucleotides sequenced from the corresponding library.

Target DNA Recovery and Enrichment Efficiency

To assess the coverage of the enriched target regions, we used the software segemehl 0.1.7 (Hoffmann et al. 2009) for mapping all raw sequencing reads to the assembled and contamination-filtered contigs. The mapping results were exported in the SAM file format, which was then imported for further analysis in tablet 1.14.10.20 (Milne et al. 2013). Tablet allowed us to conveniently calculate the number of reads that mapped to each specific contig. Exploiting this information, we calculated the actual average base-coverage depth of those contigs that contain a 250-bp-long bait-binding sequence section using the formula $C_t = N_t \times L_t \times S_t{}^{-1}$, in which N_t is the number of reads that mapped to a given contig containing the target DNA, L_t is the length (250 bp) of the reads that mapped to the contig containing the target DNA, and S_t is the length (in bp) of the contig containing the target DNA. We analogously calculated the average base-coverage depth of contigs that do not contain target DNA (C_n). Finally, we compared C_t with C_n, and calculated the ratio $C_t \times C_n{}^{-1}$ as one measure of target DNA enrichment degree.

To further assess the extent to which target DNA was enriched, we calculated the average base-coverage depth, C_g, that one would expect the sequenced and assembled fragments of the genome of a given species to exhibit if no enrichment had taken place. C_g was calculated using the formula $C_g = N_g \times L_g \times S_g{}^{-1}$, in which N_g is the total number of sequenced reads, L_g is the length (250 bp) of all sequenced reads, and S_g is the genome size (haploid nuclear DNA content in bp; Lander and Waterman 1988) of the investigated species. C_g was consequently only calculated in species whose genome size is reliably known. This is the case for *Am. compressa* (374 Mbp; Niehuis O, unpublished data), *Ap. mellifera* (235 Mbp; Ardila-Garcia et al. 2010), *N. vitripennis* (312 Mbp; Beukeboom et al. 2007), and *H. saltator* (330 Mbp; Bonasio et al. 2010). Finally, we compared C_t with C_g, and calculated the ratio $C_t \times C_g{}^{-1}$ as second measure of target DNA enrichment degree.

To shed light on the relationship between bait-to-target nucleotide sequence similarity and the relative target DNA base-coverage depth, we calculated the lowest observed distance between baits of a given bait set and the target DNA per CDS region. This has been calculated for *Di. pictus*, *I. mexicana*, and *Sph. funerarius* (in-group species whose target DNA sequence was known to us from the transcript library DNA sequences) and for *Ap. mellifera*, *H. saltator*, and *N. vitripennis* (outgroup species whose genome is sequenced) using a custom C++ program. We did not consider values referring to *C. peltarius* in this analysis due to the low overall recovery of target genes from sequencing the library of this species. Furthermore, we only considered bait-to-target DNA distance values that are based on MSAs, in which the entire length of the target DNA was known for each bait. The relative base coverage of target loci was obtained by dividing the base coverage of each target locus by the total number of nucleotides sequenced per library.

Supplementary Material

Supplementary files S1–S5 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

The authors are grateful to Leo W. Beukeboom, Jürgen Gadau, Jürgen Liebig, Robin Moritz, Dieter Schulten, and Silke Stoll for providing samples of *Ampulex compressa*, *Apis mellifera*, *Harpegnathos saltator*, and *Nasonia vitripennis*. M.O. and O.N. further acknowledge the German Research Foundation (DFG) for supporting this research (OH81/9-1; NI 1387/1-1). C.B. is a "Ramon y Cajal" fellow supported by the Spanish Ministry of Science and Education (MEC) (RYC-2014-15615). The authors are furthermore grateful to the 1KITE initiative, which produced the data that made this research possible, as well as to Elise Laetz and two anonymous reviewers for help in further improving the manuscript. All analyzed species were collected before October 2014. The study is part of the doctoral thesis of M.S. M.M. holds a company providing DNA sequencing services. K.L. is an employee of this company.

References

Albert TJ, Molla MN, Muzny DM, Nazareth L, Wheeler D, Song X, Richmond TA, Middle CM, Rodesch MJ, Packard CJ, et al. 2007. Direct selection of human genomic loci by microarray hybridization. *Nat Methods*. 4:903–905.

Ardila-Garcia AM, Umphrey GJ, Gregory TR. 2010. An expansion of the genome size dataset for the insect order Hymenoptera, with a first test of parasitism and eusociality as possible constraints. *Insect Mol Biol*. 19:337–346.

Bashiardes S, Veile R, Helms C, Mardis ER, Bowcock AM, Lovett M. 2005. Direct genomic selection. *Nat Methods*. 2:63–69.

Bejerano G, Pheasant M, Makunin I, Stephen S, Kent WJ, Mattick JS, Haussler D. 2004. Ultraconserved elements in the human genome. *Science* 304:1321–1325.

Beukeboom LW, Kamping A, Louter M, Pijnacker LP, Katju V, Ferree PM, Werren JH. 2007. Haploid females in the parasitic wasp *Nasonia vitripennis*. *Science* 315:206.

Bi K, Vanderpool D, Singhal S, Linderoth T, Moritz C, Good JM. 2012. Transcriptome-based exon capture enables highly cost-effective comparative genomic data collection at moderate evolutionary scales. *BMC Genomics* 13:403.

Blumenstiel B, Cibulskis K, Fisher S, DeFelice M, Barry A, Fennell T, Abreu J, Minie B, Costello M, Young G, et al. 2010. Targeted exon sequencing by in-solution hybrid selection. *Curr Protoc Hum Genet*. Chapter 18, Unit 18.4.

Bodi K, Perera AG, Adams PS, Bintzler D, Dewar K, Grove DS, Kieleczawa J, Lyons RH, Neubert TA, Noll AC, et al. 2013. Comparison of commercially available target enrichment methods for next-generation sequencing. *J Biomol Tech*. 24:73–86.

Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 30:2114–2120.

Bonasio R, Zhang G, Ye C, Mutti NS, Fang X, Qin N, Donahue G, Yang P, Li Q, Li C, et al. 2010. Genomic comparison of the ants *Camponotus floridanus* and *Harpegnathos saltator*. *Science* 329:1068–1071.

Bragg JG, Potter S, Bi K, Moritz C. 2015. Exon capture phylogenomics: efficacy across scales of divergence. *Mol Ecol Res*. Advance Access published August 20, 2015, doi:10.1111/1755-0998.12449

Brandley MC, Bragg JG, Singhal S, Chapple DG, Jennings CK, Lemmon AR, Lemmon EM, Thompson MB, et al. 2015. Evaluating the performance of anchored hybrid enrichment at the tips of the tree of life: a phylogenetic analysis of Australian *Eugongylus* group scincid lizards. *BMC Evol Biol*. 15:62.

Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. 2008. BLAST+: architecture and applications. *BMC Bioinformatics* 10:421.

Cosart T, Beja-Pereira A, Chen S, Ng SB, Shendure J, Luikart G. 2011. Exome-wide DNA capture and next generation sequencing in domestic and wild species. *BMC Genomics* 12:347.

Crawford NG, Faircloth BC, McCormack JE, Brumfield RT, Winker K, Glenn TC. 2012. More than 1000 ultraconserved elements provide evidence that turtles are the sister group of archosaurs. *Biol Lett*. 8:783–786.

Ebersberger I, Strauss S, von Haeseler A. 2009. HaMStR: profile hidden markov model based search for ortholog ESTs. *BMC Evol Biol*. 9:157.

Faircloth BC, Branstetter MG, White ND, Brady SG. 2014. Target enrichment of ultraconserved elements from arthropods provides a genomic perspective on relationships among Hymenoptera. *Mol Ecol Res*. 15:489–501.

Faircloth BC, McCormack JE, Crawford NG, Harvey MG, Brumfield RT, Glenn TC. 2012. Ultraconserved elements anchor thousands of genetic markers spanning multiple evolutionary timescales. *Syst Biol*. 61:717–726.

Fisher S, Barry A, Abreu J, Minie B, Nolan J, Delorey TM, Young G, Fennell TJ, Allen A, Ambrogio L, et al. 2011. A scalable, fully automated process for construction of sequence-ready human exome targeted capture libraries. *Genome Biol*. 12:R1.

Guschanski K, Krause J, Sawyer S, Valente LM, Bailey S, Finstermeier K, Sabin R, Gilissen E, Sonet G, Nagy ZT, et al. 2013. Next-generation museomics disentangles one of the largest primate radiations. *Syst Biol*. 62:539–554.

Hancock-Hanser BL, Frey A, Leslie MS, Dutton PH, Archer FI, Morin PA. 2013. Targeted multiplex next-generation sequencing: advances in techniques of mitochondrial and nuclear DNA sequencing for population genomics. *Mol Ecol Res*. 13:254–268.

Hawkins MT, Hofman CA, Callicrate T, McDonough MM, Tsuchiya MT, Gutiérrez EE, Helgen KM, Maldonado JE. 2015. In-solution hybridization for mammalian mitogenome enrichment: pros, cons and challenges associated with multiplexing degraded DNA. *Mol Ecol Res*. Advance Access published August 24, 2015, doi:10.1111/1755-0998.12448

Hodges E, Rooks M, Xuan Z, Bhattacharjee A, Gordon DB, Brizuela L, McCombie WR, Hannon GJ. 2009. Hybrid selection of discrete genomic intervals on custom-designed microarrays for massively parallel sequencing. *Nat Protoc*. 4:960–974.

Hodges E, Xuan Z, Balija V, Kramer M, Molla MN, Smith SW, Middle CM, Rodesch MJ, Albert TJ, Hannon GJ, et al. 2007. Genome-wide *in situ* exon capture for selective resequencing. *Nat Genet*. 39: 1522–1527.

Hoffmann S, Otto C, Kurtz S, Sharma CM, Khaitovich P, Vogel J, Stadler PF, Hackermüller J. 2009. Fast mapping of short sequences with mismatches, insertions and deletions using index structures. *PLoS Comput Biol*. 5:e1000502.

Hofreiter M, Paijmans JL, Goodchild H, Speller CF, Barlow A, Fortes GG, Thomas JA, Ludwig A, Collins MJ. 2015. The future of ancient DNA: technical advances and conceptual shifts. *BioEssays* 37:284–293.

Hugall AF, O'Hara TD, Hunjan S, Nilsen R, Moussalli A. 2015. An exon-capture system for the entire class Ophiuroidea. *Mol Bio Evol*. 33:281–294.

Jones MR, Good JM. 2016. Targeted capture in evolutionary and ecological genomics. *Mol Ecol*. 25:185–202.

Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol*. 30:772–780.

Lander ES, Waterman MS. 1988. Genomic mapping by fingerprinting random clones: a mathematical analysis. *Genomics* 2:231–239.

Lemmon AR, Emme SA, Lemmon EM. 2012. Anchored hybrid enrichment for massively high-throughput phylogenomics. *Syst Biol*. 61:727–744.

Li C, Hofreiter M, Straube N, Corrigan S, Naylor GJ. 2013. Capturing protein-coding genes across highly divergent species. *Biotechniques* 54:321–326.

Li M, Ma B, Wang L. 2002. On the closest string and substring problems. *J ACM*. 49:157–171.

Liu S, Wang X, Xie L, Tan M, Li Z, Su X, Zhang H, Misof B, Kjer KM, Tang M, et al. 2016. Mitochondrial capture enriches mito-DNA 100 folds enabling PCR-free mitogenomics biodiversity analysis. *Mol Ecol Res*. 16:470–479.

Mann T, Humbert R, Dorschner M, Stamatoyannopoulos J, Noble WS. 2009. A thermodynamic approach to PCR primer design. *Nucleic Acids Res*. 37:e95.

Mamanova L, Coffey AJ, Scott CE, Kozarewa I, Turner EH, Kumar A, Howard E, Shendure J, Turner DJ. 2010. Target-enrichment strategies for next generation sequencing. *Nat Methods*. 7:111–118.

McCormack JE, Harvey MG, Faircloth BC, Crawford NG, Glenn TC, Brumfield RT. 2013. A phylogeny of birds based on over 1,500 loci collected by target enrichment and high-throughput sequencing. *PLoS One* 8:e54848.

McCormack JE, Hird SM, Zellmer AJ, Carstens BC, Brumfield RT. 2013. Applications of next-generation sequencing to phylogeography and phylogenetics. *Mol Phylogenet Evol*. 66:526–538.

Meyer M, Kircher M. 2010. Illumina sequencing library preparation for highly multiplexed target capture and sequencing. *Cold Spring Harb Protoc*. 2010:pdb.prot5448.

Milne I, Stephen G, Bayer M, Cock PJ, Pritchard L, Cardle L, Shaw PD, Marshall D. 2013. Using Tablet for visual exploration of second-generation sequencing data. *Brief Bioinform*. 14:193–202.

Misof B, Liu S, Meusemann K, Peters RS, Donath A, Mayer C, Frandsen PB, Ware J, Flouri T, Beutel RG, et al. 2014. Phylogenomics resolves the timing and pattern of insect evolution. *Science* 346:763–767.

Muñoz-Torres MC, Reese JT, Childers CP, Bennett AK, Sundaram JP, Childs KL, Anzola JM, Milshina N, Elsik CG. 2011. Hymenoptera Genome Database: integrated community resources for insect species of the order Hymenoptera. *Nucleic Acids Res*. 39:D658–D662.

Needleman SB, Wunsch CD. 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J Mol Biol*. 48:443–453.

Ng SB, Turner EH, Robertson PD, Flygare SD, Bigham AW, Lee C, Shaffer T, Wong M, Bhattacharjee A, Eichler EE, et al. 2009. Targeted capture and massively parallel sequencing of 12 human exomes. *Nature* 461:272–276.

Nygaard S, Zhang G, Schiøtt M, Li C, Wurm Y, Hu H, Zhou J, Ji L, Qiu F, Rasmussen M, et al. 2011. The genome of the leaf-cutting ant *Acromyrmex echinatior* suggests key adaptations to advanced social life and fungus farming. *Genome Res*. 21:1339–1348.

Öttingen Hv. 1938. Erfahrungen über das Arbeiten mit Thysanopteren. *Arb Phys Augew Ent Berlin-Dahlem.* 5:178–182.

Paijmans JL, Fickel J, Courtiol A, Hofreiter M, Förster DW. 2016. Impact of enrichment conditions on cross-species capture of fresh and degraded DNA. *Mol Ecol Res*. 16:42–55.

Peñalba JV, Smith LL, Tonione MA, Sass C, Hykin SM, Skipwith PL, McGuire JA, Bowie RC, Moritz C. 2014. Sequence capture using PCR-generated probes: a cost-effective method of targeted high-throughput sequencing for nonmodel organisms. *Mol Ecol Res*. 14:1000–1010.

Peng Y, Leung HC, Yiu SM, Chin FY. 2012. IDBA-UD: a *de novo* assembler for single-cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics* 28:1420–1428.

Petersen M, Meusemann K, Donath A, Dowling D, Liu S, Peters RS, Podsiadlowski L, Vasilikopoulos A, Zhou X, Misof B, Niehuis O. 2015. Orthograph 0.5.6. Available from: http://mptrsen.github.io/Orthograph.

Prum RO, Berv JS, Dornburg A, Field DJ, Townsend JP, Lemmon EM, Lemmon AR. 2015. A comprehensive phylogeny of birds (Aves) using targeted next-generation DNA sequencing. *Nature* 526:569–573.

Rogers SO, Bendich AJ. 1985. Extraction of DNA from milligram amounts of fresh, herbarium and mummified plant tissues. *Plant Mol Biol*. 5:69–76.

Smith CD, Zimin A, Holt C, Abouheif E, Benton R, Cash E, Croset V, Currie CR, Elhaik E, Elsik CG, et al. 2011. Draft genome of the globally widespread and invasive Argentine ant (*Linepithema humile*). *Proc Natl Acad Sci U S A*. 108:5673–5678.

Suchan T, Pitteloud C, Gerasimova NS, Kostikova A, Schmid S, Arrigo N, Pajkovic M, Ronikier M, Alvarez N. 2016. Hybridization capture using RAD probes (hyRAD), a new tool for performing genomic analyses on collection specimens. *PLoS One* 11:e0151651.

Suyama M, Torrents D, Bork P. 2006. PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic Acids Res*. 34:W609–W612.

Vinner L, Mourier T, Friis-Nielsen J, Gniadecki R, Dybkaer K, Rosenberg J, Langhoff JL, Cruz DF, Fonager J, Izarzugaza JM, et al. 2015. Investigation of human cancers for retrovirus by low-stringency target enrichment and high-throughput sequencing. *Sci Rep*. 5:13201.

Wallace RB, Shaffer J, Murphy RF, Bonner J, Hirose T, Itakura K. 1979. Hybridization of synthetic oligodeoxyribonucleotides to phi chi 174 DNA: the effect of single base pair mismatch. *Nucleic Acids Res*. 6:3543–3557.

Waterhouse RM, Zdobnov EM, Tegenfeldt F, Li J, Kriventseva EV. 2010. OrthoDB: the hierarchical catalog of eukaryotic orthologs in 2011. *Nucleic Acids Res*. 39:D283–D288.

Weinstock GM, Robinson GE, Gibbs RA, Weinstock GM, Robinson GE, Worley KC, Evans JD, Maleszka R, Robertson HM, Weaver DB, et al. 2006. Insights into social insects from the genome of the honeybee *Apis mellifera*. *Nature* 443:931–949.

Werren JH, Richards S, Desjardins CA, Niehuis O, Gadau J, Colbourne JK, Beukeboom LW, Desplan C, Elsik CG, Grimmelikhuijzen CJ, et al. 2010. Functional and evolutionary insights from the genomes of three parasitoid *Nasonia* species. *Science* 327:343–348.