2013

Evaluation of a New Method for Large-Scale and Gene-targeted Next Generation DNA Sequencing in Nonmodel Species

Ted Cosart
The University of Montana

EVALUTATION OF A NEW METHOD FOR LARGE-SCALE AND GENE-TARGETED NEXT GENERATION DNA SEQUENCING IN NONMODEL SPECIES

By

Ted Cosart

BA, University of Montana, Missoula, Montana, 1983
MS, University of Montana, Missoula, Montana, 2006

Dissertation

presented in partial fulfillment of the requirements
for the degree of

Doctor of Philosophy
in the Individualized, Interdisciplinary Graduate Program

The University of Montana
Missoula, Montana

August, 2013

Approved by:

Sandy Ross, Associate Dean of The Graduate School
Graduate School

Dr. Jesse Johnson, Co-Chair
Computer Science

Dr. Gordon Luikart, Co-Chair
Flathead Biological Station

Dr. Jeffrey Good
Division of Biological Sciences

Dr. William Holben
Division of Biological Sciences

Dr. Stephen Porcella
Rocky Mountain Laboratories, National Institute of Allergy and Infectious Diseases

Dr. Alden Wright
Computer Science

Cosart, Ted, 2013

Evaluation of a New Method for Large-Scale and Gene-targeted Next Generation DNA Sequencing in Nonmodel Species

Chairperson or Co-Chairperson: Gordon Luikart, Ph.D.
Chairperson or Co-Chairperson: Jesse Johnson, Ph.D.

The efficient method called exon capture provides for sequencing genes genome-wide, targeting candidate genes, and sampling specific exons within genes. Although developed for model species with available whole genome sequences, the method can capture exons in nonmodel species using the genomic resources of a related model species. How close the relatives must be for effective exon capture is not known. The work herein demonstrates cross-taxa capture in ungulates, using the domestic cow genome as a reference. It also describes a computer program designed for collecting exon sequences for exon capture, allowing users to set per-gene and overall base pair (bp) limits, and to prefer internal or external exons. Cross-taxa exon capture was tested with subject-reference divergence times from 0 to ~60 million years. Sequencing success decreased with increasing subject-reference phylogenetic divergence. With the domestic cow genome as reference, American bison exons, at 1-2 million years (MY) of divergence, were captured as successfully as those of a domestic cow. The cow and bison captures each yielded sequence from ~80% of the 3.6 million bp targeted. Two bighorn sheep, 7 mule deer, and 4 pigs at about 20, 30, and 60 MY of divergence from the cow, respectively, yielded averages of ~70%, ~60%, and ~55% of the targeted bp. A gene family with many closely related, duplicated loci was expected to show reduced success compared to the whole collection. This prediction was supported, as 63 exons in the MHC gene family sequences yielded 62% fully sequenced in the cow, and 32%, 20%, and 4% for the bighorn, deer, and pigs, respectively. A comparison of two sequence alignment programs showed that Stampy, designed for high sample-reference divergence, was dramatically better than BWA, designed for low divergence, only in the pig capture, in which Stampy yielded ~30% more bp than did BWA. A universal ungulate exon capture array could be developed using the 8,999 exons that were fully sequenced in all species, including the pig at ~60 MY. As this method helps us understand the genetic basis of evolutionary processes, so it can contribute to an informed study and stewardship of our ecological endowment.

Acknowledgements

There are many who were crucial to this collaborative work. My acknowledgements will be of necessity incomplete. I thank the National Science Foundation, and all those who led and taught in the integrated graduate training program, Montana-Ecology of Infectious Diseases. M-EID provided instruction in biology and mathematics that was invaluable preparation for this work.

My committee members were all important to the work presented here. Dr. Gordon Luikart had the crucial insight into the potential of cross-taxa exon capture for nonmodel wildlife species, and was the chief instigator of all the next-generation genomics work in which I participated. He was also a guide through many biological concepts, and gave support and shared his expertise freely and with enthusiasm.

Dr. Jesse Johnson has been a valued advisor and guide throughout my time in graduate school, as well as a mentor in computer science and mathematics.

Dr. Bill Holben supported me both through the M-EID program and welcomed me into his lab as I began my studies.

Dr. Alden Wright has also been a longtime mentor in computer science and mathematics, and his enthusiasm for genomics has been infectious and motivating.

My genomics education and this research depend crucially on Dr. Steve Porcella, who welcomed me into a volunteership at Rocky Mountain Laboratories. He and the Genomics Unit were, to say the least, generous with time, resources, and advice. The research central to my dissertation would not have been possible without them. I am particularly grateful to Dr. Craig Martens for mentoring me in the analysis of next-generation sequence data.

I am also grateful to Albano Beja-Pereira, both for his collaboration and support through the Centro de Investigação em Biodiversidade e Recursos Genéticos (CIBIO). I thank all my coauthors and collaborators upon whom the genomics in this dissertation so heavily depends.

Finally I thank my parents, Carolyn Cosart and the late Don Cosart. Besides their unflagging support, my brothers and I have also benefited incalculably from their belief in education and, more generally, the value of intellectual pursuits.

Table of Contents

List of Figures

List of Tables

Chapter 1: Introduction

"These advances [in high-throughput DNA sequencing] have at last brought a truly genomic perspective to the study of adaptive evolutionary change." (Radwan and Babik 2012)

With the work described here I hope to contribute to genomics, the study of genomes or many genes, by broadening applications of a promising gene sequencing method called *exon capture*. The method allows for massively parallel, genome-wide gene sequencing for species with large (billion+ base-pair) genomes, thus far mostly the mammals. It can also precisely target individual genes and even, within genes, selected exonic subsequences. Its economies of scale enable sequencing many genes and genomic regions in many individuals simultaneously (Bansal *et al.* 2011; Rivas *et al.* 2011).

Its use, however, has been largely restricted to species with relatively extensive genomic resources. It typically targets most of the transcribed gene sequences in the genome (collectively called the *exome*). My collaborators and I used the existing genomic resources of a model species (the domestic cow) to sample and sequence thousands of genes in evolutionarily divergent species with few genomic resources of their own (such as mule deer and bighorn sheep). We hope that our results will encourage geneticists who study genomically obscure (nonmodel) organisms to use exon capture to address genetic questions intractable or unimaginable only 5-10 years ago.

High-throughput, gene-targeted DNA sequencing is valuable to many fields in biology, with applications in human disease, agriculture, wildlife conservation, and evolutionary history (O'Roak *et al.* 2012; Bruford *et al.* 2003; Allendorf *et al.* 2010; Fu *et al.* 2012). This work assesses the success of exon capture when used, atypically, to sequence genes in species with few genomic resources, targeting only about 10% of the exome, itself only about 1-3% of the genome. The "capture" in exon capture refers to enriching genomic DNA for exons (with some flanking sequence), while discarding billions of base pairs of non-exonic genome.

The majority of exons are protein coding sequences and adjacent regions that regulate their expression (some exons are non-protein coding templates for RNA molecules). For many geneticists exons are the "high value" part of the genome. These sequences can reveal functional variation and help us understand how genes influence phenotypes, fitness, and adaptation (Hodges *et al.* 2007; Bruneaux *et al.* 2013).

The following assessment of exon capture of partial exomes in nonmodel species centers on the bioinformatic selection of target exon sequences for capture, and the bioinformatic analysis of the short DNA fragments called *reads*, produced by next-generation sequencers. In order to select and compile exon sequences genome wide and from candidate genes for the synthesis of exon capture baits I developed a computer program (Chapter 2) that collects exon sequences from across a published genome. It employs user-specified parameters to sample exons for the design of exon capture arrays.

I then test how well the current, freely-available, and widely-used bioinformatics tools for read analysis (Li and Durbin 2009; Lunter and Goodson 2011; Li *et al.* 2009; McKenna *et al.* 2010) perform for exon capture in two empirical studies (Chapters 3 and

4). These studies yielded gene sequences and gene sequence variation in nonmodel ungulates, sampling small portions (~10%) of their exomes for discovery of DNA marker loci useful in population genomics.

DNA markers for population genetics/genomics

Population geneticists look for genetic markers that are polymorphic, and that discriminate between individuals and populations. Tallied among many individuals, markers can provide metrics of genetic diversity within and between individuals, demes, and larger populations.

Our increasing ability to assay genetic markers in any species continues to "change the way we view nature" (Schlotterer 2004), especially when markers include both neutral and adaptive gene loci (Hansen *et al.* 2012). Markers are used to compute heterozygosity, assess allele frequency differences between populations (e.g., via the Fst statistic), identify haplotypes, discover non-random marker associations (gametic disequilibrium), and to associate these metrics with phenotypes, environmental variables, and fitness (Hohenlohe *et al.* 2010; Yi *et al.* 2010). These approaches, requiring many markers genome-wide, have been called population genomics (Hohenlohe *et al.* 2010).

In the 1990's nuclear microsatellite DNA loci and mitochondrial DNA sequences were the markers most often and widely used to study genetic diversity within and between populations (Morin *et al.* 2004). With their high mutation rates, microsatellites can reveal evolutionarily recent genetic change and divergence among populations. They are, however, susceptible to back mutation and homoplasy (alleles identical by convergent mutation rather than inheritance). They are generally less informative about the inherited genetic basis of phenotype than are single nucleotide polymorphisms (SNPs). SNPs are base differences at single nucleotide sites in genomic DNA. They are the most abundant polymorphisms in the genome. Unlike microsatellites, SNPs routinely occur in and near gene coding sequences. As such, they are more likely than microsatellites to be responsible for differences in the proteins translated from genes. Population geneticists have increasingly turned to SNPs for many applications, from measuring population genetic structure (Liu *et al.* 2005), to the association of genetic variation with adaptive traits (Renaut *et al.* 2011).

To be useful in measuring variation, and to assess structure within and among populations, SNP discovery requires sampling enough individuals, as a rule at least 10, to minimize ascertainment bias (Clark *et al.* 2005; Morin *et al.* 2004). Although costs have declined since the advent of next-generation sequencing, SNP discovery has been sufficiently expensive that many studies have suffered from ascertainment bias (Akey *et al.* 2003; R. Nielsen 2004). The bias can appear when a SNP assay chip (a microarray featuring selected SNP locations) contains markers discovered in only a few individuals, likely missing many rare alleles and/or alleles in populations not sampled. When these assay chips are used on many individuals, SNP data is biased toward those loci and genome regions with high heterozygosity (even allele frequencies) in the populations of origin of the individuals used in the initial SNP discovery (Morin *et al.* 2004).

Fortunately, costs continue to decline for both DNA enrichment and sequencing.

Efficiencies in enrichment for exon capture have been gained recently through refined techniques for pooling (multiplexing) DNA samples during capture (Bansal *et al.* 2011). Marker discovery from a genome-wide scan (as well as in selected genes of interest) of many individuals is an increasingly attractive method for sampling SNPs in populations.

RADs and transcriptomes

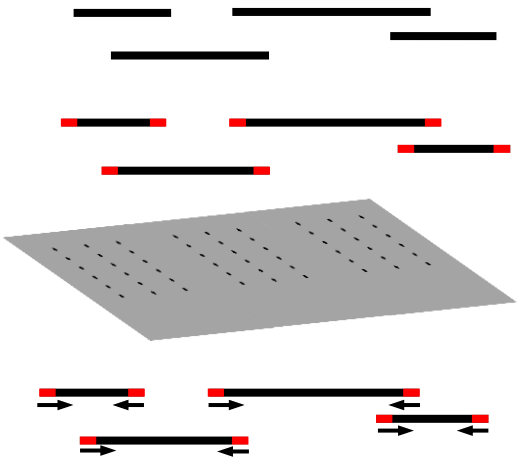
While the research described here is centered on exon capture, there are other massively-parallel methods that target a subset of the genome for high-throughput sequencing.

Restriction-site-associated DNA (RAD) sequencing (Rowe, Renaut, and Guggisberg 2011) analyzes a reduced representation of the genome sequence through use of restriction enzymes that cut the genomic DNA at loci with specific DNA sequence (Davey *et al.* 2011). Fragments are sequenced adjacent to cut locations. RAD sequencing can affordably discover thousands of SNPs throughout a genome. It is amenable to species without genomic resources, since the alignment of the sequence reads de novo can be seeded by grouping identical reads (Catchen *et al.* 2011). However, SNPs discovered through de novo assembly often lack the gene and coding information used to associate genetic variation with functional differences.

Transcriptome sequencing by high throughput methods such as RNA-seq (Wang, Gerstein, and Snyder 2009) can also provide an exon-rich sample of the genome sequence. RNA-seq reads, representing gene transcripts, can be aligned to a reference genome. In addition to quantifying relative gene expression in different tissues, the alignment of the reads can be used for SNP discovery in exon sequences (Cánovas *et al.* 2010). However, RNA is less stable than DNA, and is thus more difficult to sample, especially when collecting from wildlife in the field, from many individuals. Further, correct alignment to a reference genome of RNA-seq is complicated by reads that include splice junctions (representing neighboring exons separated by introns in the genome, but joined in the transcript by splicing), as well as the presence of the poly(A) (polyadenylated) tail of the original RNA molecule (Wang, Gerstein, and Snyder 2009).

Exon capture

Compared to both RAD and RNA-seq, exon capture is more flexible and precise in the size and number of genomic regions to be sequenced. Investigators can choose any portion of the exome, synthesize oligonucleotide probes (typically between 60-120 base pairs) based on the exon sequences of interest. These molecular probes, called *baits,* “capture” the exon sequences from genomic DNA samples by a hybridization reaction of randomly fragmented genomic DNA with the baits (Fig. 1.1). The reaction takes place either on the surface of a microarray (as done in Chapter 3), or in a solution (the method used in Chapter 4), with oligonucleotide baits (DNA fragments) attached to magnetized beads. Small DNA sequences of about 6 bases, called *bar codes*, can be attached to the fragments before capture, so that many individuals can be pooled for capture in one hybridization reaction with retention of sample identity. The baits serve to select DNA fragments representing only the targeted region, and the majority of genomic sequence is discarded.



Exon Capture Basics:

-Fragment a DNA sample into small pieces.

-Attach sequencing primers (red) to each fragment.

-Hybridize (capture) targeted DNA fragments (exon sequences) to probes (baits).

-Wash off non-target DNA. Note: nearly 99% of the DNA is non-target (non-exon) sequence

-Sequence (one or both ends) of all gene fragments ‘captured’ by the baits.

Figure 1.1: Illustration of exon capture technology. It allows targeted-sequencing of 1000s of functional genes. (Illustration after Hodges *et.al* 2007).

Next generation sequencing

After exon capture, the exon-enriched DNA is sequenced on a next-generation platform. Per base-pair cost of DNA sequencing has plummeted with wide adoption of next-generation sequencers, from about $1000.00 per million base pairs (Mb) in 2004, to as little as $0.09 per Mb in 2012 (reported by the National Human Genome Research Institute at http://www.genome.gov/sequencingcosts, using data for sequencing centers funded by the Institute).

The machines in widest use simultaneously sequence fragments of DNA by, first, replicating each fragment to create discrete colonies of single-stranded clones. Clonal colonies are then sequenced as complementary strands are built, base by base, from ends of clones. Each colony produces a sequence read. Currently the most prolific platforms (e.g., the Illumina HiSeq), can produce several billion reads in a single sequencer run. Read length is controlled by limiting the number of times nucleotides are added to colonial complementary strands. Fragments often exceed read length, so that the end-sequencing, even from both ends (producing “paired end” reads), results in only partial sequencing of the template molecule. The common method used to read the sequence is an optical measurement of photoluminescence induced during base incorporation.

Compared to single end reads, paired end reads can be aligned to a reference sequence with more confidence. The pairs represent two ends of the same fragment. Fragment lengths have a mean and standard deviation known from the molecular preparation of the sample, or inferred from read-mapping to a reference. Confidence in the alignment of a pair increases when individual alignments of read pairs to a reference genome together bound a span of reference approaching the mean fragment length. Less certainly, and requiring evidence of similar placements by other reads, a pair spanning a length in the tails of the distribution can identify deletions and insertions in the sample versus the reference. Paired-end sequences are used in the experiment described in Chapter 4.

High throughput sequencers are prodigious. The widely used Illumina HiSeq2000 instrument can sequence over 50 billion base pairs per day (Caporaso *et al.* 2012). If the number of base pairs targeted (i.e., the number of genomic positions baited) is limited to several million rather than the billions in the typical mammalian whole genome, and, further, if pooled samples are used (with attached oligonucleotide indices retaining sample identity), a next-generation sequencer can sequence genes in many individuals at once.

Genomics technology transferred to nonmodel species

Model species have benefited from intensive genomics and bioinformatic tool development. These are the organisms of high interest in medicine and agriculture, such as humans, mice, cows, corn, and those with a central position in the history of genetic research, such as *Drosophila sp.* and *Arabidopsis thaliana.* Many have multiple full genome sequences and large databases of gene transcripts and genetic markers. Plentiful genomic data serve as reference for alignment and assembly of new sequences, and provide accuracy metrics for variant discovery. The depth of genomic data lowers the cost and increases the accuracy of new genomics studies in additional individuals of the model species, so that rich genomic resources can self-perpetuate.

As next-generation efficiencies have lowered the costs of whole genome sequencing, more species' genomes have been sequenced. However, whole genome sequencing project totals for eukaryotic classifications suggest that, to date, there is no wholesale, pan-phylogenetic effort to produce complete, large eukaryotic genome sequences. The vast majority of eukaryotic families, genera and species have no representative with a genome sequencing project (Table 1.1). There is a continuing need to find affordable gene sequencing methods for species with no reference genome.

For vertebrate genomes the Genome 10K project aims to coordinate the sequencing of complete genomes for about one species in each vertebrate genus, to enable genomics in many species (Genome 10K Community of Scientists 2009). As of this writing the project awaits further reductions in whole-genome sequencing costs before getting fully underway (information at http://genome10k.soe.ucsc.edu, accessed Aug., 2013).

Table 1.1: Eukaryotes with genome projects. From the U.S. Dept. of Energy, Joint Genome Institute, the Genomes Online Database, http://www.genomesonline.org, the number of Eukaryote types with genome sequencing projects as a percentage of phylogenetic classification. Accessed Jan., 2013.

Subdivision	Total Types	With genome project	Percentage with project
Phylum	56	35	62.50%
Class	182	97	53.30%
Order	1,037	280	27.00%
Family	6,689	521	7.79%
Genus	54,319	855	1.57%
Species	218,222	1,217	0.56%

To test exon capture in nonmodel species, collaborators and I used the cow genome sequence to create baits for exon capture in several ungulate species (see the summary for chapters 3 and 4, below). These are represented by both model species (e.g., the domestic cow, sheep, and pig) and nonmodel species (bison, African buffalo, bighorn sheep, and mule deer). Some domesticated ungulates have whole genome sequences which can serve as references, the cow (Elsik, Tellam, and Worley 2009), sheep (Archibald *et al.* 2010a), and pig (Archibald *et al.* 2010b). The relative abundance of genomics data for these species motivated our exon capture study in Chapter 4, in which we sampled ungulates with varying degrees of genetic divergence from the cow using baits designed from cow exon sequences, and read alignment to the cow genome sequence. We aligned our most divergent species (the domestic pig) to both the taurine cow and pig genomes.

Exon target selection for array design

Exon sequences for bait design are usually selected from a reference genome sequence with gene annotations. Conservation of exon sequences in mammals (Modrek and Lee 2003; Thomas *et al.* 2003), suggests that exons on the genome sequences of one mammal can be used to capture many homologous exon sequences from those of another.

There are other sources of sequences for exon bait arrays. While a nonmodel species may have little or no genomic sequence available, many have publicly available gene transcript sequences, often in collections of expressed sequence tags (ESTs), whose sequences represent the ends of mRNA gene transcripts. Sequence alignment programs designed to align transcripts (which may include splice junctions, as noted above) to genomic DNA sequences can extract individual exon sequences from ESTs and other sources of gene transcripts by aligning them to a genome of a related species. Restricting exon selection to those represented only in transcripts, however, may rule out candidate genes of interest, if transcripts are missing from the genes. Transcripts can bias exon selection to genes with high expression in certain tissues.

Our choice in the studies that follow was to use the relatively complete set of gene annotations offered by the cow genome, as the basis for a capture array for use on several ungulates.

From reads to genotypes

Next-generation sequencing projects routinely produce hundreds of millions of short reads, typically ranging from 75 to 400 base pairs (bp) long. Each nucleotide in a read is known as a *base call,* apropos of the uncertainty in the sequencing process. The reads represent sequences at the ends of fragments of randomly fragmented genomic template DNA, end-sequenced as described above. Those representing overlapping pieces of the same genomic sequence can be stacked to produce aligned columns of bases (Fig. 1.2). The alignments reveal the original, targeted sequence by a consensus of bases at the aligned positions, subject to errors from the sequencer, and gaps caused by repeat regions in the targeted regions that are too long to be disambiguated by the reads.

The alignment of reads for consensus genotyping (see section below) is usually accomplished by one of two approaches, either *de novo assembly* (Zerbino and Birney 2008; Salzberg *et al.* 2012), in which reads are aligned with each other without recourse to sequence data outside the reads, or *mapping*, whereby reads are individually aligned to a pre-existing DNA reference sequence (Fig. 1.2), often a full genome sequence (Trapnell and Salzberg 2009). Some analysis protocols combine these two approaches. In both cases the enormous volume of reads from even a small part of a single next-generation sequencer run presents a computational problem so large that exclusive reliance on dynamic programming algorithms that guarantee optimal sequence alignments (under a given system of scoring for mismatches and gaps) is impractical.

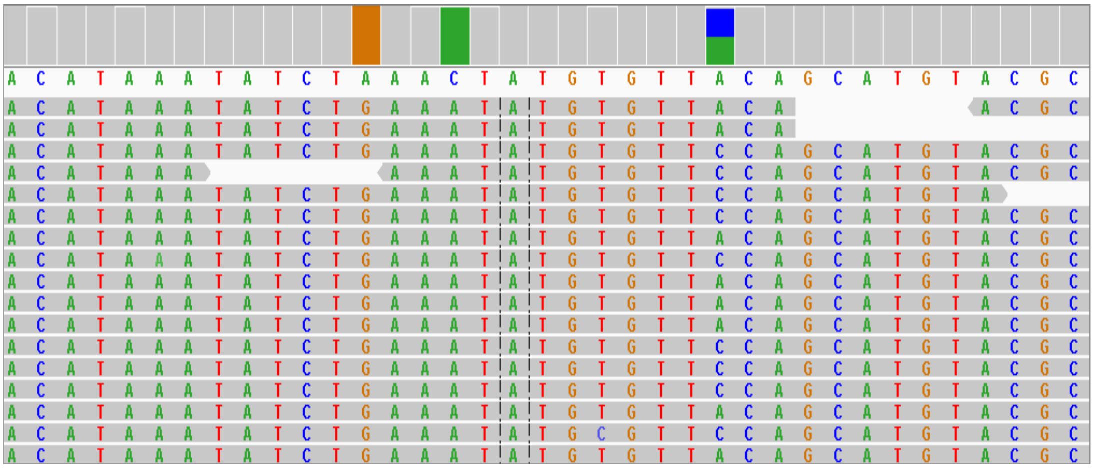

A C A T A A A T A T C T A A A C T A T G T G T T A C A G C A T G T A C G C

Figure 1.2: Excerpted alignment of 100 bp bighorn sheep reads. Read bases are on gray background, aligned to the cow reference genome (top, bases on white). The non-gray rectangles at the top show where the sheep consensus differs from the cow reference. Note the one SNP candidate (heterozygous position) indicated by the dual-colored rectangle, below which bases A and C are aligned to the same position. Solid colored rectangles (top) show where the reads indicate a homozygous difference versus the reference base. This image was generated by the Integrated Genomics Viewer, available at http://www.broadinstitute.org/igv.

In the most recent software, mapping and genotyping programs both use quality scores to estimate accuracy. Also, the sequencers themselves calculate a quality score for each base. Scoring in base-calling, mapping, and genotyping emulates the code devised for the Phred base calling program (Ewing and Green 1998), and is sometimes called Phred-like when not applied to base calling. In the Phred encoding a score *q* encodes the probability *p* that a base is incorrect as $q = -10 \log_{10}(p)$ so that, for example, a base call with error probability 1/1000 has a Phred-like score of 30.

Types of read mappers

The need for algorithms that balance accuracy and speed has yielded, in the case of read mappers, two strategies in widest use, *hash-table* based mappers (e.g., Stampy, Lunter and Goodson 2011), and those based on *suffix arrays* (e.g. BWA, Li and Durbin 2009). These differ in their methods of indexing sequence information in computer memory.

Hash tables index positions in either the reference (e.g., Lunter and Goodson 2011), or the set of reads (e.g., Li, Ruan, and Durbin 2008). Hash tables allow the computer to find positions in the indexed sequences with a few operations, generally computing an integer based on the nucleotide sequence which matches that in the matched positions of the indexed reference or read set. Hash-table based mappers have been empirically shown to rate high in accuracy relative to software based on other indexing methods (Lunter and Goodson 2011; Pattnaik *et al.* 2012; Nielsen *et al.* 2011).

For mapping reads with few differences to their reference the suffix array lookup methods are much faster than hash-table mappers. Also, in their most efficient implementations, they use appreciably less computer memory (Li and Durbin 2009). They gain their speed and memory efficiencies by using text compression. A compressed representation of the reference in memory allows single alignment to reference substrings that occur in multiple genomic locations, versus multiple locations in a hash table index. One widely used suffix-array mapper, BWA (Li and Durbin 2009), while appreciably faster than most hash-table-based mappers at low read-reference divergence (e.g., 2 mismatches per 100 bases), slows considerably when tuned to allow high read-reference divergence (e.g., more than 4 or 5 mismatches per 100 bases).

Quantifying uncertainty in mapping

Having found (heuristically) the best alignment of the read to the reference, many current mappers provide a *mapping quality,* a Phred-like score for the probability that the alignment placement is wrong. For example, for paired-end reads, the Stampy read aligner models the alignment error as a likelihood of a read pair (*r1,r2*) mapping to reference genome loci (*x0,y0*) as,

$$L(r1,r2,x0y0) = P_r\ (r1|x0)\ P_r(r2|y0)\ P_d(y0\text{-}x0)\ P_u(x0).$$

P_r is the likelihood of the alignment, and incorporates probabilities of read errors, single nucleotide variations from the reference base, and indels. To calculate read error probabilities, mappers use the Phred-like base quality scores that sequencers provide for each base of a read. P_d models the insert size. The insert size is the number of bases bounded by the read pair, representing the length of the fragment that was sequenced, in this case at both ends. Stampy approximates the insert size distribution for paired-end reads as it aligns the first few hundred pairs. P_u is the uniform distribution over the reference genome. This likelihood is then used to compute a posterior probability that the wrong locus was chosen for alignment of the read. For concision let $L_{rxy} = L(r1,r2,x,y)$**,** then,

$$1 - P(x0,y0|r1,r2) = 1 - \ L_{rx0y0}\ /\ \textstyle\sum_{(x,y)\in \mathrm{C}} L_{rxy}\ \ \mathrm{X}\ \ \sum_{(x,y)\in C} L_{rxy}\ /\ \sum_{(x,y)\in \Omega} L_{rxy}\ ,$$

with *C* enumerating all candidate positions found by the algorithm, and *Ω* enumerating all pairs of loci on the reference. The last factor is not feasible to calculate, and is replaced by a probability that the correct alignment position was not considered (details are in the supplement to Lunter and Goodson 2011).

Consensus genotyping

With aligned reads, it is possible to posit a genotype at each base-pair position covered by the alignment. When using a reference sequence to align reads, genotyping computer programs assign one base, a gap, or an insertion at each reference position to which one or more reads are aligned. The reference position consists of a chromosome name (e.g., chromosome 3), and a position, i.e., an integer i representing the i^{th} base pair on the chromosome, position 1 being the first base on the 5′ end of the reference strand.

The number of different bases inferred at a nucleotide position (for each study individual) is constrained by ploidy. In the usual case with mammals, an alignment of reads for one individual at any one nucleotide position in the genome represents one diploid genotype. The genotyper assigns two bases, identical if the individual is inferred to be homozygous at the position, two different bases if the individual is inferred to be heterozygous.

Quantifying uncertainty in genotyping

Roughly, genotypes are assigned according to the commonest base or bases in an aligned stack of reads (Fig. 1.2). As Nielsen *et al.* 2011 notes, in alignments with deep coverage, (and especially in the early days of next-generation read alignment to references), genotype calling has been often based on a majority-rule of the count of bases aligned at a given site, preceded by discarding all aligned bases whose base-quality scores were below a threshold (typically, a Phred score of 20). In this scheme positions are called heterozygous when counts for the two most frequently occurring bases at the aligned position are out of balance by no more than a threshold ratio (e.g. 80/20).

Because the majority-rule methods require relatively deep coverage (estimated generally as coverage over 20X), and do not quantify accuracy using base or mapping error probability, most recent mappers use a probabilistic analysis expressing the probability of an incorrect genotype inference. Generally a genotype likelihood is computed, $p(X|G)$, with X as the data in the reads at a given genomic site, and G the genotype at that position. With a prior posited $p(G)$, Bayes formula is used to arrive at the posterior $p(G|X)$ (Nielsen *et al.* 2011). Rescaling the quality scores, and assuming their independence, the likelihood $p(X|G)$ can be the product of the individual probabilities $p(X_i|G)$, with X_i the data from the i^{th} read.

Incorporating reads from multiple samples in a single genotyping likelihood calculation can increase accuracy, for example, in the assignment of prior probabilities for each possible genotype. In the case of an individual sample, with no data available except that of the reads, the prior is often posited as uniform for all genotypes. With multiple samples, the GATK genotyper (UnifiedGenotyper, McKenna *et al.* 2010), for example, uses a Bayesian genotype likelihood model that incorporates data from all samples to compute the most likely genotypes for each. Further information outside the individual base calls and their Phred scores, adds to accuracy. Nielsen *et al.* (2011) found high genotyping accuracy using linkage disequilibrium information. When genotypes at multiple sites are linked, reliable information at some sites can be used to calculate genotype likelihoods at linked sites. The studies that follow did not employ linkage disequilibrium information, infeasible in cases in which the genome of the sample species is not available. The study in Chapter 4, however, did use the GATK's multi-sample genotyping feature.

Multiple metrics for each genotype call

For each genotype inference, besides a single Phred-like quality score, genotypers such as the UnifiedGenotyper, and the genotyper provided in Samtools (Li *et al.* 2009), generate multiple quantities for each position. Examples include the root mean square of mapping qualities of all reads aligned to the site being genotyped, the relative likelihood of genotypes alternative to the most likely, and many others. In the UnifiedGenotyper and the Samtools genotyper, the data are presented in the Variant Calling Format (Danecek *et al.* 2011), a standard data file format for genotyping based on reads mapped to a reference genome.

False positive and false negative genotypes, generally arising when inferring SNPs and indels, but also resulting from de-novo assembly of reads for inferring whole genome sequences, continue to challenge analyses of next-generation sequencing. Work is underway to establish protocols for genotyping mapped reads under different experimental conditions of sample number, depth of coverage, and available genomic data for the species being sequenced (DePristo *et al.* 2011; Martin *et al.* 2010; Li 2011). The multiplicity of metrics attending genotyping testifies to the lack of a simple, single metric of uncertainty that is reliable under all circumstances.

Chapter overviews

Chapter 2

Chapter 2 describes a computer program, ExonSampler, which collects exon sequences from a reference genome, genome-wide. It can choose exons evenly spaced across all chromosomes for efficient genome-wide scans. It also can collect exon sequences based on a list of gene abbreviations (e.g., from candidate genes such as the immune system genes TLR4, IFNG, etc.). This software was designed to automate and customize the collection of partial exomes for design of exon capture baits. For example, in the capture described in Chapter 3, the program was used to collect ~16,000 exons representing ~10% of a mammalian exome. It can set limits on several collection parameters to limit the size of the collection (e.g., up to 1 kb per gene) and to prefer some kinds of exons over others, for example, in each gene, to collect the 5′ upstream exon near regulatory sequences before collecting other exons, to meet specific experimental designs. ExonSampler allows researchers to design exon capture to their own specifications without the cost and time of software development, especially for those who have no ready access to bioinformatics infrastructure.

Chapter 3

Chapter 3 presents a published proof of concept of exon capture with reduced target size and reference-sample divergence, with successful genotyping of one *Bos taurus*, one *Bos indicus*, and one *Bison bison* individual using the *Bos taurus* reference genome, and a target size of about 3 million base pairs (~10% of the exome). As a proof of concept of available genotyping tools to genotype divergent species, our study showed that our wildlife species, the bison, as well as the non-taurine cow (*B. indicus*), could be captured using taurine cow baits, the resulting reads aligned and genotyped for SNPs using a taurine cow reference. In bison we report 2,426 putative SNPs genome-wide, with 339 SNPs in 96 candidate genes. Results here should encourage researchers in conservation and molecular ecology, studying nonmodel subjects, that exon capture can be a viable method of targeted genetic marker discovery using a related model species genome.

Chapter 4

Chapter 4 quantifies how the genetic divergence of a sampled genome (e.g., sheep) from a domestic cow reference genome affects the success of sequencing, SNP genotyping, and SNP discovery following exon capture. The study evaluated two kinds of mapping software, one designed for close sample-reference genome sequence similarity (Li and Durbin 2009), the other designed for higher divergence between sample and reference genomes (Lunter and Goodson 2011). This study also targeted a subset of the exome (about 3.6 million base pairs), including exons in ~350 candidate genes. The *Bos taurus* reference was again the basis for probe design and mapping. Six divergent ungulate species provided a wide range of genetic divergence from the taurine cow. Divergence ranged from near zero in a taurine cow, to ~1-2 million years (MY), in the *Bison bison*, ~1-3 MY in *Syncerus caffer* (African buffalo), ~20MY in *Ovis canadensis* (bighorn sheep), ~40 MY in *Odocoileus hemionus* (mule deer), to ~60 million years in *Sus scrofa* (wild boar) and *Sus domesticus* (domestic pig).

Results from this study further clarify the feasibility and limitations of cross-species exon capture described in Chapter 3. Promisingly, targeting a reduced exome to increase the number of individuals that can be sequenced for a given cost, gives high coverage and high quality genotypes in a large proportion of the targeted exons. This applied even to our ungulate species phylogenetically divergent from the cow reference by 10s of millions of years. Success in this study suggests wide possibilities for cross-species exon capture for marker discovery in species with few gnomic resources, using well-tested molecular methods and bioinformatic software.

Supplementary work and appended publications

My work in the analysis of next-generation sequencing owes much to collaboration and instruction preceding the exon capture studies, as well supplementary activities during the exon capture analyses. The collaborations noted below provided me with introductions to molecular biology, phylogenetics and bioinformatics, with practice in the latter two. These were essential activities not only in developing a necessary biological background for the work central to this dissertation, but also to understand first-hand the centrality of collaboration when integrating computer science with biology in using the new sequencing methods.

Microbial community DNA analysis

Collaboration with microbial ecologists at the Holben Lab at the University of Montana introduced me to DNA sequence alignment and its importance in ecology. Microbial ecologists there were investigating solutions to problems in measuring microbial community richness and diversity by PCR amplification of 16s RNA gene sequences from composite soil DNA samples. My contribution to the effort was creating two computer programs that implemented the novel analyses of my collaborators, contributing to three publications (Morales *et al.* 2008, 2009, 2010). These programs are publicly available online at http://holben-lab.dbs.umt.edu/links.php.

The DAM program (*D*OTUR *A*RB *M*atching) extracts groups of 16s rDNA sequences as processed by two programs in wide use among microbial ecologists, (i) ARB, which generates distance matrices and constructs phylogenetic trees based on sequence similarity to references, and (ii) DOTUR, which bins sequences according to distance values given by similarity matrices. Our DAM program isolates ARB-generated sequence groups in the DOTUR bins and produces a new DOTUR-like file of bins containing only the ARB-specified sequences. DOTMAN (*D*OTUR *MAN*ipulation) generates sequence files (in the FASTA format) based on a range of sequence similarity values, as well as a range of bin sizes, from a DOTUR (or DAM) list of bins.

The programs were used to test the feasibility of a “universal” cutoff value of 16s rDNA sequence similarity for phylum-level binning (Morales *et al.* 2009). Such a value had been suggested by former publications. Tests with DAM and DOTMAN showed highly variable sequence similarity cutoffs for phylum level discrimination. We also published a description of the two programs and their functions (Morales *et al.* 2008).

A third study with collaborators at the Holben Lab linked greenhouse gas emission in soils with the abundance of selected bacterial genes (Morales, Cosart, and Holben 2010). My contribution in this study was the multivariate analysis of the gas emission and gene abundance data in different soil types. I used principal components analysis to visualize the relationship of the gas emissions to soil type, gene abundances to soil type, and each gas measurement and gene abundance variable's contribution to the variance in the soil type responses. Co-inertia analysis visualized how similarly the gas emission values and abundances, in separate PCA analyses, explained the variances in the soil type responses.

The three publications described here are appended to this dissertation.

Volunteership at the NIH Rocky Mountain Laboratories

As a trainee in the Montana-Ecology of Infectious Diseases Program (an Integrative Graduate Education and Research Traineeship program funded by the National Science Foundation), I fulfilled an internship requirement by volunteering at the Rocky Mountain Laboratories (RML), a National Institute of Allergies and Infectious Diseases (NIAID) laboratory, in the Genomics Unit of the Research Technologies Branch. The unit provides NIAID researchers with microarray and sequencing services, including bioinformatic analyses.

At RML I learned from both molecular biologists and bioinformatics staff about sequencing using multiple next-generation sequencing platforms, ABI Solid, Illumina, and the Roche 454 platform. Multiple projects on all of the platforms has given RML biologists and bioinformatics researchers extensive experience in the problems peculiar to each platform. Weekly reviews of current projects with the sequencing group provided invaluable exposure to the collaboration between the molecular biologists and bioinformatics staff, as they made decisions about which platform best suited a particular sequencing project, confronting both informatics and biological problems offered by a given organism and the required amount of sequencing.

My direct involvement included quantitative evaluations of a prospective read mapping program, its correctness and efficiency relative to mappers in the existing bioinformatics pipeline. Programming required for this analysis seeded my collection of python programming and scripting used in the analysis of exon capture, especially in processing the high volume of data in the multi-ungulate exon capture (Chapter 4). Interactions with the read mapping program's developers acquainted me with problems such as the trade-offs between speed and thoroughness when finding alignments of millions of reads on a large genome, and helped me understand some current methods of meeting the computational challenges.

I also participated in the SNP analysis of whole microbial genome sequencing, involving SNP discovery in hundreds of genomes. Through a group analysis of hundreds of genome-wide mappings, including participation by a biologist and programmer on staff at the company who manufactured the sequencing instrument used in the experiment, I was introduced to the problems surrounding variant detection, as the bioinformatics group searched for proper SNP filtration criteria in order to achieve a final set of high-confidence SNPs on which to base phylogenies.

My experience at RML included the multi-ungulate exon capture detailed in Chapter 4. The ungulate DNA was sequenced at the Genomics Unit's sequencing facilities, on the Illumina HiSeq platform. Direct consultation and collaboration with the genomics staff provided insight into sequencing run itself, and, especially, help interpreting sequencer metrics affecting read qualities.

References

Akey, Joshua M., Kun Zhang, Momiao Xiong, and Li Jin. 2003. “The Effect of Single Nucleotide Polymorphism Identification Strategies on Estimates of Linkage Disequilibrium.” Molecular Biology and Evolution 20 (2) (February 1): 232–242. doi:10.1093/molbev/msg032.

Allendorf, Fred W., Paul A. Hohenlohe, and Gordon Luikart. 2010. “Genomics and the Future of Conservation Genetics.” Nat Rev Genet 11 (10) (October): 697–709. doi:10.1038/nrg2844.

Archibald, A L, N E Cockett, B P Dalrymple, T Faraut, J W Kijas, J F Maddox, J C McEwan, *et al.* 2010. “The Sheep Genome Reference Sequence: a Work in Progress.”

Animal Genetics 41 (5) (October): 449–453. doi:10.1111/j.1365-2052.2010.02100.x.

Archibald, Alan, Lars Bolund, Carol Churcher, Merete Fredholm, Martien Groenen, Barbara Harlizius, Kyung-Tai Lee, *et al.* 2010. “Pig Genome Sequence - Analysis and Publication Strategy.” BMC Genomics 11 (1): 438. doi:10.1186/1471-2164-11-438.

Bansal, Vikas, Ryan Tewhey, Emily M. LeProust, and Nicholas J. Schork. 2011. “Efficient and Cost Effective Population Resequencing by Pooling and In-Solution Hybridization.” PLoS ONE 6 (3) (March): 1–6. afh.

Bruford, Michael W., Daniel G. Bradley, and Gordon Luikart. 2003. “DNA Markers Reveal the Complexity of Livestock Domestication.” Nature Reviews Genetics 4 (11) (November 1): 900–910. doi:10.1038/nrg1203.

Bruneaux, Matthieu, Susan E. Johnston, Gábor Herczeg, Juha Merilä, Craig R. Primmer, and Anti Vasemägi. 2013. “Molecular Evolutionary and Population Genomic Analysis of the Nine-spined Stickleback Using a Modified Restriction-site-associated DNA Tag Approach.” Molecular Ecology 22 (3): 565–582. doi:10.1111/j.1365-94X.2012.05749.x.

Cánovas, Angela, Gonzalo Rincon, Alma Islas-Trejo, Saumya Wickramasinghe, and Juan F. Medrano. 2010. “SNP Discovery in the Bovine Milk Transcriptome Using RNA-Seq Technology.” Mammalian Genome 21 (11-12) (December): 592–598. doi:10.1007/s00335-010-9297-z.

Caporaso, J. Gregory, Christian L. Lauber, William A. Walters, Donna Berg-Lyons, James Huntley, Noah Fierer, Sarah M. Owens, *et al.* 2012. “Ultra-high-throughput Microbial Community Analysis on the Illumina HiSeq and MiSeq Platforms.” The ISME Journal 6 (8): 1621–1624. doi:10.1038/ismej.2012.8.

Catchen, Julian M., Angel Amores, Paul Hohenlohe, William Cresko, and John H. Postlethwait. 2011. “Stacks: Building and Genotyping Loci De Novo From Short-Read Sequences.” G3: Genes|Genomes|Genetics 1 (3) (August 1): 171–182. doi:10.1534/g3.111.000240.

Clark, Andrew G, Melissa J Hubisz, Carlos D Bustamante, Scott H Williamson, and Rasmus Nielsen. 2005. “Ascertainment Bias in Studies of Human Genome-wide Polymorphism.” Genome Research 15 (11) (November): 1496–1502. doi:10.1101/gr.4107905.

Danecek, Petr, Adam Auton, Goncalo Abecasis, Cornelis A. Albers, Eric Banks, Mark A. DePristo, Robert Handsaker, *et al.* 2011. “The Variant Call Format and VCFtools.” Bioinformatics (June 7). doi:10.1093/bioinformatics/btr330.

Davey, John W., Paul A. Hohenlohe, Paul D. Etter, Jason Q. Boone, Julian M. Catchen, and Mark L. Blaxter. 2011. “Genome-wide Genetic Marker Discovery and Genotyping

Using Next-generation Sequencing." Nature Reviews Genetics 12 (7) (July 1): 499–510. doi:10.1038/nrg3012.

DePristo, Mark A, Eric Banks, Ryan Poplin, Kiran V Garimella, Jared R Maguire, Christopher Hartl, Anthony A Philippakis, *et al.* 2011. "A Framework for Variation Discovery and Genotyping Using Next-generation DNA Sequencing Data." Nat Genet 43 (5) (May): 491–498. doi:10.1038/ng.806.

Elsik, Christine G., Ross L. Tellam, and Kim C. Worley. "The Genome Sequence of Taurine Cattle: A Window to Ruminant Biology and Evolution." Science (New York, N.Y.) 324, no. 5926 (April 2009): 522–528. doi:10.1126/science.1169588.

Ewing, Brent, and Phil Green. "Base-Calling of Automated Sequencer Traces Using Phred. II. Error Probabilities." Genome Research 8, no. 3 (March 1, 1998): 186–194.

Fu, Wenqing, Timothy D. O'Connor, Goo Jun, Hyun Min Kang, Goncalo Abecasis, Suzanne M. Leal, Stacey Gabriel, *et al.* 2012. "Analysis of 6,515 Exomes Reveals the Recent Origin of Most Human Protein-coding Variants." Nature. doi:10.1038/nature11690.

Genome 10K Community of Scientists. 2009. "Genome 10K: A Proposal to Obtain Whole-Genome Sequence for 10,000 Vertebrate Species." Journal of Heredity 100 (6) (November 1): 659 –674. doi:10.1093/jhered/esp086.

Hansen, Michael M., Isabelle Olivieri, Donald M. Waller, Einar E. Nielsen, and The GeM Working Group. 2012. "Monitoring Adaptive Genetic Responses to Environmental Change." Molecular Ecology 21 (6): 1311–1329. doi:10.1111/j.1365-94X.2011.05463.x.

Hodges, Emily, Zhenyu Xuan, Vivekanand Balija, Melissa Kramer, Michael N Molla, Steven W Smith, Christina M Middle, *et al.* 2007. "Genome-wide in Situ Exon Capture for Selective Resequencing." Nature Genetics 39 (12) (December): 1522–1527. doi:10.1038/ng.2007.42.

Hohenlohe, Paul A., Susan Bassham, Paul D. Etter, Nicholas Stiffler, Eric A. Johnson, and William A. Cresko. 2010. "Population Genomics of Parallel Adaptation in Threespine Stickleback Using Sequenced RAD Tags." PLoS Genetics 6 (2) (February): 1–23. doi:10.1371/journal.pgen.1000862.

Li, H., J. Ruan, and R. Durbin. 2008. "Mapping Short DNA Sequencing Reads and Calling Variants Using Mapping Quality Scores." Genome Research 18 (11): 1851.

Li, Heng. 2011. "A Statistical Framework for SNP Calling, Mutation Discovery, Association Mapping and Population Genetical Parameter Estimation from Sequencing Data." Bioinformatics 27 (21) (November 1): 2987 –2993.

doi:10.1093/bioinformatics/btr509.

Li, Heng, and Richard Durbin. 2009. “Fast and Accurate Short Read Alignment with Burrows–Wheeler Transform.” Bioinformatics 25 (14) (July 15): 1754 –1760. doi:10.1093/bioinformatics/btp324.

Li, Heng, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, and Richard Durbin. 2009. “The Sequence Alignment/Map Format and SAMtools” 25 (16) (August 15): 2078–2079. doi:10.1093/bioinformatics/btp352.

Li, Heng, and Nils Homer. 2010. “A Survey of Sequence Alignment Algorithms for Next-generation Sequencing.” Briefings in Bioinformatics 11 (5): 473 –483. doi:10.1093/bib/bbq015.

Liu, Nianjun, Liang Chen, Shuang Wang, Cheongeun Oh, and Hongyu Zhao. 2005. “Comparison of Single-nucleotide Polymorphisms and Microsatellites in Inference of Population Structure.” BMC Genetics 6 (Suppl 1) (December 30): S26. doi:10.1186/1471-2156-6-S1-S26.

Lunter, Gerton, and Martin Goodson. 2011. “Stampy: A Statistical Algorithm for Sensitive and Fast Mapping of Illumina Sequence Reads.” Genome Research 21 (6) (June 1): 936–939. doi:10.1101/gr.111120.110.

Martin, E. R., D. D. Kinnamon, M. A. Schmidt, E. H. Powell, S. Zuchner, and R. W. Morris. 2010. “SeqEM: An Adaptive Genotype-calling Approach for Next-generation Sequencing Studies.” Bioinformatics 26 (22) (November 15): 2803 –2810. doi:10.1093/bioinformatics/btq526.

McKenna, Aaron, Matthew Hanna, Eric Banks, Andrey Sivachenko, Kristian Cibulskis, Andrew Kernytsky, Kiran Garimella, *et al.* 2010. “The Genome Analysis Toolkit: A MapReduce Framework for Analyzing Next-generation DNA Sequencing Data.” Genome Research 20 (9): 1297 –1303. doi:10.1101/gr.107524.110.

Modrek, Barmak, and Christopher J. Lee. 2003. “Alternative Splicing in the Human, Mouse and Rat Genomes Is Associated with an Increased Frequency of Exon Creation And/or Loss.” Nature Genetics 34 (2) (May 5): 177–180. doi:10.1038/ng1159.

Morales, S. E., T. Cosart, J. V. Johnson, and W. E. Holben. “Supplemental Programs for Enhanced Recovery of Data from the DOTUR Application.” Journal of Microbiological Methods 75, no. 3 (2008): 572–575.

Morales, Sergio E., Theodore F. Cosart, Jesse V. Johnson, and William E. Holben. “Extensive Phylogenetic Analysis of a Soil Bacterial Community Illustrates Extreme Taxon Evenness and the Effects of Amplicon Length, Degree of Coverage, and DNA

Fractionation on Classification and Ecological Parameters.” Applied and Environmental Microbiology 75, no. 3 (February 1, 2009): 668 –675. doi:10.1128/AEM.01757-08.

Morales, Sergio E, Theodore Cosart, and William E Holben. 2010. “Bacterial Gene Abundances as Indicators of Greenhouse Gas Emission in Soils.” ISME Journal (February 25). http://dx.doi.org/10.1038/ismej.2010.8.

Morin, Phillip A., Gordon Luikart, Robert K. Wayne, and the SNP workshop group. 2004. “SNPs in Ecology, Evolution and Conservation.” Trends in Ecology & Evolution 19 (4) (April): 208–216. doi:10.1016/j.tree.2004.01.009.

Nielsen, R. 2004. “Population Genetic Analysis of Ascertained SNP Data.” Human Genomics 1 (3): 218–224.

Nielsen, Rasmus, Joshua S. Paul, Anders Albrechtsen, and Yun S. Song. 2011. “Genotype and SNP Calling from Next-generation Sequencing Data.” Nature Reviews Genetics 12 (6) (June): 443–451. doi:10.1038/nrg2986.

O’Roak, Brian J., Laura Vives, Wenqing Fu, Jarrett D. Egertson, Ian B. Stanaway, Ian G. Phelps, Gemma Carvill, *et al.* 2012. “Multiplex Targeted Sequencing Identifies Recurrently Mutated Genes in Autism Spectrum Disorders.” Science (November 15). doi:10.1126/science.1227764.

Pattnaik, Swetansu, Srividya Vaidyanathan, Durgad G. Pooja, Sa Deepak, and Binay Panda. 2012. “Customisation of the Exome Data Analysis Pipeline Using a Combinatorial Approach.” PLoS ONE 7 (1) (January 6): e30080. doi:10.1371/journal.pone.0030080.

Radwan, Jacek, and Wiesław Babik. 2012. “The Genomics of Adaptation.” Proceedings of the Royal Society B: Biological Sciences 279 (1749) (December 22): 5024–5028. doi:10.1098/rspb.2012.2322.

Renaut, Sébastien, Arne W. Nolte, Sean M. Rogers, Nicolas Derome, and Louis Bernatchez. 2011. “SNP Signatures of Selection on Standing Genetic Variation and Their Association with Adaptive Phenotypes Along Gradients of Ecological Speciation in Lake Whitefish Species Pairs (Coregonus Spp.).” Molecular Ecology 20 (3): 545–559. doi:10.1111/j.1365-294X.2010.04952.x.

Rivas, Manuel A., Mélissa Beaudoin, Agnes Gardet, Christine Stevens, Yashoda Sharma, Clarence K. Zhang, Gabrielle Boucher, *et al.* 2011. “Deep Resequencing of GWAS Loci Identifies Independent Rare Variants Associated with Inflammatory Bowel Disease.” Nature Genetics 43 (11): 1066–1073. doi:10.1038/ng.952.

Rowe, H. C., S. Renaut, and A. Guggisberg. 2011. “RAD in the Realm of Next-generation Sequencing Technologies.” Molecular Ecology 20 (17): 3499–3502.

doi:10.1111/j.1365-294X.2011.05197.x.

Salzberg, Steven L., Adam M. Phillippy, Aleksey Zimin, Daniela Puiu, Tanja Magoc, Sergey Koren, Todd J. Treangen, *et al.* 2012. “GAGE: A Critical Evaluation of Genome Assemblies and Assembly Algorithms.” Genome Research 22 (3) (March 1): 557–567. doi:10.1101/gr.131383.111.

Schlötterer, Christian. “The Evolution of Molecular Markers — Just a Matter of Fashion?” Nature *Reviews Genetics* 5, no. 1 (January 2004): 63–69. doi:10.1038/nrg1249.

Thomas, J. W., J. W. Touchman, R. W. Blakesley, G. G. Bouffard, S. M. Beckstrom-Sternberg, E. H. Margulies, M. Blanchette, *et al.* 2003. “Comparative Analyses of Multi-species Sequences from Targeted Genomic Regions.” Nature 424 (6950) (August 14): 788–793. doi:10.1038/nature01858.

Trapnell, Cole, and Steven L Salzberg. 2009. “How to Map Billions of Short Reads onto Genomes.” Nature Biotechnology 27 (5) (May): 455–457. doi:10.1038/nbt0509-455.

Wang, Zhong, Mark Gerstein, and Michael Snyder. 2009. “RNA-Seq: a Revolutionary Tool for Transcriptomics.” Nature Reviews Genetics 10 (1) (January): 57–63. doi:10.1038/nrg2484.

Yi, Xin, Yu Liang, Emilia Huerta-Sanchez, Xin Jin, Zha Xi Ping Cuo, John E. Pool, Xun Xu, *et al.* 2010. “Sequencing of 50 Human Exomes Reveals Adaptation to High Altitude.” Science 329 (5987) (July 2): 75–78. doi:10.1126/science.1190371.

Zerbino, D. R., and E. Birney. 2008. “Velvet: Algorithms for de Novo Short Read Assembly Using de Bruijn Graphs.” Genome Research 18 (5) (February 21): 821–829. doi:10.1101/gr.074492.107.

Chapter 2: A Computer Program for Genome-Wide and Candidate Gene Exon Sampling for Targeted Next-Generation Sequencing

This chapter is in preparation for submission for publication with co-authors[1].

Abstract

The computer program ExonSampler automates the sampling of thousands of exon sequences from publicly available reference genome sequences and gene annotation databases. It was designed to provide exons for the promising, next-generation sequencing method called *exon capture*. The exon sequences can be sampled by using a list of gene name abbreviations (e.g., IFNG, TLR1), or by sampling exons from genes spaced evenly across chromosomes. It provides a list of genomic coordinates (a bed file), as well as a set of sequences in fasta format. User-adjustable parameters for sampling (collecting) exons include a minimum and maximum acceptable exon length, maximum number of exonic base pairs (bps) to sample per gene, and maximum total bp for the entire collection. It allows for partial sampling of very large exons. It can preferentially sample upstream (5´) exons, downstream (3´) exons, both external exons, or all internal exons. It is written in the Python programming language using its free, web-distributed libraries. We describe the use of ExonSampler to collect exon sequences from the domestic cow (*Bos taurus*) genome for the design of an exon capture microarray to sequence exons from related species: the domestic cow, *Bos indicus*, and wild bison (*Bison bison*), which have no genome sequences available. We collected ~10% of the exome (~3 million bp), including 155 candidate genes, and ~16,000 exons evenly spaced genome-wide. We prioritized 5´ exons to facilitate discovery and genotyping of SNPs near upstream gene regulatory DNA sequences, which control gene expression and are often under natural selection.

Collecting exons for capture

ExonSampler automates the collection of exon sequences from nucleotide sequence databases. While useful for any project requiring exon sequences we designed the program to provide sequences needed to make oligonucleotide probes for the laboratory method called *exon capture*, which enriches genomic DNA for targeted exon sequences

[1] A. Beja-Pereira, Centro de Investigação em Biodiversidade e Recursos Genéticos, Universidade do Porto, and G. Luikart, Flathead Lake Biological Station, Division of Biological Sciences, University of Montana. I also wish to acknowledge the help of Jesse Johnson, who commented substantially on the manuscript and advised on program design.

through probe-hybridization. The exon-enriched DNA sample is sequenced on a next-generation sequencing platform. Exon capture, often targeting all known exons (collectively, the exome), is used increasingly as a cost-efficient assay of genetic variation, commonly single nucleotide polymorphisms (SNPs), across entire eukaryotic genomes (S. B. Ng *et al.,* 2009; Hodges *et al.*, 2007; Cosart *et al*. 2011;). We found no existing program that met our goal of sampling only part of an exome, limiting total base pairs while automatically sampling sequences from thousands of genes across all chromosomes and spaced as evenly as possible, to facilitate genome wide scan experiments (Nadeau *et al*., 2012, Ng *et al.,* 2010).

For an exon capture from cows and bison (Chapter 3; Cosart *et al.*, 2011), the program collected 15,583 exons, totaling ~2.9 million base pairs (Mb) from the BTau4.0 *Bos taurus* genome sequence (Elsik *et al.*, 2009). Sequences came from 2,522 genes out of ~10,000 or so RefGene annotations from RefSeq mRNA's (Pruitt *et al.*, 2011), aligned to the genome by the BLAT program (Kent, 2002) at UCSC's Genome Browser Database (Fujita *et al.*, 2010). We added 48 candidate gene annotations from NCBI's Entrez gene database (Maglott *et al.*, 2006), fetched over the internet, using code not part of the current program.

Guided by our selection parameters, the ExonSampler provided exons sufficient to capture and sequence thousands of putative SNPs, genome wide. These included SNPs in selected candidate genes in all three individuals of our sampled species: taurine and zebu cows, and an American bison. ExonSampler provided a genome-wide target area (total exon bps collected) large enough for significant SNP discovery, while small enough (about 0.1% of the total genomic bps in the cow, and ~15% of the ~20Mb refGene-annotated exome) for cost-effective sequencing in many individuals. Sequencing genetic markers in many individuals reduces ascertainment bias in population genetic studies (Morin *et al.*, 2004) while revealing rare alleles.

The program allows users to balance the total exonic base pairs targeted with the number of samples to be sequenced by setting a limit on total base pairs for the exon collection and/or the number of base pairs per gene (Table 2.1). Testing and verification of the program included using the BLAST alignment program (McGinnis and Madden, 2004) to verify that sequences produced by the program were at the correct genomic coordinates, and use of the UCSC genome browser to verify that coordinates produced by the program were correctly associated with exon intervals.

Table 2.1: Exon parameters used by ExonSampler. The middle column gives the values we used in our study (Cosart *et al.* 2011). Asterisks indicate the use of BLAST to ensure a collection of sequences without high similarity to each other. As used in our study (Cosart *et al.*, 2011), each exon was blasted to the current collection before being added to the collection, and was discarded if too similar (see thresholds in the table). See "Optional features" below, for the current implementation, which uses the whole genome for blast verification of sequence uniqueness.

Parameter	**Our chosen value**	**Note**
Total base pairs (bps) to be collected (approximate)	3,000,000	About 10% to 15% of the exome of cattle (and many mammals).
Max. bps per gene	1,500	We avoided sampling large numbers of bps per gene to maximize the number of genes and chromosomal locations that could be sampled.
Min. exon length (bps)	40	This was imposed only on the non-candidate genes collected. It is an arbitrary value, to avoid baiting for small coding sequences. We imposed no minimum exon size for our candidate genes.
Max. exon length (bps)	1,500	If a gene had no exon under 1,500 bps, *ExonSampler* collected half of the maximum exon length (750 bps) at each end of one of the large exons, preferring to sample the upstream (5′) exon.
Exon preference	Upstream	The upstream (5′) exon was collected before any others. Other choices are downstream, external, internal, or all exons (i.e. no preferential collection)
Max. BLAST align length	40	No sequences are rejected on the basis of an alignment length under this value.*
Max. BLAST percent identity	90	Exons are discarded if the alignment exceeds the maximum alignment length and the percent identity exceeds this threshold.*
Max. BLAST bit score	N/A	Not used in Cosart *et al.* Exon sequences are discarded if they align to different loci than their own with length exceeding the value of "Max. BLAST align length" and with a bit score exceeding this value.
Min. BLAST e-value	N/A	Not used in Cosart *et al.* Alignments of 2 different loci below this value, and exceeding the other thresholds, are discarded.

Input: genome sequences and gene annotations.

The program uses exon annotation information from refGene.txt files, and chromosome sequences, both freely downloadable from the annotation directories for various genomes at the UCSC Genome Bioinformatics site http://hgdownload.cse.ucsc.edu/downloads.html. Each refGene.txt file is associated with one genome sequence build. ExonSampler uses the chromosome sequences on which the annotations are based to provide sequences corresponding to the exon intervals, in fasta file format. The program also requires a genome sequence in the fasta file format, also available at the UCSC site. If named genes are to be collected, the program requires a file listing gene name abbreviations (e.g., IFNG, BOLA-DOB, PRNP, etc.), one abbreviation to a line. Input files are described in documentation accompanying the program, which describes installing the program, data, and the format of input and output files.

Execution: collecting exon sequences and information

A run of the program uses one of two sampling strategies to obtain exons, parameterized by the user in a configuration file:

1. List-based Sampling: Genes are chosen using a list of gene symbols. For each gene symbol in the user-supplied list file, exons are collected for gene annotations (one or more) with a perfectly matching symbol. Any symbols without matches in the annotation file are recorded in a log.

2. Even Sampling: Genes are chosen by an even sampling over the whole genome. On a given chromosome, the genes are collected serially, selecting that gene whose midpoint is nearest the midpoint of the largest contiguous stretch of unsampled base pairs on the chromosome (Fig. 2.1). Chromosomes are visited by turns and a gene selected on each visit. They are sampled proportionately to their base-pair length.

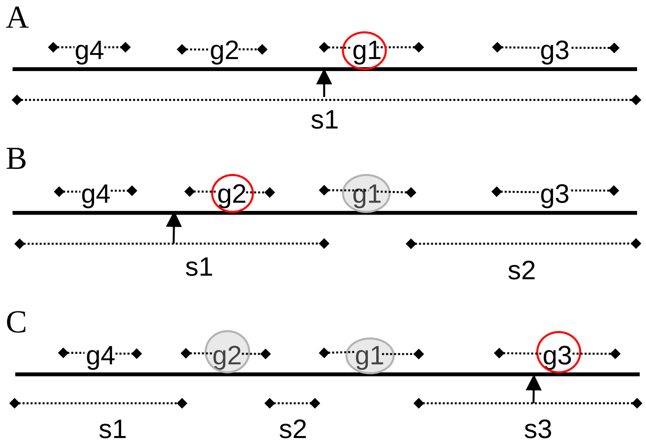


Figure 2.1: Algorithm for gene selection for collecting exons evenly across chromosomes. The program visits chromosomes multiple times, proportionally to chromosome length, and selects genes nearest to the middle of largest, unsampled spans. (A) The first selection on the hypothetical, unsampled chromosome above is gene 1 (g1, circled in red), whose midpoint is closest to the chromosome's midpoint (s1). (B) g1's selection divides the chromosome into two spans, s1 and s2, with s1 the longest. g2 is the gene sampled on the second visit, since its midpoint is closest to the midpoint of s1. (C) Of the three spans resulting from selections above, s3 is the longest. On the 3rd visit, g3 is selected, since its midpoint is closest to that of s3. Gene selection and exon sampling continues on each chromosome accordingly until base pair limits are reached or every gene has been selected.

In any session the user can indicate a preferred exon type, categorized as *upstream* (the gene's 5′ external exon), *downstream* (3′ external exon), *external* (5′ and 3′), *internal* exons, or *all* (indicating no preference). Exons of the preferred type are put at the head of a list as potential additions to the collection for a particular gene. Nonpreferred types are collected after the maximal number of preferred types is collected, without exceeding any exonic or per-gene base pair limit. Within preferred or nonpreferred groups, exons are shuffled to randomize the order in which they are considered for addition to the collection. As each exon is considered, its length (in bps) is added to a running total for the gene. Exons with lengths that are over the user-set maximum exon size, and exons that put the total over the per-gene bp limit, are skipped. In genes with no exon meeting exon size limits, if at least one exon exceeds the user-set maximum, the first such exon in the randomized exon order is sampled from both ends, taking half of the maximum bps-per-exon from each end. Collection completes when either all genes have been sampled, or adding a new exon takes the total base pairs collected over the user-set limit.

Output: exon positions and sequences

For each session, ExonSampler generates four files:

1. A coordinate file in the bed file format, giving each exon's chromosome, start position, stop position, an exon name composed of the gene name, an number *n* indicating that the gene was the n^{th} gene of that name (abbreviation) collected, a number indicating the exon's position in the annotation, numbered from the upstream to the downstream end, and the total exons annotated for the gene.
2. A fasta file, giving the genomic DNA sequence for each exon. The fasta ID lines in this file also contain the genomic coordinates and the exon name.
3. A log file that lists every exon that was under consideration for addition to the collection. This file also notes each exon's strand orientation and, if an exon was rejected, the reason for rejection (e.g. the exon is below or above the exon length limits).
4. A message file is also provided for each session, listing the type of sampling session used, and the parameters set by the user, along with summary numbers such as total base pairs collected.

Optional Blast Alignment

The program will optionally perform a BLAST sequence alignment (McGinnis and Madden, 2004) of the collected exons to a BLAST database of the genome sequences from which the exons were collected. The program discards collected exons that have at least one BLAST hit, other than the self-alignment of the exon to its source location on the genome, at or above a minimum sequence percent identity, alignment length, and bit score, and below a given e-value. This is meant to filter out ambiguous or similar sequences that produce reads that have 2 or more maximally-scoring alignments on the genome. This is often the case for reads that represent recent gene duplication. Exons discarded by the BLAST test are recorded in the log file.

Program results for an example study

In making an exon collection of target sequences for the design of our bovine exon capture microarray (Cosart *et al.*, 2011) we used both even sampling across all chromosomes and a list of candidate genes of special interest (Fig. 2.2, and listed in Table S1 in supplementary material). For even sampling over the chromosomes we set the program's parameters to prefer the upstream exon (to identify SNPs in/near the upstream regulatory region), and to collect no more than 1,500 bps per gene (to distribute sequencing across many genes and likely discover SNPs in each gene). In this selection process, if a candidate gene had no exon of an allowed length, but had at least one exon over the maximum length, *ExonSampler* split the exon creating an upstream 750 bp sequence and downstream 750 bps. We collected all exons in the candidate gene list, with no restrictions on exon size or type.

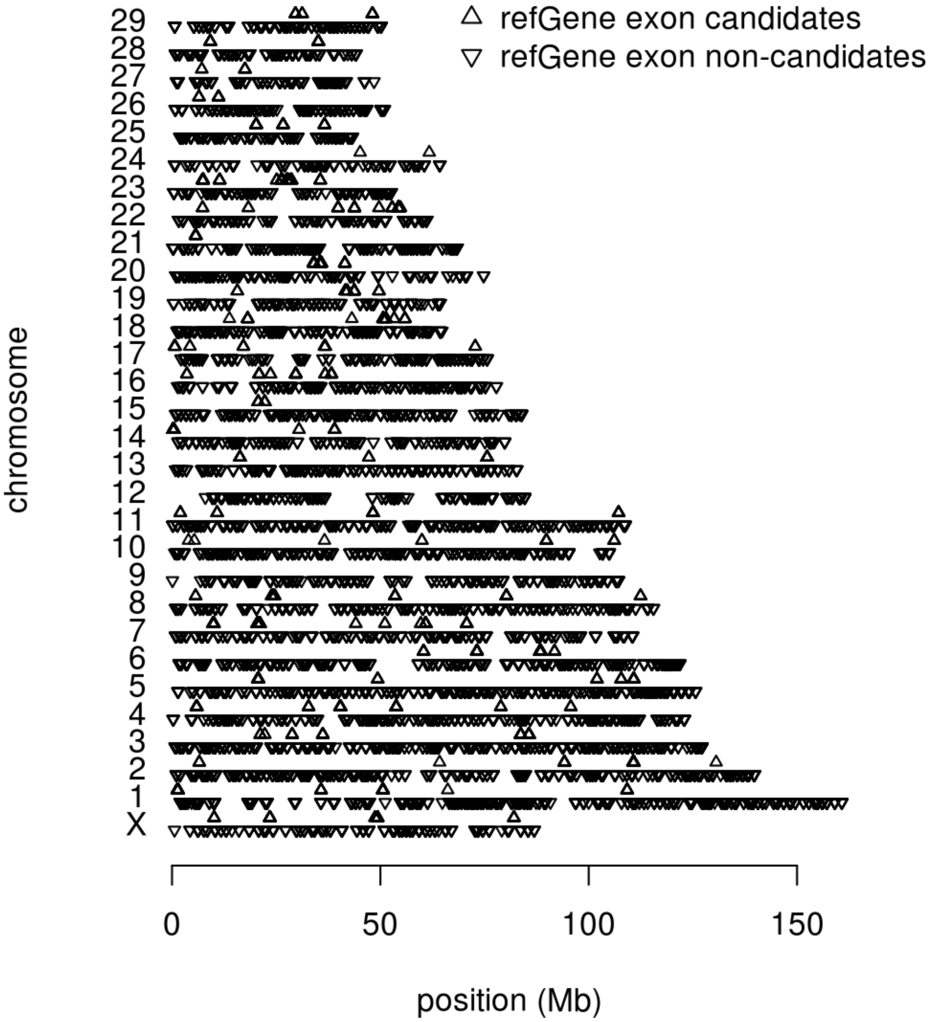


Figure 2.2: Gene locations for the 15,583 exons collected by ExonSampler. The source genome was the cow, btau 4.0. Exons were collected with the program parameters as described above. The collection totaled 2,522 genes, of which 155 are candidate genes (upward-pointing triangles), that is, genes chosen for their known or suspected association with adaptive traits (listed in supplementary Table S2.1). The whole collection totals 2,880,061 base pairs. As detailed in the legend, plotted symbols candidate versus non-candidate genes.

The resulting collection has exons spread evenly across all the published chromosome sequences (Fig. 2.2). The distribution of the lengths of the sampled exons had a similar median length, but a lower mean length, than the entire exome in the refGene annotations, likely reflecting the end-sampling of exons in genes with no exons under the maximum exon bp limit (Table 2.2).

Table 2.2: Distribution of lengths (bp) of exons collected by ExonSampler. These are compared to the exon length distribution for all annotated exons in the refGene annotations. All numbers are counts of base pairs.

	All refGene Exons	**Sampled Exons**
Maximum	12,590	4,095
Third Quartile	189	181.5
Mean	229.5	183.8
Median	129	126
First Quartile	90	89
Minimum	3	11

The per-gene bp limit also reduced the mean number of exons-per-gene from 8.3 for the whole RefGene annotation to 6.2 for the ExonSampler's collection. In the even-sampling session most of the exons of a given gene were collected if the gene had between 1 and 10 exons (Fig. 2.3). For genes with larger numbers of exons, generally about 10 exons were collected (Fig. 2.3). 114 exons were discarded (their rejection recorded in the programs log file output) for their high BLAST similarity to already-collected exons, out of 15,697 total exons selected.

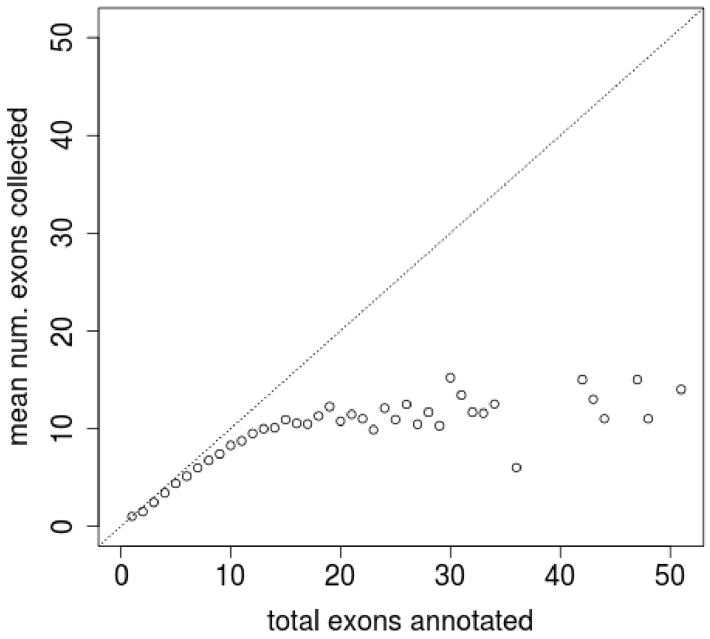


Figure 2.3: Exons per gene, collected versus annotated. The plateau of mean exons collected near 10-12 shows the effect of the 1,500bp per-gene limit. Y axis values are means of the total exons collected for all genes sampled with the same number of total exons annotated (x-axis value).

Future development

Enhancements planned for the program include adding NCBI seq_gene.md files, available at the NCBI Genomes site (http://www.ncbi.nlm.nih.gov/genome) as sources for annotations. Other enhancements planned are to collect flanking (noncoding) sequence of a user-specified size to be included upstream and/or downstream of collected exon sequences. Exon capture targeting widened flanks can assay (1) bps flanking genes (e.g., to measure linkage disequilibrium due to selective sweeps, or to capture upstream regulatory regions) and (2) bps between genes to allow for sampling of neutral DNA or development of markers in gene-poor chromosomal regions.

References

Cosart,T. *et al.* (2011) Exome-wide DNA Capture and Next Generation Sequencing in Domestic and Wild Species. BMC genomics, 12, 347.

Elsik,C.G. *et al.* (2009) The Genome Sequence of Taurine Cattle: A window to ruminant biology and evolution. Science (New York, N.Y.), 324, 522–528.
Fujita,P.A. *et al.* (2010) The UCSC Genome Browser database: update 2011. Nucleic Acids Research, 39, D876–D882.

Hodges,E. *et al.* (2007) Genome-wide in situ exon capture for selective resequencing. Nat Genet, 39, 1522–1527.

McGinnis,S. and Madden,T.L. (2004) BLAST: at the core of a powerful and diverse set of sequence analysis tools. Nucleic Acids Research, 32, W20.

Nadeau, N. J. *et al.* (2012). Genomic islands of divergence in hybridizing Heliconius butterflies identified by large-scale targeted sequencing. Phil. Trans. R. Soc. B**,** 367**,** 343–353

Ng,S.B. *et al.* (2009) Targeted capture and massively parallel sequencing of 12 human exomes. Nature, 461, 272–276.

Ng, S. B. *et al.* (2010) Exome sequencing identifies the cause of a mendelian disorder. Nat Genet, 42**,** 30–35.

Pruitt,K.D. *et al.* (2011) NCBI Reference Sequences (RefSeq): current status, new features and genome annotation policy. Nucleic Acids Research, 40, D130–D135.

Supplementary materials

Table S2.1: Candidate gene symbols and brief descriptions. The candidate genes whose exons ExonSampler collected for our cow and bison exon capture (Cosart *et al*. 2011).

Abbreviation	Name and/or description
ALCAM	activated leukocyte cell adhesion molecule
ADRB2	adrenergic_ beta-2-_ receptor_ surface
B2M	beta-2-microglobulin
BOLA1	bolA homolog 1 (E. coli)
BOLA3	bolA homolog 3 (E. coli)
CPE	carboxypeptidase E
CSN1S1	casein alpha s1
CSN1S2	casein alpha-S2
CSN2	casein beta
CSN3	casein kappa
CASP8	caspase 8_ apoptosis-related cysteine peptidase
CAMP	cathelicidin antimicrobial peptide
CAV3	caveolin 3
CEBPA	CCAAT/enhancer binding protein (C/EBP)_ alpha
CD14	CD14 molecule
CD2	CD2 molecule
CD40LG	CD40 ligand (TNF superfamily_ member 5_ hyper-IgM syndrome)
CD40	CD40 molecule_ TNF receptor superfamily member 5
CD69	CD69 molecule
CCL2	chemokine (C-C motif) ligand 2
CCR4	chemokine (C-C motif) receptor 4
CCR5	chemokine (C-C motif) receptor 5
CCR7	chemokine (C-C motif) receptor 7
CCR9	chemokine (C-C motif) receptor 9
CXCR3	chemokine (C-X-C motif) receptor 3
CXCR4	chemokine (C-X-C motif) receptor 4
CXCR6	chemokine (C-X-C motif) receptor 6
CSF1	colony stimulating factor 1 (macrophage)
CSF1R	colony stimulating factor 1 receptor_ formerly McDonough feline sarcoma viral (v-fms) oncogene homolog

CSF2	colony stimulating factor 2 (granulocyte-macrophage)
CSF3	colony stimulating factor 3 (granulocyte)
C6	complement component 6
CRH	corticotropin releasing hormone
DEFB	defensin_ beta
DGAT1	diacylglycerol O-acyltransferase homolog 1 (mouse)
DRD1	dopamine receptor D1
DRD2	dopamine receptor D2
FAS	Fas (TNF receptor superfamily_ member 6)
FASLG	Fas ligand
FEZF2	FEZ family zinc finger 2
FLT3LG	fms-related tyrosine kinase 3 ligand
GDF9	growth differentiation factor 9
GH	growth hormone
GHR	growth hormone receptor
HGF	hepatocyte growth factor (hepapoietin A; scatter factor)
HNRNPU	heterogeneous nuclear ribonucleoprotein U (scaffold attachment factor A)
IGHMBP2	immunoglobulin mu binding protein 2
IGFBP3	insulin-like growth factor binding protein 3
IFNAR2	interferon (alpha_ beta and omega) receptor 2
IFN-tau-c1	interferon tau c1
IFNAR1	interferon_ alpha; receptor
IFN1@	interferon_ alpha_ leukocyte
IFNB1	interferon_ beta 1_ fibroblast
IFNG	interferon_ gamma
IFNW1	interferon_ omega 1
IFN-a	interferon-alpha
IL1A	interleukin 1_ alpha
IL1B	interleukin 1_ beta
IL10	interleukin 10
IL10RB	interleukin 10 receptor_ beta
IL11RA	interleukin 11 receptor_ alpha
IL12RB2	interleukin 12 receptor_ beta 2
IL12A	interleukin 12A (natural killer cell stimulatory factor 1_ cytotoxic lymphocyte maturation factor 1_ p35)
IL12B	interleukin 12B (natural killer cell stimulatory factor 2_ cytotoxic lymphocyte maturation factor 2_ p40)
IL13	interleukin 13
IL15	interleukin 15
IL17A	interleukin 17A
IL18	interleukin 18 (interferon-gamma-inducing factor)

IL2	interleukin 2
IL2RA	interleukin 2 receptor_ alpha
IL2RG	interleukin 2 receptor_ gamma (severe combined immunodeficiency)
IL21	interleukin 21
IL27RA	interleukin 27 receptor_ alpha
IL3	interleukin 3
IL4	interleukin 4
IL4R	interleukin 4 receptor
IL5	interleukin 5
IL6	interleukin 6 (interferon_ beta 2)
IL7	interleukin 7
IL8	interleukin 8
IL8RA	interleukin 8 receptor_ alpha
IL8RB	interleukin 8 receptor_ beta
IRAK1	interleukin-1 receptor-associated kinase 1
KDR	kinase insert domain receptor (a type III receptor tyrosine kinase)
KITLG	KIT ligand
LGB	lactoglobulin_ beta
LTF	lactotransferrin
LEP	leptin (obesity homolog_ mouse)
LEPR	leptin receptor
LIF	leukemia inhibitory factor (cholinergic differentiation factor)
LIPE	lipase_ hormone-sensitive
LTA	lymphotoxin alpha (TNF superfamily_ member 1)
LYSMD1	LysM_ putative peptidoglycan-binding_ domain containing 1
LYSMD2	LysM_ putative peptidoglycan-binding_ domain containing 2
BoLA	major histocompatibility complex_ class I_ A
BOLA	major histocompatibility complex_ class I_ A
HLA-A	major histocompatibility complex_ class I_ A
BOLA-DMA	major histocompatibility complex_ class II_ DM alpha-chain_ expressed
BOLA-DMB	major histocompatibility complex_ class II_ DM beta-chain_ expressed
BOLA-DQA5	major histocompatibility complex_ class II_ DQ alpha 5
BOLA-DQB	major histocompatibility complex_ class II_ DQ beta
BOLA-DRA	major histocompatibility complex_ class II_ DR alpha
BOLA-DYA	major histocompatibility complex_ class II_ DY alpha
MAL	mal_ T-cell differentiation protein
MBL2	mannose-binding lectin (protein C) 2_ soluble (opsonic defect)
MGP	matrix Gla protein
MC1R	melanocortin 1 receptor (alpha melanocyte stimulating hormone receptor)

MC4R	melanocortin 4 receptor
MC5R	melanocortin 5 receptor
MET	met proto-oncogene (hepatocyte growth factor receptor)
MEF2A	myocyte enhancer factor 2A
MSTN	myostatin
BOLA-NC1	non-classical MHC class I antigen
NFKBIL2	nuclear factor of kappa light polypeptide gene enhancer in B-cells inhibitor-like 2
NOD2	nucleotide-binding oligomerization domain containing 2
OSTF1	osteoclast stimulating factor 1
PGLYRP1	peptidoglycan recognition protein 1
PIM1	pim-1 oncogene
POU1F1	POU class 1 homeobox 1
PRNP	prion protein (p27-30) (Creutzfeldt-Jakob disease_ Gerstmann-Strausler-Scheinker syndrome_ fatal familial insomnia)
PRL	prolactin
PRLR	prolactin receptor
PRKAA1	protein kinase_ AMP-activated_ alpha 1 catalytic subunit
PDHB	pyruvate dehydrogenase (lipoamide) beta
SCRG1	scrapie responsive protein 1
STAT5A	signal transducer and activator of transcription 5A
SLC11A1	solute carrier family 11 (proton-coupled divalent metal ion transporters)_ member 1
SPA17	sperm autoantigenic protein 17
SFN	stratifin
SP-A	surfactant protein A
SFTPD	surfactant_ pulmonary-associated protein D
TIRAP	toll-interleukin 1 receptor (TIR) domain containing adaptor protein
TLR1	toll-like receptor 1
TLR10	Toll-like receptor 10
TLR2	toll-like receptor 2
TLR3	toll-like receptor 3
TLR4	toll-like receptor 4
TLR5	Toll-like receptor 5
TLR6	toll-like receptor 6
TLR7	toll-like receptor 7
TLR8	toll-like receptor 8
TLR9	toll-like receptor 9
TICAM2	toll-like receptor adaptor molecule 2

TGFB2	transforming growth factor_ beta 2
TGFB3	transforming growth factor_ beta 3
TNF	tumor necrosis factor (TNF superfamily_ member 2)
TNFRSF1A	tumor necrosis factor receptor superfamily_ member 1A
TNFRSF1B	tumor necrosis factor receptor superfamily_ member 1B
ZPBP	zona pellucida binding protein
ZP2	zona pellucida glycoprotein 2 (sperm receptor)
ZP3	zona pellucida glycoprotein 3 (sperm receptor)
ZP4	zona pellucida glycoprotein 4
BOLA-DOB	HLA class II histocompatibility antigen, DO beta chain precursor
BoLA-DRB3	major histocompatibility complex, class II, DRB3
BSPH1	bovine seminal plasma protein homolog 1-like

Chapter 3, Published Article: Exome-wide DNA capture and next generation sequencing in domestic and wild species

Ted Cosart[1,2,3]*, **Albano Beja-Pereira**[3]*, **Shanyuan Chen**[3], **Sarah B Ng**[4], **Jay Shendure**[4] and **Gordon Luikart**[3,5]

- * Corresponding authors: Ted Cosart ted.cosart@umontana.edu - Albano Beja-Pereira albanobp@fc.up.pt

Author Affiliations

[1] Department of Computer Science, University of Montana, Missoula, MT, USA

[2] Montana-Ecology of Infectious Diseases Program, The University of Montana, Missoula, MT, USA

[3] Centro de Investigação em Biodiversidade e Recursos Genéticos (CIBIO), Universidade do Porto, Rua Padre Armando Quintas 7, Campus Agrário de Vairão, 4485-661 Vairão, Portugal

[4] Department of Genome Sciences, University of Washington, Seattle, WA, 98195, USA

[5] Flathead Lake Biological Station and Division of Biological Sciences, University of Montana, Polson, MT 59860, USA

For all author emails, please log on.

BMC Genomics 2011, **12**:347 doi:10.1186/1471-2164-12-347

The electronic version of this article is the complete one and can be found online at: http://www.biomedcentral.com/1471-2164/12/347

Received: 23 November 2010

Accepted: 5 July 2011

Published:5 July 2011

Abstract

Background

Gene-targeted and genome-wide markers are crucial to advance evolutionary biology, agriculture, and biodiversity conservation by improving our understanding of genetic processes underlying adaptation and speciation. Unfortunately, for eukaryotic species with large genomes it remains costly to obtain genome sequences and to develop genome resources such as genome-wide SNPs. A method is needed to allow gene-targeted, next-generation sequencing that is flexible enough to include any gene or number of genes, unlike transcriptome sequencing. Such a method would allow sequencing of many individuals, avoiding ascertainment bias in subsequent population genetic analyses.

We demonstrate the usefulness of a recent technology, exon capture, for genome-wide, gene-targeted marker discovery in species with no genome resources. We use coding gene sequences from the domestic cow genome sequence (*Bos taurus*) to capture (enrich for), and subsequently sequence, thousands of exons of *B. taurus*, *B. indicus*, and *Bison bison* (wild bison). Our capture array has probes for 16,131 exons in 2,570 genes, including 203 candidate genes with known function and of interest for their association with disease and other fitness traits.

Results

We successfully sequenced and mapped exon sequences from across the 29 autosomes and X chromosome in the *B. taurus* genome sequence. Exon capture and high-throughput sequencing identified thousands of putative SNPs spread evenly across all reference chromosomes, in all three individuals, including hundreds of SNPs in our targeted candidate genes.

Conclusions

This study shows exon capture can be customized for SNP discovery in many individuals and for nonmodel species without genomic resources. Our captured exome subset was small enough for affordable next-generation sequencing, and successfully captured exons from a divergent wild species using the domestic cow genome as reference.

Background

Our understanding of the molecular, genetic basis of adaptations and phenotypic differentiation among individuals will advance quickly thanks to new molecular techniques. This understanding is crucial given that accelerating environmental change and human population growth are increasingly threatening natural populations of fish and wildlife as well as increasing demands for agricultural production in domesticated species. This makes it urgent in many wild and domestic species to investigate the genetic basis of fitness, adaptation, and disease resistance [1], and to discover adaptive genes and speciation genes, i.e., the "loci of evolution" [2].

Understanding the genetic basis of phenotypes generally requires genotyping thousands of gene-targeted loci, genome-wide. Despite the declining costs of next generation DNA sequencing (summarized in [3]), it remains costly enough to prohibit analyzing large portions of genomes in numerous individuals as is required for population studies (e.g. population genomics, [4]). Fortunately, with coding gene sequences (the exome) comprising a mere 2% of the typical eukaryotic genome, and the development of techniques for isolating exome DNA, re-sequencing coding portions genome-wide can be done at a reasonable per-sample cost, locating thousands of informative gene markers. Because exon sequences are relatively conserved we hypothesized that most exons from one species (e.g. with a sequenced genome) could be used to capture exons from another species for use in next generation sequencing for SNP discovery.

Exon capture enriches for exon DNA by simultaneous hybridization of fragmented genomic DNA from the study individual to many thousands of oligonucleotide probes (e.g. refs. [5,6]) that are complementary to gene-coding (exon) sequences. The captured fragments are then sequenced in parallel on next-generation sequencing platforms. Exon capture has been tested almost exclusively in model species (e.g. refs. [7-9]), typically baiting either the whole exome or a single chromosomal region. Facilitated by availability of genome sequences for the target organism, such studies leave untested the potential application of exon capture to a wider variety of organisms. Probe design for exome-wide capture requires knowledge of thousands of exon sequences. With few fully sequenced eukaryotic genome sequences available (to date, 40 complete, 425 draft whole genome sequences are found at NCBI's Entrez gene service), it would appear to be useful for only a small proportion of eukaryotic species. Even if 10,000 vertebrate genomes are eventually sequenced [10], there would still remain tens of thousands of vertebrate species without genome sequences or any genome resources.

Here we show that the exon capture method has a more general application, reporting exon capture in two livestock species, *Bos taurus* (taurine cattle) and *Bos indicus* (zebu cattle), and one wildlife species, *Bison bison* (American bison). We conducted all three captures through hybridization to sequences from the published *Bos taurus* genome [11]. We baited a small genome-wide fraction of the exome, sampling exons in about 10% of the taurine genome's estimated minimum total of 22,000 genes [11]. Our results demonstrate that genetic divergence between a reference genome and individuals queried does not prohibit exome-wide identification of candidate SNPs and differences (e.g., substitutions) in nonmodel species. This suggests the method can be applied to many domestic and wildlife species lacking sequenced genomes. Further, we found that baiting a small fraction of the exome yielded thousands of candidate heterozygous SNPs.

Results and discussion

We sequenced genomic DNA from our three individuals, enriched for 16,131 exons (~ 3 million base pairs) by hybridization to probes on a microarray. Reference exon sequences came from sampling an average of 6 exons from each of 2,367 genes spread evenly across the 29 autosomes and the X chromosome. We also chose 203 candidate genes with known associations with disease susceptibility and other important traits. For all candidate genes, the entire exon sets were targeted for capture.

Illumina Genome Analyzer sequencing of the enriched DNA, followed by mapping of the 36 base-pair, single end sequence reads and consensus genotyping with Maq software [12], yielded high-confidence nucleotide base calls (see Methods for our base calling criteria) comprising 77% of our targeted exonic positions in the taurine, 80% in the zebu and 82% in the bison (Figure 1). The called single-nucleotide genotypes differed from the reference across the genome at positions totalling 11,061 in the bison, 5,524 in the zebu, and 3,854 in the taurine (Figure 2a).

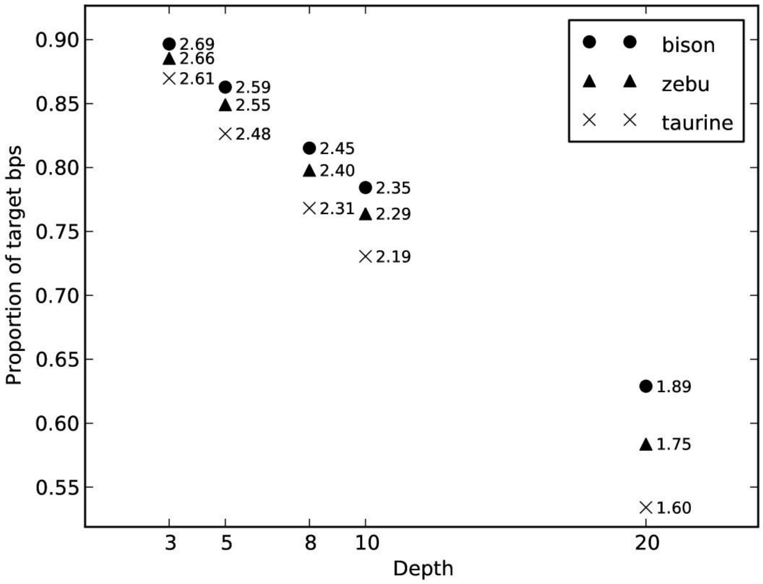


Figure 3.1. Proportion of targeted exonic base pairs with a consensus genotype. All have a Phred-like quality score of at least 30. Total base pair counts, in millions, are plotted at selected minimum depths of coverage. Our estimates of exonic fixed differences and SNPs are based on consensus genotypes with coverage of at least 8 ×.

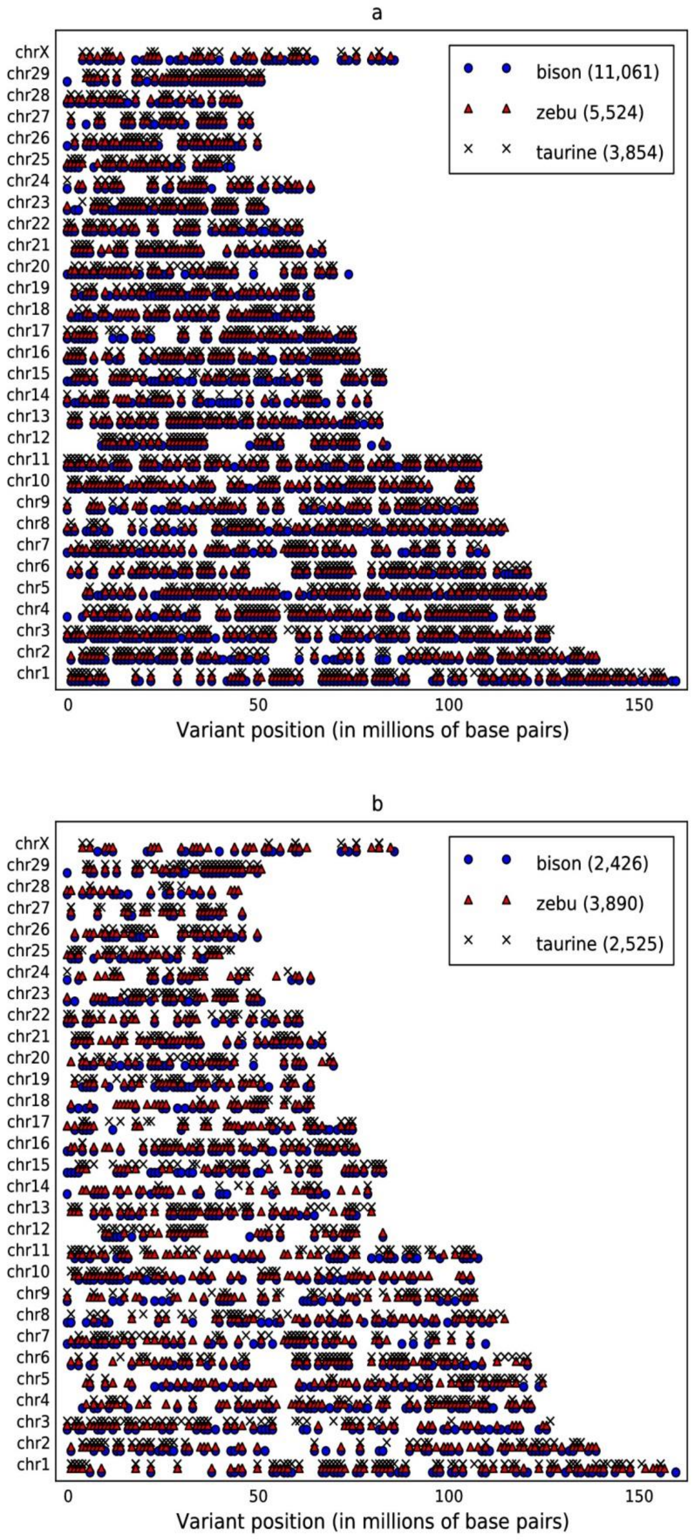

a
chrX
chr29
chr28
chr27
chr26
chr25
chr24
chr23
chr22
chr21
chr20
chr19
chr18
chr17
chr16
chr15
chr14
chr13
chr12
chr11
chr10
chr9
chr8
chr7
chr6
chr5
chr4
chr3
chr2
chr1
● bison (11,061)
▲ zebu (5,524)
× taurine (3,854)
0
50
100
150
Variant position (in millions of base pairs)
b
chrX
chr29
chr28
chr27
chr26
chr25
chr24
chr23
chr22
chr21
chr20
chr19
chr18
chr17
chr16
chr15
chr14
chr13
chr12
chr11
chr10
chr9
chr8
chr7
chr6
chr5
chr4
chr3
chr2
chr1
● bison (2,426)
▲ zebu (3,890)
× taurine (2,525)
0
50
100
150
Variant position (in millions of base pairs)

Figure 3.2. Chromosomal positions. a, all consensus base differences from the reference taurine genome, and **b**, heterozygous SNPs only. Both maps show consensus bases with at least 8 × coverage and Phred-like quality of at least 30. Numbers in the legends give totals for variants for each species.

As a percentage of total targeted nucleotides with high confidence base calls, 0.5% of the bison calls differed from the reference taurine, compared to 0.2% for each of the two *Bos* species. The higher percentage of differences in the bison is expected in light of its one to two million years of genetic separation from the taurine cattle (*B. taurus*) [13]. The divergence between the two species in the target region of the exome estimated by our results, about 5 differences in every thousand bases, is likely conservative, given the limitations of mapping software in accounting for real base differences versus incorrect sequencer base calls (discussed in Methods and in [12]).

In our 203 candidate genes, we identified 339 putative heterozygous SNPs among 96 genes in the bison, 598 heterozygous SNPs in 123 genes in the zebu, and 372 in 92 genes in the taurine. It is encouraging that from only one individual zebu, for example, we find high-confidence SNP calls in 60% of our 203 candidate genes of interest for future research. For all targeted base pairs, 2,525 heterozygous positions were called in our taurine, 3,890 in the zebu, and 2,426 in the bison (Figure 2b, Table 1). Concordance of some of our called single-base differences with published SNPs is indicated by the 545 (about 14%) of our taurine variant calls matching in position and all but one allele among 1.8 million NCBI dbSNP [14] records positioned on the same reference genome used in our study. As expected we found lower dbSNP concordance in our non-taurus individuals; about 11% among our zebu's called differences were matched in dbSNP and 4% of our bison's SNP calls had matches (Table 1).

Table 3.1. Variant Summary.
For each individual, total consensus bases different from the reference, for heterozygous SNPs, fixed differences, and concordance with 1.8 million entries in NCBI's dbSNP database. Total differences from the reference are also given as a percentage of total genotyped bases. dbSNP position matches are also given as a percentage of total differences. dbSNP allele mismatches give the number of alleles that differed from a dbSNP allele while matching its position; for example, in the bison, of the 483 positions matching SNPs in dbSNP, 10 showed alleles different than those listed at dbSNP.

	Bison	Zebu	Taurine
Heterozygous SNPs	2,426	3,890	2,525
Fixed differences	8,635	1,634	1,329
Total differences	11,061 (0.45%)	5,524 (0.23%)	3,854 (0.16%)
Total genotyped bases	2,447,500	2,395,651	2,306,566
dbSNP position matches	483 (4.37%)	594 (10.75%)	545 (14.14%)
dbSNP allele mismatches	10	4	1

Conclusions

Our results demonstrate two novel strategies for exon capture: (1) Sampling a small but genome-wide subset of the exome for discovery of thousands of putative SNPs, and (2) successful bait and capture across relatively divergent genomes. Result (1) reduces the cost of sequencing the capture products, making genome-wide SNP discovery more affordable. Exon capture with a subset of exons can complement large genotyping projects (e.g. in [15]) by facilitating discovery of thousands of SNPs based on assaying many individuals to avoid ascertainment bias in population genetic inferences [16]. Further, it allows genotyping of both candidate genes and genome-wide loci, combining the strengths of the candidate gene and genome scan approaches commonly used to identify adaptive and economically important loci.

Result (2) makes feasible these analyses in natural populations of divergent species with lesser-known genomes and from diverse environments worldwide, e.g. domestic and wild bovids from Siberia to the tropics. The conservation of exon sequences appears sufficient for the method to enable genome-wide studies based on probing across taxa as phylogenetically divergent as American bison and taurine cattle. Future research should test increased divergence between organisms referenced and baited to see how wide a taxonomic distance the method can bridge.

With success across many taxa while targeting a high value part of the exome small enough for affordable next-generation sequencing of many individuals, exon capture can be a powerful application of high-throughput genomics to the genetic analysis of populations, even in species with enormous genomes but no whole-genome reference. It has exciting potential to reveal in unprecedented detail the genetic basis of evolution, including adaptive differentiation and speciation.

Methods

Genomic DNA extraction

Three female individuals, each from *Bos taurus* (Portugal), *Bos indicus* (India), and *Bison bison* (USA) were used for this study. We used genomic DNA samples stored for many years in our labs (at the University of Porto and the University of Montana). The samples from cattle have been used in several published works related to the population genetics of cattle. The cattle biological tissue source from which the genomic DNA was isolated was ear skin (<2 mm2), extracted by DNeasy Blood & Tissue Kit (Qiagen). The bison sample was from lymph node tissue obtained from an abattoir with Tissue Use Approval provided by the Institutional Animal Care and Use Committee (identification number TU01-11GLDBS-040511) at the University of Montana. The obtaining of genomic data for this work did not involve experimental procedures or manipulation of living animals.

Selecting exon sequence targets

16,131 exon sequences were selected from the Btau 4.0, *Bos taurus* genome sequence [11], as annotated by the alignment of mRNA's from the NCBI RefSeq database [17] by the BLAT program [18], the alignment available at the UCSC genome browser web site [19]. Complete exon sets were collected for 203 genes selected by name. Most of these were found to be annotated as above, the few remaining annotations found through NCBI's Entrez genome site [20]. Other than those collected for the 203 selected genes, exon sequences came from an exome-wide sampling by iterating many times over the chromosome sequences, choosing one gene annotation on visiting each chromosome.

Longer chromosomes were visited more often proportionally to their total base pair (bp) count. As each chromosome was visited in turn, the exon sequences were collected from the gene whose midpoint coordinate was closest to the (currently) largest, contiguous non-sampled span of the chromosome sequence.

To meet our goal of sampling about 2,000 genes and keeping the total bases to about three million we collected no more than 1,500 exon base pairs per gene, except for the 203 named genes. To look for sequence variation in regulatory regions of genes, for all genes we collected the exon containing the 5' UTR, then chose randomly from among the remaining exons until adding an exon brought the total base pair count above 1,500. If a gene had only exons longer than 1,500 bps, we sampled 750 from each end of the 5' terminal exon. For genes other than those 203 for which all exons were collected, we collected no exons with fewer than 40 bps. As exon sequence candidates were chosen, the BLAST program [21] was used to remove any exon with at least 40 contiguous base pairs showing more than 90% identity with a subsequence in an exon already collected.

Targeted capture by hybridization

Hybridization probes for a microarray (Agilent, 244K aCGH format) were designed as previously described [5]. A single array was used per individual and hybridization performed as previously described.

Sequencing

Sequencing of post-enrichment shotgun libraries was carried out on Illumina Genome Analyzers (GA) I and II, one lane per individual on each Analyzer, as single-end 36 bp reads, following the manufacturer's protocols and using the standard sequencing primer. Image analysis and base calling was performed by the Genome Analyzer Pipelines with default parameters, but with no pre-filtering of reads by quality. In the reads produced by the GAII lanes, quality values were estimated directly by the Illumina software. A recalibration of the base qualities from the GAI lanes was performed during mapping as described below. Sequencing reads are being submitted to the NCBI Short Read Archive under accession SRA037397.1.

Mapping of sequencer reads

We used Maq software version 0.7.1 [12] to map the reads to the reference *Bos taurus* genome sequence and compute consensus genotypes at all positions covered by a uniquely mapped read. We used Maq's "map" command with default parameters, except when testing the bison reads using the "map" command's parameter "-m" (detailed below in the section, Calling single-base differences to the reference).

Reads produced by the GAII were mapped twice. Before a final mapping preliminary mappings were filtered by in-house programs to create a final collection of reads, under the following criteria:

1. Reads not uniquely mapped were discarded.

2. Reads that mapped off-target, so that no base in the read was aligned to a targeted exon base pair, were discarded.

3. Reads representing likely polymerase chain reaction (PCR) duplicates were removed by discarding, in any group of reads that mapped identically at position and strand, all but the read with the highest sum of base qualities.

The final mapping of the reads produced by the GAI was preceded by two preliminary mappings, both of which involved the same steps (1-3 above) performed for the GAII. For GAI reads, however, the filtered set of reads produced by the first mapping was used solely to recalibrate (with an in-house program) Illumina base quality scores, in order to estimate a correction performed by Illumina software supplied with the GAII but missing from the GAI. The recalibration treated all mismatches in the (filtered, on-target) initial mappings as sequencer error, under the assumption that the great majority of mismatches were errors in the reads. An error rate was calculated as the ratio of mismatches to matches for all mapped bases with a given sequencer-generated base quality score. The sequencer-generated base quality scores were then replaced with the (generally lower) quality based on the calculated error rate. This calibration was done separately for our taurine and zebu individuals. After finding a severe reduction in quality scores when the error rates were calculated based on the bison reads, the final bison quality recalibration was based on an average of the error rates found for the two *Bos* species, under the assumption that the relative wealth of mismatches between the bison and the reference likely reflected an abundance of real differences, and in total would significantly overestimate sequencing error rates. All of the GAI reads, with recalibrated base qualities, were then mapped twice using the procedure described above for the GAII reads.

After recalibration and removal of likely PCR duplicates, uniquely mapped reads for the bison totalled 11,384,125, for the zebu 11,432,216, and the taurine 7,154,561. Of these, bison on-target reads totalled 2,653,386 (23% of uniquely mapped reads), for the zebu, 2,320,339 on-target (20%), and the taurine, 2,105,157 (29%).

As a final note on mapping, it was found that duplicate mappings, using the same MAQ map command (with default parameters), and the same reads and reference data yielded slightly different results. Most of the differences in mappings were a single point difference in mapping score for a read on one execution versus the duplicate execution. An inquiry to the authors of [12] has been made and a more precise accounting of the differences is in progress.

Calling single-base differences to the reference

Consensus genotyping by Maq of targeted exon positions covered by the mapped reads identified both candidate homozygous differences from the reference sequence and heterozygous SNPs. Analyses of differences to the reference were based on consensus genotypes with at least 8 × coverage and a Maq Phred-like consensus quality score of 30 or more. Emulating methods in [5] by removing likely PCR duplicate reads and recalibrating base qualities as described above, we then chose our minimum coverage and depth values for high-confidence genotype calls based on the findings in [5] that Maq-based genotype calls with at least these coverage and depth values were in high concordance with genotypes inferred by several alternative resequencing methods. We also tested variant calls by Sanger resequencing DNA from our three individuals in regions in five of our 203 candidate genes in which our exon capture analysis found likely variants (Table 2). In these regions Sanger-sequence-based genotypes were in concordance with 19 variant calls for the bison, 12 for the zebu, and 4 for the taurine. Neither the bison nor zebu showed any false positives for the regions, while our taurine individual showed 5 false positives (Table 2). Further indication of lower accuracy in the taurine is seen in the deflated transition-to-transversion ratio (2.94) in heterozygous positions not found in dbSNP versus the ratio for those found in dbSNP (3.74). For the zebu, transition-to-transversion ratios are 3.17 for heterozygous positions not found in dbSNP, and 3.03 for heterozygous positions matched in dbSNP. Because our bison individual had only 22 matched heterozygous positions in dbSNP, its transition-to-transversion ratio of 2.14 for positions matched in dbSNP, versus 2.86 for unmatched positions, is probably a poor indicator of an error rate in variant calling.

Table 3.2. Sanger Sequence Calls vs. Maq. Variant calls in several Sanger sequenced fragments in and near exons in the genes indicated, compared with Maq consensus bases. All variant calls in the Sanger sequenced regions for the bison and zebu are in agreement with the Maq consensus bases. For the taurine five positions (rows in bold) are called as variant by the Maq consensus but not by the Sanger sequenced fragments.

sample	chrom	sanger start	sanger stop	position	reference	sanger	maq	gene
bison	chr6	88531917	88532399	88532280	A	G	G	*CSN3*
bison	chr6	88531917	88532399	88532296	T	C	C	*CSN3*
zebu	chr6	88531917	88532399	88532296	T	Y	Y	*CSN3*
zebu	chr6	88531917	88532399	88532332	C	M	M	*CSN3*
zebu	chr6	88531917	88532399	88532339	A	R	R	*CSN3*
zebu	chr6	88531917	88532399	88532393	G	R	R	*CSN3*
taurine	**chr6**	**88531917**	**88532399**	**88532293**	**C**	**C**	**Y**	***CSN3***
taurine	**chr6**	**88531917**	**88532399**	**88532296**	**T**	**T**	**C**	***CSN3***
taurine	**chr6**	**88531917**	**88532399**	**88532393**	**G**	**G**	**A**	***CSN3***
bison	chr4	95689756	95690201	95690049	T	C	C	*LEPTIN*

zebu	chr4	95689756	95690201	95690049	T	C	C	*LEPTIN*
taurine	chr4	95689756	95690201	95690049	T	C	C	*LEPTIN*
bison	chr10	3941758	3942115	3941786	T	C	C	*TICAM2*
bison	chr10	3941758	3942115	3941805	A	G	G	*TICAM2*
bison	chr10	3941758	3942115	3941921	A	G	G	*TICAM2*
bison	chr10	3941758	3942115	3941934	G	A	A	*TICAM2*
bison	chr10	3941758	3942115	3941946	A	G	G	*TICAM2*
zebu	chr10	3941758	3942115	3941921	A	G	G	*TICAM2*
zebu	chr10	3941758	3942115	3941946	A	R	R	*TICAM2*
zebu	chr10	3941758	3942115	3941963	C	Y	Y	*TICAM2*
taurine	chr10	3941758	3942115	3941921	A	R	R	*TICAM2*
bison	chr17	4284137	4284804	4284160	T	A	A	*TLR2*
bison	chr17	4284137	4284804	4284210	A	G	G	*TLR2*
bison	chr17	4284137	4284804	4284358	C	Y	Y	*TLR2*
bison	chr17	4284137	4284804	4284655	T	C	C	*TLR2*
bison	chr17	4284137	4284804	4284747	C	T	T	*TLR2*
zebu	chr17	4284137	4284804	4284160	T	W	W	*TLR2*
zebu	chr17	4284137	4284804	4284210	A	G	G	*TLR2*
zebu	chr17	4284137	4284804	4284652	G	K	K	*TLR2*
zebu	chr17	4284137	4284804	4284655	T	Y	Y	*TLR2*
taurine	chr17	4284137	4284804	4284210	A	R	R	*TLR2*
taurine	**chr17**	**4284137**	**4284804**	**4284639**	**G**	**G**	**R**	***TLR2***
taurine	chr17	4284137	4284804	4284652	G	K	K	*TLR2*
taurine	**chr17**	**4284137**	**4284804**	**4284655**	**T**	**T**	**Y**	***TLR2***
bison	chr8	112427182	112427427	112427204	C	T	T	*TLR4*
bison	chr8	112427182	112427427	112427213	C	T	T	*TLR4*
bison	chr8	112427182	112427427	112427326	A	C	C	*TLR4*
bison	chr8	112431812	112432152	112431927	G	A	A	*TLR4*
bison	chr8	112434757	112435132	112435011	A	C	C	*TLR4*
bison	chr8	112434757	112435132	112435120	C	A	A	*TLR4*

The Maq mapping software uses a base variation (mutation) rate between reference and reads (the default is 0.001) in its mapping algorithm. Further, Maq's alignment scores are based on the probability of error in mismatches between read and reference (details are in [12]). Therefore, a true per-base variation rate for our bison versus the taurine reference is likely higher than that suggested by our percent of total differences given by the alignment (0.45%, or about 5 per 1000 bases). The relatively long evolutionary distance between the bison's genome and that of the taurine likely increases the chance, compared with the reads for the two cows, of the bison reads being incorrectly mapped. To test the effect of the base variation rate on Maq analysis of the bison reads, we re-mapped the reads after raising the variation rate (using Maq's mapping "-m" parameter) from the default 1/1000 (0.001) used in our initial analysis to 0.002, 0.003 ... up to 0.007. While, against expectations, increasing the mutation rate was associated with drops in the number of total differences called at our depth/quality threshold (for heterozygous differences, the largest drop was a loss of 40 calls at mutation rate 0.003, less than those called at 0.002), 95% of the 2,426 heterozygous SNPs called at rate 0.001 were shared by all 7 mappings, and, including fixed differences, 99% of single-base variants were called identically at all mutation rates. The high concordance suggests that, despite a likely bias in mapping against divergent exon sequences, there are bison exome sequences genome-wide among our ~ 16,000 exons sufficiently similar to those in the cow for identification of thousands of likely variant bases (Table 1).

Sanger sequencing for verifying variant calls

Several exonic fragments from five genes (*CSN3*, *LEP*, *TICAM2*, *TLR2*, and *TLR4*) were re-sequenced using conventional Sanger sequencing, for verifying Maq-based variant calls. The primers were designed for amplifying those exons with >150 base pairs, using Primer3 online Web interface (http://frodo.wi.mit.edu/primer3/ webcite). The primer sequences are provided in Table 3. PCR reactions were performed in a 20 µl volume containing 10× PCR Buffer, 1.5 to 3 mM $MgCl_2$ (upon primers), 0.2 mM dNTPs, 1 µM each primer, 0.4 U Platinum® *Taq* DNA Polymerase (Invitrogen), and approximately 30 ng genomic DNA. The PCR mixture underwent 15 min at 94°C, 35 cycles of 30 s at 94°C, 30 s at 58 to 64°C (upon primers), and 35 s at 72°C, and final 10 min at 72°C on

GeneAmp PCR System 9700 (*Applied Biosystems*, Foster *City*, CA,*USA*). PCR products were purified and sequenced for both strands, at High-Throughput Genomics Unit (HTGU), Department of Genome Sciences, University of Washington (http://www.htseq.org/ webcite). Sequence trace files were checked and aligned using software package DNASTAR v7.1 (DNASTAR Inc., Madison, WI, USA).

Table 3.3. Primers used for PCR amplification and Sanger sequencing.

Gene	Forward (5' to 3')	Reverse (5' to 3')
CSN3	AGAAATAATACCATTCTGCAT	GTTGTCTTCTTTGATGTCTCCTTAGAG
LEP	GATTCCGCCGCACCTCTC	CCTGTGCAAGGCTGCACAGCC
TICAM2	TCCTCTTCTGACTCGGATCTTT	CCAAGTTCTGTAAATGCTGTCTGC
TLR2-f1	TGGGTCTGGGCTGTCATCAT	AAGAGATGTTTCCCCAAGTGTTTT
TLR2-f2	GACCTGCAGAGGTGTGTGAA	TGAAAAATGGAAAGTGTGCAA
TLR4-f1	CGGGGAGAGACGACACTACA	TGTTTGCAAATGAACCTAACCA
TLR4-f2	TCTTTGCTCGTCCCAGTAGC	AAGTGAATGAAAAGGAGACCTCA
TLR4-f3	GGAGACCTAGATGACTGGGTTG	GGGGCATTTGATGTAGAACTTT

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

The experiment was conceived by GL, designed by ABP, GL, JS, and SC. It was performed by JS, SBN and SC. Data analysis was performed by TC with guidance from JS and GL. All authors contributed to the writing of the paper. All authors read and approved the final version of the manuscript.

Acknowledgements

We thank Ruolan Qiu and Choli Lee for technical assistance with capture and sequencing. This study was funded by the Portuguese Foundation for Science and Technology (FCT) project PTDC/CVT/099782/2008 and PTDC/CVT/68907/2006. T.C.

was supported by the National Science Foundation and CIBIO/UP. S.C. is supported by a FCT grant SFRH/BPD/46082/2008. G.L was partially supported by the National Science Foundation (DEB 074218), the Walton Family Foundation, CIBIO/UP, and the Portuguese science foundation grants (PTDC/BIA-BDE/65625/2006; and PTDC/CVT/69438/2006).

References

1. Ellegren H, Sheldon BC: **Genetic basis of fitness differences in natural populations.** *Nature* 2008, **452:**169-175.
2. Hoekstra HE, Coyne JA: **The locus of evolution: evo devo and the genetics of adaptation.** *Evolution* 2007, **61:**995-1016.
3. Shendure J, Ji H: **Next-generation DNA sequencing.** *Nature Biotechnology* 2008, **26:**1135-1145.
4. Luikart G, England PR, Tallmon D, Jordan S, Taberlet P: **The power and promise of population genomics: from genotyping to genome typing.***Nat Rev Genet* 2003, **4:**981-994.
5. Ng SB, Turner EH, Robertson PD, Flygare SD, Bigham AW, Lee C, Shaffer T, Wong M, Bhattacharjee A, Eichler EE, Bamshad M, Nickerson DA, Shendure J: **Targeted capture and massively parallel sequencing of 12 human exomes.** *Nature* 2009, **461:**272-276.
6. Gnirke A, Melnikov A, Maguire J, Rogov P, LeProust EM, Brockman W, Fennell T, Giannoukos G, Fisher S, Russ C: **Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing.** *Nature Biotechnology* 2009, **27:**182
7. Bau S, Schracke N, Kränzle M, Wu H, Stähler PF, Hoheisel JD, Beier M, Summerer D: **Targeted next-generation sequencing by specific capture of multiple genomic loci using low-volume microfluidic DNA arrays.** *Analytical and Bioanalytical Chemistry* 2009, **393:**171-175.
8. Wang H, Chattopadhyay A, Li Z, Daines B, Li Y, Gao C, Gibbs R, Zhang K, Chen R: **Rapid identification of heterozygous mutations in Drosophila melanogaster using genomic capture sequencing.** *Genome Research* 2010.
9. Raca G, Jackson C, Warman B, Bair T, Schimmenti LA: **Next generation sequencing in research and diagnostics of ocular birth defects.** *Molecular Genetics and Metabolism* 2010, **100:**184-192.
10. Genome 10K Community of Scientists: **Genome 10K: A Proposal to Obtain Whole-Genome Sequence for 10,000 Vertebrate Species.** *Journal of Heredity* 2009, **100:**659-674.
11. The Bovine Genome Sequencing and Analysis Consortium, Elsik CG, Tellam RL, *et al.*: **The Genome Sequence of Taurine Cattle: A Window to Ruminant Biology and Evolution.** *Science* 2009, **324:**522-528.

12. Li H, Ruan J, Durbin R: **Mapping short DNA sequencing reads and calling variants using mapping quality scores.** *Genome Research* 2008, **18:**1851.
13. Hedrick PW: **Conservation genetics and North American bison (Bison bison).** *Journal of Heredity* 2009, **100:**411.
14. Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, Sirotkin K: **dbSNP: the NCBI database of genetic variation.** *Nucleic acids research* 2001, **29:**308.
15. The Bovine HapMap Consortium: **Genome-Wide Survey of SNP Variation Uncovers the Genetic Structure of Cattle Breeds.** *Science* 2009, **324:**528-532
16. Morin PA, Luikart G, Wayne RK, the SNP workshop group: **SNPs in ecology, evolution and conservation.** *Trends in Ecology & Evolution* 2004, **19:**208-216.
17. Pruitt KD, Tatusova T, Maglott DR: **NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins.** *Nucleic Acids Res* 2007, **35:**D61-D65.
18. Kent WJ: *BLAT-The BLAST-Like Alignment Tool. Volume 12.* Cold Spring Harbor Lab; 2002.
19. Kuhn RM, Karolchik D, Zweig AS, Wang T, Smith KE, Rosenbloom KR, Rhead B, Raney BJ, Pohl A, Pheasant M, Meyer L, Hsu F, Hinrichs AS, Harte RA, Giardine B, Fujita P, Diekhans M, Dreszer T, Clawson H, Barber GP, Haussler D, Kent WJ: **The UCSC Genome Browser Database: update 2009.** *Nucl Acids Res* 2009, **37:**D755-761.
20. Maglott D, Ostell J, Pruitt KD, Tatusova T: **Entrez Gene: gene-centered information at NCBI.** *Nucleic acids research* 2006.
21. McGinnis S, Madden TL: **BLAST: at the core of a powerful and diverse set of sequence analysis tools.** *Nucleic Acids Research* 2004, **32:**W20.

Chapter 4: Next-generation Sequencing of Thousands of Genes in Divergent Nonmodel Taxa using Exon Capture

Following revisions and consent from co-authors, the following will be submitted for publication with my collaborators[2]. A possible target journal is Genome Research.

Abstract

Genome-wide sequencing of numerous functional genes in nonmodel species will allow researchers to addresses novel questions in conservation and evolutionary biology. We sequenced 24,525 exons in 18 individuals spanning 6 ungulate species using genomic DNA enriched for exons by hybridization to an array of DNA probes designed from the domestic cow (*Bos taurus*) reference genome sequence. The Illumina sequencer reads were aligned to the cow genome, and the aligned bases were genotyped for each individual. The percentage of quality filtered reads whose alignment overlapped at least one target base was on average 20% for all mappings for all samples, with both the high of 28% and the low of 8% in pig individuals, the species most divergent from the cow. Captured cow sequences aligned and completely covered 65% of the 24,525 targeted exons, with at least 20X depth of coverage and a high consensus quality score (Phred 50 or greater). In the closely related bison, 63% of the exons were completely sequenced to the same quality thresholds. Two African buffalo and two bighorn sheep yielded ~52% of exons. The average yield among 7 deer was 42%, and 38% among the 4 pigs. The total number of putative heterozygous sites discovered in exons per individual ranged from a maximum of 4,418 in one deer to a low of 1,112 in a pig. Among 7 deer 12,645 putative, exonic SNPs were located in 6,574 targeted exons, from 3,668 genes of the 5,935 targeted. In 63 exons targeted in the MHC gene family we found 67 putative exonic SNPs in one African buffalo (the best performance) and only 8 in a pig (the lowest yield). Overall, our results indicate that, using available mapping and genotyping tools, we can sequence thousands of exons, including those in large gene families such as the MHC, in species that diverged tens of millions of years from the genome of a reference species.

[2] Stephen Amish, Fish and Wildlife Genomics Group, Division of Biological Sciences, University of Montana, Emily Latch, Biological Sciences, University of Wisconsin, Milwaukee, Dan Bruno, Rocky Mountain Laboratories, National Institute of Allergy and Infectious Diseases (NIAID), Stacy Ricklefs, Rocky Mountain Laboratories, NIAID, Sarah Anzik, Rocky Mountain Laboratories, NIAID, Craig Martens, Rocky Mountain Laboratories, NIAID, Albano Beja-Pereira, Centro de Investigação em Biodiversidade e Recursos Genéticos, Universidade do Porto, Steve Porcella, Rocky Mountain Laboratories, NIAID, Gordon Luikart, Fish and Wildlife Genomics Group, Flathead Lake Biological Station, Division of Biological Sciences, University of Montana.

Introduction

Exon capture enriches genomic DNA samples for the protein-coding sequences that comprise a high-value genomic target in searches for adaptive or functional genetic variation (Vasemägi et al. 2005; Hodges et al. 2007; Nadeau et al. 2012). The exome represents about 1-3% of the ~3 billion base pairs in a typical mammalian genome. Reduction of the targeted area to a small part of the vast genome reduces sequencing cost per individual and thereby facilitates population-scale genomics studies (Luikart et al. 2003), including discovery of single nucleotide polymorphisms (SNPs) without ascertainment bias (Morin et al. 2004).

Generalized as "targeted resequencing," for any chromosomal region(s), exon capture can help address a wide range of research questions in evolutionary biology such as: Which genetic variants are associated with complex traits such as disease risk (Price et al. 2010)? Which gene causes a monogenic, Mendelian disorder (Ng et al. 2009a)? Which candidate genes are targets of selection and influence fitness and local adaptation (Allendorf et al. 2010; Good et al. 2013; Cheviron and Brumfield 2012; Hohenlohe 2013)?

Many of these questions are important in both conservation and evolutionary biology, in which species of interest often have few genomic resources compared to the fully sequenced, richly annotated genomes of a few organisms of intense genetic research, such as the human, mouse, domestic cow, chicken and corn (Lander et al. 2001; Chinwalla et al. 2002; Elsik et al. 2009; Hillier et al. 2004; Schnable et al. 2009).

Despite plummeting DNA sequencing costs afforded by next-generation sequencing, the short sequence fragments produced by high-throughput sequencing platforms make assembly of whole mammalian genome sequences problematic (Salzberg et al. 2012). For many wildlife species, saliently mammals, a dearth of genome sequence information makes it difficult to address questions about the genetic basis of fitness, adaptation, and phenotype variation.

Exon capture was developed with and is most often performed on human DNA samples and human genome reference sequences (Hodges et al. 2007; Ng et al. 2009b; Gnirke et al. 2009; Tennessen et al. 2012; Sanders et al. 2012; Do et al. 2012). Conventionally, a full and well-annotated genome sequence provides exon sequences for synthesizing the DNA probes, ~60-120 base pairs (bp), to capture exon-containing sequences by probe-hybridization to fragmented whole genome samples. One or both ends of the captured exon fragments are sequenced by a next-generation DNA sequencer, producing millions of short sequences (reads), of ~75-400 bp. The annotated genome sequence also serves as a reference on which to map the reads. The overlapping placement of reads on the reference sequence provides a basis for consensus sequences for targeted exons in each individual.

Capture in divergent, nonmodel species

An alignment of 28 vertebrate genomes has shown that over 90% of human coding exons can be aligned to genomes in eight placental mammals and the platypus

(Miller et al. 2007b). Even genomic sequences in five fish species aligned to 60%-70% of human coding exons. Generally, Miller et al. (2007b) found that, across vertebrates, coding sequence alignment scores indicated high sequence conservation. This argues that exon capture methods can be applied to many species without a reference genome sequence. Existing genome sequences can provide probes for exon capture in divergent species (Nadeau et al. 2012; Burbano et al. 2010; Cosart et al. 2011).

The relationship of exon capture success to phylogenetic divergence between the subject taxon (e.g., a nonmodel species) and the taxon of the reference genome, however, has not been well characterized. It would help those who would use exon capture in a species with no reference sequence, to know whether the phylogenetically nearest available genome can serve for both probe design and mapping, and whether the software tools used generally in species-species exon capture can serve to map and genotype sequence reads from divergent species.

Gene families (and duplicated genes) are of special interest for research on fitness and adaptation, but also are a special challenge for exon capture. For example, the major histocompatibility complex (MHC) genes influence disease resistance (Spurgin and Richardson 2010) and mate choice (Milinski 2006). Dozens of duplicated MHC genes exist (Spurgin and Richardson 2010), and are often the source of novel functions or adaptations (Bielawski and Yang 2004). However, sequence reads from gene families with many closely related, duplicated loci can be difficult to map to their correct genome position when sequences at one locus are highly similar to those at another (Schaschl et al. 2006). This between-locus mapping difficulty might be exacerbated with cross-species exon capture as conducted here.

Our goal is to quantify the success of exon capture as it relates to increasing levels of phylogenetic divergence between the reference and subject (nonmodel species) genomes (Table 4.1).

Table 4.1: Ungulate divergence from the reference. For each sampled species, approximate million years (MY) of reproductive isolation from *Bos taurus*, and total number of individuals sampled. The reference column cites literature supporting the approximate divergence times.

Common Name	Species	MY	Reference	Num. Samples
Taurine cow	*Bos taurus*	0	n.a.	1
American bison	*Bison bison*	1-2	Hedrick 2009	1
African buffalo	*Syncerus caffer*	1-3	Ritz et al. 2000	2
Bighorn sheep	*Ovis canadensis*	20	Randi et al. 1991	2
Mule deer	*Odocoileus hemionus*	27-38	Guha et al. 2007	8
Domestic pig	*Sus domesticus*	60	Ursing et al. 2000	2
Wild boar	*Sus scrofa*	60	Ursing et al. 2000	2

We chose artiodactyls because they are a group of high interest for conservation, evolution, and agriculture, including bison, elk, mule deer, and wild and domestic cow, sheep and pigs (Bruford et al. 2003; Andersson and Georges 2004). Artiodactyl full genome sequences are available for divergent species (cattle, pigs, and domestic sheep). We targeted 24,525 exons (among 5,935 genes), located throughout the genome. To quantify success using a gene family with potentially many alleles, we included 63 exons in 14 genes from the MHC.

Our three main objectives were to 1) Quantify the number (and proportion) of exon targets successfully captured and sequenced in divergent taxa by mapping sequence reads to the taurine reference genome, 2) Determine the number (and proportion) of heterozygous single nucleotide variants inferred by the capture, sequencing, and mapping to the taurine genome, and 3) Measure concordance of read alignments and base calls from the divergent species on the taurine genome versus alignments to a phylogenetically similar reference genome (e.g., bighorn sheep to a domestic sheep genome, and wild boar or pig to the domestic *Sus scrofa* genome). We quantified objective (1) using two kinds of read-mapping computer programs, (i) BWA (Li and Durbin 2009), a fast aligner designed for relatively little divergence between sample and reference, and (ii) Stampy (Lunter and Goodson 2011) designed for higher divergence.

Results

Sequence read alignment

The exon sequencing success reported below is informed by the relative numbers of sequencing reads, 100-bp, paired-end, obtained for each species, and, of these, how many were aligned to targeted exons. Of trimmed and non-PCR-duplicated reads (see Methods), using the Stampy mapper, the percentage that uniquely aligned to the cow surpassed 90% in all species except for the pigs, which averaged about 60% (Fig. 4.1). The percentage of trimmed, PCR-filtered reads, however, whose alignment overlapped at least one target base (on-target reads), was on average 20% for the combined Stampy and BWA mappings for all samples, with a high of 28% for a Stampy mapping of pig 3, and a low of 8% for the BWA mapping for pig 1. These percentages are low compared to ~56% reported in human exome capture using similar technologies (Asan et al., 2011). Another study achieved over 90% of read bases, in uniquely mapped reads, mapped on or within 250 bp of targeted exons (Blumenstiel et al. 2010). This measurement in the Stampy mappings for our samples averaged 34%, with the high 54% in pig 4, and the low 27% in sheep 2.

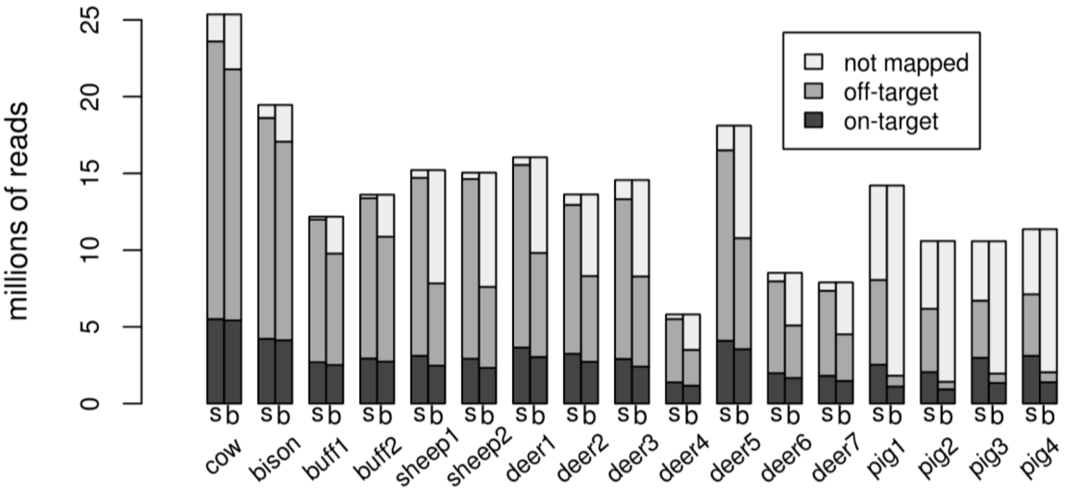


Figure 4.1: Totals of sequence reads. Bar heights give totals of sequence reads with putative PCR duplicates removed, and with low quality bases and sequencer adaptor sequences trimmed (see Methods). Proportions of reads are indicated as noted in the legend. "Not mapped" are sequences that had no single best placement on the reference genome sequence. Off-target reads are those with a single, best alignment outside of targeted exons. On-target reads are those with a single best placement in which at least one base aligns to a targeted exon sequence. "S" below a bar indicates totals for Stampy mappings, "b" indicates BWA.

Mapper comparisons: BWA versus Stampy

Overall, for the cow, bison, and African buffalo, the two mappers (BWA and Stampy) produced about equal numbers of exons sequenced (Fig. 4.2, and see Methods, "Measuring sequencing success"). Stampy yielded more sequenced exons for the bighorn, deer and pigs. We relaxed the BWA parameters to allow for more than the default mismatches (substitutions and gaps) between read and reference for the deer and pigs (discussed in Methods, in the section "Mapping"). The difference in number of sequenced exons produced by the two mappers is most dramatic in our four pigs, where BWA mappings resulted in an average of 5,880 exons fully sequenced (to 20X/Q50), while Stampy produced an average of 9,262.

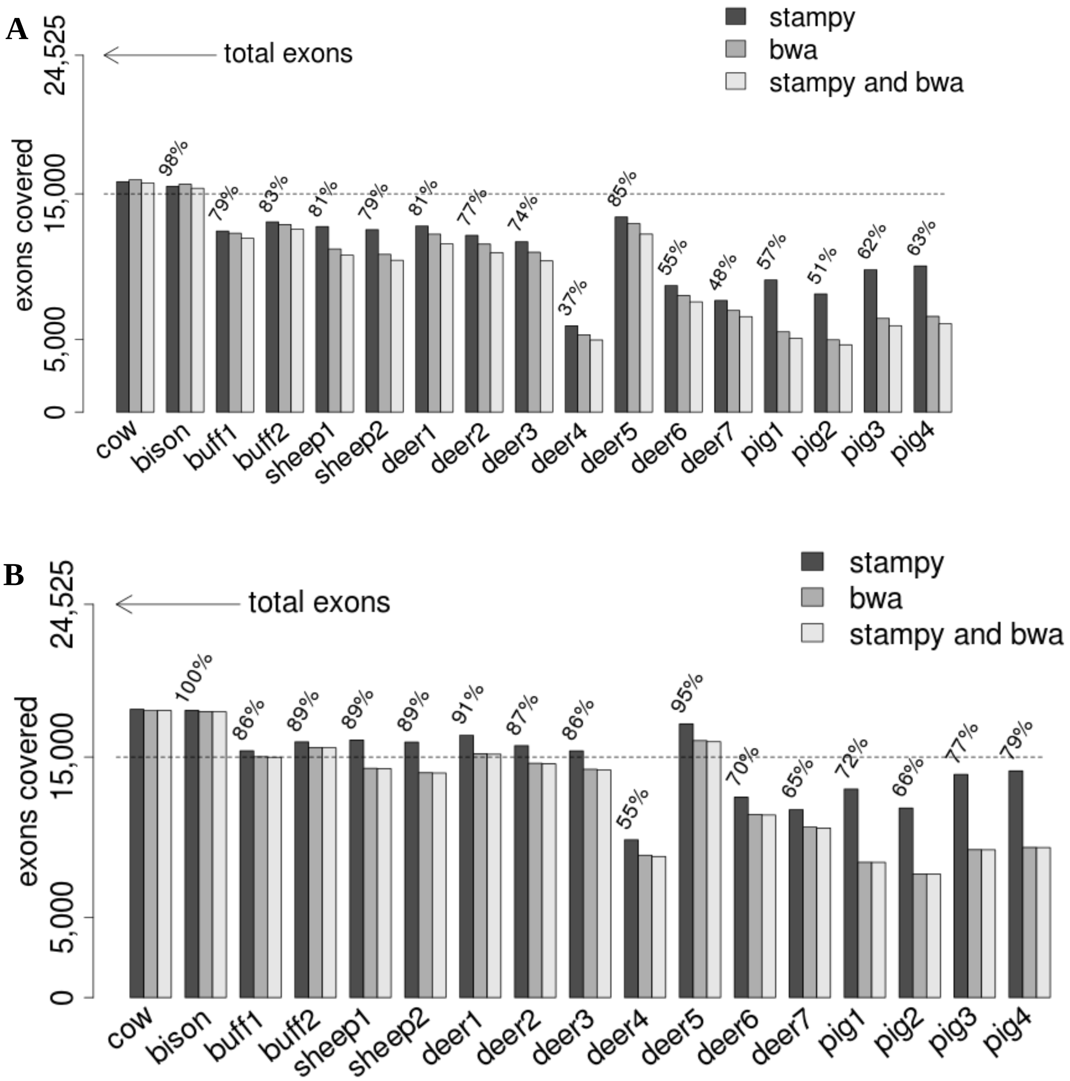


Figure 4.2: Number of exons sequenced. Success of each mapper for the 24,525 exons targeted, with (A) 100%, and (B) at least 60% of bases having at least 20X depth of coverage and with a Phred consensus quality score of at least 50, (abbreviated 20X/Q50). See Methods for details on the choice of these relatively stringent thresholds. Percentages above the Stampy bar give the totals as a percentage of the total for the cow. The white bar labeled “stampy and bwa” gives the total exons sequenced to these criteria by both mappers. BWA mismatch allowance parameters were relaxed only for the deer and pigs (see Methods).

For 63 exons from 14 targeted MHC genes (Table 4.2), BWA produced more complete (20X/Q50) exon sequences than did Stampy in the cow, bison, African buffalo 1, and deer 1 (Fig. 4.3). Overall, few MHC exons were sequenced completely in pigs. In pig 2 BWA produced only one completely sequenced MHC exon, Stampy none. Stampy produced more complete 20X/Q50 MHC exon sequences in the divergent species. For example, in deer Stampy produced an average of 12.1 exons with complete coverage whereas BWA produced on average 10.3.

Table 4.2: Targeted MHC genes. Exons were targeted in all annotations of the same gene, if they passed the similarity test requirement during target selection (see Methods).

Abbreviation	**Name/Description**	**Number annotated**
BOLA	MHC class I A	2
BOLA1	BolA homolog 1 (*E. coli*)	2
BOLA3	BolA homolog 3 (*E. coli*)	2
BOLA-DMA	MHC class II DM alpha-chain expressed	1
BOLA-DMB	MHC class II DM beta-chain expressed	1
BOLA-DOB	MHC class II DO beta	1
BOLA-DQA2	MHC class II DQ alpha 2	1
BOLA-DQA5	MHC class II DQ alpha 5	1
BOLA-DQB	MHC class II DQ beta	1
BOLA-DRA	MHC class II DR alpha	1
BOLA-DRB2	MHC class II DR beta 2	1
BoLA-DRB3	MHC class II DR beta-chain	1
BOLA-DYA	MHC class II DY alpha	1
BOLA-NC1	Non-classical MHC class I antigen	1

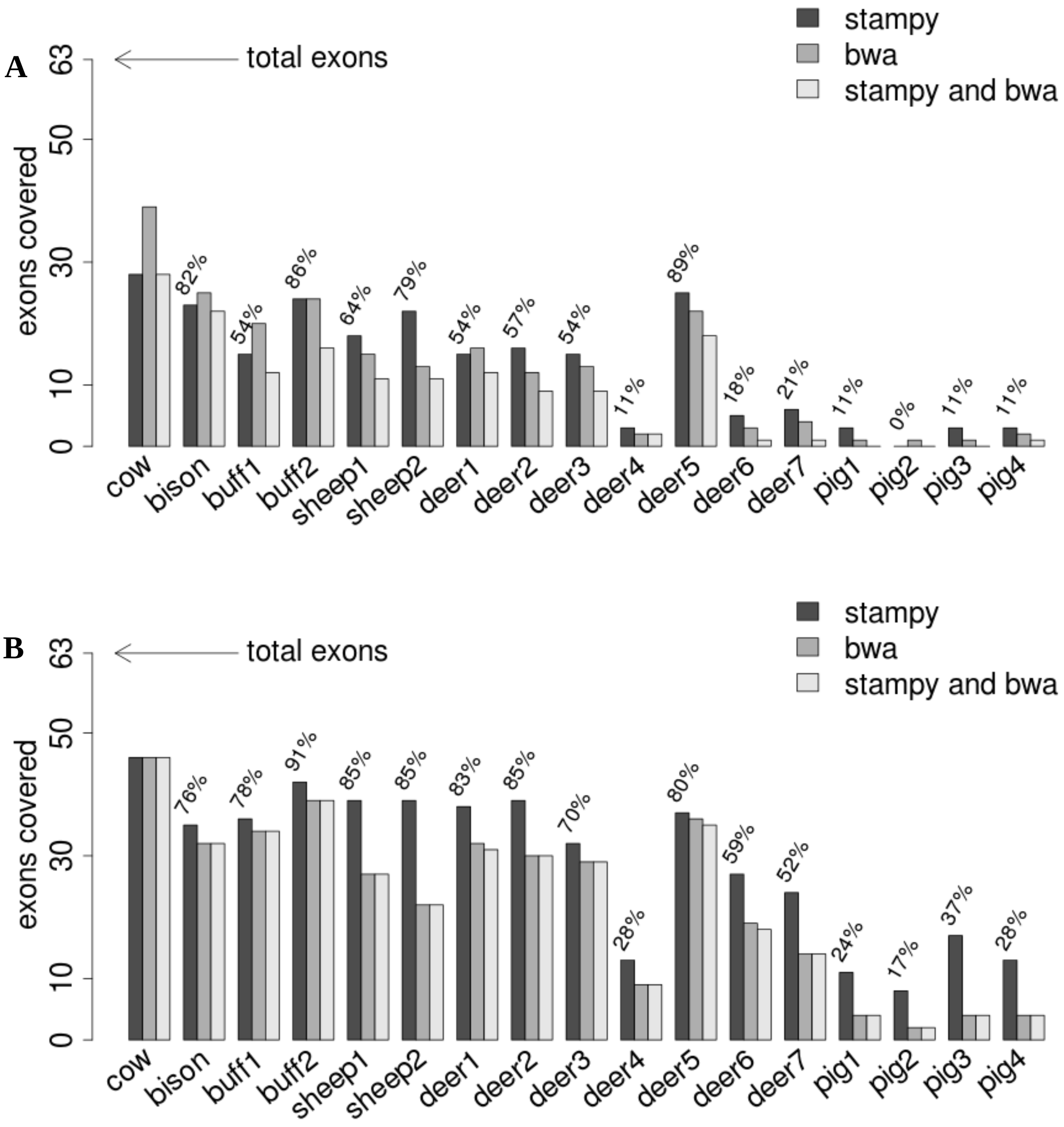


Figure 4.3: MHC exon coverage. Of 63 exons in 14 MHC genes, the number with (A) 100%, and (B) at least 60% of bases covered to at least 20X and with a Phred consensus score of at least 50. Percentages above the Stampy bars give plotted values as a percentage of the total for the cow. Average length for these MHC exons is 212.5 bp. Base pairs total 13,386. BWA mismatch allowance parameters were relaxed only for the deer and pigs (detailed in Methods).

Effect of divergence on exome-wide sequencing

The success of exon sequencing declined with divergence from the taurine reference genome sequence. Counting the total number of exon targets with all exon bases sequenced, the cow showed the highest total of 15,835 exons (65% of the 24,525 exons targeted, Fig. 4.2a). The bison yielded 15,519 completely sequenced exons (98% of the cow's total). Averaging for the species with more than one sample, the African buffalo produced 12,759 (81% of the cow's total), bighorn, 12,647 (80%), deer, 10,337 (65%), and pigs, 9,261 (58%). There were 8,999 exons completely sequenced in at least one individual for each of the five species sampled. Among individuals, deer 4 had the lowest total at 5,929 exons.

The negative effect of divergence was less pronounced in totals for exons with at least 60% of all bases sequenced (at 20X/Q50, Fig. 4.2b). As expected the highest success comes from the domestic cow sample, totaling 17,975 exons (73% of the 24,525 exons targeted). The bison yielded 17,914 exons (about equal to the cow). Averaging as above, the African buffalo produced 15,667 (87% of the cow's total), bighorn, 15,997 (89%), deer, 14,083 (78%), and pigs, 13,219 (74%). There were 13,131 exons sequenced at 60% or more bases (at 20X/Q50) in at least one individual of every species sampled.

For all 3.6 million base pairs (Mb) targeted by the exon capture array, the cow sample had the highest Stampy-aligned total of 2.83 Mb (78 %) that were sequenced (at 20X/Q50, Fig. 4.4). The bison produced 2.80 (99% of the cow's total). Averaging for the multisampled species, African buffalo yielded 2.50 Mb (88% of the cow's total), bighorn, 2.55 Mb (90%), deer, 2.27 Mb (80%), and pigs, 2.02 Mb (71%).

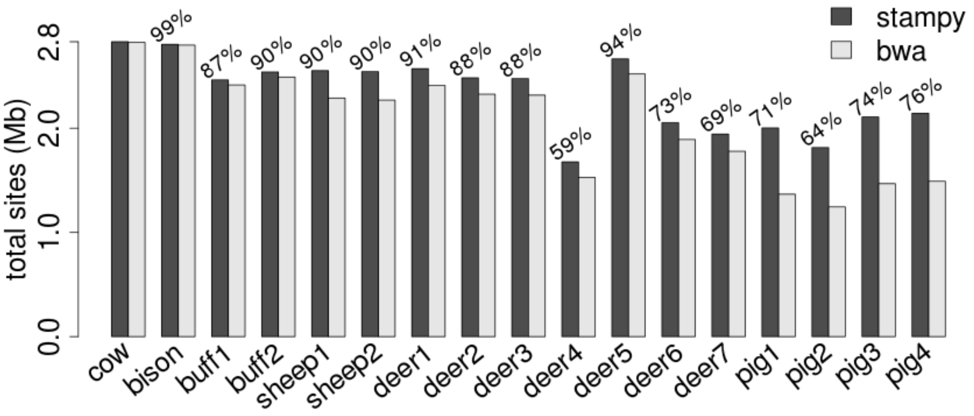


Figure 4.4: Sequencing success in total base pairs. The number of nucleotide sites sequenced to 20X/Q50 in the ~3.6 Mb of targeted exons. Numbers above the Stampy bar for non-cows give 20X/Q50 bases as a percentage of the total for the cow. BWA mismatch allowance parameters were relaxed only for the deer and pigs (see Methods).

Our capture and sequencing also produced data for sites flanking the targeted exons. Of the 11.6 Mb that flank 250 bp either side of each targeted exon, the cow sample had the highest count for Stampy aligned positions with 4.28 Mb (at 20X/Q50), and a deer individual had the low with 1.00 Mb (Supplementary Fig. 4-S1).

Candidate gene sequencing

Among our 24,525 targeted exons are 2,542 exons in 349 candidate genes that we chose for their function or association with speciation, reproduction, or role in disease susceptibility. Successful exon sequencing in these was 3.8% lower than success for the entire gene collection of 24,525 total exons (Supplementary Fig. 4-S4). With a mean exon length of 220 bp for candidate genes, and 148 bp for the whole 24,525, this result is contrary to the finding that longer exons are associated with greater depth of coverage (see section below, Coverage relationship to GC content and target length).

For candidate gene exons with 100% bases sequenced (at 20X/Q50), the cow yielded 1,532 exons (60% of the 2,542 exons targeted). The bison yielded 1,485 candidate gene exons (58% of the 2,542 exons targeted). Averaging for the species with more than one sample, the African buffalo produced 1,258 (50%), bighorn, 1,225 (48%), deer, 1,019 (40%), and pigs, 770 (30%).

Gene family sequencing

Exon capture for our 63 exons from 15 genes in the MHC family (Fig. 4.3) yielded 62% of the MHC exons fully sequenced for the cow sample, comparable to the 64% for all exons for the cow (and higher than the 60% for the candidate genes). All other species showed reduced percentages of fully sequenced (i.e., 20X/Q50) MHC exons compared to the full set of 25,424 targeted exons. Using BWA counts for the bison and African buffalo (which showed better performance than did Stampy for the MHC exons), the percentage of exons sequenced dropped by 38% in the bison, and, on average, 32% in the two African buffalo. Using the higher Stampy-based counts for the bighorn, deer, and pigs, the average reductions were 38%, 54%, and 91%, respectively, compared to the cow. There were two MHC exons fully sequenced in at least one individual of every species sampled. If the pigs are excluded, there are 14 MHC exons fully sequenced (at 20X/Q50) in a least one individual from every species.

Heterozygous calls and SNP discovery

In exons, the number of heterozygous sites per thousand bp (genotyped at 20X/Q50) in an individual ranged from 0.52 (pig 4) to 2.26 (deer 4, Fig 4.5a). In flanking sites (adjacent to exons) the range was 0.26 (pig 1) to 2.53 (African buffalo 2, Fig. 4.5b). Percentage of exons with at least one heterozygous site ranged from ~3% in pig 1 to ~11% in buffalo 1 (Fig. 4.6a). Percentage of flanking regions with at least one heterozygous position ranged from under 1% in pig 1 to ~11% in buffalo 2 (Fig. 4.6b).

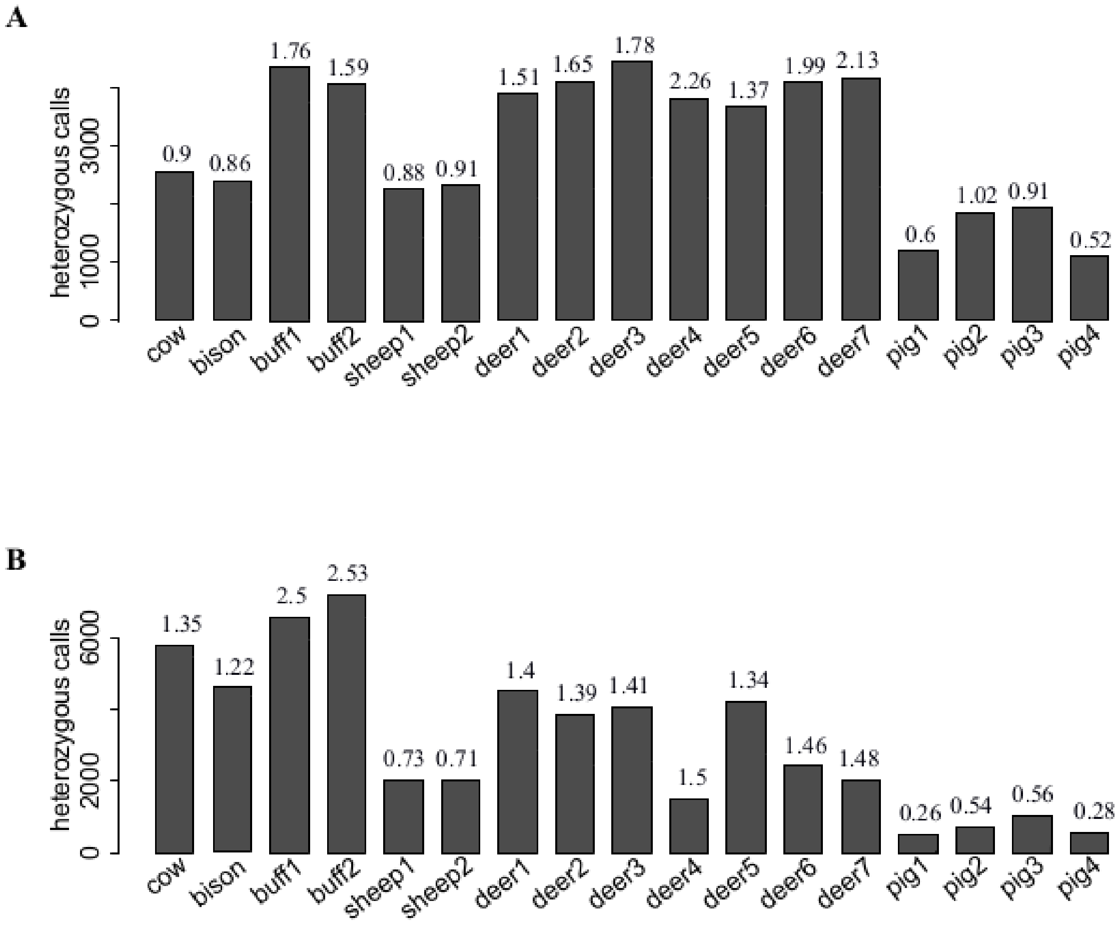


Figure 4.5: Total heterozygous calls. For the Stampy mapping, total single nucleotide heterozygous sites called in (A) the ~3.6 Mb of targeted exons and (B) ~11.6 Mb of 250-bp exon flanks. Numbers above the bars give the heterozygous positions inferred per thousand 20X/Q50 sites. Note that the inflation in the rate in flanks versus targets for the cow, bison, and buffalo, is replaced by a deflation in the rate for the more divergent species (see Discussion).

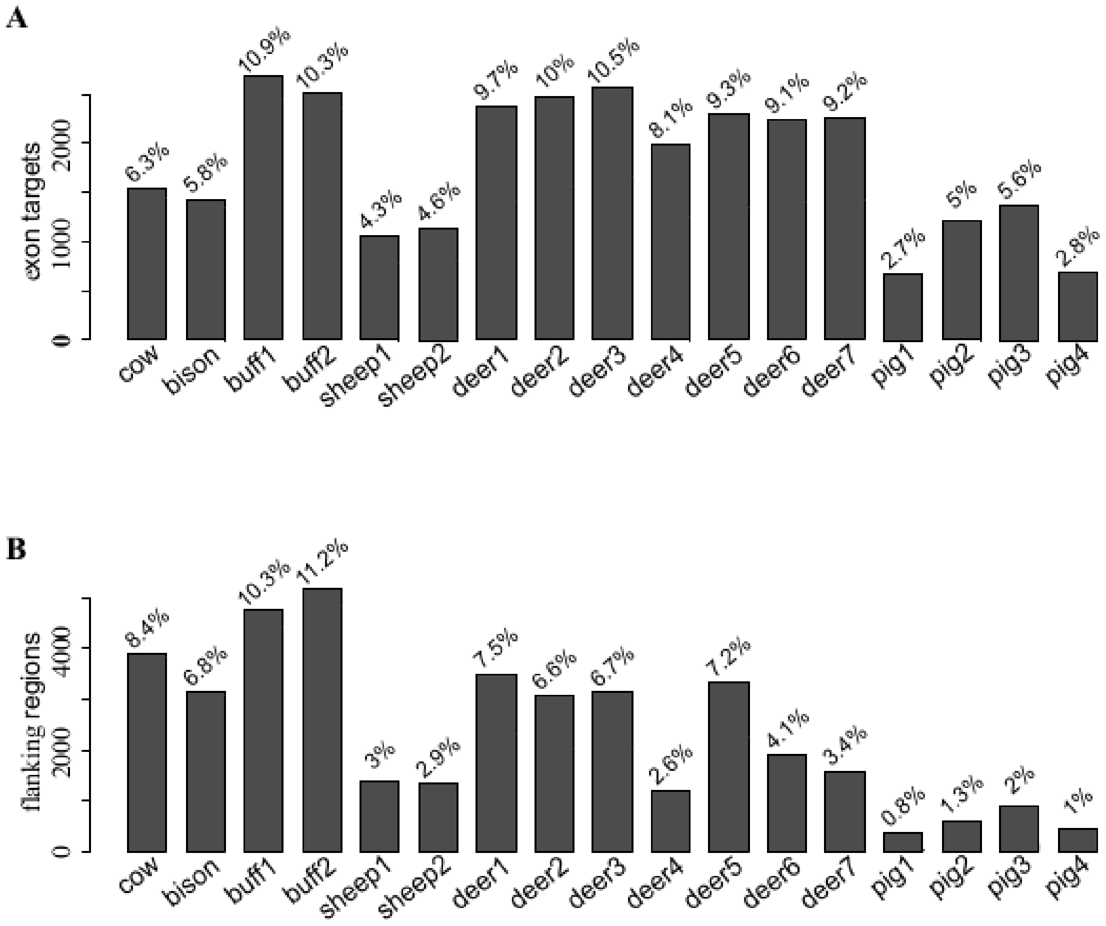


Figure 4.6: Regions with heterozygous calls. For Stampy mappings of (A) the 24,525 total exons targeted, and (B) the 46,538 (unique) flanking regions (of at most 250 bps), the totals with at least one heterozygous position. See Methods for variant filtering criteria. Percentages of total targets and flanking regions with at least one heterozygous position are plotted above the bars.

The sample size of seven deer offers an opportunity to assess the usefulness of cow-based exon capture for SNP discovery in population samples from a species divergent from the taurine reference genome. We found 14,657 SNPs with at least 4 individuals genotyped (i.e., genotype calls for at least four individuals passed the filtration criteria), and with at least 2 observations of the minor allele (e.g., 2 heterozygotes or one homozygote). Of these, 6,800 were exonic SNPs (the others were in exon flanks), spread over 4,147 exons in 2,753 genes. Thirty SNPs were genotyped such that 3 alleles were called among the 4 or more individuals.

Of the 2,542 exons in our 347 candidate genes, the cow sample yielded at least one filtered, heterozygous call in each of 219 exons (8.6%). In the bison, heterozygous calls were made in 194 exons (7.6%), on average, for the two African buffalo, 324.5 exons (12.7%), for the two bighorn, 117 exons (4.6%), the 7 deer, 251.4 exons (9.89%), and the four pigs, 104.3 exons (4.1%).

In the 63 MHC exons, the number with at least one heterozygous call (averaged for species with multi-individuals) are, for the cow, 20 exons (31.7%), the bison, 11 exons (17.5%), African buffalo, 21.5 exons (34.1%), bighorn, 11 exons (17.5%), deer, 12.6 exons (20.1%), and pigs, 6 exons (9.5%, Fig. 4-S3).

False heterozygous calls on the X chromosome in males

Since males inherit a single X chromosome, heterozygous base calls (nucleotide sites) for males on the X represent false heterozygotes. Our cow individual was a male, as were all of our deer samples except deer number 2. Our taurine reference genome had no Y chromosome sequence. Stampy-based mapping of our male cow showed heterozygous genotype calls at 107 sites, 0.1% of the 74,130 exonic nucleotide positions covered at 20X/Q50 on the X chromosome. In the 7 male deer, total heterozygous genotype calls on the X ranged from 95 to 137 for each deer. The one female deer had 89 heterozygous calls.

Most of these (false) heterozygous calls were clustered together within relatively few genes. The largest 3 clusters of heterozygous calls within genes on the male cow X chromosome (in the Stampy alignment) accounted for 79% of the 107 total heterozygous sites. These included 32 calls in gene UBA1, known to have a paralog, UBE1Y, on the Y chromosome in humans (Murtagh et al. 2012). There were 25 calls in exons in gene PDHA1, 17 in gene EIF2S3. Both PDHA1 and EIF2S3 are known to have been retrocopied onto autosomes in humans (McLysaght 2008). There are 11 calls in ZFX, known in bovines to have a paralog, ZFY, on the Y chromosome (Poloumienko 2004). Another potential source for false mappings to the ZFY is the ubiquity of the zinc finger domain in the genome. The remaining genes, with the number of associated heterozygous calls are CA5B (10 calls), EIF1AX (6), NONO (2), GPM6B (2), and MMGT1 (1). The BWA alignment produced its 25 heterozygous calls in 4 genes, most numerously in ZFX, with 9 calls in one exon. Among genes on the X chromosome with BWA-aligned heterozygous calls, only one, NXF3 (1 call), was not represented in the Stampy-based genotyping.

The Stampy alignment on the X chromosome for our six male deer showed relatively high false heterozygous call counts in these genes as well, with counts for individuals ranging from 13-22 calls in EIF2S3, 7-15 calls in ZFX, and 10-34 calls in UBA1. The female showed no heterozygous calls in EIF2S3 and ZFX, and 12 heterozygous calls in the UBA1 gene. Not seen in the cow but with high call counts in the deer (including the female) is MCTS1, with 8 calls for the female, and 9-16 for the males. MCTS1 is known to be the origin of a retrogene (MCTS2) in mice and humans (Cowley and Oakey 2010).

Further evidence of mappings onto the X of reads from non-X loci comes from the ratio of mean coverage on the X chromosome versus the autosomes. With their single copy of the X, males have half the expected mean coverage on the X versus the autosomes. The male cow's ratio of coverage on the X to that on the autosomes is 0.70, which is higher that the expectation of 0.5. The ratio ranges from 0.77 to 0.95 in the male deer. These percentages are similar in both mappers. Further, even the female deer (deer 2) and female bighorn, (sheep 1) show ratios above the expected 1.00, with ratios of ~1.34 and ~1.17 for both mappers for deer and bighorn, respectively.

Comparison of mapping to the cow reference versus a less divergent reference genome

Since mapping accuracy is more likely when the sample-reference divergence is small, some measure of the reliability of our Stampy-based divergent mappings to the cow genome can be had by comparing them to mappings under relatively low divergence. We compared the success of mapping of reads (to homologous loci) between the cow and pig genome sequences, and between the cow and domestic sheep genome (Harris 2007; Fujita et al. 2010). As detailed in Methods, the Liftover program identified homologous exon sequences between the cow and pig, and, also, separately, between the cow and sheep genome sequences.

First we considered mapping consistency by looking for mapping concordance in the homologous sequences. We considered homologous sequences to show high mapping concordance when at least 80% of reads that aligned to either the sequence on the cow genome (by Stampy) or the homologous sequence on the less-divergent genome (by BWA, pig reads to the pig genome, in one case, bighorn sheep reads to the domestic sheep, in the other case) were aligned to both homologous sequences. Table 4.3 shows that 40% of the Liftover-selected exons with cow-pig homology were mapped with high concordance. The same measure for the bighorn mappings showed that the bighorn-cow and bighorn-domestic sheep mapping concordance was 56%.

In the pig, we also tested consistency in genotyping between the Stampy, cow-genome alignment and the BWA, pig-genome alignment. In homologous sequences, selecting exons whose most-common base was genotyped identically in both mappings, we found 1,342 such sites that were called as having at least two alleles. Of these 1,054 sites were found to be heterozygous in both, while 261, ~20%, were called as heterozygous only in the Stampy cow-alignment. Only 27, ~3%, were called as heterozygous solely in the BWA pig-genome alignment. This shows that the cow-based, Stampy mapping aligned reads to a given site with a SNP more often (by 20%) than did the pig-based mapping. Although all sites between the two mappings for this test agreed on the most common allele, the test did not determine whether, for sites for which both mappings found multiple alleles, the alternate alleles were of the same base.

Table 4.3: Liftover results. Genome-genome alignment totals and mapping concordance for pig reads mapped in the cow, versus pig reads mapped to the Sus scrofa genome. Similarly the bighorn was mapped to the cow and the domestic sheep (*Ovis aries*) genomes. Asymmetric liftovers refer to those in which the target and its liftover differed in length by more than 20%. 80% concordance refers to targets for which at least 80% of all sequence reads that mapped to either or both genomes were mapped in both. A target, either liftover or that of the original cow, is called unmapped if less than 10 reads were aligned to it.

	Targets lifted over	**Asymmetric liftovers**	**80% concordance**	**Unmapped, liftover**	**Unmapped, cow**	**Unmapped in both**
Pigs	20,223	142	8,003	770	255	1,532
Bighorn sheep	16,110	34	8,944	641	46	973

Coverage relationship to GC content and target length

Depth of coverage at a genomic position (i.e., the number of reads at a nucleotide site) was associated with exon length and also the proportion of guanine and cytosine bases (GC ratio) in exons (Fig. 4.7). Though we filtered our exon collection so that only 0.2% of our ~56,000 baits had GC percentages lower than 25% or over 75% (see Methods), we nonetheless saw a relationship between exon coverage and exon GC percentage (content). All species showed deepest coverage for exons with GC percentages from approximately 40% to 50% (Fig. 4.7b), reflecting a bias noted elsewhere as common to several exon capture protocols (Asan et al. 2011). Coverage was highest for exons with length of ~600-900 bp. This high coverage peak effect was most pronounced in the least divergent species, the cow and bison (Fig. 4.7a).

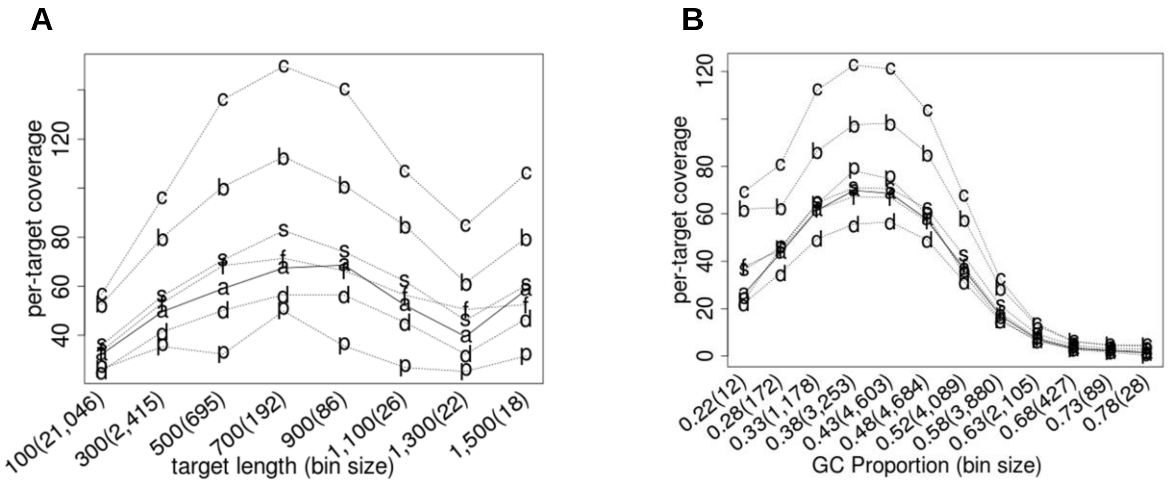


Figure 4.7: Sequence depth vs. GC and target length. For the Stampy mapping, per target (exon) coverage depth (number of read bases aligned to a target position) as a function of (A) target length (total bps) and (B) GC content (proportion of GC bases per target). GC and target length values are binned, and the mean of coverage depth plotted for each bin. Plotted letters are for (c)ow, (b)ison bu(f)falo, (s)heep, (d)eer, (p)igs, and (a)ll species (solid line). Sparse bins (less than 10 values) omitted include, in (A), 25 exons over 2000 bps in length and, in (B), 5 exons with GC proportion ~ 0.82. Coverage depth is here represented for each target by the mean value for each species, and one plot for the mean of all samples. Further, for a given individual, each depth of coverage value for each exon target is the mean depth for all of the base pairs in the sequence. X-axis numbers give values for bin midpoints, and are paired with parenthesized numbers giving the bin size. The increase in coverage depth in (A) for the bin of largest exons is mostly due to deep coverage of the single exon, 1,545 bp, of the COX1 gene on the mitochondrial chromosome. The coverage for the binned exons centered at 1,500 bp is reduced by an average of 8.7x for each species when this gene is excluded.

Discussion

Success of exon capture declined with increasing divergence between the sampled species and the cow reference genome, as expected. However, even with ~60 million years of divergence between pig and cow, over half of our targeted ~3.6 Mb were genotyped with high depth and quality scores (Fig. 4.4).

Also encouraging are the results with the deer samples, productive despite their relatively high divergence of ~30 million years from the cow (Guha et al. 2007). Four of the eight samples yielded over 68% of the targeted bp (at 20X/Q50 depth and quality). The other four yielded about half the targeted bp (50%). Overall, among seven individuals, half to 2/3 of the targeted exon base pairs were sequenced to a depth and quality sufficient for variant (SNP) discovery. Further, they yielded thousands of putative heterozygous genotypes (SNPs), despite our use of cow exon sequences for the capture probes, selected without regard to relative exon sequence divergence across our species.

Cross-species exon capture and availability of genome sequences

With the decreasing costs of whole genome sequencing using next-generation platforms, and efforts such as the 10,000 genomes project (Genome 10K Community of Scientists 2009), soon many vertebrate genera could have a species with a reference genome sequence with thousands of gene annotations. This would facilitate exon capture using the mapping and genotyping tools currently in use for the most common capture and alignment applications such those used in human exon capture, with the same species providing both sampled and referenced genomes.

For sample/reference divergence values of 10+ million years, using standard and widely available molecular protocols and bioinformatic tools, we sequenced and genotyped 50% to 65% of our targeted exonic bases at 20X/Q50. This, and high costs of whole genome assembly and annotation, suggests that cross-taxa exon capture is (and will remain) an attractive solution for genome-wide genetic marker discovery and genotyping in taxa with few genomic resources. For example, the work of Salzberg et al. (2012) suggests that assembling whole mammalian genome sequences from next-generation data alone presents difficulties that may inhibit their rapid proliferation.

Our most divergent species, the pigs (with a divergence of ~60 million years), yielded the lowest numbers of exon bases sequenced to 20X/Q50, as expected. However, the pigs provided many thousands of complete exon sequences. On average 9,261 of the targeted 24,525 exons were completely sequenced for a pig individual. They also yielded over two million of the 3.6 million targeted, exonic bases. The effect of the pig's relatively high divergence from the cow, computed at an average 8 differences in 100 bases between one-to-one orthologs genome wide (See Methods, “Read mapping and genotyping”), looks to be most telling in the MHC gene family, where each pig individual produced at most a handful of the 63 targeted MHC exons.

Which mapper is best, if any?

For the pig in particular, we saw a substantial increase in exon capture and sequencing success from using a mapper (Stampy) designed for read-alignment with relatively high sample-to-reference divergence (Figs 4.1 and 4.2). On average, for the pigs, Stampy-based genotyping showed ~620,000 (~30%) more inferred base calls in exons over the BWA-based genotyping, even when BWA was run allowing 10% mismatches between read and reference, an allowance larger than that recommended by BWA's creators.

A cost of allowing a high mismatch rate in alignments could be relatively high false positive SNP calling, even with high coverage and quality thresholds, as employed here. Our Liftover comparison (see Methods) of a pig-to-pig-genome BWA mapping with a pig-to-cow-genome Stampy mapping (see Results) showed that ~20% of sites inferred as heterozygous in the cow-based, Stampy alignment, were inferred as homozygous in the BWA pig-based mapping. While these heterozygous calls unique to the cow-based alignment in the comparison may represent true heterozygosity simply missed in the pig-to-pig alignment, it is prudent to consider these as false inferences, and as such to suspect an increased tendency toward false alignments in the Stampy cow based mapping, compared to the BWA pig-based mapping.

It may be that an alternative method without mapping may be preferable at the ~60 MY divergence of pig and cow. As described below, in "Alternative methods of genomic DNA enrichment," reference sequences based on a transcriptome assembly can serve as a species to species map for exon capture sequence reads.

MHC gene gamily sequencing

A smaller percentage of MHC exons were 100% sequenced (at 20X/Q50), compared to the total set of 24,525 exons targeted exome-wide, for all species except the cow. Only the pigs showed dramatically reduced success for MHC compared to the exome-wide capture when the threshold is lowered to 60% of exon bases aligned at 20X/Q50 (Figs. 4.1 and 4.3). For example, deer showed only 16% reduced MHC success (compared to the 62% reduced success for pigs). These reductions suggest genotyping and SNP discovery might be of relatively limited usefulness in gene families with many loci and high allelic diversity (like MHC) for highly divergent mappings (in this case ~60 million years between sampled pigs and the cow reference).

Nonetheless our results suggest that in the bovids and cervids at least 35% of the exons from this highly variable gene family can be sequenced (Fig. 4.3a). We have also found 13 of the targeted 63 exons in MHC genes that were sequenced to 60% of bases at 20X/Q50 in at least one individual of each species, including the pig. This is important because it suggests that, though success will be reduced, many MHC exons can be sequenced from most ungulates, including divergent species.

The difficulty of exon capture from the MHC family likely stems from its divergent haplotypes (Spurgin and Richardson 2010; Klein and O'hUigin 1993). This could result in reads from a single locus mapping to more than one locus on a given genome, as well as the converse case, reads from multiple loci in the sampled genome mapping to a single locus in the reference. Other gene families will likely present less of a challenge in sequencing exons to a depth and quality needed for genotyping because most families have fewer loci (less than the ~170 loci in ungulates), fewer alleles, and perhaps fewer highly divergent alleles per locus.

Heterozygous sites and sample/reference divergence

The proportion of heterozygous sites (heterozygosity) per individual did not show an obvious trend related to phylogenetic divergence from the cow reference (Fig. 4.5). An exception, perhaps, is seen in the highly divergent pigs, which, overall, show the smallest totals for both total called bases at 20X/Q50 (Fig. 4.4), and total heterozygous calls; the average SNP rate for 20X/Q50 exonic bases for the four individuals was 0.0007, Figs. 4.5 and 4.6). Observed heterozygosity in European and Asian domestic pig breeds, as well as wild boars, has been estimated to range from 0.33 to 0.70 (Zhang and Plastow, 2011).

The pigs aside, the proportion of heterozygous sites in an individual (for a given species) likely reflects relative levels of inbreeding and effective population sizes for the species or population. Our 2 African buffalo individuals, for example, showed high SNP proportions compared to the other species, with a SNP rate for 20X/Q50 exonic bases of ~0.016 (Fig. 4.5). African buffalo populations in three Serengeti regions show high genetic variation, with observed heterozygosity at 15 microsatellites averaging 0.70, 0.67, and 0.75 (Ernest et al. 2012).

A comparatively low proportion of heterozygous sites in the bison (Fig. 4.5), may reflect the severe loss of genetic variation during a near extinction event, ca. 1900, due mostly to over hunting (Hedrick 2009). Deer had a relatively high individual heterozygosity, which is not surprising because deer populations have very large effective population sizes, long distance gene flow, and genome-wide introgression between distinct species (Latch et al. 2011).

Overall these proportions of heterozygous sites in individuals are likely underestimates of true rates for the targeted regions, as the stringency of the filter for eliminating false positive SNP calls likely discarded many true SNPs (heterozygous sites), especially when they were called near indels (see Methods).

The differences in counts of heterozygous sites (per thousand bases) between exonic (targeted) positions versus flanking region positions suggest an effect of phylogenetic divergence on SNP discovery: the flanking heterozygosity counts exceed those for exonic counts for the cow, bison, and African buffalo (Fig. 4.5), but are lower than exonic counts in the more divergent bighorn, deer, and pigs. As noted in Methods, our filter rejected heterozygous base calls near indel calls, which are more numerous in the flanking regions versus the exons. They are especially numerous for bighorn, deer and pigs in flanking regions. For Stampy alignments, on which the heterozygous calls are based, the mean number of indel calls per reference site genotyped at 20X/Q50, in exons (and flanks), are, for cow, 0.0001 (0.0004), bison 0.0003 (0.0012), African buffalo, 0.0008 (0.0038), bighorn, 0.0015 (0.0077), deer, 0.0020 (0.0101), and pig, 0.0052 (0.0367).

A 'universal' ungulate array for SNP discovery, genotyping, and population genomics

With most ungulates lacking reference genomes, it would be useful for population geneticists (and phylogeneticists) to have a set of exons with a high likelihood of capture, for SNP discovery and genotyping. Our results indicate that a "universal ungulate array" is possible, with the number of exons varying according to the stringency applied to selecting the exons. A least stringent collection could be based on the 13,131 exons (~2

Mb in 4,466 genes) with at least 60% of bases sequenced to 20X/Q50 in at least one individual in each of our species sampled. A smaller (more stringent) collection could be limited to the 8,999 exons (~1.1 Mb, in 3,546 genes) that were completely sequenced to 20X/Q50 in at least one individual in each species. Of these, an even more conservative collection can be assembled, containing only those 5,660 exons (~0.6 Mb in 2,884 genes) that meet a ceiling on coverage, rejecting any base calls with coverage that is over twice the mean coverage for the individual at the targeted, exon sites (see Methods, on single nucleotide variant filtration). Such a ceiling can obviate false mappings due to gene duplicates, e.g., aligning together at one locus reads from paralogs.

These collections of exons are distributed widely across all of the cow chromosomes. For those covered to at least 60% of their base pairs, in at least one of each species we sampled, the per-chromosome exon count shows a minimum of 220 exons covered on the cow's chromosome 25 and a maximum of 873 on the chromosome 1 (the largest chromosome at ~160 Mb). For the set of completely covered exons, the minimum is 142 exons on the chromosome 25 and the maximum, 645 exons on chromosome 1. For the smallest (conservative) set of 5,660 exons (with the coverage ceiling applied) the minimum is 97 on chromosomes 23, 25, and 29, and the maximum is 396 exons on chromosome 1.

This genome-wide distribution of exons facilitates scanning for markers associated with phenotypic traits when no gene is currently a candidate. Gene positioning (synteny) is generally highly conserved among mammals (Rettenberger et al. 1995), so that their dispersal across the cow genome suggests a similar dispersal in the other species.

Tables listing gene abbreviations and exon starts on the cow reference genome for these exon collections are available in supplementary materials.

Variability among deer individuals in sequencing success

Our deer samples produced a wide range of difference in mapping and sequencing success, showing within-species differences close to that between species (Figs 4.1-4.3). We reported on 7 of 8 deer samples, excluding our least successful sample. The excluded sample had a very low sequence success compared to the others, indicating poor sample quality. However, we do report three of the remaining 7 individuals, deer 4, 6, and 7, with consistently lower sequencing success rates than those seen in the other deer (Figs 4.1-4.3). The reason for this group's relatively poor performance is unknown. We note that these three individuals were captured in the only 3-plexed hybridization reaction (Table 4.5, in Methods). However, we have no evidence that this was the cause of their poor performance. All other samples, except the cow, were hybridized to the bait (capture) array in 2-plexed reactions. The cow individual's DNA was hybridized without other samples in the reaction.

Picard tools read alignment metric software (Handsaker et al. 2009) showed that our least productive deer sample produced fewer total Stampy aligned bases at base quality 20 or more than did any other sample (including the four pigs, which are phylogenetically more divergent from the cow than are the deer). Deer 4 produced fewer Stampy aligned bases at Q20 or greater than all samples except our excluded deer. Deer

4, 6, and 7, also had the fewest total reads of all the samples, before adapter and base quality trimming.

Flanking site sequencing

Flanking base pairs (not in exons) provided more sequence data and SNPs than did exons (e.g., Fig 4-S1). Genotyping 250 bp flanks (11.6 Mb), totals in the cow and bison provided 50% more sequence than did exon sites. At least two individuals of each species that was sampled in more than a single individual produced 20X/Q50 genotyped sites in exon flanks numbering at least 90% of the total for exon sites. Flanking sites are generally less conserved than exons and thus can provide more polymorphic markers (per thousand bp of sequence). Despite likely increased SNP filtering (i.e., rejection) from increased indel calling in the flanks, they still provided substantial numbers of heterozygous calls and thus putative SNP markers (Fig. 4.5b).

Limitations and future research

An important qualification for mapping and genotyping of divergent species in this study is the absence of a more certain test for detecting genotyping errors, especially for calling single nucleotide polymorphisms. As noted above, the Liftover test, using the pooled pigs and stringent SNP filtering, revealed that ~20% of the sites called as heterozygous, were called as such only in the Stampy, cow-based mapping (they were called as homozygous in the pig-based mapping), suggesting a high false positive SNP rate in the divergent, cow-based mapping.

The percentage of inferred false heterozygous calls in the male X chromosomes (see Results) likely do not provide a reliable rate for false positive heterozygous calls on the autosomes. Though the distribution of target lengths shows generally that those on the X chromosome (median 129 bp, standard deviation, 142 bp) are not dramatically larger than those in the autosomes (median 115.5 bp, standard deviation, 139), for targets with at least one SNP the concentration of SNPs is notably higher in the X for the male cow, with 4.28 heterozygous calls per exon, and only an average 1.6 for the autosomes. In male deer, the rate ranges from 2.39 to 3.08 for the X chromosome, and 1.61 to 1.89 for rates averaged over the autosomes. For the female (deer 2), the rate in the X chromosome is 1.75, and the average rate over the autosomes is 1.66. These rates suggest that the false heterozygous calls in the males on the X chromosome could be partly attributable to peculiarities of the X as the source, for example, of retrocopied genes and pseudoautosomal regions, and are unlikely to reflect false positive rates in the autosomes.

It is also possible that the build assembly of the X chromosome, could contribute to false mappings. The genome build assembly statistics for btau4.0 (at http://www.ncbi.nlm.nih.gov/assembly) show that the chromosome X assembly had the lowest scaffold N50, at 1,086,000 bp, compared to the median for the chromosome assemblies at 2,251,000 bp. Further, an assessment of chromosomal positions of SNPs based on pair-wise linkage disequilibrium estimation has been used to test the integrity of the btau4.0 assembly, finding inconsistencies in SNP positions in the X chromosome sequence that suggest a general problem with its assembly, while finding a “high level of integrity” for the assembly as a whole (Khatkar et al. 2010).

A useful follow up to this work would be to quantify genotyping error rates from exon capture by genotyping exon capture-inferred genotypes at SNP loci using independent sanger sequencing or SNP genotyping assays (or SNP chips).

A further filtering criterion may be needed, limiting insert sizes, that is, the length (in bp) of the region bounded by the alignment to the cow genome of paired end reads. In the pigs and deer, ~0.5% to about 1.8% were at least 10 thousand base pairs, whereas the mean size in the alignments was reported to be about 250 bp. Though relatively rare these suspected false mappings could contributed to false positive SNP calls.

Alternative methods of genomic DNA enrichment

An alternative to exon capture for genome-wide SNP discovery when the subject has no close relative with a genome sequence is restriction-site associated DNA sequencing (RAD) (Miller et al. 2007a; Rowe et al. 2011). This method can be scaled, to some extent, as to the number of loci sequenced, by the choice of restriction enzyme (e.g. a 6 versus 8 cutter). However, unlike exon capture, RAD is not gene-targeted and thus is less amenable to targeting candidate genes, to test for genetic associations with fitness or phenotypes (Bruneaux et al. 2013; Luikart et al. 2003). Exon capture is also far more flexible in the number of loci sequenced, e.g., from 100s to millions (Cosart et al. 2011; Rivas et al. 2011; Gnirke et al. 2009; Hancock-Hanser et al. 2013). A disadvantage of exon capture, compared to RADs, is the cost of an array (e.g., > $100 per array). However, the cost of exon capture is declining as the methods for simultaneous (multiplex) capture of many individuals in a single reaction are being refined (Bansal et al. 2011).

Especially for species whose nearest relative with a reference genome sequence is at or beyond the ~60 MY represented by our pig-to-cow capture, an exonic reference can be constructed from the species own transcriptome. With a nearest reference sequence genome (of mice and rats) representing ~70 million years of divergence, Bi et al. (2012) developed their own chipmunk exonic reference by sequencing whole transcriptomes. One disadvantage of this approach is the need for two rounds of sequencing and read assembly or alignment, one for the transcriptome, a second for exon capture, while its advantage is its effectiveness in species genetically distant from any species with a reference genome.

Conclusions

Our capture and sequencing produced 8,999 exons (~1.1 Mb) sequenced at high depth and quality, in 3,546 genes from at least one individual from each of the five divergent ungulate species. This represents 37% of our targeted exons, and 54% of our candidate genes. This number exceeds 13,000 (54% of targeted exons), if we require only 60% of each exon to be sequenced at 20X/Q50. These exons are candidate targets for development of a universal ungulate array to sequence thousands of genes in all ungulate species. Future whole exome sequencing of a wide range of taxa similar to ours could provide an even larger pool of exons for such an array.

Our results suggest that cross-species exon capture can sequence, at high depth and quality, thousands of exons in species 10s of millions of years divergent from the reference genome. Further, successful SNP discovery is seen in captures among 7 deer, yielding 12,645 putative exonic SNPs in 6,574 targeted exons, from 3,668 genes.

Exon capture allows precise targeting of a subset of the exome. This decreases the costs of population genomic studies. Cost and analysis time is also reduced by its optimized, well tested, and freely available mapping and genotyping tools. Exon capture has growing potential as a tool for gene-centered marker discovery and genotyping, for those addressing to-date intractable questions about the genomic basis of adaptation in nonmodel mammals with few genomic resources.

Methods

Exon target selection and probe design

We collected 24,525 exon sequences from the btau4.0, *Bos taurus,* genome sequence (Elsik et al. 2009), as annotated by the alignment of mRNAs from the NCBI RefSeq database (Pruitt et al. 2011) by the BLAT program (Kent 2002). The genome sequence, alignments, and annotation are available at the UCSC genome browser web site (Fujita et al. 2010). Of the 24,525 exon targets, 2,542 were selected from 349 candidate genes associated with speciation, reproduction, and disease resistance. Annotations for some of these came from NCBI's Entrez Genome site (Maglott et al. 2010). Noncandidate gene exons were collected with the aim of sampling from as evenly as possible across the 29 autosomes and the X chromosome. One mitochondrial gene, COX1 (cytochrome c oxidase subunit I), was also targeted. For sampling of genes evenly across the chromosomes, we used the exon selection software described in Cosart et al. (2011). In seeking a balance between the number of genes sampled and the total size of the collection, we limited the total base pairs collected per gene to 3,000 bp from candidate genes, 1,000 bp from noncandidates.

For each gene, the upstream external exon (the 5', with its UTR) was collected first, if it was under the base pair length limit. More exons were then added in random order. As exons were collected, any bringing the current total above the per-gene bp limit were skipped and another of the remaining tested for inclusion. This was repeated until no exon could be added without exceeding the limit. If all exons for a gene exceeded the per-gene bp limit, a single exon, preferably the upstream external exon, was sampled from its ends, half the per-gene bp limit from each end. Of our 24,525 exons, ~130 were end-sampled.

After an initial collection of exons totaling about 5.5 Mb, probe design included partial filtering of exons that had baits with CG content at or below 25%, or at or above 65%. Probe filtering left 127 exons at or below 25% and 1,806 at or above 65%. Probe design also resulted in 1X tiling for exons of length 120 bps or less (only one probe for the target, rather than 2 probes for each base pair, the design for exons longer than 120bp). For these shorter exons in candidate genes, we used two identical probes instead of one. In order to meet the limit of about 56,000 probes, 1X-tiled exons from noncandidate genes were randomly discarded until the probe limit was met. For all exons over 120 bp, at least three, 120-bp, overlapping probes targeted the exon. The overlapping scheme allowed for most of the exonic base pairs being represented on two different probes. On the ends of exons longer than 120, of length *l,* there were 120 – [120 - (*l modulo* 120)] bases covered by only a single probe. The resulting set of targets totals about 3.6 Mb.

Sample Information

Sample information is given in Table 4.4. The degree of multiplexing in the exon capture hybridization reactions is given in Table 4.5, which also shows the number of reads produced by the sequencer that passed its filter.

Table 4.4: Ungulate species and sample information. Sex of samples notated with an asterisk were inferred by the ratio of mean read coverage on the X chromosome to the mean read coverage over all autosomes, computed for both Stampy and BWA mappings. Samples inferred to be males had ratios under 0.70, and samples inferred to be females had ratios within 0.01 of 1.0. Note that for samples whose sex was established before sequencing, mean ratios of X to autosome coverage was 0.81 for the 7 males, and 0.99 for the 6 females. Entries notated with a dagger symbol are those designated females with ratios below 1.0 (both have ratio 0.61).

Sample Name	**Common Name**	**Species**	**Location**	**Sex**
BbSP23	bison	*Bison bison*	Montana	Male*
BtGRA9907	Cow	*Bos Taurus*	unknown	Male
OcGT653	sheep 1	*Ovis Canadensis*	Wyoming	Female
OcTAH033	sheep 2	*Ovis Canadensis*	Colorado	Male*
OhBTD01	deer 1	*Odocoileus hemionus columbianus*	Oregon	Male
OhBTD02	deer 2	*Odocoileus hemionus columbianus*	Oregon	Female
OhBTD03	deer 3	*Odocoileus hemionus columbianus*	Oregon	Male
OhBTD04	deer 4	*Odocoileus hemionus columbianus*	Oregon	Male
OhMD01	deer 5	*Odocoileus hemionus hemionus*	Oregon	Male
OhMD02	deer 6	*Odocoileus hemionus hemionus*	Washington	Male
OhMD03	deer 7	*Odocoileus hemionus hemionus*	Washington	Male
OhMDX	deer 8	*Odocoileus hemionus hemionus*	Washington	Male
ScR019	buffalo 1	*Syncerus caffer*	South Africa	Female*
ScR23	buffalo 2	*Syncerus caffer*	South Africa	Female*
SdDPIP60382	pig 1	*Sus scrofa domesticus*	Europe	Female†
SdSUSCN1	pig 2	*Sus scrofa domesticus*	China	Female
SsWBCN2	pig 3	*Sus scrofa*	China	Female†
SsWBPIP29c	pig 4	*Sus scrofa*	Europe	Female

Table 4.5. Exon capture hybridization quantities. *N*-plex values for the hybridization reaction indicate that the sample was one of *n* samples in the capture pool. Values for total reads passing the sequencer filter is a count of the reads that the Illumina read filter marked as passed.

Name	*n*-plex in capture	DNA concentration in library (ng/ul)	Species in capture	Total reads passing sequencer filter
bison	2	40.7	bison, deer	21,191,017.00
cow	1	39.3	Cow	27,138,148.00
sheep 1	2	49.5	Sheep	17,969,594.00
sheep 2	2	49.1	Sheep	16,791,181.00
deer 1	2	48.8	Deer	17,993,105.00
deer 2	2	49.7	Deer	15,048,996.00
deer 3	2	45.1	Deer	10,857,633.00
deer 4	2	50.2	bison, deer	16,370,290.00
deer 5	3	45.4	deer	6,502,041.00
deer 6	2	41.04	deer	20,056,258.00
deer 7	3	48	deer	9,443,069.00
deer 8	3	42.9	deer	8,930,130.00
buffalo 1	2	40.3	buffalo	13,343,022.00
buffalo 2	2	41	buffalo	14,935,366.00
pig 1	2	47.1	pig	16,067,560.00
pig 2	2	46.9	pig	11,919,777.00
pig 3	2	46.7	pig	11,979,120.00
pig 4	2	55.5	pig	12,850,661.00

Exon capture and sequencing[3]

SureSelect sequencing libraries were prepared according to the manufacturer's instructions (Agilent) with the following modifications. Five µg of genomic DNA in 120 µl TE-buffer was fragmented to a median size of 200 bp using the Covaris-S2 instrument (Covaris) with the following settings: duty cycle 10%, intensity 5, cycles per burst 200, and mode frequency sweeping for 180 s at 4°C. The fragmentation efficiency was evaluated by capillary electrophoresis on DNA1000 chips (Agilent) and the concentration of the DNA was estimated by PicoGreen assay (Invitrogen).

Library preparation of 1.5 ug of genomic DNA followed the TruSeq protocol (Illumina). The adapter ligated and size selected DNA was amplified by PCR. Twenty-

[3] Quantities and supplier names in this description of the library preparation and capture hybridization is pending confirmation by one of my collaborators.

four µl of DNA, 5 µl Illumina primer mix, and 25 µl Phusion master mix (Finnzymes) were amplified as follows: 30 s at 98°C, 14 cycles of: 10 s at 98°C, 30 s at 65°C, and 30 s at 72°C, then 5 min at 72°C. The reaction product was purified using AmPureXP beads and eluted into 30 µl EB. The quality of the PCR products was assessed by capillary electrophoresis (Bioanalyzer, Agilent) and the concentration of the DNA was estimated by qPCR using PicoGreen assay.

SureSelect hyb #1, #2, #3, and #4 reagents (Agilent) were mixed to prepare the hybridization buffer. The adapter ligated DNA fragments were concentrated in a DNA120 SpeedVac concentrator (Thermo Electron) to 500 ng in 3.4 µl. The blocker mix was modified for use with the TruSeq library kit. In addition, bovine Cot1DNA was added to remove repetitive elements. SureSelect block #1, #2, and #3 reagents (Agilent) were added to the 500 ng of DNA. The hybridization buffer and the DNA blocker mix were incubated for 5 min at 95°C and then for 10 min at 65°C in a thermal cycler (MJ Research). RNase block (Agilent) was added to the SureSelect oligo capture library (Agilent). The capture library was incubated for 2 min at 65°C. First the hybridization buffer, and then the DNA blocker mix were added to the capture library and the mixture was incubated for 24 hours at 65°C in a thermal cycler (MJ Research). Fifty µl of streptavidin coated Dynabeads M-280 (Invitrogen) were washed three times with 200 µl SureSelect binding buffer (Agilent) and resuspended in 200 µl of the binding buffer. The hybridization mixture was added to the bead suspension and incubated for 30 min at RT with mixing. The beads were washed with 500 µl SureSelect wash buffer #1 (Agilent) for 15 min at RT, and three times with 500 µl SureSelect wash buffer #2 (Agilent) for 10 min at 65°C. DNA was eluted with 50 µl SureSelect elution buffer (Agilent) for 10 min at RT. Fifty µl of SureSelect neutralization buffer (Agilent) was added to the eluted DNA. The reaction product was purified with SPRI beads eluting in 45 µl H2O. One PCR reaction with 14 µl of the elution product, 1 µl primer 1.1 (Illumina), 1 µl primer 2.1, 10 µl Herculase II reaction buffer (Agilent), 0.5ul 25mM dNTP mix, 1 µl Herculase II Fusion DNA polymerase (Agilent), and 22.5 µl H_2O were performed. The PCR conditions were as follows: 2 min at 98°C, 14 cycles of: 20 s at 98°C, 30 s at 60°C, and 30 s at 72°C, then 5 min at 72°C. The PCR reaction was purified with SPRI beads and eluted into 30 µl H2O. The quality of the sequencing libraries was verified by capillary electrophoresis (Bioanalyzer, Agilent) and the concentration was estimated by qPCR (KappaBioscience assay).

Eighteen indexed samples were combined into three unique pools and clustered to individual flowcell lanes using TruSeq Paired-End Clustering kit on a cBot at 4 pm titration (Illumina). Sequencing was performed by the manufacturer's recommendation on a HiSeq 2000 using 100-cycles of chemistry on each end of the fragments and 7-cycles to read the barcode.

Read mapping and genotyping

We trimmed, filtered, and mapped Illumina 100 base pair, paired end reads as follows:

Adapter trimming

Shorter fragments sometimes yield ligated adapter sequence on the 3' end of the read. These were trimmed with an in-house script. This script also trimmed from the leftmost "N" (undetermined) base pair to the 3' end. It discarded reads whose length after trimming was under 30 base pairs.

Quality filtration and trimming

The fastx toolkit (The Hannonlab 2010), removed reads whose percentage of phred-scaled base quality scores at 20 or more comprised less than 85% of the total scores. The fastx toolkit also scanned each read from its 3' end, and discarded bases until finding one with a score of 20 or more. The trimmer also discarded reads trimmed to a length under 30 bases, and, combined with adapter trimming and quality filtration, reduced total reads in the samples by an average of 21%, the maximum loss at 27%, the minimum at 19%.

Mapping

Reads were aligned using two mapping programs: BWA (Li and Durbin 2009), versions 0.5.9rc1 through 0.6.2-r126 and Stampy (Lunter and Goodson 2011) version v1.0.13.

BWA is a fast aligner that uses relatively little memory. Its creators note it is designed for a "low base error rate" of under 3% (see the BWA manual at http://bio-bwa.sourceforge.net/bwa.shtml#6). The error rate includes mismatches between read and reference caused by phylogenetic divergence between the sampled genome and referenced genome. In addition to a fast lookup strategy to find reference matches to read sequences (see details in Li and Durbin 2009), BWA also owes some of its high efficiency to a seeding method that uses the leading *n* bases in its initial search of the reference for candidate alignments, with *n*=32 in the default. For Illumina reads generally, base qualities are higher on the leftmost end of the read. We employed the default seed length for all BWA alignments, and default mismatch allowances, which is a maximum of 2 mismatches in the seed, and, for 100-bp reads, a maximum of 5 mismatches total, including gaps.

Stampy is designed to be sensitive to mismatches between the reference and the reads. In the version we used, a hybrid mode employs BWA to align reads with few mismatches, and then its own (slower) algorithm for the others. Stampy's algorithm looks for candidate alignments using multiple 15-bp overlapping sequences throughout the read, adding sensitivity over the 32-bp seed-enabled BWA, in locating candidate alignments on the reference genome sequence.

After read trimming and filtration, when both of the paired-end reads from the same fragment remained, they were mapped as paired. Singleton reads (whose mates were lost to trimming or filtering) were mapped as single-end reads, and pooled with the paired-end reads for genotyping.

Initial mappings of all samples with BWA's default parameters were used to establish base mismatch rates for each species in the targeted regions (Table 4.6). These rates were chosen by averaging the mismatch rates in a BWA alignment (with default parameters) for all individuals of a given species. The rates for each individual were provided by Picard software summary metrics (Handsaker et al. 2009).

It should be noted that, for the pigs, we made a subsequent calculation of the difference between homologous gene sequences between the cow, bosTau6, genome (Schatz et al. 2009, a later build than the bosTau4 on which from which our exon sequences were drawn), and the Sscrofa10.2 pig genome sequence (Groenen et al. 2012). The calculation was based on sequences assigned as one-to-one orthologs by the Ensembl database (Flicek, et al., 2013), and showed that the average proportion of exon bases mismatching between the blast-aligned orthologous sequences was 0.08, so that our selection of 0.05 (Table 4.6) was likely an underestimate of the true mismatch rate.

Table 4.6: Substitution rates used in the Stampy mapper. Each is based on the mismatch rate of a BWA alignment, using default parameters, to the targeted exons.

Species	Substitution rate
Bos taurus	0.0055
Bos bison	0.0100
Syncerus caffer	0.0220
Ovis canadensis	0.0250
Odocoileus hemionus	0.0380
Sus domesticus	0.0500
Sus scrofa	0.0500

The initial BWA mapping was also the final BWA mapping for the cow, bison, African buffalo, and bighorn, as their estimates of sequence mismatch rate were under 3% (Table 4.6). The BWA default maximum mismatch rate in the seed is ~6% (2 bases in the first 32) and the default allowance is 5% for total mismatches in 100 bp reads. Though these settings seemed sufficient for all species with mismatch rates below 3%, it may be that site count comparisons of BWA with Stampy would show African buffalo and bighorn counts as more even between the two mappers, if we had relaxed these BWA parameters for these two species (Figs 4.1-4.3).

For the deer, BWA was used with parameters "-k 3" and "-n 8," to allow for 3 mismatches (substitutions or gaps) in the seed, and 8 in the alignment overall. For the pig, BWA was used with parameters, "-k 4" and "-n 10." These settings do allow for mismatches in alignments with more than these thresholds, but they also mean that potential alignments with total mismatches exceeding them may be missed. It also should be noted that the nondefault settings of "-k 4" and "-n 10", in particular, had severe performance costs, and that BWA's documentation recommends avoiding setting mismatch parameters that allow for many beyond the default (see the FAQ at http://bio-bwa.sourceforge.net).

In mapping all of the samples, Stampy was used in its hybrid mode, with parameters as instructed in the documentation. For the Stampy alignment proper, we used per-species substitution rates as given by the mismatch rates taken from default

BWA mappings (Table 4.6). These were set using Stampy's "-r" parameter.

Post mapping quality control

After mapping, reads inferred to be created from PCR-duplicated fragments were removed from the mapping using the Samtools software (Li et al. 2009). False SNP calls can occur around indel mappings from read misalignments. To reduce such calls the local realignment tool of the GATK toolkit (McKenna et al. 2010) was used to identify and realign reads in suspect regions.

Genotyping

The GATK toolkit's UnifiedGenotyper program (version 1.6-7-g2be5704) called genotypes at targeted positions and a maximum of 250-bp flanking positions (250 bps unless chromosome ends or adjacent exons interfered), using both the SNP and INDEL likelihood models to call both types of variants. The output mode was for all sites (not just variants). The Phred-scaled threshold for confident variants was set at 50 (the "stand_call_conf" parameter). The documentation for the genotyper and its related programs recommends a Phred-scaled consensus quality threshold of 30 for projects with average coverage at or over 10X (http://www.broadinstitute.org/gatk/guide/topic?name=best-practices). However, given the span of divergence in our individuals, we used the more conservative value of 50. The consensus quality score is a log-transformed probability *p* that the inference of variation or reference-match at a site is an error, with Q50 indicating *p*=0.00001.

Single nucleotide variant (SNV) filters

For variant calling the GATK UnifiedGenotyper was used a second time for species with more than one sample (see Table 4.1). This second round used the same thresholds as above for genotyping, but was performed on samples combined by species so that, for example, the four pig samples are genotyped together, while still retaining individual values for genotype and depth and quality (that is, genotype quality, described below). Genotyping all individuals of a species together served both to increase the confidence of consensus base calls at a given position, when many individuals agreed on alleles, and to conveniently count the number of chromosomes with a common variant at a given position.

For calling both heterozygous positions and homozygous differences from the reference, we required a minimum genotype quality (GQ) of 30, the Phred scaled probability *p* that the genotype is wrong, conditioned on the site being variant compared to the reference. Q30 encodes that *p* = 0.001. Confidence in the site's having a matching base or variant(s) to the reference is given by the consensus quality, which we again required to be 50 (Q50) or greater. We required coverage of 20X for individual heterozygous calls. Of the 2 bases in a heterozygous call, the base with least depth of coverage was required to have at least 20% of the total depth for the two bases. Coverage and GQ values are for individual samples, while the consensus quality was calculated on the combined alignments of a given species (for example, the combined coverage of 7 deer, or 4 pigs). For all SNVs, we accepted calls with no more than twice the average individual read depth (averaged across all targeted sites), following recommendations for SNP calling in the manual for the Samtools software (http://samtools.sourceforge.net/samtools.shtml). Further, no SNV was accepted if it was within 3 base pairs of an insertion or deletion, or a member of a cluster of 5 SNVs within any contiguous 10 base positions.

Measuring sequencing success

For each sampled individual, we measured success of exon sequencing by the number of taurine reference exons, and individual taurine exon base pairs (bp), to which the individual's sequence reads have been aligned. For a nucleotide site to be included as having been sequenced, we require that a minimum of 20 sequence reads (bases) are aligned to the site (20X coverage), and that the genotyper computes a Phred-like consensus quality score of at least 50 for the genotype inferred at the site. 20X coverage and a consensus quality of 50 are stringent thresholds compared to those that have been used to call heterozygous positions in exon capture with less sample reference divergence, e.g. 8X and Q30 (Ng et al. 2009b; Cosart et al. 2011).

We measure the proportion of an exon that is successfully sequenced for a sampled individual as the proportion of the total bases in cow reference (targeted) exon that are aligned (at 20X/Q50) with the individual's sequence reads. If the exon under analysis was one of the ~130 end-sampled exons (see section above, "Exon selection and probe design," we counted it as a complete exon if the targeted end was completely sequenced to 20X/Q50. These coverage and quality thresholds are notated throughout as "20X/Q50."

We quantified success as both (i) the total number of exons sequenced as well as (ii) the total numbers of base pairs sequenced in exons. To quantify (i), we used two criteria: (1) 100% of bases sequenced in each exon, and (2) at least 60% of bases sequenced. A somewhat arbitrary choice, 60% means a majority of the base pairs in the exon are genotyped at our stringent 20X/Q50 standard, sufficient to identify markers (SNPs) in many important genes.

Using Liftover for the concordance tests

We tested accuracy in mapping and SNP calling from a comparison of mappings based on the cow genome to mapping of the same reads to a reference phylogenetically closer to the sampled individuals. Although errors in mapping and genotyping are seen in any analysis, the phylogenetically closer mapping should be equally or more accurate than our divergent mapping, and can serve, at least approximately, as a standard against which to compare the more divergent cow-based mapping.

Given that two genome sequences have been aligned (Harris 2007; Fujita et al. 2010), the Liftover computer program (Fujita et al. 2010) provides aligned genomic coordinates (chromosome, start base, end base) between the two. The UCSC genome browser (Fujita et al. 2010) has aligned the cow genome, btau4.0 (Roepstorff et al. 2006) both to the pig genome, susScr2 (Archibald et al. 2010a), and the domestic sheep genome, oviAri1 (Archibald et al. 2010b).

We downloaded the Liftover program from the UCSC Genome Browser site at http://hgdownload.cse.ucsc.edu/goldenPath/bosTau4/liftOver. Liftover with default parameters gave pig genome and sheep genome coordinates aligned to our cow genome exon coordinates. With default parameters the program produces alignments only when at least 95% of bases in an interval in the source genome maps to the target genome. It also stipulates that the source interval does not align to multiple regions.

For concordance tests in the pig the four pig individual Stampy mappings to the cow were pooled. BWA mappings, with default parameters, of the pig reads to the pig genome, were also pooled. The same procedure was used for the two bighorn samples and the sheep genome.

Mapping concordance between the Stampy cow alignment and the Liftover-aligned homologous exons (pig or sheep) was computed for each pair of aligned exons by two counts for each homologous pair of exons. First we counted (by read ID) the number of reads that mapped within the exon intervals (that is, reads that overlapped the interval by at least one bp), in at least one of the pair of alignments. Of these a second count was done of those reads that mapped within the intervals of both of the alignments. This second count, as a percentage of the total in the first count, was considered a measure of concordance in the mapping.

We also tested concordance in positions of heterozygous calls between exons in the cow genome and pig genome by making one set of exon consensus sequences from pooled pig alignments to the cow genome (the Stampy alignment as described above) and another set of consensus sequences from the pig genome (BWA with default parameters). Both of these sets were derived from genotype calls in the 20,223 exons that were successfully lifted-over from the cow to the pig genome (Table 4.3). GATK's Unified Genotyper produced genotypes for positions in the cow and lifted-over pig exon intervals, for the pooled 4 pig individuals, so that the genotyper treated the pooled reads as a single sample.

To create consensus sequences, genotyped positions were divided into separate sequences when called genotypes were separated by 10 or more positions that could not be called. Genotype gaps smaller than 10 were filled in with N's. For positions genotyped as having more than one allele the base used in the consensus sequence was that represented in the largest number of reads.

BLAST alignments (McGinnis and Madden 2004) paired the pig-based consensus sequences with cow-based consensus sequences. Though homologous positions were likely represented in many of the alignments, the positions most likely to be homologous were those in BLAST alignments of matching length and perfect identity (all bases matching, so that. in cases of heterozygous genotypes, the base with the highest coverage matched that in the other, aligned consensus sequence).

Restricting comparisons to positions with 20X/Q50, the alignment of the 2 groups of consensus sequences produced 7,770 such pairs. In these we recorded sites for which the genotyper found more than one allele (heterozygous) in the aligned bases of the pooled sequences, including these totals: sites called heterozygous in homologous positions in both sets, sites called as heterozygous only in the cow-based consensus, and those found only in the pig-based consensus. Note that this test did not establish whether, for homologous sites called as heterozygous in both mappings, that all bases were in agreement, but only that the most common base was the same in both mappings.

References

Allendorf FW, Hohenlohe PA, Luikart G. 2010. Genomics and the future of conservation genetics. Nat Rev Genet 11: 697–709.

Andersson L, Georges M. 2004. Domestic-animal genomics: deciphering the genetics of complex traits. Nature Reviews Genetics 5: 202–212.

Archibald A, Bolund L, Churcher C, Fredholm M, Groenen M, Harlizius B, Lee KT, Milan D, Rogers J, Rothschild M, et al. 2010a. Pig genome sequence-analysis and publication strategy. BMC genomics 11: 438.

Archibald AL, Cockett NE, Dalrymple BP, Faraut T, Kijas JW, Maddox JF, McEwan JC, Hutton Oddy V, Raadsma HW, Wade C, et al. 2010b. The sheep genome reference sequence: a work in progress. Anim Genet 41: 449–453.

Asan, Xu Y, Jiang H, Tyler-Smith C, Xue Y, Jiang T, Wang J, Wu M, Liu X, Tian G, et al. 2011. Comprehensive comparison of three commercial human whole-exome capture platforms. *Genome Biology* **12**: R95.

Bansal V, Tewhey R, LeProust EM, Schork NJ. 2011. Efficient and Cost Effective Population Resequencing by Pooling and In-Solution Hybridization. PLoS ONE 6: 1–6.

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.

Bielawski JP, Yang Z. 2004. A Maximum Likelihood Method for Detecting Functional Divergence at Individual Codon Sites, with Application to Gene Family Evolution. Journal of Molecular Evolution 59: 121–132.

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. In Current Protocols in Human Genetics (eds. J.L. Haines, B.R. Korf, C.C. Morton, C.E. Seidman, J.G. Seidman, and D.R. Smith), John Wiley & Sons, Inc., Hoboken, NJ, USA.

Bruford MW, Bradley DG, Luikart G. 2003. DNA markers reveal the complexity of livestock domestication. Nature Reviews Genetics 4: 900–910.

Bruneaux, Matthieu, Susan E. Johnston, Gábor Herczeg, Juha Merilä, Craig R. Primmer, and Anti Vasemägi. 2013. Molecular Evolutionary and Population Genomic Analysis of the Nine-spined Stickleback Using a Modified Restriction-site-associated DNA Tag Approach. Molecular Ecology 22: 565–582.

Burbano HA, Hodges E, Green RE, Briggs AW, Krause J, Meyer M, Good JM, Maricic T, Johnson PLF, Xuan Z, et al. 2010. Targeted Investigation of the Neandertal Genome by Array-Based Sequence Capture. Science 328: 723–725.

Cheviron ZA, Brumfield RT. 2012. Genomic insights into adaptation to high-altitude environments. Heredity 108: 354–361.

Chinwalla AT, Cook LL, Delehaunty KD, Fewell GA, Fulton LA, Fulton RS, Graves TA, Hillier LW, Mardis ER, McPherson JD, et al. 2002. Initial sequencing and comparative analysis of the mouse genome. Nature 420: 520–562.

Cosart T, Beja-Pereira A, Chen S, Ng S, Shendure J, Luikart G. 2011. Exome-wide DNA Capture and Next Generation Sequencing in Domestic and Wild Species. BMC genomics 12: 347.

Cowley M, Oakey RJ. 2010. Retrotransposition and genomic imprinting. Briefings in Functional Genomics 9: 340–346.

Do R, Kathiresan S, Abecasis GR. 2012. Exome sequencing and complex disease: practical aspects of rare variant association studies. Human Molecular Genetics 21: R1–R9.

Elsik CG, Tellam RL, Worley KC. 2009. The Genome Sequence of Taurine Cattle: A window to ruminant biology and evolution. Science (New York, NY) 324: 522–528.

Ernest EM, Haanes H, Bitanyi S, Fyumagwa RD, Msoffe PL, Bjørnstad G, Røed KH. 2012. Influence of habitat fragmentation on the genetic structure of large mammals: evidence for increased structuring of African buffalo (Syncerus caffer) within the Serengeti ecosystem. Conservation Genetics 13: 381–391.

Flicek P, Ahmed I, Amode MR, Barrell D, Beal K, Brent S, Carvalho-Silva D, Clapham P, Coates G, Fairley S, et al. 2012. Ensembl 2013. Nucleic Acids Research 41: D48–D55.

Fujita PA, Rhead B, Zweig AS, Hinrichs AS, Karolchik D, Cline MS, Goldman M, Barber GP, Clawson H, Coelho A, et al. 2010. The UCSC Genome Browser database: update 2011. Nucleic Acids Research 39: D876–D882.

Genome 10K Community of Scientists. 2009. Genome 10K: A Proposal to Obtain Whole-Genome Sequence for 10 000 Vertebrate Species. Journal of Heredity 100: 659 – 674.

Gnirke A, Melnikov A, Maguire J, Rogov P, LeProust EM, Brockman W, Fennell T, Giannoukos G, Fisher S, Russ C, et al. 2009. Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. Nature biotechnology 27: 182.

Good JM, Wiebe V, Albert FW, Burbano HA, Kircher M, Green RE, Halbwax M, André C, Atencia R, Fischer A, Pääbo S. 2013. Comparative Population Genomics of the Ejaculate in Humans and the Great Apes. Molecular Biology and Evoluton 30:964–976.

Groenen MAM, Archibald AL, Uenishi H, Tuggle CK, Takeuchi Y, Rothschild MF, Rogel-Gaillard C, Park C, Milan D, Megens H-J, et al. 2012. Analyses of pig genomes provide insight into porcine demography and evolution. Nature 491: 393–398.

Guha S, Goyal SP, Kashyap VK. 2007. Molecular phylogeny of musk deer: A genomic view with mitochondrial 16S rRNA and cytochrome b gene. Molecular Phylogenetics and Evolution 42: 585–597.

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. Molecular Ecology Resources 13: 254–268.

Handsaker B, Wysoker A, Tibbetts K, Fennell T. 2009. Picard. http://picard.sourceforge.net/.

The Hannonlab. 2010. FASTX-Toolkit. Hannon Lab, Cold Spring Harbor Laboratory http://hannonlab.cshl.edu/fastx_toolkit/index.html.

Harris RS. 2007. Improved pairwise alignment of genomic DNA. Ph.D., The Pennsylvania State Univerisity.

Hedrick PW. 2009. Conservation genetics and North American bison (Bison bison). Journal of Heredity 100: 411.

Hillier LW, Miller W, Birney E, Warren W, Hardison RC, Ponting CP, Bork P, Burt DW, Groenen MAM, Delany ME, et al. 2004. Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. Nature 432: 695–716.

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. Nature Genetics 39: 1522–1527.

Hohenlohe PA, Day MD, Amish SJ, Miller MR, Kamps-Hughes N, Boyer MC, Muhlfeld CC, Allendorf FW, Johnson EA, Luikart G. 2013. Genomic patterns of introgression in rainbow and westslope cutthroat trout illuminated by overlapping paired-end RAD sequencing. Molecular Ecology 22: 3002–3013.

Kent WJ. 2002. BLAT-The BLAST-Like Alignment Tool. Cold Spring Harbor Lab.

Khatkar MS, Hobbs M, Neuditschko M, Sölkner J, Nicholas FW, Raadsma HW. 2010. Assignment of chromosomal locations for unassigned SNPs/scaffolds based on pair-wise linkage disequilibrium estimates. BMC Bioinformatics 11: 171.

Klein J, O'hUigin C. 1993. Composite origin of major histocompatibility complex genes. Current Opinion in Genetics & Development 3: 923–930.

Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, et al. 2001. Initial sequencing and analysis of the human genome. Nature 409: 860–921.

Latch EK, Kierepka EM, Heffelfinger JR, Rhodes OE. 2011. Hybrid swarm between divergent lineages of mule deer (Odocoileus hemionus). Molecular Ecology 20: 5265–5279.

Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics 25: 1754 –1760.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics 25: 2078–2079.

Luikart G, England PR, Tallmon D, Jordan S, Taberlet P. 2003. The power and promise of population genomics: from genotyping to genome typing. Nat Rev Genet 4: 981–994.

Lunter G, Goodson M. 2011. Stampy: A statistical algorithm for sensitive and fast mapping of Illumina sequence reads. Genome Res 21: 936–939.

Maglott D, Ostell J, Pruitt KD, Tatusova T. 2010. Entrez Gene: gene-centered information at NCBI. Nucleic Acids Research 39: D52–D57.

McGinnis S, Madden TL. 2004. BLAST: at the core of a powerful and diverse set of sequence analysis tools. Nucleic Acids Research 32: W20.

McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, et al. 2010. The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. Genome Research 20: 1297 –1303.

McLysaght A. 2008. Evolutionary steps of sex chromosomes are reflected in retrogenes. Trends in Genetics 24: 478–481.

Milinski M. 2006. The Major Histocompatibility Complex, Sexual Selection, and Mate Choice. Annual Review of Ecology, Evolution, and Systematics 37: 159–186.

Miller MR, Dunham JP, Amores A, Cresko WA, Johnson EA. 2007a. Rapid and cost-effective polymorphism identification and genotyping using restriction site associated DNA (RAD) markers. Genome Research 17: 240–248.

Miller W, Rosenbloom K, Hardison RC, Hou M, Taylor J, Raney B, Burhans R, King DC, Baertsch R, Blankenberg D, et al. 2007b. 28-Way vertebrate alignment and conservation track in the UCSC Genome Browser. Genome Res 17: 1797–1808.

Morin PA, Luikart G, Wayne RK, The SNP workshop group. 2004. SNPs in ecology, evolution and conservation. Trends in Ecology & Evolution 19: 208–216.

Murtagh VJ, O'Meally D, Sankovic N, Delbridge ML, Kuroki Y, Boore JL, Toyoda A, Jordan KS, Pask AJ, Renfree MB, et al. 2012. Evolutionary history of novel genes on the tammar wallaby Y chromosome: Implications for sex chromosome evolution. Genome Res 22: 498–507.

Nadeau NJ, Whibley A, Jones RT, Davey JW, Dasmahapatra KK, Baxter SW, Quail MA, Joron M, ffrench-Constant RH, Blaxter ML, et al. 2012. Genomic islands of divergence in hybridizing Heliconius butterflies identified by large-scale targeted sequencing. Phil Trans R Soc B 367: 343–353.

Ng SB, Buckingham KJ, Lee C, Bigham AW, Tabor HK, Dent KM, Huff CD, Shannon PT, Jabs EW, Nickerson DA, et al. 2009a. Exome sequencing identifies the cause of a mendelian disorder. Nature Genetics 42: 30–35.

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

Poloumienko A. 2004. Cloning and comparative analysis of the bovine, porcine, and equine sex chromosome genes ZFX and ZFY. Genome 47: 74–83.

Price AL, Kryukov GV, De Bakker PIW, Purcell SM, Staples J, Wei L-J, Sunyaev SR. 2010. Pooled Association Tests for Rare Variants in Exon-Resequencing Studies. Am J Hum Genet 86: 832–838.

Pruitt KD, Tatusova T, Brown GR, Maglott DR. 2011. NCBI Reference Sequences (RefSeq): current status, new features and genome annotation policy. Nucleic Acids Research 40: D130–D135.

Randi E, Fusco G, Lorenzini R, Toso S, Tosi G. 1991. Allozyme divergence and phylogenetic relationships among Capra, Ovis and Rupicapra (Artyodactyla, Bovidae). Heredity 67: 281–286.

Rettenberger G, Klett C, Zechner U, Kunz J, Vogel W, Hameister H. 1995. Visualization of the conservation of synteny between humans and pigs by heterologous chromosomal painting. Genomics 26: 372–378.

Ritz LR, Glowatzki-Mullis M-L, MacHugh DE, Gaillard C. 2000. Phylogenetic analysis of the tribe Bovini using microsatellites. Animal Genetics 31: 178–185.

Rivas MA, Beaudoin M, Gardet A, Stevens C, Sharma Y, Zhang CK, Boucher G, Ripke S, Ellinghaus D, Burtt N, et al. 2011. Deep resequencing of GWAS loci identifies independent rare variants associated with inflammatory bowel disease. Nature Genetics 43: 1066–1073.

Roepstorff A, Lin SC, Chen KL, Chung YY, Lin WY, Hsu S. 2006. The Genome Sequence of Taurine Cattle: A Window to Ruminant Biology and Evolution. Neuroimage 32: 1850.

Rowe HC, Renaut S, Guggisberg A. 2011. RAD in the realm of next-generation sequencing technologies. Molecular Ecology 20: 3499–3502.

Salzberg SL, Phillippy AM, Zimin A, Puiu D, Magoc T, Koren S, Treangen TJ, Schatz MC, Delcher AL, Roberts M, et al. 2012. GAGE: A critical evaluation of genome assemblies and assembly algorithms. Genome Res 22: 557–567.

Sanders SJ, Murtha MT, Gupta AR, Murdoch JD, Raubeson MJ, Willsey AJ, Ercan-Sencicek AG, DiLullo NM, Parikshak NN, Stein JL, et al. 2012. De novo mutations revealed by whole-exome sequencing are strongly associated with autism. Nature 485: 237–241.

Schaschl H, Wandeler P, Suchentrunk F, Obexer-Ruff G, Goodman SJ. 2006. Selection and recombination drive the evolution of MHC class II DRB diversity in ungulates. Heredity 97: 427–437.

Schatz MC, Puiu D, Hanrahan F, Pertea G, Van Tassell CP, Sonstegard TS, et al. 2009. A whole-genome assembly of the domestic cow, Bos taurus. Genome Bioloby 10: R42

Schnable PS, Ware D, Fulton RS, Stein JC, Wei F, Pasternak S, Liang C, Zhang J, Fulton L, Graves TA, et al. 2009. The B73 Maize Genome: Complexity, Diversity, and Dynamics. Science 326: 1112–1115.

Spurgin LG, Richardson DS. 2010. How pathogens drive genetic diversity: MHC, mechanisms and misunderstandings. Proc R Soc B 277: 979–988.

Tennessen JA, Bigham AW, O'Connor TD, Fu W, Kenny EE, Gravel S, McGee S, Do R, Liu X, Jun G, et al. 2012. Evolution and Functional Impact of Rare Coding Variation from Deep Sequencing of Human Exomes. Science 337: 64–69.

The Hannonlab. 2010. FASTX-Toolkit. Hannon Lab, Cold Spring Harbor Laboratory http://hannonlab.cshl.edu/fastx_toolkit/index.html.

Ursing BM, Slack KE, Arnason U. 2000. Subordinal artiodactyl relationships in the light of phylogenetic analysis of 12 mitochondrial protein-coding genes. Zoologica Scripta 29: 83–88.

Vasemägi A, Nilsson J, Primmer CR. 2005. Expressed Sequence Tag-Linked Microsatellites as a Source of Gene-Associated Polymorphisms for Detecting Signatures of Divergent Selection in Atlantic Salmon (Salmo salar L.). Mol Biol Evol 22: 1067–1076.

Zhang C, Plastow G. 2011. Genomic Diversity in Pig (Sus scrofa) and its Comparison with Human and other Livestock. Curr Genomics **12**: 138–146.

Supplementary figures

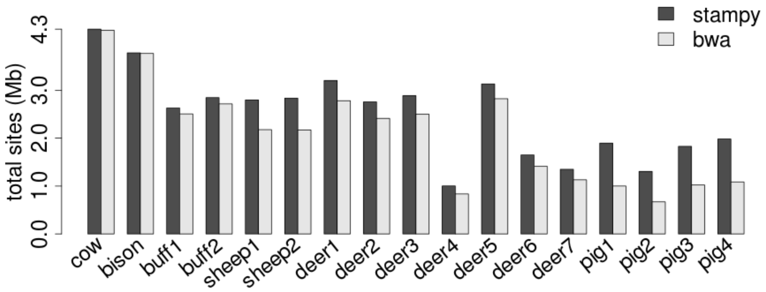


Figure 4-S1: Sequening in flanks. Number of nucleoide sites sequenced to 20X/Q50 in ~11.6 Mb of flanking nucleotide sites 250 bp either side of targeted exons. BWA mismatch allowance paramaters were relaxed only for the deer and pig (see Methods).

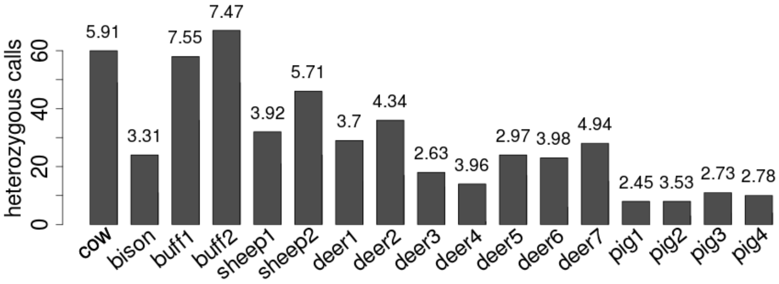


Figure 4-S2: MHC total heterozygous calls. For the Stampy mapping, for the 63 exons in 14 MHC genes total sites called as heterozygous. These were filtered as described in Methods. Numbers above the bars give the heterozygous positions inferred per thousand 20X/Q50 sites.

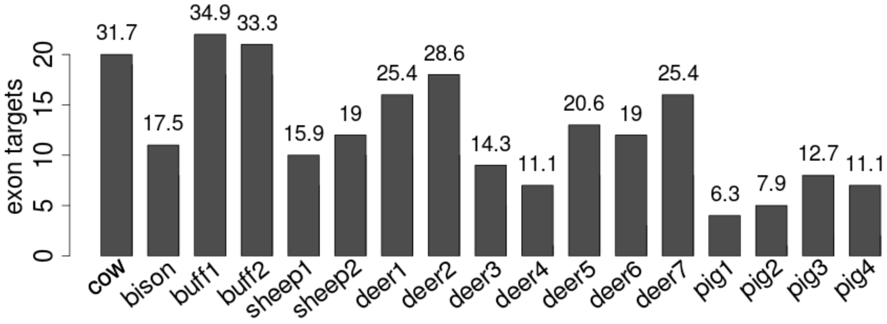


Figure 4-S3: MHC exons with heterozygous calls. For the Stampy mapping, for 63 exons in 14 MHC genes, total For the Stampy mapping, for 63 exons in 14 MHC genes, total exons with at least one heterozygous base call. These were filltered as described in Methods. Numbers above the bars give the percentage of the MHC exons with at least one heterozygous base call.

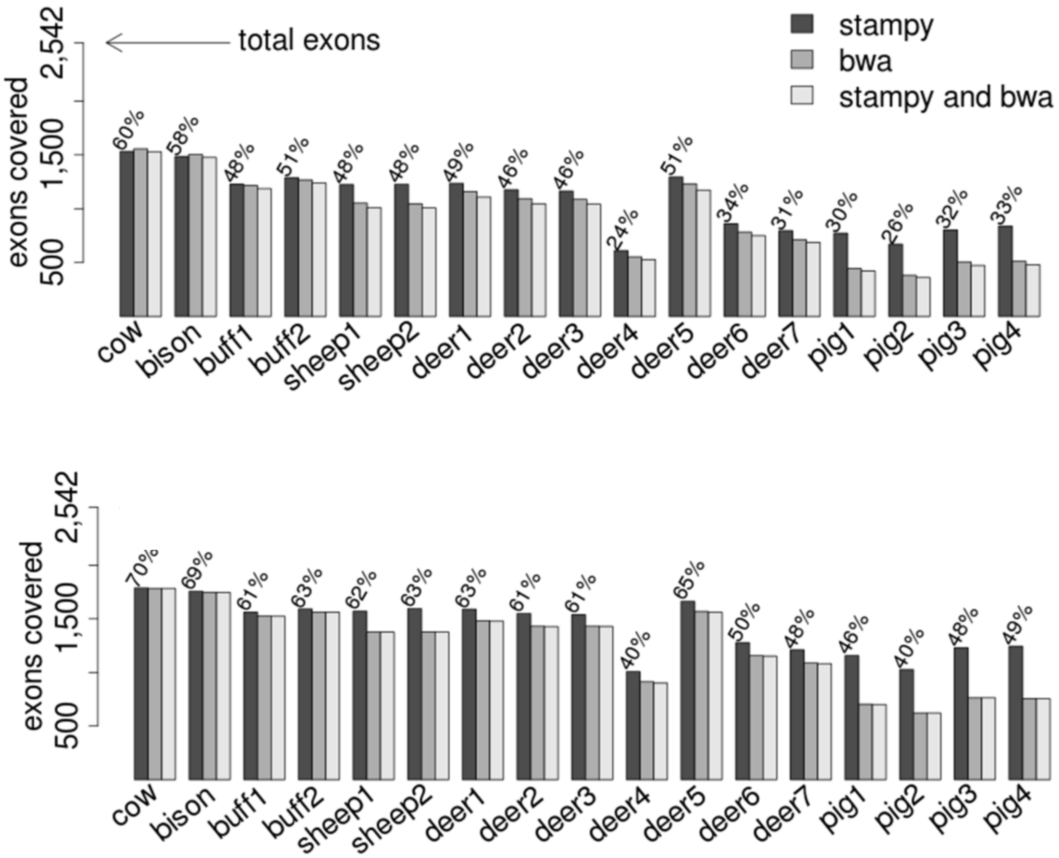


Figure 4-S4: Sequencing success in candidate genes. Of 2,542 exons in 342 candidate genes, the number with (A) 100%, and (B) at least 60% of bases covered to at least 20X and with a Phred consensus score of at least 50. BWA mismatch allowance parameters were relaxed only for the deer and pigs (see Methods).

Appendix

These appended works are previously published. References for the publications are as follows:

Morales, Sergio E. et al. “Extensive Phylogenetic Analysis of a Soil Bacterial Community Illustrates Extreme Taxon Evenness and the Effects of Amplicon Length, Degree of Coverage, and DNA Fractionation on Classification and Ecological Parameters.” *Applied and Environmental Microbiology* 75.3 (2009): 668 –675. *Highwire 2.0*. Web. 8 Jan. 2012.

Morales, S. E. et al. “Supplemental Programs for Enhanced Recovery of Data from the DOTUR Application.” *Journal of Microbiological Methods* 75.3 (2008): 572–575. Print.

Morales, Sergio E, Theodore Cosart, and William E Holben. “Bacterial Gene Abundances as Indicators of Greenhouse Gas Emission in Soils.” *ISME J* (2010): n. pag. *Nature*. Web. 7 Mar. 2010.

APPLIED AND ENVIRONMENTAL MICROBIOLOGY, Feb. 2009, p. 668–675 Vol. 75, No. 3
0099-2240/09/$08.00+0 doi:10.1128/AEM.01757-08

Extensive Phylogenetic Analysis of a Soil Bacterial Community Illustrates Extreme Taxon Evenness and the Effects of Amplicon Length, Degree of Coverage, and DNA Fractionation on Classification and Ecological Parameters$^{\nabla}$†

Sergio E. Morales,[1] Theodore F. Cosart,[2,3] Jesse V. Johnson,[2,3] and William E. Holben[1,3]*

Microbial Ecology Program, Division of Biological Sciences,[1] Department of Computer Science,[2] and Montana—Ecology of Infectious Diseases Program,[3] The University of Montana, Missoula, Montana

Received 30 July 2008/Accepted 2 November 2008

To thoroughly investigate the bacterial community diversity present in a single composite sample from an agricultural soil and to examine potential biases resulting from data acquisition and analytical approaches, we examined the effects of percent G+C DNA fractionation, sequence length, and degree of coverage of bacterial diversity on several commonly used ecological parameters (species estimation, diversity indices, and evenness). We also examined variation in phylogenetic placement based on multiple commonly used approaches (ARB alignments and multiple RDP tools). The results demonstrate that this soil bacterial community is highly diverse, with 1,714 operational taxonomic units demonstrated and 3,555 estimated (based on the Chao1 richness estimation) at 97% sequence similarity using the 16S rRNA gene. The results also demonstrate a fundamental lack of dominance (i.e., a high degree of evenness), with 82% of phylotypes being encountered three times or less. The data also indicate that generally accepted cutoff values for phylum-level taxonomic classification might not be as applicable or as general as previously assumed and that such values likely vary between prokaryotic phyla or groups.

Efforts to describe bacterial species richness and diversity have long been hampered by the inability to cultivate the vast majority of bacteria from natural environments. New methods to study bacterial diversity have been developed in the last two decades (32), many of which rely on PCR-based procedures and phylogenetic comparison of 16S rRNA gene sequences. However, PCR using complex mixtures of templates (as in the case of total microbial community DNA) is presumed to preferentially amplify certain templates in the mixture (23) based on their primary sequence, percent G+C (hereafter GC) content, or other factors, resulting in so-called PCR bias. Moreover, the amplification of template sequences depends on their initial concentration and tends to skew detection toward the most abundant members of the community (23). To further complicate matters, subsequent random cloning steps on amplicon mixtures are destined to result in the detection of numerically dominant sequences, especially where relative abundance can vary over orders of magnitude. Indeed, any analysis based on random encounter is destined to primarily detect numerically dominant populations. This is especially of concern where limited sampling is performed on highly complex microbial communities exhibiting mostly even distribution of populations with only a few showing any degree of dominance, as typically perceived for soils (17). These artifacts and sampling limitations represent major hurdles in bacterial community diversity analysis, since the vast majority of bacterial diversity probably lies in "underrepresented minority" populations (24, 30). This is important because taxa that are present only in low abundance may still perform important ecosystem functions (e.g., ammonia-oxidizing bacteria). Of special concern is that biases in detection might invalidate hypothesis testing on complex communities where limited sampling is performed (5).

Recently, there has been a concerted effort toward addressing problems impeding comprehensive bacterial diversity studies (7, 13, 24, 26, 28). In recent years, studies have increased sequencing efforts, with targeted 16S rRNA gene sequence libraries approaching 2,000 clones (11) and high-throughput DNA-sequencing efforts (e.g., via 454 pyrosequencing and newer-generation high-throughput approaches) of up to 149,000 templates from one or a few samples (25, 30). These technological advances have come as researchers recognize that massive sequencing efforts are required to accurately assess the diversity of populations that comprise complex microbial communities (29, 30). Alternatively, where fully aligned sequence comparisons need to be made, novel experimental strategies that allow more-comprehensive detection of underrepresented bacterial taxa can be applied. One such approach involves the application of prefractionation of total bacterial community genomic DNA based on its GC content (hereafter GC fractionation) prior to subsequent molecular manipulations of total community DNA (14). This strategy has been successfully applied in combination with denaturing gradient

* Corresponding author. Mailing address: Microbial Ecology Program, Division of Biological Sciences, University of Montana, Missoula, MT 59812-1006. Phone: (406) 243-6163. Fax: (406) 243-4184. E-mail: bill.holben@mso.umt.edu.

† Supplemental material for this article may be found at http://aem.asm.org/.

$^{\nabla}$ Published ahead of print on 14 November 2008.

gel electrophoresis (13) and 16S rRNA gene cloning (2, 21) to study microbial communities. This approach separates community genomic DNA, prior to any PCR, into fractions of similar percent GC content, effectively reducing the overall complexity of the total community DNA mixture by physical separation into multiple fractions. This facilitates PCR amplification, cloning, and detection of sequences in fractions with relatively low abundance in the community, thereby enhancing the detection of minority populations (13). Collectively, this strategy reduces the biases introduced by PCR amplification and random cloning of the extremely complex mixtures of templates of different GC content, primary sequence, and relative abundance present in total environmental genomic DNA.

Any large molecular survey that relies on sequencing further requires the analysis of large amounts of data that must be catalogued into phylogenetically relevant groups. This is usually done using high-throughput methods like RDP Classifier or Sequence Match (6) or a tree-based method like Greengenes (8) or ARB (18). Two major pitfalls that are encountered using these former approaches are the presence of huge numbers of unclassified sequences in databases and the lack of representative sequences from all phyla. This leads to most surveys having large portions of their phylotypes designated as unclassified. The latter tree-based approaches, although better suited for classification schemes, are also dependent on having a comprehensive database with well-classified sequences for reproducible results. This reproducibility becomes especially important when trying to compare data across different studies, especially those that utilize different approaches and study systems.

In the current study, we analyzed an extensive (~5,000 clones) partial 16S rRNA gene library from a single soil sample that was generated using very general primers and GC-fractionated DNA. Total DNA was extracted from soil at a cultivated treatment plot at the National Science Foundation Long Term Ecological Research (NSF-LTER) site at the Kellogg Biological Station (KBS) in mid-Michigan (http://www.kbs.msu.edu/lter). To test the effect of GC fractionation on recovery of 16S rRNA gene sequences, we conducted a direct comparison with a nonfractionated library generated from the same soil sample. Using the GC-fractionated library, we also calculated several measures of bacterial diversity and examined the effects of sampling size and sequence length on Shannon-Weaver diversity index, Simpson's reciprocal index (1/*D*, where *D* is the probability that two randomly selected individuals from a sample belong to the same species), evenness, and Chao1 richness estimation. The results show that GC fractionation is a powerful tool to help mitigate limitations of random PCR- and cloning-based analyses of total microbial community diversity, resulting in the recovery of underrepresented taxa and, in turn, reducing the sampling size needed for accurate estimations of bacterial richness. The results also provided evidence for the need to expand the typical scale of sequence-based survey efforts, particularly in environments where evenness abounds or where minority bacterial populations may have important effects on community function and processes. We suggest that there is a need for the establishment of standardized approaches for the analysis of sequence data from community diversity studies in order to maximize data comparisons across independent studies and show examples of software programs developed to facilitate comparative analysis of large sequence datasets.

MATERIALS AND METHODS

Study site and sample collection. Samples were collected from the KBS LTER Row-Crop Agriculture site in mid-Michigan (for an overview of that project see http://lter.kbs.msu.edu/). The current study examined the bacterial community in the replicate plots of Treatment 1 at the main experimental site, which is representative of canonical agricultural practice in the upper Midwest. The treatment consisted of conventional wheat, corn, and soybean annual rotations receiving standard levels of chemical inputs, with chisel plowing. Soil was classified as a fine-loamy, mixed, mesic Typic Hapludalfs. For this bacterial population survey, five randomly positioned, 0- to 20-cm soil cores were taken from each of six treatment replicates in July, 2004, at the height of the growing season. Each replicate treatment sample was sieved through 2-mm mesh and mixed thoroughly, providing six replicate samples. All soil samples for this study were stored on dry ice or at −70°C immediately after soil processing (i.e., sieving and mixing) prior to bacterial community DNA extraction.

DNA manipulations. Total microbial community DNA was extracted and purified from the samples by using the large-scale direct lysis method developed by Holben (12). Equal amounts of DNA (10 μg) from each replicate sample were pooled to provide a representative sample from this treatment regimen that was subsequently fractionated based on the percent GC content of the DNA of the component populations of the community as originally described by Holben and Harris (14). Following centrifugation, the gradients were fractionated into 15 separate fractions representing percent GC contents ranging from 20 to 80% (the full range observed in the domain *Bacteria*) and the amount and percent GC content of the DNA at each position in the gradient were determined as described elsewhere (1). The DNA in individual fractions was desalted by using PD-10 columns (Amersham Pharmacia Biotech, Piscataway, NJ) with the manufacturer's recommended protocol. Each individual fraction was then PCR amplified independently for creation of the 16S rRNA gene clone library.

PCR conditions employed the primer pair 536f (5′-CAGCMGCCGCGGTAATWC-3′) and 907r (5′-CCGTCAATTCMTTTRAGTTT-3′) (13) and used the optimal reaction and amplification conditions described by Ishii and Fukui (16) for reducing PCR bias, namely, 50-μl volumes containing 10 pg of template DNA, 1× *Taq* buffer, 200 μM of each deoxynucleoside triphosphate, 25 pmol of each primer, and 1.25 U of *Taq* polymerase amplified for 21 cycles of 94°C for 1 min, 45°C for 1 min, and 72°C for 2 min. PCR products were cloned by using the plasmid vector pT7Blue-3 and a Perfectly Blunt cloning kit (Novagen, Inc., Madison, WI) according to the manufacturer's instructions. Plasmid clones were purified from 2-ml cultures of *Escherichia coli* incubated overnight at 37°C with shaking using Qiagen mini-prep kits (Qiagen, Valencia, CA) as recommended by the manufacturer. Restriction analysis using EcoRI was performed to ensure that plasmids contained correctly sized inserts. Plasmid DNA was sequenced by using the universal primer T7 and standard dideoxy sequencing conditions.

Phylogenetic placement and tree creation based on clone libraries. All 16S rRNA gene sequences were manually trimmed of vector and primer sequence prior to alignment and analysis. Trimmed sequences were subsequently checked for chimeric character and other anomalies by using Pintail (3), and suspect sequences were excluded from further analysis, leaving 4,889 sequences to be analyzed. Multiple Fasta files were created and independently aligned in ARB (18). Alignments were performed in ARB using the Fast Aligner and at least three reference sequences for each clone from the 16S rRNA gene database PT server containing 51,024 reference sequences (http://www.arb-home.de/downloads.html). Sequences from the current study were integrated into the annotated tree based on parsimony.

Assignment to similarity-based OTUs and species richness estimators. Prior to assignment into *o*perational *t*axonomic *u*nits (OTUs), ARB-generated 16S sequence alignments were used to create Jukes-Cantor corrected distance matrices and exported. These matrices were used as input for the DOTUR program (26), which was used to calculate Simpson's and Shannon-Weaver diversity indices, Chao1 richness estimates, and OTU bins using default settings.

Comparison of GC-fractionated to nonfractionated data was performed by creating a master sequence library containing both fractionated and nonfractionated sequence libraries. Approximately 500 (487 and 490, respectively) sequences were compared for fractionated and nonfractionated libraries by comparing ~33 sequences obtained from each of the 15 GC-based fractions of the total community to a library of 490 sequences randomly cloned from nonfractionated total community DNA from the same sample. The sequences obtained were aligned in ARB and then run through the DOTUR program. DOTUR data files were then used as input for the SONS program (27), which was used to compare OTU representation within each library.

Identification of phylum-specific taxonomic bins and OTU composition. To identify distance score cutoff values for individual phyla, we developed the DAM (*D*OTUR-*A*RB *m*atching) program (19), available at (http://dbs.umt.edu/research_labs/holbenlab/links.php). This allowed comparison between ARB-generated group lists and DOTUR list files created from the total data set of 4,889 sequences. The DAM program was employed to match a query list of sequence identifications (hereafter, IDs) from ARB to OTUs as determined by the DOTUR program, allowing for a user-specified range of DOTUR distance values. Querying against a DOTUR list file for each distance value in range, the program extracted only OTUs that contained one or more of the query IDs. Results were written to a file formatted as a DOTUR list file, with each line listing the DOTUR distance value, the number of matched OTUs for the prescribed distance, and a list of each bin's contents. For this study, DAM results provided the percent sequence similarity at which an ARB-generated phylum list was contained in a single DOTUR OTU.

In order to identify sequences belonging to specific OTUs, a new program, DOTMAN (for "*DOT*UR *man*ipulation"; available at http://dbs.umt.edu/research_labs/holbenlab/links.php), was created (19). DOTMAN queries selected OTUs (based on DOTUR bins) against a sequence database, generating FASTA files from a user-given file. To accomplish this, the program is given a range of DOTUR distance values, a DOTUR list file, and a file in FASTA format containing sequences corresponding to the IDs in the list file. For each distance value *d*, DOTMAN makes one FASTA file for each of the *n* largest OTUs. *n* is set by the user and is less than or equal to the total number of OTUs for a distance *d*.

Sample size simulations. To explore the effects of sampling size on ecological parameters (Chao1 richness estimation, Shannon-Weaver indices, and dominance), we used EcoSim700 null model software for ecology (version 7.0) to analyze data created from the first 500, 2,000, 3,390, and 5,000 sequences contained in our library. Input files were created from OTUs that clustered with 97% similarity and were subsequently used as the data matrix for running the program.

Nucleotide sequence accession numbers. All sequences used in this paper have been deposited in the GenBank database (accession no. EU352912 to EU357802).

RESULTS

Effect of sample size on observed and estimated richness. Environmental rRNA gene libraries vary considerably in size but typically are of 500 sequences or less (4, 20). Although it has been shown that small sample sizes are useful for providing a "snapshot" of the predominant species (29) and they have been employed in theoretical estimates of bacterial species richness (20), there is little empirically derived data actually demonstrating the effect of sample size on ecological parameters, such as richness estimation, dominance, diversity indices, or evenness. To better understand sampling size-induced errors and to better estimate bacterial diversity in soil, we paired the additional resolving power of GC fractionation with the general utility of 16S rRNA gene clone libraries in a microbial community survey of a single soil type.

The effect of sample size was tested by creating datasets from the first 500, 2,000, 3,390, and 5,000 sequenced clones in our GC-fractionated library. Subsequent removal of anomalous and nonbacterial sequences produced sets of 487, 1,962, 3,322, and 4,889 sequences, respectively. These datasets were analyzed based on "bins" created as a function of 16S sequence similarity. Since 16S sequences are not necessarily linked to a whole-genome evolutionary or ecological context, the values chosen for binning are arbitrary and only serve the purpose of creating objectively derived bins that cluster data into a reasonable number of taxonomically related groups (10, 22). In order to facilitate comparison to prior bacterial community diversity studies, the data were grouped at multiple levels of similarity (Table 1), but discussion in this report is focused primarily on the widely utilized 97% sequence similarity level.

A 5.1-fold increase in the number of OTUs and a 3.5-fold increase in the richness estimation were observed (at 97% sequence similarity) from the smallest to the largest data set (Table 1). Shannon diversity index values increased approximately 1.2-fold across this same span, with the Simpson's reciprocal index (1/*D*) increasing from 202.19 to 341.67, representing a 1.7-fold increase. In contrast, evenness estimates decreased from 0.966 to 0.906 between the smallest and largest data sets, presumably indicating that sampling was approaching a minimal saturation point where low-abundance sequences (unique in the smaller datasets) were being detected more than once.

The largest library, containing 4,889 sequences, represented the most complete survey of aligned 16S rRNA gene sequences from a single composite soil sample and was composed of 1,714 OTUs identified at 97% sequence similarity (Table 1). Projections based on this large data set predict that 3,555 different OTUs were actually present in this soil sample (Table 1) and that a GC-fractionated clone library of well over 10,000 sequences would be required to begin bordering an asymptote in the rarefaction curve.

At 97% sequence similarity, a Shannon-Weaver score of 6.75 was calculated, much greater than the values of 4.35 and 4.68 previously estimated for an Amazon and a Scottish soil, respectively (26). Further, the vast majority of bacterial taxa in the soil were present in very low numbers, producing an extremely high evenness estimate of 0.906, while only a few OTUs exhibited any numerical predominance (Table 1). Our data firmly validate the increasingly common perception (as does a recent report; see reference 29) that numerous taxa present in comparably low overall abundance comprise the bulk of the soil bacterial community.

To compare common community diversity measures as a function of different sample sizes, we used EcoSim700 null model software to create species richness estimates, Shannon-Weaver diversity indices, and dominance curves. The results revealed underestimations in all three parameters when using the smaller datasets (Fig. 1). All parameters tested followed a conserved and overlapping general trend with increasing sample size, but the smaller data sets lacked sufficient sequence coverage to indicate an asymptote or to reflect end results comparable to those obtained from the larger data sets.

Community composition. To examine taxonomic representation within the community, we explored two commonly used methods of taxonomic placement for 16S rRNA gene sequence data. Sequences were first analyzed using the Classifier (version 1.0; taxonomical hierarchy release 6.0) and the SeqMatch tools (6) of the RDP. Individual sequences were considered classified only if both programs showed agreement at the phylum level. Unclassified sequences were assigned a potential placement based on Classifier. Using this method, 3,233 (66%) of the sequences were classified (Table 2) into 17 known phyla. These same sequences were also classified using ARB (18) by placement into an ARB-generated phylogenetic tree of 51,024 classified sequences. With this approach, a 33% increase in placement to known phyla was obtained, with 4,854 (99%) sequences assigned to 25 known phylogenetic groups. It is

TABLE 1. Effect of sample size on similarity-based OTUs, Shannon-Weaver diversity index, evenness, and richness estimation

Sequence sample size, % similarity level	No. of unique OTUs	Shannon-Weaver index	Evenness	Richness estimate[a]	No. of sequences represented in top 10 OTUs									
					OTU_1	OTU_2	OTU_3	OTU_4	OTU_5	OTU_6	OTU_7	OTU_8	OTU_9	OTU_{10}
First 500[b]														
100	461	6.10	0.995	5,851	6	3	3	2	2	2	2	2	2	2
97	335	5.62	0.966	1,020	12	8	7	7	7	7	5	5	5	5
70	30	2.94	0.866	32	59	45	43	41	40	31	27	24	24	23
55	6	0.78	0.434	6	379	51	43	6	5	3				
47	2	0.13	0.188	2	473	14								
First 2,000[c]														
100	1,662	7.33	0.986	12,163	22	9	8	8	7	6	5	5	5	5
97	928	6.39	0.935	2,126	44	33	24	24	24	22	19	18	16	12
70	67	3.34	0.794	80	174	156	139	134	132	110	100	97	96	91
55	14	1.69	0.641	20	608	597	297	224	91	82	29	12	10	8
38	6	0.04	0.025	12	1,950	8	1	1	1	1				
First 3,390[d]														
100	2,680	7.73	0.980	15,015	40	17	15	15	11	10	8	8	8	7
97	1,319	6.59	0.918	2,991	76	54	54	41	38	31	27	27	27	26
70	84	3.48	0.785	88	309	243	237	210	200	182	165	159	150	147
55	14	1.62	0.615	14	1,284	1,064	387	205	85	85	60	54	46	28
38	2	0.02	0.032	2	3,311	11								
First 5,000[e]														
100	3,789	8.04	0.976	20,790	54	25	22	17	15	14	13	12	12	10
97	1,714	6.75	0.906	3,555	99	81	81	63	62	61	46	39	38	38
70	102	3.60	0.778	119	474	345	297	256	248	236	233	215	211	210
55	18	1.98	0.685	20	1,402	1,026	729	504	453	231	221	175	58	37
38	5	0.03	0.017	5	4,873	13	2	2	1					

[a] Based on full biased corrected Chao1 richness estimates.
[b] Based on 487 starting sequences.
[c] Based on 1,962 starting sequences.
[d] Based on 3,322 starting sequences.
[e] Based on 4,887 starting sequences and 2 archeal sequences used as references.

worth noting that the classification of certain groups was comparable using both methods (Table 2), but in groups with low sequence representation within databases (e.g., refer to *Chlorobi*, *Acidobacteria*, *Thermomicrobia*, *Fibrobacteres*, and candidate divisions of Table 2), the ARB-based approach allowed for more-consistent assignment of bacteria at the phylum level.

Since the analysis reported herein was performed, a new release (34) of the Classifier tool has been made available (version 2.0; taxonomical hierarchy release 7.8). Reanalysis of our data set with this new release produced taxonomic placements that were nearly identical to those obtained with ARB for classified sequences. Despite this, Classifier was still unable to classify 1,013 (21%) of the sequences in this library.

Using DOTUR, a total of 1,405 OTUs (at 97% sequence similarity), comprising 82% of all identified OTUs, were represented three or fewer times in this 4,889-sequence library. When the data were reanalyzed to include all OTUs represented 19 or fewer times (half the value of the 10th most predominant OTU), 99% of all OTUs in the study were included in this category. This represents 83% of all sequences in the full library. In order to provide some phylogenetic context to the predominant OTU bins generated by DOTUR using the 97% similarity cutoff, we analyzed the 10 most predominant taxa, which were represented by only 99 (OTU_1; *Gammaproteobacteria*), 81 (OTU_2; *Acidobacteria*), 81 (OTU_3; *Gammaproteobacteria*), 63 (OTU_4; *Thermomicrobia*), 62 (OTU_5; *Betaproteobacteria*), 61 (OTU_6; *Acidobacteria*), 46 (OTU_7; *Thermomicrobia*), 39 (OTU_8; *Alphaproteobacteria*), 38 (OTU_9; *Gammaproteobacteria*), and 38 (OTU_{10}; *Betaproteobacteria*) sequences out of 4,889 (Tables 1 and 2).

Effect of sequence length on community analysis. The region of the 16S rRNA gene used to generate the clone library in the current study is approximately 400 bp in length, spanning between *E. coli* positions 518 and 927 and encompassing two hypervariable regions (V4 and V5). Further, the highly conserved regions representing primers 536f and 907r (15) were removed prior to analysis because the minor degeneracies built into these primers potentially introduce errors into the sequences analyzed.

To test the effect that using this smaller (versus full-length) but highly variable region had on data analysis, we created a 1,184-sequence library from (nearly) full-length sequences in the ARB database. These reference sequences covered all of the phyla detected and were selected as having the greatest similarity to the sequences within our own library, thus serving as proxies to the sequences obtained in the current study. These reference sequences were analyzed separately as both full-length and truncated sequences (by trimming to match the 536-to-907 region, excluding primers) to create distance matrices at 97% sequence similarity which were used as input for the DOTUR program (26). Fairly modest differences were observed for the truncated and full-length sequences, with 911 and 1,031 OTUs identified, respectively (Table 3). Likewise, the Shannon-Weaver indices derived from truncated and full-

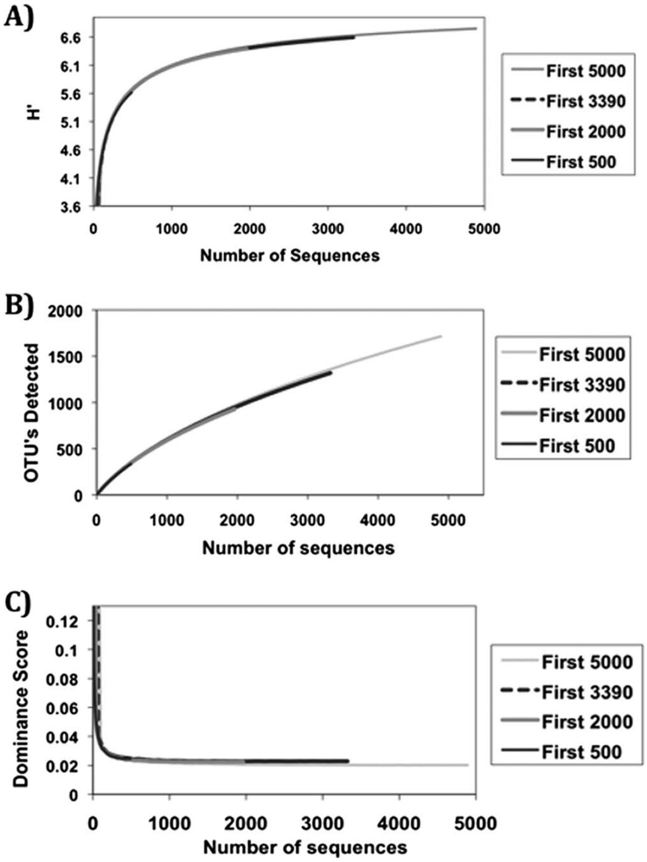


FIG. 1. The effects of sampling size on estimated diversity (A), number of OTUs detected (B), and evenness (C). Iterative plots of estimated Shannon-Weaver diversity (H′), OTUs detected, and dominance score were generated for the first 500, 2,000, 3,390, and 5,000 sequences in the partial 16S gene sequence library to demonstrate the effects of sample size on each parameter. Note that the lines in each panel directly overlap as a result of this iterative process and that the trajectory of each estimation curve is extended as sample size increases.

length sequences were slightly different, being 6.65 and 6.86, respectively. A substantial effect on Chao1 richness estimation was detected however, with an almost-twofold increase in estimated OTUs when full-length sequences were used for the analysis, presumably reflecting the additional fineness of phylogenetic resolving power afforded by the additional sequence information.

We further tested the effect of sequence length on outcome by comparing taxonomic placements based on full-length versus truncated sequences and found that comparable results were obtained for both. In 1,166 of the 1,184 cases (98.5%), congruent taxonomic assignments were obtained with the truncated sequences, while in only 16 cases (1.5%) did the additional sequence information result in a different taxonomic assignment (Table 3). Phylum-level classification based on ARB-based tree generation was highly reproducible independent of fragment length (Table 4). Collectively, this suggests that the ~400-bp V4-V5 region examined for our survey, which is readily obtained from a single dideoxy sequence reaction, is sufficient to provide reliable phylogenetic placement at phylum and higher-order levels. The effect of sequence length on finer-level placement (genus and species) was not examined, being outside the context of the current study.

Assignment of cutoff values for phylogenetic clusters. Previous publications have suggested that sequences sharing >60 to 80% identity likely belong to the same phylum (26, 29). Using this guideline, we empirically assessed the feasibility of employing such a "universal" cutoff value for phylum-level discrimination and the effect of using truncated versus full-length sequences to determine cutoff values. To accomplish this, we developed and applied the DAM program, which matches a list of query sequences (belonging to a discrete group [i.e., phylum-level cluster] as determined by ARB) to a

TABLE 2. Taxonomic classification based on multiple methods

Taxon	ARB[a]	Total no. of sequences		% Sequence similarity[d]
		Classified[b]	Unclassified[c]	
Phylum				
Acidobacteria	955	88	41	38
Actinobacteria	491	452	39	38
Bacteroidetes	453	450	17	59
CD OD1	68	0	0	49
CD OP10	43	18	0	38
CD OP11	14	6	6	27
CD OP3	12	0	0	53
CD TM6	3	0	0	83
CD TM7	17	11	2	56
CD WS1	2	0	0	99
CD WS3	34	17	5	38
Chamydiae	2	2	10	88
Chlorobi	22	0	1	70
Chloroflexi	27	32	18	55
Cyanobacteria	12	6	57	38
Defferibacteres	0	0	2	
Deinococcus-Thermus	1	0	3	100
Dictyoglomi	0	0	21	
Fibrobacteres	11	0	0	60
Firmicutes	15	16	800	48
Gemmatimonadetes	251	195	33	38
Lentisphaerae	0	0	1	
Nitrospira	59	47	0	38
Planctomycetes	243	174	22	38
Proteobacteria	1,690	1,631	506	21
Spirochaetes	3	1	5	38
Thermodesulfobacteria	0	0	4	
Thermomicrobia	323	0	10	38
Thermotogae	0	0	1	
Verrucomicrobia	103	89	50	43
Unclassified	35	1,654		
Class				
Alphaproteobacteria	374	368	13	21
Betaproteobacteria	485	475	0	43
Deltaproteobacteria	348	278	22	38
Epsilonproteobacteria	1	0	0	99
Gammaproteobacteria	478	468	7	38
Unclassified *Proteobacteria*	4			

[a] Classification based on ARB-generated tree. CD, candidate division.

[b] Sequences were considered classified if assigned to the same phylum using both SeqMatch and Classifier of the RDP.

[c] Unclassified sequences were assigned to likely phylum based on Classifier results.

[d] Based on ARB phylum level classification.

TABLE 3. Effect of fragment length on similarity-based OTU number, Shannon-Weaver diversity index, and richness estimation

Sequence length, % similarity level	No. of unique OTUs	Shannon-Weaver index	Richness estimate[a]
Full length[b]			
100	1,183	7.08	350,169
97	1,031	6.86	7,452
70	54	3.01	67
55	6	0.67	6
46	2	0.01	2
Truncated[c]			
100	1,166	7.05	61,646
97	911	6.65	4,175
70	80	3.41	93
55	15	1.68	18
46	5	0.47	6

[a] Based on full biased corrected Chao1 richness estimates.
[b] Based on 1,184 full-length sequences.
[c] Based on 1,184 truncated sequences. Truncations were created by deleting the upstream base pair region from the *E. coli* consensus position 536 and downstream of consensus position 906.

distance matrix-determined OTU group encompassing all sequences in a given query list (as created by the DOTUR program). This allowed determination of the percent sequence similarity at which groupings of sequences were identified as discrete OTUs. For this exercise, phylum-level groups with both large and smaller numbers of sequences were compared again using the ARB-derived full-length and truncated sequences described above, as well as the KBS-LTER study data set. The results showed that no single, consistent consensus value can be used as a phylum-level cutoff point across all taxa (Tables 1 and 4). With our own large data set, we annotated 102 separate groups at 70% sequence similarity, greatly exceeding estimates of 36 to 52 extant bacterial phyla suggested by Rappé and Giovannoni (24). At 55% sequence similarity, 18 groups were defined, in line with the number of phyla expected to be in soil (Table 1).

Comparison of fractionated versus nonfractionated DNA libraries. In order to clearly test the effect of GC fractionation on the recovery of low-abundance sequences in the complex mixture of bacterial community DNA, a direct comparison was made between 16S rRNA gene clone libraries generated with and without the use of GC fractionation. No substantive differences in phylum or genus level community composition were detected (see the supplemental material). However, when these aligned sequences were analyzed using the DOTUR and EcoSim programs, a species detection (also known as rarefaction) curve of OTUs detected at 97% sequence similarity indicated a higher rate of recovery of new phylotypes for the GC-fractionated library (Table 5; see the supplemental material). In addition, the values for the Shannon-Weaver diversity index, evenness, and Chao1 richness estimation were all higher for the GC-fractionated DNA (Table 5).

TABLE 4. Effect of sequence length on taxonomic placement and distance based on ARB alignment

Taxon[a]	Full-length		Truncated	
	% Sequence similarity	No. of sequences	% Sequence similarity	No. of sequences
Phylum				
Acidobacteria	58	176	46	175
Bacteroidetes	52	106	46	107
CD OD1	57	20	46	20
Gemmatimonadetes	75	42	51	41
Planctomycetes	46	88	39	88
Class				
Betaproteobacteria	61	141	54	144

[a] CD, candidate division.

TABLE 5. Effect of GC fractionation on similarity-based OTU numbers, Shannon-Weaver diversity indices, and richness estimates

Fractionation status	No. of unique OTUs	Shannon-Weaver index	Evenness	Richness estimate	% of shared OTUs[c]
GC fractionated[a]	335	5.62	0.966	1,020	64
Not GC fractionated[b]	301	5.45	0.954	780	74

[a] Based on 487 starting sequences.
[b] Based on 490 starting sequences.
[c] OTUs identified in both libraries.

To compare community composition and classification results for the above-mentioned libraries, the data were analyzed using the SONS program (27), which was designed to compare OTUs between libraries in order to establish patterns of community membership and structure based on sequence comparisons. This analysis indicated that GC fractionation facilitated the detection of a higher number of OTUs, both shared and unique, from the same soil bacterial community (Table 5; see the supplemental material).

Community composition. We relied on a tree-based approach utilizing an ARB-annotated (18) sequence library into which our sequences were placed for assignment of phylogenies. Essentially all of the sequences in the study were assigned into 25 known phyla by this approach, with just 35 of the 4,898 sequences not assignable to any known phylum or group (Table 2). The most predominant phylum in this soil was the *Proteobacteria*, which comprised 35% of the sequence library, followed by the *Acidobacteria* with 20% of the total. Six other phyla, including *Actinobacteria*, *Bacteroidetes*, *Thermomicrobia*, *Gemmatimonadetes*, *Planctomycetes*, and *Verrucomicrobia*, averaged 7% representation. The remaining phyla were represented by numbers of phylotypes totaling <2% of the total library.

DISCUSSION

Based on the results presented herein, this agricultural soil bacterial community was empirically demonstrated to be a highly complex assemblage with extremely broad evenness. Such a community composition requires vast sequencing efforts to even approach onefold coverage of richness and to obtain reliable results for traditional ecological parameters originally developed for the analysis of many metazoan communities. One way to mitigate sample size requirements for complete coverage of community diversity is to reduce sample complexity and disparity in abundance between taxa by prefractionation of community DNA, using methods such as GC fractionation. Using this method, 1,714 OTUs were detected at a sequence similarity level cutoff of 97% (representing a new OTU for every 2.9 sequences acquired), with an estimated 3,555 OTUs present. These values are potentially underestimations due to the focus on an ~400-bp hypervariable region within the 16S gene, with the corrected richness estimation for full-length sequences approaching 6,500 OTUs (based on the twofold increase detected in our data) (Table 3). Compared to the results of other, conventional 16S rRNA gene clone library-based soil studies (26), our library exhibits an ~1.6-fold increase in the Shannon-Weaver diversity index, most likely due to the 50-fold increase in sample size and DNA prefractionation approach employed. This is the highest index reported to date for a bacterial community and presumably reflects the additional resolution afforded by the unique combination of existing and novel approaches employed.

While it may seem intuitive even in the absence of empirical data as presented here, the comparison of different-sized libraries from the same sample clearly demonstrates that for highly complex bacterial communities, such as those typically found in surface soils, rich sampling of 16S rRNA gene sequences (i.e., several thousand) is necessary to obtain a robust measure or estimation of community diversity parameters. This is especially true where even near saturation of sampling curves is not feasible or is seemingly impossible due to large numbers of taxa exhibiting high degrees of evenness, or where theoretical estimates based on sample sizes under 1,000 do not appear to be accurate (e.g., asymptotic behavior is not yet apparent in a sampling curve). The importance of at least approaching sampling saturation is supported by a recent publication indicating that surveys missing or ignoring a small subset (e.g.<10%) of species result in minimal loss of information but that more-extensive gaps substantially increase rates of information loss (33).

To directly compare the effectiveness of GC fractionation for sampling coverage, we compared our results to a nonfractionated 500-clone library from the same soil sample, which produced lower recovery of OTUs, as well as a lower Shannon diversity index and less evenness. The main benefit of GC fractionation prior to PCR amplification is the reduction in DNA complexity within each fraction which allows underrepresented sequences to be detected more readily than in a random survey. This resulted in a higher recovery rate for minority species and more-even detection of total diversity, thereby reducing the required survey size needed to approach complete coverage of the entire bacterial community.

The low and variable levels of sequence similarity required to sort this large group of sequences into phylum-level bins comparable to those suggested by other soil microbiological studies suggests that having a universally applied phylum-level cutoff is impractical and would not apply across the full range of known bacterial taxa. Additionally, the sample size (number of sequences within a given group) showed no correlation with the percent sequence similarity required for clustering, suggesting that there are actual differences in the degree of 16S sequence variation between different phyla. This observation potentially represents different evolutionary strategies between phyla at the molecular level where ribosomal-gene sequence conservation is concerned.

When full-length sequence data were compared to those for the 400-bp region of focus in our library, a 1.4-fold increase in sequence similarity at the phylum level was observed. We suggest that this is explained by the hypervariable nature of the 536f-907r-sequenced region as mentioned above, especially given that the conserved primer regions at each end were removed prior to analysis. The inclusion of nearly full-length sequences in the comparison would introduce several additional highly conserved areas into the analyses and thus lower the overall variation observed between longer 16S rRNA gene sequences. Contrary to what was observed for percent sequence similarity, phylum-level classification based on ARB-based tree generation was highly reproducible independent of fragment length, as shown in Table 4.

Based on the data presented here, we suggest that GC fractionation or other prefractionation approaches for reducing complexity within total community DNA prior to PCR and cloning are useful for DNA-based phylogenetic surveys of microbial community diversity. We further suggest that, even with prefractionation of community DNA, 16S rRNA gene clone libraries of at least 2,000 sequences are required to achieve reliable results for estimating ecological parameters, such as richness, evenness, and diversity, for complex bacterial communities such as those typically found in surface soils. The results also validate the use of the 536f-907r primer set for rapid and relatively inexpensive analysis of total bacterial diversity based on single, unidirectional sequence reads that support binning into a reasonable number of OTUs. This strategy provided sufficient resolution for the analyses described herein. However, analysis of full-length or nearly full-length sequences is highly recommended where phylogenetic placement at the genus or near-species level is desired. The determination of evolutionary relatedness between organisms requires the use of large stretches of genetic information. This is especially true for highly conserved genes, such as the 16S rRNA gene (10).

Wherever possible, phylogenetic surveys should use large library sizes and scrutinize data using multiple taxonomic tools. As part of our study, we used methods from the study of Thompson et al. (31) (Clustal W alignments) and MUSCLE software (9), which produced datasets with similar numbers of OTUs, Chao1 richness estimates, and other diversity parameters. However, phylogenetic trees generated from those approaches did not produce coherent clustering with phylogenetic assignments using RDP tools (not shown). In contrast, phylogenetic trees generated using ARB alignments were reproducible and provided consistent phylogenetic placement with the RDP toolset.

The continued use of nucleic acid sequence-based phylogenetic approaches will yield more information, providing additional insights into the effectiveness and validity of current phylogenetic classification strategies and whether they reflect fundamental biological properties. Continued evolution of this general approach should come with the development of a common platform for data acquisition and analysis, which would allow for microbial community comparisons across multiple studies. Special focus should be given to a universal set of rules for assigning bacterial phylogenies. Although our data clearly suggest that there is no universal cutoff value for phylum assignment, it does not provide enough insight to suggest a specific number of phyla in our sample based strictly on sequence similarity, nor does it suggest phylum-specific cutoff values which might come from a more-complete integration of all data in reliably assigned phylotypes present in extant databases. The fact that sequences in the current study were only reliably affiliated to higher-order phylogenetic groups (phylum level and higher) highlights the need to develop a clearer definition for bacterial phylogenetic assignments at the genus and species level that are based on more than just single 16S rRNA gene sequences. In closing, we suggest that additional studies are needed to explore the extent of taxonomic variance within and between phylogenetic groups to provide additional ecological and biological context that will underpin bacterial community diversity studies into the future.

ACKNOWLEDGMENTS

Funding for this project was provided by the United States Department of Agriculture, National Research Initiative (USDA-CSREES grant no. 2004-03501).

Soil samples for this project were graciously provided by the KBS-LTER. We also gratefully acknowledge the Ribosomal Database Project (RDP-II) for aid in the chimera check process, as well as Linda Schimmelpfennig and Tara Westlie for technical assistance in performing GC fractionations, cloning, and sequence processing. We thank Jared Rapp for aiding in computer-based data analysis.

REFERENCES

1. **Apajalahti, J. H. A., A. Kettunen, M. R. Bedford, and W. E. Holben.** 2001. Percent G+C profiling accurately reveals diet-related differences in the gastrointestinal microbial community of broiler chickens. Appl. Environ. Microbiol. **67:**5656–5667.
2. **Apajalahti, J. H. A., H. Kettunen, A. Kettunen, W. E. Holben, P. H. Nurminen, N. Rautonen, and M. Mutanen.** 2002. Culture-independent microbial community analysis reveals that inulin in the diet primarily affects previously unknown bacteria in the mouse cecum. Appl. Environ. Microbiol. **68:**4986–4995.
3. **Ashelford, K. E., N. A. Chuzhanova, J. C. Fry, A. J. Jones, and A. J. Weightman.** 2005. At least 1 in 20 16S rRNA sequence records currently held in public repositories is estimated to contain substantial anomalies. Appl. Environ. Microbiol. **71:**7724–7736.
4. **Ashelford, K. E., N. A. Chuzhanova, J. C. Fry, A. J. Jones, and A. J. Weightman.** 2006. New screening software shows that most recent large 16S rRNA gene clone libraries contain chimeras. Appl. Environ. Microbiol. **72:**5734–5741.
5. **Brose, U., N. D. Martinez, and R. J. Williams.** 2003. Estimating species richness: sensitivity to sample coverage and insensitivity to spatial patterns. Ecology **84:**2364–2377.
6. **Cole, J. R., B. Chai, R. J. Farris, Q. Wang, A. S. Kulam-Syed-Mohideen, D. M. McGarrell, A. M. Bandela, E. Cardenas, G. M. Garrity, and J. M. Tiedje.** 2007. The ribosomal database project (RDP-II): introducing myRDP space and quality controlled public data. Nucleic Acids Res. **35:**D169–D172.
7. **Curtis, T.** 2006. Microbial ecologists: it's time to "go large". Nature **4:**488.
8. **DeSantis, T. Z., P. Hugenholtz, N. Larsen, M. Rojas, E. L. Brodie, K. Keller, T. Huber, D. Dalevi, P. Hu, and G. L. Andersen.** 2006. Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. Appl. Environ. Microbiol. **72:**5069–5072.
9. **Edgar, R. C.** 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. **32:**1792–1797.
10. **Hanage, W. P., C. Fraser, and B. G. Spratt.** 2006. Sequences, sequence clusters and bacterial species. Philos. Trans. R. Soc. B **361:**1917–1927.
11. **Hartmann, M., and F. Widmer.** 2006. Community structure analyses are more sensitive to differences in soil bacterial communities than anonymous diversity indices. Appl. Environ. Microbiol. **72:**7804–7812.
12. **Holben, W. E.** 1997. Isolation and purification of bacterial community DNA from environmental samples, p. 431–436. *In* C. J. Hurst et al. (ed.), Manual of environmental microbiology, 1st ed. ASM Press, Washington, DC.
13. **Holben, W. E., K. P. Feris, A. Kettunen, and J. H. A. Apajalahti.** 2004. GC fractionation enhances microbial community diversity assessment and detection of minority populations of bacteria by denaturing gradient gel electrophoresis. Appl. Environ. Microbiol. **70:**2263–2270.
14. **Holben, W. E., and D. Harris.** 1995. DNA-based monitoring of total bacterial community structure in environmental samples. Mol. Ecol. **4:**627–631.
15. **Holben, W. E., P. Williams, M. A. Gilbert, M. Saarinen, L. K. Särkilahti, and J. H. Apajalahti.** 2002. Phylogenetic analysis of intestinal microflora indicates a novel mycoplasma phylotype in farmed and wild salmon. Microb. Ecol. **44:**175–185.
16. **Ishii, K., and M. Fukui.** 2001. Optimization of annealing temperature to reduce bias caused by a primer mismatch in multitemplate PCR. Appl. Environ. Microbiol. **67:**3753–3755.
17. **Janssen, P. H.** 2006. Identifying the dominant soil bacterial taxa in libraries of 16S rRNA and 16S rRNA genes. Appl. Environ. Microbiol. **72:**1719–1728.
18. **Ludwig, W., O. Strunk, R. Westram, L. Richter, H. Meier, Yadhukumar, A. Buchner, T. Lai, S. Steppi, G. Jobb, W. Forster, I. Brettske, S. Gerber, A. W. Ginhart, O. Gross, S. Grumann, S. Hermann, R. Jost, A. Konig, T. Liss, R. Lussmann, M. May, B. Nonhoff, B. Reichel, R. Strehlow, A. Stamatakis, N. Stuckmann, A. Vilbig, M. Lenke, T. Ludwig, A. Bode, and K. H. Schleifer.** 2004. ARB: a software environment for sequence data. Nucleic Acids Res. **32:**1363–1371.
19. **Morales, S. E., T. Cosart, J. V. Johnson, and W. E. Holben.** 25 July 2008. Supplemental programs for enhanced recovery of data from the DOTUR application. J. Microbiol. Methods **75:**572–575.
20. **Narang, R., and J. Dunbar.** 2004. Modeling bacterial species abundance from small community surveys. Microb. Ecol. **47:**396–406.
21. **Nusslein, K., and J. M. Tiedje.** 1998. Characterization of the dominant and rare members of a young Hawaiian soil bacterial community with small-subunit ribosomal DNA amplified from DNA fractionated on the basis of its guanine and cytosine composition. Appl. Environ. Microbiol. **64:**1283–1289.
22. **Oren, A.** 2004. Prokaryote diversity and taxonomy: current status and future challenges. Philos. Trans. R. Soc. B **359:**623–638.
23. **Polz, M. F., and C. M. Cavanaugh.** 1998. Bias in template-to-product ratios in multitemplate PCR. Appl. Environ. Microbiol. **64:**3724–3730.
24. **Rappé, M. S., and S. J. Giovannoni.** 2003. The uncultured microbial majority. Ann. Rev. Microbiol. **57:**369–394.
25. **Roesch, L. F. W., R. R. Fulthorpe, A. Riva, G. Casella, A. K. M. Hadwin, A. D. Kent, S. H. Daroub, F. A. O. Camargo, W. G. Farmerie, and E. W. Triplett.** 2007. Pyrosequencing enumerates and contrasts soil microbial diversity. ISME J. **1:**283–290.
26. **Schloss, P. D., and J. Handelsman.** 2005. Introducing DOTUR, a computer program for defining operational taxonomic units and estimating species richness. Appl. Environ. Microbiol. **71:**1501–1506.
27. **Schloss, P. D., and J. Handelsman.** 2006. Introducing SONS, a tool for operational taxonomic unit-based comparisons of microbial community memberships and structures. Appl. Environ. Microbiol. **72:**6773–6779.
28. **Schloss, P. D., and J. Handelsman.** 2004. Status of the microbial census. Microbiol. Mol. Biol. Rev. **68:**686–691.
29. **Schloss, P. D., and J. Handelsman.** 2006. Toward a census of bacteria in soil. PLoS Comput. Biol. **2**(7)**:**e92.
30. **Sogin, M. L., H. G. Morrison, J. A. Huber, D. M. Welch, S. M. Huse, P. R. Neal, J. M. Arrieta, and G. J. Herndl.** 2006. Microbial diversity in the deep sea and the underexplored "rare biosphere". Proc. Natl. Acad. Sci. USA **103:**12115–12120.
31. **Thompson, J. D., D. G. Higgins, and T. J. Gibson.** 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. **22:**4673–4680.
32. **Tiedje, J. M., S. Asuming-Brempong, K. Nusslein, T. L. Marsh, and S. J. Flynn.** 1999. Opening the black box of soil microbial diversity. Appl. Soil Ecol. **13:**109–122.
33. **Vellend, M., P. L. Lilley, and B. M. Starzomski.** 2007. Using subsets of species in biodiversity surveys. J. Appl. Ecol. **45:**161–169.
34. **Wang, Q., G. M. Garrity, J. M. Tiedje, and J. R. Cole.** 2007. Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. Appl. Environ. Microbiol. **73:**5261–5267.

Journal of Microbiological Methods 75 (2008) 572–575

Contents lists available at ScienceDirect

Journal of Microbiological Methods

journal homepage: www.elsevier.com/locate/jmicmeth

Note

Supplemental programs for enhanced recovery of data from the DOTUR application

Sergio E. Morales [a], Theodore Cosart [b,c], Jesse V. Johnson [b,c], William E. Holben [a,c,*]

[a] Microbial Ecology Program, Division of Biological Sciences, The University of Montana, Missoula, MT, USA
[b] Department of Computer Science, The University of Montana, Missoula, MT, USA
[c] Montana-Ecology of Infectious Diseases Program, The University of Montana, Missoula, MT, USA

ARTICLE INFO

Article history:
Received 16 April 2008
Received in revised form 4 July 2008
Accepted 21 July 2008
Available online 25 July 2008

Keywords:
16S rRNA gene
ARB
DOTUR
Phylogenetic analysis

ABSTRACT

In order to retrieve phylogenetic information from distance matrices generated from large-scale clone libraries, and to explore OTU distribution among them, we have developed downstream applications for use with the already available DOTUR program. These programs enhance and ease data extraction, providing phylogeny to the already generated distance data.

The past ten years have seen an explosion in microbial ecology research based on 16S rRNA gene sequences. Concurrent with that movement has come the development of tools for analyzing ever-increasing data sets of rRNA gene sequences (Cole et al., 2003, 2005, 2007; Desantis et al., 2006; Lozupone and Knight 2005; Wang et al., 2007). Two of the most commonly used tools are DOTUR (Schloss and Handelsman 2005) and ARB (Ludwig et al., 2004). ARB properly aligns sequence data and generates distance matrices that can be transformed by DOTUR into operational taxonomic unit (OTU) composition data: groups of sequences binned together under given similarity parameters. OTU groups are then used to make collector's and rarefaction curves for sampling coverage estimations, richness estimators, and diversity indices. Although the two programs represent powerful tools for data analysis, some information is not easily extracted from DOTUR output files. Rank abundance and OTU distribution files are generated, but these are not linked to sequences, and are not ordered by abundance (in the case of OTU distribution). Sequence identity is left out of the analysis unless individual identification tags (names) from DOTUR are manually linked to their corresponding DNA sequence prior to using some phylogenetic assignment tool. This is a time-intensive task which is often skipped, but which provides critical information regarding OTU bin composition. The ability to use alternative methods to validate group phylogeny depends on being able to track specific OTUs to their sequences. Although manual searching and matching is feasible for small libraries, new studies analyzing thousands of sequences make this task intractable.

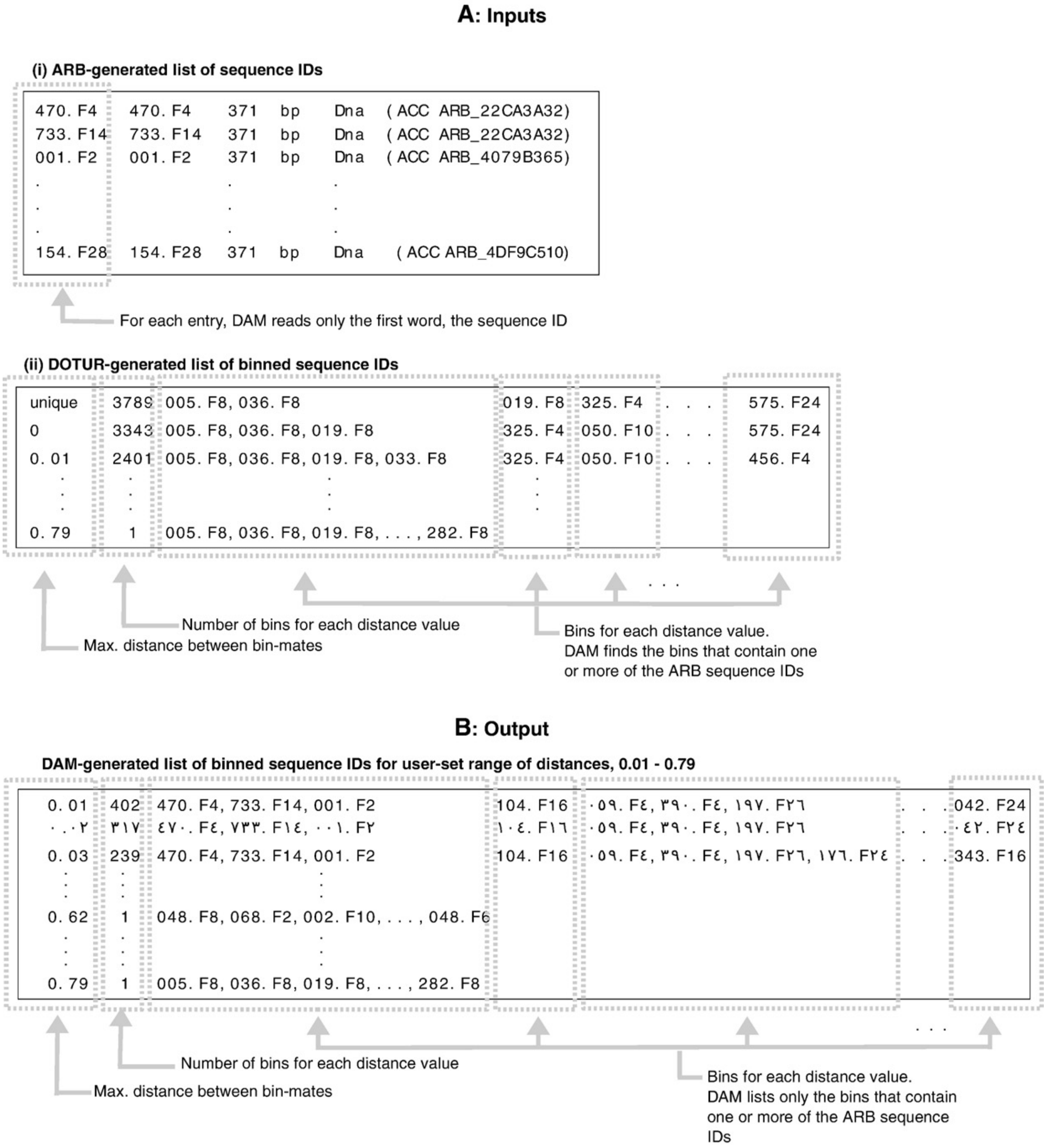
Another concept not easily explored is the cohesive organization of OTU placements at different phylogenetic levels and how they relate to tree topology in well-annotated trees. Although programs like ARB allow users to align their own sequences to reference sequences and insert them into annotated trees, the DOTUR program bins them into OTUs using arbitrary cutoffs not linked to any validated hierarchy (Schloss and Handelsman, 2005). To date, multiple 16S rRNA gene based studies have used this approach without rigorously assessing its appropriateness or validity.

To address both of these concerns, we developed two simple programs that are freely available from the authors at: http://dbs.umt.edu/research_labs/holbenlab/links.php. The DAM (*D*OTUR — *A*RB *M*atching) application matches a list of sequence IDs to bins as given by the DOTUR program for some range of DOTUR distance values (Fig. 1). This allows the user to identify the phylogenetic distance at which all sequences within the provided list are grouped as a single OTU. Input for the program (Fig. 1A) includes: (i) A DOTUR list file comprised of rows of space or tab delimited entries, (ii) A list file with one sequence ID on each line, and (iii) A configuration file (not shown). The program was created to match to DOTUR bins a list of sequences that represent a cluster gathered from the output of the ARB program, but any list of sequence IDs can be matched so long as each is present in all the DOTUR list lines and they are in a file with the proper format.

Having stored the bin information in the DOTUR list file, the list of target sequence IDs (TIDs) to be matched to the bins, and a user-specified range of distance values for each distance value in range, DAM first keys each TID to a bin in the DOTUR list by finding it's match in the DOTUR bin information. Each bin that contains at least one of

[*] Corresponding author. Mailing address: Microbial Ecology Program, Division of Biological Sciences, The University of Montana, Missoula, Montana, 59812-1006, USA. Tel.: +1 406 243 6163; fax: +1 406 243 4184.
E-mail address: bill.holben@mso.umt.edu (W.E. Holben).

0167-7012/$ – see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.mimet.2008.07.017



A: Inputs

(i) ARB-generated list of sequence IDs

```
470. F4    470. F4    371   bp    Dna   ( ACC  ARB_22CA3A32)
733. F14   733. F14   371   bp    Dna   ( ACC  ARB_22CA3A32)
001. F2    001. F2    371   bp    Dna   ( ACC  ARB_4079B365)
.                     .           .
.                     .           .
.                     .           .
154. F28   154. F28   371   bp    Dna    ( ACC ARB_4DF9C510)
```

For each entry, DAM reads only the first word, the sequence ID

(ii) DOTUR-generated list of binned sequence IDs

```
unique   3789  005. F8, 036. F8                          019. F8  325. F4  . . .   575. F24
0        3343  005. F8, 036. F8, 019. F8                 325. F4  050. F10 . . .   575. F24
0. 01    2401  005. F8, 036. F8, 019. F8, 033. F8        325. F4  050. F10 . . .   456. F4
  .        .                     .                         .
  .        .                     .                         .
  .        .                     .                         .
0. 79      1   005. F8, 036. F8, 019. F8, . . . , 282. F8
```

Number of bins for each distance value
Max. distance between bin-mates
. . .
Bins for each distance value.
DAM finds the bins that contain one
or more of the ARB sequence IDs

B: Output

DAM-generated list of binned sequence IDs for user-set range of distances, 0.01 - 0.79

```
0. 01   402  470. F4, 733. F14, 001. F2                104. F16  ٠٥٩. F٤, ٣٩٠. F٤, ١٩٧. F٢٦            . . . 042. F24
٠.٠٢    ٣١٧  ٤٧٠. F٤, ٧٣٣. F١٤, ٠٠١. F٢                ١٠٤. F١٦  ٠٥٩. F٤, ٣٩٠. F٤, ١٩٧. F٢٦            . . . ٠٤٢. F٢٤
0. 03   239  470. F4, 733. F14, 001. F2                104. F16  ٠٥٩. F٤, ٣٩٠. F٤, ١٩٧. F٢٦, ١٧٦. F٢٤ . . . 343. F16
  .      .                     .
  .      .                     .
0. 62    1   048. F8, 068. F2, 002. F10, . . . , 048. F6
  .      .                     .
  .      .                     .
0. 79    1   005. F8, 036. F8, 019. F8, . . . , 282. F8
```

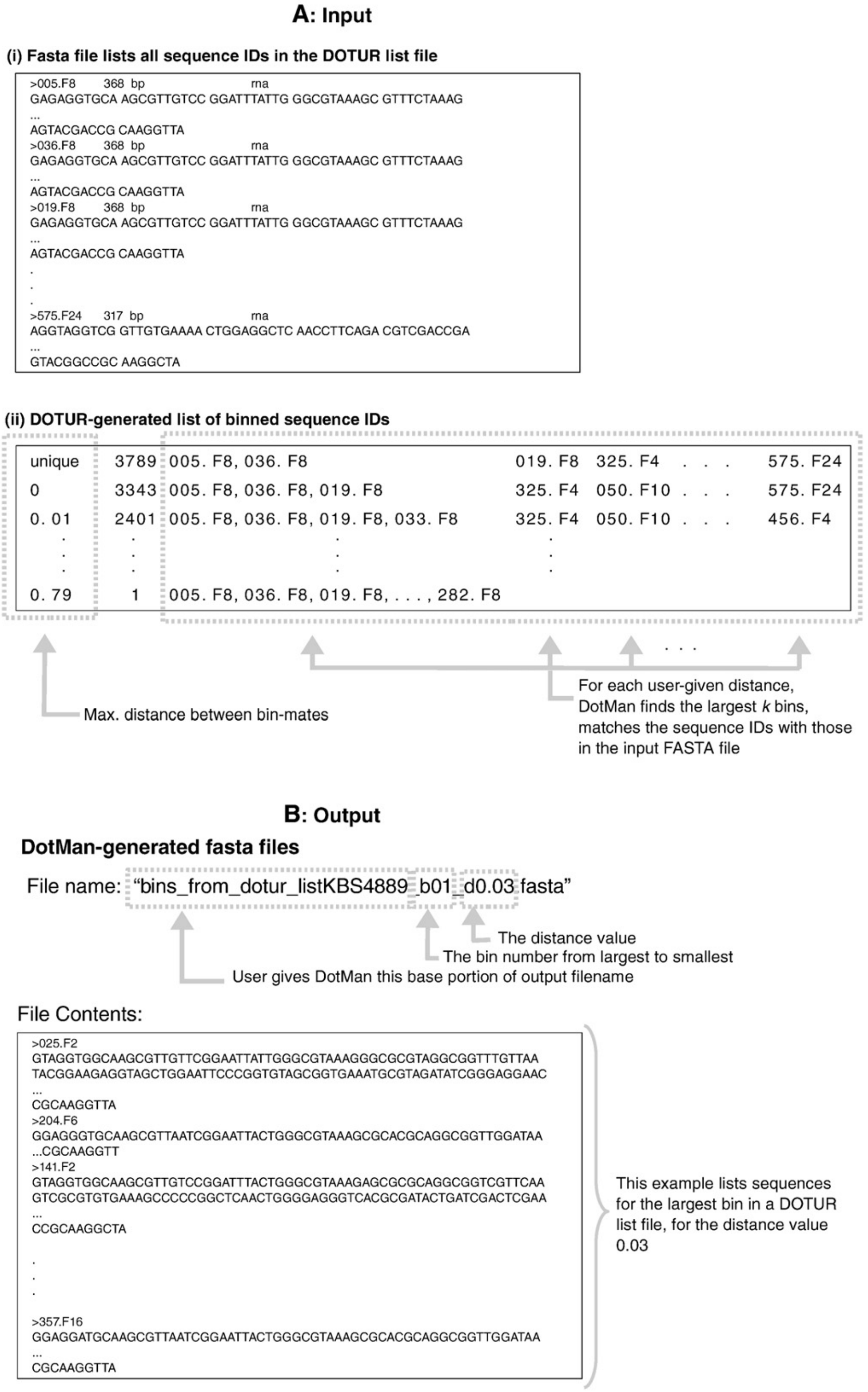
Number of bins for each distance value
Max. distance between bin-mates
. . .
Bins for each distance value.
DAM lists only the bins that contain
one or more of the ARB sequence
IDs

Fig. 1. Examples of the format of input files and the output file from the DAM program. The data shown comes from a 16S rRNA gene sequence survey from an agricultural soil in Michigan. DAM also reads a configuration file (not shown) listing the user-specified settings, an example of which is given with the program's help file.

the TIDs is then copied from the DOTUR file and added to a list of bins for the given distance. The output, then, is a filtered version of the original DOTUR list file, in which bins are listed only for the specified range of distance values, and for each distance only bins containing at least one of the TIDs in the sequence ID list file.

DAM output is a file formatted similarly to a DOTUR list file (Fig. 1B). For each DOTUR distance value, there is a line in the file giving the distance value itself and the set of bins found in the DOTUR file that account for all of the sequences given by the list file, followed by a list of bins and their contents.

The second program, DotMan (*DOT*UR *Man*ipulation Program), creates FASTA files from DOTUR bins (Fig. 2). For a list of DOTUR distance values, it makes one FASTA file for each of the k largest bins listed at each distance. Inputs for the program (Fig. 2A) include: (i) A DOTUR list file comprised of rows of space- or tab-delimited entries, (ii) A file of sequences and their identification tags in FASTA format matching those contained in the DOTUR list, and (iii) A configuration file that provides the program with the names of the above data files, a base name for the FASTA files, a range of distance values over which to create FASTA files, and a list of distance values, each of which is the basis for a series of FASTA files (not shown). Dotman reads, in order, the DOTUR list file, the FASTA file, the user-specified *k* value and then the list of distance values. Then, for each specified distance value, the program first orders the DOTUR bins by population from largest to smallest. For each of the *k* largest bins in the list, it matches each sequence ID in the bin to an entry in the FASTA file, assembling one FASTA file for each bin. In each output file, The ID entry is the sequence ID as given in the DOTUR bin, and the sequence itself is formatted as



A: Input
(i) Fasta file lists all sequence IDs in the DOTUR list file

```
>005.F8      368  bp                       rna
GAGAGGTGCA AGCGTTGTCC GGATTTATTG GGCGTAAAGC GTTTCTAAAG
...
AGTACGACCG CAAGGTTA
>036.F8      368  bp                       rna
GAGAGGTGCA AGCGTTGTCC GGATTTATTG GGCGTAAAGC GTTTCTAAAG
...
AGTACGACCG CAAGGTTA
>019.F8      368  bp                       rna
GAGAGGTGCA AGCGTTGTCC GGATTTATTG GGCGTAAAGC GTTTCTAAAG
...
AGTACGACCG CAAGGTTA
.
.
.
>575.F24     317  bp                       rna
AGGTAGGTCG GTTGTGAAAA CTGGAGGCTC AACCTTCAGA CGTCGACCGA
...
GTACGGCCGC AAGGCTA
```

```
unique   3789  005. F8, 036. F8                          019. F8   325. F4   . . .    575. F24
0        3343  005. F8, 036. F8, 019. F8                 325. F4   050. F10  . . .    575. F24
0. 01    2401  005. F8, 036. F8, 019. F8, 033. F8        325. F4   050. F10  . . .    456. F4
  .        .                    .                          .
  .        .                    .                          .
  .        .                    .                          .
0. 79      1   005. F8, 036. F8, 019. F8, . . . , 282. F8
```

. . .
Max. distance between bin-mates
For each user-given distance, DotMan finds the largest k bins, matches the sequence IDs with those in the input FASTA file
B: Output
DotMan-generated fasta files
File name: "bins_from_dotur_listKBS4889_b01_d0.03 fasta"
The distance value
The bin number from largest to smallest
User gives DotMan this base portion of output filename
File Contents:

```
>025.F2
GTAGGTGGCAAGCGTTGTTCGGAATTATTGGGCGTAAAGGGCGCGTAGGCGGTTTGTTAA
TACGGAAGAGGTAGCTGGAATTCCCGGTGTAGCGGTGAAATGCGTAGATATCGGGAGGAAC
...
CGCAAGGTTA
>204.F6
GGAGGGTGCAAGCGTTAATCGGAATTACTGGGCGTAAAGCGCACGCAGGCGGTTGGATAA
...CGCAAGGTT
>141.F2
GTAGGTGGCAAGCGTTGTCCGGATTTACTGGGCGTAAAGAGCGCGCAGGCGGTCGTTCAA
GTCGCGTGTGAAAGCCCCCGGCTCAACTGGGGAGGGTCACGCGATACTGATCGACTCGAA
...
CCGCAAGGCTA

.
.
.

>357.F16
GGAGGATGCAAGCGTTAATCGGAATTACTGGGCGTAAAGCGCACGCAGGCGGTTGGATAA
...
CGCAAGGTTA
```

This example lists sequences for the largest bin in a DOTUR list file, for the distance value 0.03

Fig. 2. Examples of the format of DotMan input files and an example FASTA file from its output. These examples come from the same soil study noted in Fig. 1. As with the DAM program, DotMan also reads a configuration file (not shown) with user-specified settings. An example of the configuration file is given in the program's help file.

given in the original FASTA file (Fig. 2B). If the number of bins *n* in the DOTUR list for a given distance value is less than *k*, DotMan writes one file for each of the *n* bins, ordered from largest to smallest. When selecting the i*th* bin of the k largest bins, DOTUR bin order is not necessarily preserved when selecting among bins of equal size. This means, for example, that when requesting the 5 largest bins from a set of 10 bins of equal size, the FASTA files produced will not necessarily represent the 5 leftmost bins as ordered in the DOTUR list file.

Both programs are written in C++ as command-line programs. They are available either as Windows-executables or source-code packages for compiling on Linux or the MacIntosh using OSX. Recent studies in our lab have used both programs, allowing us to identify specific groups of sequences found to be numerically dominant within our clone library (Morales et al., 2008)and demonstrating that “universal” cutoff values for binning at the phylum level were not observed within the studied site . A sample output file from that study is used in Fig. 1. The figure shows the distance score (similarity score is calculated as 1-distance score) needed to bin all Acidobacteria sequences into a single OTU (62% distance or 38% sequence similarity). These tools should be useful to anyone interested in identifying specific subgroups of OTUs at given taxonomic levels of resolution, or wanting to corroborate OTU bins by way of tools like the RDP Classifier (Wang et al., 2007) or BLAST (Altschul et al., 1990; Tatusova and Madden 1999).

References

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J., 1990. Basic local alignment search tool. Journal of Molecular Biology 215 (3), 403–410.

Cole, J.R., Chai, B., Farris, R.J., Wang, Q., Kulam-Syed-Mohideen, A.S., Mcgarrell, D.M., Bandela, A.M., Cardenas, E., Garrity, G.M., Tiedje, J.M., 2007. The Ribosomal Database Project (RDP-II): Introducing MyRDP Space And Quality Controlled Public Data. Nucleic Acids Research 35 (Suppl_1), D169–D172.

Cole, J.R., Chai, B., Farris, R.J., Wang, Q., Kulam, S.A., Mcgarrell, D.M., Garrity, G.M., Tiedje, J.M., 2005. The Ribosomal Database Project (RDP-II): sequences and tools for high-throughput rRNA analysis. Nucleic Acids Research 33 (Suppl_1), D294–D296.

Cole, J.R., Chai, B., Marsh, T.L., Farris, R.J., Wang, Q., Kulam, S.A., Chandra, S., Mcgarrel, D.M., Schmidt, T.M., Garrity, G.M., Tiedje, J.M., 2003. The Ribosomal Database Project (RDP-II): previewing a new autoaligner that allows regular updates and the new prokaryotic taxonomy. Nucleic Acids Research 31 (1), 442–443.

Desantis, T.Z., Hugenholtz, P., Larsen, N., Rojas, M., Brodie, E.L., Keller, K., Huber, T., Dalevi, D., Hu, P., Andersen, G.L., 2006. Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. Applied And Environmental Microbiology 72 (7), 5069–5072.

Lozupone, C., Knight, R., 2005. Unifrac: a new phylogenetic method for comparing microbial communities. Applied And Environmental Microbiology 71 (12), 8228–8235.

Ludwig, W., Strunk, O., Westram, R., Richter, L., Meier, H., Kumar, Y., Buchner, A., Lai, T., Steppi, S., Jobb, G., Förster, W., Brettske, I., Gerber, S., Ginhart, A.W., Gross, O., Grumann, S., Hermann, S., Jost, R., König, A., Liss, T., Lüßmann, R., May, M., Nonhoff, B., Reichel, B., Strehlow, R., Stamatakis, A., Stuckmann, N., Vilbig, A., Lenke, M., Ludwig, T., Bode, A., Schleifer, K.H., 2004. ARB: A Software Environment For Sequence Data. Nucleic Acids Research 32 (4), 1363–1371.

Morales, S.E., Cosart, T.F., Johnson, J.V., & Holben, W.E., 2008, “Phylogenetic Analysis Of A Complex Bacterial Community Reveals Effects Of Amplicon Length, Degree Of Coverage And DNA Fractionation”, In Review.

Schloss, P.D., Handelsman, J., 2005. Introducing DOTUR, a computer program for defining operational taxonomic units and estimating species richness. Applied And Environmental Microbiology 71 (3), 1501–1506.

Tatusova, T.A., Madden, T.L., 1999. BLAST 2 sequences, a new tool for comparing protein and nucleotide sequences. FEMS Microbiology Letters 174, 247–250.

Wang, Q., Garrity, G.M., Tiedje, J.M., Cole, J.R., 2007. Naive Bayesian Classifier For Rapid Assignment of rRNA Sequences Into The New Bacterial Taxonomy. Applied And Environmental Microbiology 73 (16), 5261–5267.

The ISME Journal (2010) 4, 799–808
© 2010 International Society for Microbial Ecology All rights reserved 1751-7362/10 $32.00
www.nature.com/ismej

ORIGINAL ARTICLE

Bacterial gene abundances as indicators of greenhouse gas emission in soils

Sergio E Morales[1], Theodore Cosart[2,3] and William E Holben[1,3]
[1]*Microbial Ecology Program, Division of Biological Sciences, The University of Montana, Missoula, MT, USA;* [2]*Department of Computer Science, The University of Montana, Missoula, MT, USA and* [3]*Montana—Ecology of Infectious Diseases Program, The University of Montana, Missoula, MT, USA*

Nitrogen fixing and denitrifying bacteria, respectively, control bulk inputs and outputs of nitrogen in soils, thereby mediating nitrogen-based greenhouse gas emissions in an ecosystem. Molecular techniques were used to evaluate the relative abundances of nitrogen fixing, denitrifying and two numerically dominant ribotypes (based on the $\geqslant$97% sequence similarity at the 16S rRNA gene) of bacteria in plots representing 10 agricultural and other land-use practices at the Kellogg biological station long-term ecological research site. Quantification of nitrogen-related functional genes (nitrite reductase, *nirS*; nitrous oxide reductase, *nosZ*; and nitrogenase, *nifH*) as well as two dominant 16S ribotypes (belonging to the phyla *Acidobacteria*, *Thermomicrobia*) allowed us to evaluate the hypothesis that microbial community differences are linked to greenhouse gas emissions under different land management practices. Our results suggest that the successional stages of the ecosystem are strongly linked to bacterial functional group abundance, and that the legacy of agricultural practices can be sustained over decades. We also link greenhouse gas emissions with specific compositional responses in the soil bacterial community and assess the use of denitrifying gene abundances as proxies for determining nitrous oxide emissions from soils.

The ISME Journal (2010) **4,** 799–808; doi:10.1038/ismej.2010.8; published online 25 February 2010
Subject Category: Geomicrobiology and microbial contributions to geochemical cycles
Keywords: 16S rRNA; greenhouse gases; KBS-LTER; *nifH*; *nirS*; *nosZ*

Introduction

The impact of agricultural practices on the environment has been studied extensively, leading to changes in land management policy worldwide (Tilman *et al.*, 2002). Yet, surprisingly little is known about the interactions between agroecosystem management practices and the soil microbial community, which has a key role in nutrient transformation and chemical cycling (Staley and Reysenbach, 2002). The Kellogg biological station long-term ecological research (KBS-LTER) site has hosted numerous microbiological studies (Bruns *et al.*, 1998, 1999; Broughton and Gross, 2000; Phillips *et al.*, 2000a, b; Buckley and Schmidt, 2001; Blackwood and Paul, 2003), but few studies have focused on quantitative analysis of bacterial community composition in relation to nitrogen turnover rates, specifically those related to greenhouse gas emissions. In addition, comparative quantitative analysis of specific functional or phylogenetic groups within the soil community is still limited. To date, genes encoding enzymes involved in nitrogen cycling have been targets of choice for studies focusing on functional groups (that is, guilds) of bacteria (Leininger *et al.*, 2006; Henry *et al.*, 2008). This focus is well founded as nitrogen is essential for plant growth, along with phosphate, carbon, hydrogen and oxygen. However, simultaneous comparison of the abundance of multiple N-cycle-related genes across multiple treatments and land-use types has not yet been conducted, especially when framed around an ecosystem-level process, such as, greenhouse gas emissions from soils.

The United States Department of Agriculture tracks emissions of multiple greenhouse gases related to agricultural activities and ranks them based on their global warming potential (GWP). Of all the sources of GWP in cropping systems, including CO_2 and CH_4, none are more poorly quantified than N_2O production (Robertson and Grace, 2004). This represents a tremendous knowledge gap regarding the role of N_2O in global warming, especially considering the fact that its GWP is 296-fold greater than that of CO_2 and it is frequently the major source of GWP in agricultural systems (Robertson *et al.*, 2000; Robertson and Grace, 2004; EIA, 2008). Poor quantification of

Correspondence: SE Morales, Microbial Ecology Program, Division of Biological Sciences, The University of Montana, Missoula, Montana, MT 59812–1006, USA.
E-mail: sergio.morales@mso.umt.edu

Received 25 September 2009; revised 4 January 2010; accepted 4 January 2010; published online 25 February 2010

N_2O is to a large extent linked to the challenges of measuring N_2O fluxes in the field, requiring numerous measurements with inherently high variability. This constraint has limited data collection, and, until refined, represents a rudimentary measurement of a globally important activity. The ability to construct more informative models, predictions and mitigation strategies related to greenhouse gas emissions depends on the development of new analytical approaches that are more efficient and accurate than currently available ones. As microbial populations control natural production and consumption of nitrous oxide, their abundance and activities represent a potential method for predicting gas emissions.

Using real-time quantitative PCR (qPCR), we examined the abundance of key soil microbial guilds and taxa, including nitrogen-fixing bacteria (through the *nifH* gene) and denitrifying bacteria (through the *nirS* and *nosZ* genes). Although nitrification represents a key step in the conversion of ammonia nitrogen into its gaseous forms (NO_2 and NO_3), this process is less relevant to N_2O emissions. Further, nitrifier numbers are typically low in soils and are challenging to quantify using direct qPCR as it is applied to all other genes analyzed making nitrification measures beyond the scope of this study. Quantification of two numerically predominant operational taxonomic units (OTUs) belonging to the phyla *Acidobacteria* and *Thermomicrobia* obtained at KBS (Morales and Holben, 2009) was also performed for comparison with data obtained using function-based primers. These analyses were performed across 10 different treatments based on land-use types at the KBS-LTER. Analyses in other soils using 16S rRNA gene-based PCR denaturing gradient gel electrophoresis have shown that soil type may be the strongest selector of soil microbial community structure (Wakelin *et al.*, 2008), but that study focused on large-scale rearrangements in community composition measured using a broad-scale technique. However, changes in functional or phylogenetic group abundance may go undetected when using such 16S rRNA gene-based approaches, as poor resolution between taxa due to gene conservation, nonspecific primers and other factors (for example, see Morales and Holben, 2009) can lead to mistaken conclusions about functional group abundance or population dynamics. To examine the extent to which this occurs in this system, the current study employed qPCR targets directly related to functional traits, as well as two 16S rRNA gene-based ribotypes, to compare the patterns observed based on the abundance of genes encoding key enzymatic activities with those observed using 'OTU-based primers'.

Four general hypotheses were tested. The first stated that 16S-based taxon abundance estimates would be higher than those observed for the functional genes as the former can detect multiple phylogenetic subgroups that might each harbor multiple different bacterial functional groups (for example, denitrifiers and nitrogen fixers can both be found within the same genus). Second, we hypothesized that, due to the high number of leguminous and other nitrogen-fixing symbiotic plants in certain land-use treatments at the KBS-LTER site, nitrogenase reductase gene (*nifH*) numbers would be consistently higher in all treatments involving leguminous cover crops (that is, treatments 1–4 (T1–4), soybean; and T6, alfalfa). The third hypothesis predicted that abundance of denitrification bacterial genes (*nosZ* and *nirS*) would be relatively uniform across all treatments, given the widespread distribution of this metabolic activity across the breadth of bacterial phylogenetic groups. Finally, as the balance between input and output of nitrogen gas in soils is respectively controlled by nitrogen fixers and denitrifiers, we hypothesized that differences in bacterial gene abundances of these key nitrogen cycling genes between annual, perennial and successional sites would correspond to those observed for greenhouse gas emission rates for these sites.

Materials and methods

Study site and sample collection

Samples were collected from the KBS-LTER Row-Crop Agriculture site in mid-Michigan (for an overview of that project see http://lter.kbs.msu.edu/). Our study examined the bacterial community in the replicate plots of Treatment 1–8 (T 1–8) of the main experimental site, as well as two additional successional and forest sites (Table 1). Four of the eight main site treatments are annual crop rotations (T1–4), two are perennial (T5, poplars; and T6, alfalfa), and two are successional systems under native vegetation

Table 1 Kellogg biological station LTER treatment and successional regimes

Treatment	*Crop cover*	*Notes*
T1	WCS	STD, chisel plowed
T2	WCS	STD, no-till
T3	WCS	Org red, N at planting, WCCC, H, PPC
T4	WCS	Rotary hoed, WCCC, PCC
T5	Continuous poplar	
T6	Continuous alfalfa	
T7	Native successional	Last plowed on spring of 1989
T8	Native mid-successional	Never plowed, soil organic matter historical control
DFR	Deciduous forest	Late successional site
SFR	Successional	40 to 60-year-old former agricultural field

Abbreviations: DFR, deciduous forest; LTER, long-term ecological research; N, nitrogen added; Org Red, reduced input organic; PCC, post planting cultivation; SFR, successional forest; STD, standard; WCCC, winter clover cover crop; WCS, wheat corn soy rotation.

(T7, former agricultural site left fallow following spring plowing in 1989; and T8, never plowed or cultivated). The two forested sites comprised of a set of 40 to 60-year-old successional sites (former agricultural fields, SFR (successional forest) and DFR (a deciduous forest site that has never been cut)). All soils in these treatments are classified as fine loamy, mixed and mesic Typic Hapludalfs. Sampling was carried out on 2 May 2007 by collecting five randomly positioned, 0–20 cm soil cores using a soil probe from each of the treatment replicates as is standard at the LTER site (http://lter.kbs.msu.edu/protocols/112). Each set of five samples was sieved through 2 mm mesh and then mixed thoroughly in equal proportions, providing a single composite sample for each replicate treatment plot (that is, there were either three or six replicated plots per treatment, and one composited sample was developed and examined for each replicate plot). All soil samples were stored in Whirl-Pak bags on dry ice or at −70 °C immediately after sieving and mixing until processed for bacterial community DNA extraction. All samples were processed within 20 days of sampling.

DNA extraction

Total community DNA was extracted from 0.25 g of each soil sample, in triplicate, using the MoBio PowerSoil DNA Isolation Kit (MoBio, Solana Beach, CA, USA) according to the manufacturers instructions and using sterile MilliQ water in the final elution step. All DNA samples were stored at −20 °C until used in downstream analyses.

Real-time qPCR assays

Real-time qPCR was performed using an iCycler iQ thermocycler (Bio-Rad, Hercules, CA, USA) with an ABsolute QPCR SYBR green mix (AbGene, Epsom, UK) using primers and conditions previously described (nitrogenase reductase (*nifH* gene) (Rösch and Bothe, 2005; Yergeau *et al.*, 2007); nitrite reductase (*nirS* gene) (Throback *et al.*, 2004; Yergeau *et al.*, 2007); nitrous oxide reductase (*nosZ*) (Henry *et al.*, 2006); *Thermomicrobia* group 4 (OTU-specific based on ⩾97% sequence similarity at the 16S rRNA gene) (Morales and Holben, 2009); *Acidobacteria* group 6 (OTU-specific based on ⩾97% sequence similarity at the 16S rRNA gene) (Morales and Holben, 2009)) and are summarized in Appendix Table S1. Although no function- or 16S-based primer sets are necessarily comprehensive across the spectrum of microbial diversity, they have been widely used to good effect in comparative studies (for example, between treatments), as is the case in the current study. Known template standards were made from cloned PCR products amplified from whole-genome extracts of pure bacterial isolates (see Appendix Table S1), and each standard was sequenced to confirm target identity. Primer validation analyses were performed before use as previously described (Morales and Holben, 2009).

Variance in gene abundance measurements was determined between replicate plots of the same treatment by pooling three individual DNA preparations from each replicate plot in equimolar amounts to provide a representative sample for that replicate plot. A total of 5 ng of DNA was used to compare gene abundance in each plot. For comparing the effect of different treatments on the overall abundance of each gene, DNA preparations from all replicate plots and extractions within a treatment were combined in equimolar amounts to provide a representative sample for that treatment. All qPCR reactions for any single sample were run at least in triplicate, as described above.

Correct target amplification from soil DNA was confirmed by cloning PCR fragments from T1. Triplicate standard PCR reactions were performed separately as described above for each primer pair using total community DNA from T1, which represents the canonical treatment practice for the KBS-LTER site. The resultant PCR products were purified and cloned as previously described (Morales and Holben, 2009) to confirm that specific amplification of the corresponding target had occurred.

Statistical analysis

Relationships between microbial gene abundance, successional stage, greenhouse gas flux and other environmental parameters, were determined by principal components analysis (PCA) with data matrices composed of chemical and bacterial qPCR data (Supplementary Table S2) for T1–8, SFR and DFR of the KBS-LTER, collectively representing annual, perennial and successional sites. Chemical data were extracted from Robertson *et al.* (2000) and represented gas fluxes and their respective greenhouse warming potentials, aboveground net primary productivity, NO_3-N, N mineralization potentials and soil carbon concentrations over an 8-year period. The chemical metadata set used was from 1991 to 1999 (as reported in Robertson *et al.*, 2000) because this timeframe maximized the number of variables available for analysis, as several of the measurements were not continued beyond that point. However, more recent gas emission data through 2007 (presented in Supplementary Figure S6) strongly support the suggestion that the same soil processes and relationships persisted at KBS through our sampling time and beyond, as observed differences in these parameters are consistent across treatments and the relative relationships between treatments remain constant. Further, the classification of treatments as a sink or source of gases on the basis of current or past mean gas fluxes for any treatment remains the same. Data were organized with rows representing treatments and columns representing individual variables. Principal component scores were plotted by site, abbreviated as shown in the site description and in Table 1, for the first two principal components. Parameters driving the distribution of the PCA plots were determined by querying all variables against the first three principal components. Individual chemical factors were also independently queried to qPCR results on a per gene basis by conducting pairwise correlations.

To show co-trends between genetic and chemical variables, data were also analyzed using co-inertia analysis (CIA) (Dray *et al*., 2003), computed with the MADE4 package in the R statistical software environment (Culhane *et al*., 2005). CIA is a dimensional reduction procedure designed to measure the similarity of two sets of variables (measurements), as they are associated with a single set of cases. Due to very strong correlations (>0.99) between some chemistry measurements, we were able to unclutter the CIA plot by using CH_4, N_2O and Organic C as proxies for CH_4-C, N_2O-N and Organic C (kg), respectively.

Results

Bacterial gene abundance and diversity

The efficiencies of real-time PCR assays for all targets averaged 92% (s.d. ± 4%), allowing for direct comparison of results for all targets. Statistically significant differences in gene target abundances were observed between replicate plots under the same treatment (T1) for all tested genes (Table 2). As anticipated, qPCR results obtained from pooling DNA extracted from each individual replicate plot and run as a representative sample for that treatment resulted in mean values within the standard deviation of the true replicates (Figure 1). Two 16S rRNA gene targets representing the numerically dominant bacterial groups *Thermomicrobia* and *Acidobacteria* based on sequences generated from this site (Morales *et al*., 2009) were the most abundant of all targets tested (Figure 1), supporting the first hypothesis regarding higher abundances based on phylogenetic targets (that is, ‘genus-level’ 16S rRNA genes) compared with functional gene (that is, enzyme coding) targets.

By contrast, the measured abundances of the nitrogenase gene (*nifH*) did not support our second hypothesis that predicted higher abundance in treatments containing crops with known symbiotic nitrogen fixation associations (for example, soybean rotations). Indeed, the abundance of *nifH* was generally higher in the successional treatments (SFR and DFR) than in traditional agricultural sites (Figure 1). Denitrifier abundance, as indicated by *nirS* and *nosZ* gene abundance, varied significantly between treatments, which did not support the third hypothesis predicting comparable denitrifier numbers between treatments.

Correlating bacterial gene abundance to environmental variables

Principal component analysis of annual ecosystem averages for key environmental parameters, global warming potential (GWP), bacterial qPCR results and a combined data set comprised of all variables was employed to assess relationships between bacterial gene abundance and process-level measurements. Strong clustering of samples based on aboveground plant cover type (that is, annual, perennial, successional) was observed (Figure 2), supporting the fourth hypothesis predicting differences in greenhouse gas emissions as being correlated to differences in the balance between nitrogen fixing and denitrifying bacteria. PCA based solely on qPCR results from bacterial gene targets accounted for the most variance within the first two components (~89%). Lower combined principal component 1 and 2 scores were observed for annual ecosystem averages and GWP values (~75%) and for the combined data set (~68%) (Figure 2). All three PCA plots showed general clustering of sites based on land-use type, with the exception of T7 (the early-successional site which was previously a woodpile). Loadings for principal components 1 and 2 (that is, loading vectors) were plotted to show independent variable contributions to variance between the treatments (Figure 2). Successional sites were generally associated with higher organic carbon levels and *nifH* gene abundance. Perennial plant-based treatments, as well as the early native successional plot T7, exhibited increased nitrous oxide reducer abundance (*nosZ*), total carbon levels and abundance of *Thermomicrobia* and *Acidobacteria*. Aboveground annual productivity, nitrite reducers (*nirS*) and GWP were all strongly

Table 2 *t*-Test analysis of replicate plot qPCR values for each gene target within T1.

Target	*Test value*	*Count*	*Mean*	*Std dev*	*Std Error of mean*	*t*	*df*	P-*level*
AB#6	10	24	2.63E + 05	1.36E + 05	2.79E + 04	9.44	23	<0.001
TM#4	10	23	9.97E + 04	6.91E + 04	1.44E + 04	6.92	22	<0.001
nifH	10	21	3.59E + 03	1.75E + 03	3.82E + 02	9.37	20	<0.001
nirS	10	19	4.63E + 04	1.98E + 04	4.53E + 03	10.2	18	<0.001
nosZ	10	20	8.16E + 03	4.71E + 03	1.05E + 03	7.74	19	<0.001

Abbreviations: t, test score; df, degrees of freedom; qPCR, quantitative PCR; Std dev, standard deviation.

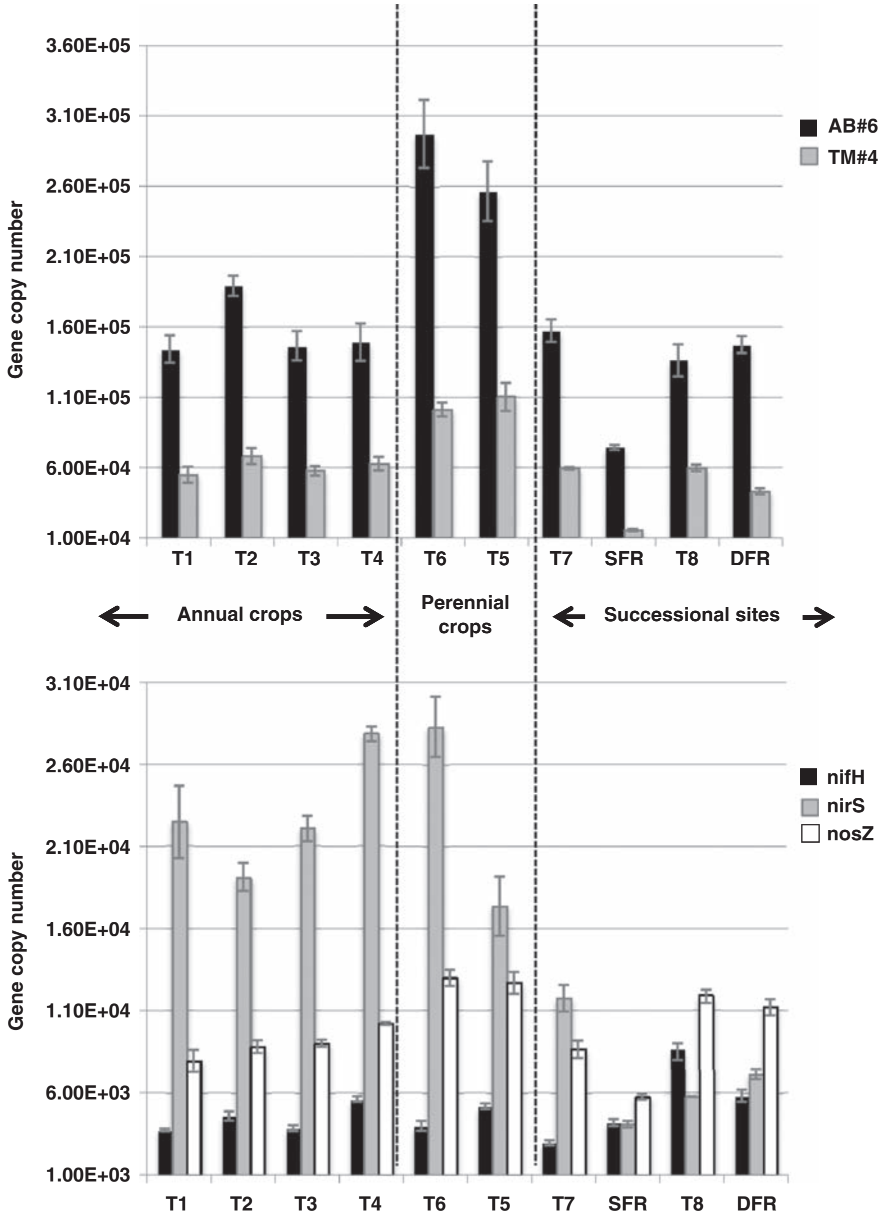


Figure 1 Specific detection of bacterial groups by quantitative PCR (qPCR). Values indicate gene copy numbers, as determined from 5 ng of DNA extracted from soil using a dilution curve with known standards. Error bars are one standard error of the mean (s.e.) of at least triplicate qPCR reactions ($n \geqslant 3$). AB#6 and TM#4: operational taxonomic units based on $\geqslant$97% sequence similarity to the 16S rRNA gene representing *Acidobacteria* group 6 and *Thermomicrobia* group 4, respectively. *nifH*, nitrogenase gene; *nirS*, nitrite reductase gene; *nosZ*, nitrous oxide reductase gene.

correlated with PC1, which was responsible for clustering of sites into annual, perennial and successional treatments, with annual sites showing the highest levels of nitrite reducers and GWP.

Although plotting the first two principal components accounted for much of the variance in the data, the third component significantly increased the percentage of variance accounted for. On the basis of the first three principal components for the qPCR data alone, ~98% of the variance was accounted for, whereas the number decreased to ~86% when only the annual ecosystem averages and GWP values were analyzed. The combined data set of all measured variables accounted for ~82% of the variance within the first three components, with *nosZ* gene abundance being significantly and negatively correlated (−0.85, $P \leqslant 0.05$) to the third component. Correlation values of individual variables with each principal component are summarized in Supplementary Table S3.

An alternative way of analyzing the data based on CIA showed a pattern similar to that in the PCA analysis (Appendix S4). The CIA gives an RV coefficient, a global measure of similarity between data sets based on a multivariate extension of the Pearson correlation coefficient, scaled from 0 (no

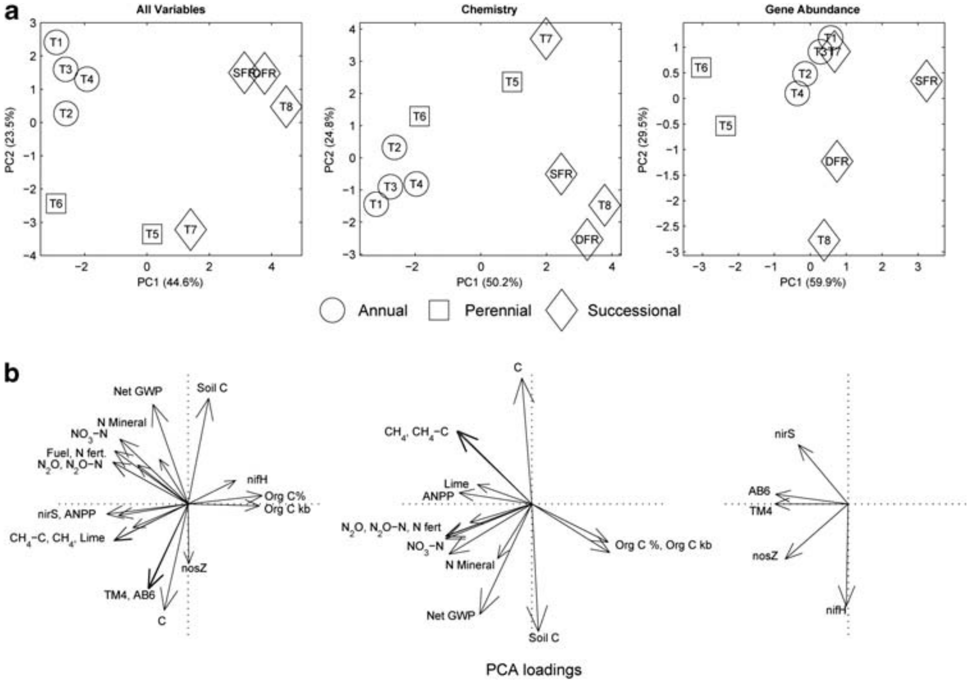


Figure 2 Principal component analysis (**a**) and factor loadings plot (**b**) for gene abundances (far right), ecosystem chemistry (center) and combined data sets (far left) summarized in Appendix S1 for different management systems at the KBS-LTER site (described in Table 1). The percentage of the variation in the samples described by the plotted principle components is indicated on the axis.

similarity) to 1.0 (identical). The genetic and chemical variables show strong similarity with an RV = 0.76. The significance of the similarity is underscored by a permutation test (Culhane *et al*., 2003), in which total co-inertia (the measure of co-variability of the two data sets) is computed after permutations to one of the two data sets (randomly chosen before each permutation). This shuffles the data, disassociating the soil sites from the genetic and chemical sample values. The test yielded 100 000 permutation-based values of total co-inertia. Under the null hypothesis that the data sets are independent, only five of these 100 000 values were as large or larger than that observed for the genetic and chemical data sets, for a *P*-value of 0.00005. The plots also provide a view of the relative strengths of relationships between genetic and chemical variables with respect to each other and to the different soil environments. Genetic and chemical variables show the strongest overall co-trend (shortest arrows) in the group of annuals (Appendix S4). This figure also indicates that, although chemical profiles in transition treatments (former agricultural sites left to undergo natural succession; namely T7 and SFR) closely resembled expected values for successional sites, genetic variables were slower to change, retaining their original signature for longer periods. This is to say, the original bacterial community signature associated to agricultural treatments is seemingly persistent (that is, apparent) after >40 years, but this cannot be unequivocally confirmed with our data as no temporal comparisons are available.

Pairwise correlations between all of the variables showed strong (>0.65) positive correlation between bacterial gene numbers and either greenhouse gas fluxes, GWP, annual productivity or carbon levels (Appendix S5). Nitrite reductase (*nirS*) gene abundance was positively correlated to greenhouse gas (N_2O-N, NO_3-N and CH_4-C) emission values, GWP values and aboveground net primary productivity, while being negatively correlated to organic carbon. Nitrogenase (*nifH*) gene abundance was positively correlated to organic carbon levels. The abundance of the *nifH* gene shared a weak negative correlation (−0.358) with the abundance of the *nirS* gene (nitrite reductase) (Appendix S5). A second denitrification gene (*nosZ*), responsible for the reduction of nitrous oxide to dinitrogen, did not exhibit the same trend as *nirS* (Figure 1). The two target genes corresponding to the numerically dominant OTUs belonging to *Thermomicrobia* and *Acidobacteria* (Morales *et al*., 2009) were the most abundant of all targets measured (Figure 1). These genes also showed a strong positive correlation (Appendix S5) across all treatments, with the highest values found in single cultivar perennial treatment plots (Figure 1).

nirS–nosZ *gene abundance as proxy for greenhouse gas (N_2O) emissions*

A simple regression analysis was conducted to compare direct measurements of nitrous oxide emissions from soils with the abundance of *nirS* gene targets minus *nosZ* gene targets (Figure 3),

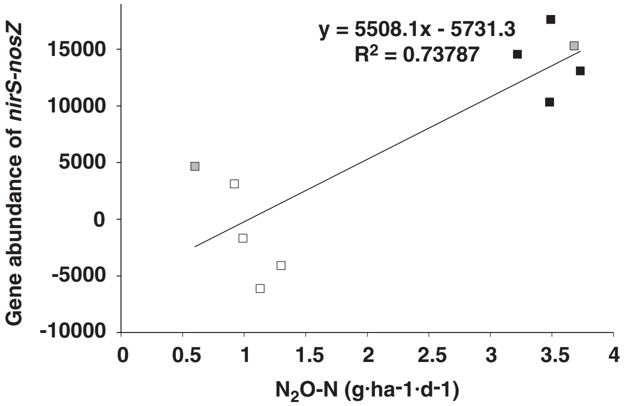


Figure 3 Simple regression analysis of nitrous oxide emission and the *nirS* gene abundance minus *nosZ* gene abundance. Annual crops (black fill); Perennial crops (gray fill); Successional sites (white fill).

which showed a strong correlation ($r^2 = 0.74$). Two discrete data clusters were observed related to direct N_2O measurements, whereas a more incremental relationship was found between site successional stage and *nirS–nosZ* gene abundance (Figure 3).

Discussion

Multivariate analysis of annual ecosystem parameter averages, global warming potential and bacterial gene abundances in the current study support the hypothesis that the legacy of agriculture (that is, resilience of treatment effects) has a stronger influence on soil biogeochemistry than current environmental parameters (that is, real-time soil conditions), as previously suggested by other studies (Buckley and Schmidt, 2001, 2003). Strong correlations were obtained that suggest a key role for bacterial activities in controlling responses between agricultural practice or land-use regimes and greenhouse gas emissions. The data also support the interpretation of long-term repercussions at the microbial community level to certain land-use practices. However, given that our study does not include a temporal sampling sequence, these interpretations are based on hypothesized ecosystem successions, as observed on successional treatment plots at the KBS-LTER site after 40–60 years of cessation of agricultural management. Those plots represent more advanced successional stages of the current agricultural treatments, and can be used as references for comparisons.

Previous research at KBS showed differences in microbial community structure between treatments (for example, indicating effects on bacterial community composition based on 16S rRNA sequences (Buckley and Schmidt, 2001, 2003); denitrifiers (Cavigelli and Robertson, 2001; Stres *et al.*, 2004); and ammonia oxidizers (Bruns *et al.*, 1999)). In addition, treatment-based differences at KBS have been reported for greenhouse gas emission and have suggested possible mitigating properties of certain treatments (Robertson *et al.*, 2000; Suwanwaree and Robertson, 2005). Related findings have also been reported for other systems, with plant species identity affecting denitrifying communities (Bremer *et al.*, 2007), and vegetation type driving separation of community structure (Chim Chan *et al.*, 2008).

We found carbon levels to be strongly correlated to the clustering of treatments based on cropping system, suggesting a strong role for carbon as a driver of bacterial community structure. Although higher emission values of the greenhouse gas nitrous oxide were positively correlated to traditional annual crop rotations, previous work has suggested that nitrate, as applied in fertilizer, does not select for denitrifiers (Tiedje, 1988). Instead, it has been proposed that denitrifiers are generally functioning as aerobic competitors for carbon, using their denitrification capabilities only under metabolically advantageous conditions (Tiedje, 1988). Although it appears that carbon has a major role in community structure in these treatments, the activity of an established community can be significantly altered by real-time events such as nitrogen deposition (Suwanwaree and Robertson, 2005). This leads us to suggest that community composition measurements (for example, DNA-based measurements of gene abundance in treatments) are good indicators of how treatment practices shape the community in the long run, whereas rRNA or mRNA measurements would be more useful to illustrate the response of the community to changing parameters in the short term (for example, diel cycles, rainfall, fertilizer application).

Although we included qPCR analysis of two predominant taxa (at approximately the sub-phylum or 'genus' level) based on 16S rRNA gene quantification to assess their ubiquity and abundance, data derived from those groups are hard to interpret in the context of biogeochemical cycling, given the lack of correlation between a specific 16S ribotype and its metabolic or catabolic capabilities. Where a direct link to a given biogeochemical reaction is desired, specific tracking of relevant functional genes is likely to be more productive. Thus, in the current study, we rely on quantification of functional genes for nitrogen cycling for correlation with process-level greenhouse gas emissions from KBS soils.

Contrary, perhaps, to common assumptions, the numbers of nitrogen fixers (as determined by *nifH* gene quantification) were found to be higher in forested or successional sites than in the agricultural fields, including those with regular soybean rotations. Although leguminous plants, which include beans, clover, alfalfa, lupine and peanuts, are among the best studied systems for nitrogen-fixing symbioses (Young and Haukka, 1996; van der Heijden *et al.*, 2006; Nandasena *et al.*, 2007; Houlton *et al.*, 2008), other non-leguminous plants including grasses

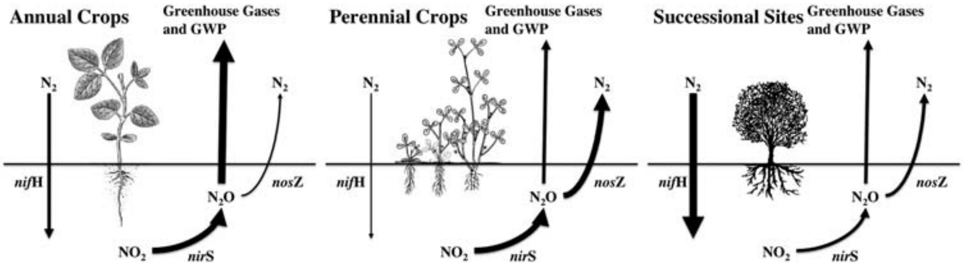


Figure 4 Generalized schematic model showing predicted changes in nitrogen flux, N-based greenhouse gas emissions and greenhouse warming potential in soils as a function of plant community differences and bacterial group abundance. Line thickness represents relative contribution of a given gene to nitrogen compound turnover rate based on data from this study and from a study by Robertson *et al.* (2000), and also summarized in Appendix S2.

(Minamisawa *et al.*, 2004; Coelho *et al.*, 2008), pine (Izumi *et al.*, 2006), wheat (Iniguez *et al.*, 2004) and alder (Ridgway *et al.*, 2004) have also been shown to posses significant populations of endophytic nitrogen-fixing bacteria. This suggests that the contribution of non-leguminous nitrogen-fixing plant symbioses and free-living nitrogen fixers is not just significant, but likely essential, in ecosystem development.

Two other nitrogen cycle-related genes encoding the denitrification enzymes nitrous oxide reductase (*nosZ*) and cytochrome cd1 nitrite reductase (*nirS*) did not exhibit similar trends in abundance across the system. Instead, what was observed was an apparent balance in the relative abundance of the two genes that can be used to predict greenhouse gas emissions and global warming potential (refer to Figures 3 and 4). To maximally relate our measurements to available metadata, we have used the same 1991–1999 data set as described by Robertson *et al.* (2000) for the KBS treatments. Whereas we note that gas emission data from 2000 to 2007 for both CH_4 and N_2O showed that emissions of these gases rose somewhat compared with the 1991–1999 timeframe (Robertson *et al.*, unpublished data (http://lter.kbs.msu.edu/datatables/28)), the relationship of these parameters between treatments has not changed. As noted above and in Supplementary Figure S6, this suggests that the same soil processes and relationships between treatments have persisted through to our sampling time, and that classification of treatments as a sink or source of greenhouse gas emissions remains the same. As the KBS is an LTER site with same conditions carefully maintained for more than two decades, it is well supported that ecosystem processes should remain comparable within and between treatments.

In the schematic model that we have developed to explain this behavior (Figure 4), the rate of greenhouse gas emissions is controlled by the interplay between different guilds within the local bacterial community. In this initial study, we focused on nitrogen cycling and show how the prevalence of bacteria involved in key steps in the cycle are related to the overall outcome in terms of key environmental parameters, and also how differences in relative abundance of individual functional groups control, or at least relate to, whether a system is a net sink or a net source of greenhouse gases.

The data presented herein showed the presence of all targeted genes, and also illustrated how soil management practices have altered the relative abundance of two predominant ribotypes and several functional genes involved in different stages of the nitrogen cycle. These were linked, or at least correlated to, measured differences in greenhouse gas emissions and global warming potential resulting from various land-use practices. Although the general roles of nitrogen fixers and denitrifiers in the nitrogen cycle have been known for decades, the application of a quantitative approach based on functional gene abundance to provide predictive power regarding the fate of nitrogenous compounds in soils is novel.

Two major observations arise from these findings. The first is that monitoring bacterial community response variables, which can exhibit greater sensitivity to environmental change than other commonly used measures (Feris *et al.*, 2009), represents a robust way to monitor geochemical dynamics**.** Second, the possibility, perhaps probability, that bacteria can respond rapidly to environmental change without major changes in community composition through altered patterns of gene expression suggests the importance of considering the contributions and responses of microbial populations in global biogeochemical cycles to such phenomena as climate change. The latter point further suggests that there is a role for both DNA- and RNA-based approaches in modern molecular microbial ecology depending on whether the investigator is looking at long-term drivers that shape community composition or at short-term response to perturbation and change, respectively.

The variability observed with our approach, as indicated by plot-to-plot replicate variability within a treatment, likely reflects an inherent property of small- or mid-scale heterogeneities in the soil environment that affect bacterial populations locally. Thorough sampling of study sites can compensate for such variability. This is readily achieved when using molecular methods by collecting multiple, small soil samples that can be pooled (composited), creating a representative sample for a site or treatment and thus its greenhouse gas production potential.

We note that our analyses focused on a single nitrite reductase gene, *nirS*. The copper-containing nitrite reductase (encoded by the *nirK* gene) was not analyzed in this study and might contribute to some of the variation not accounted for in our data. However, it has been shown that three-quarters of cultured denitrifying bacteria contain *nirS* rather than *nirK* (Zumft, 1997) and it has been found to predominate in most environments (Bothe *et al.*, 2000). It is also important to note that these experiments targeted copy numbers of genes of interest, which represent the standing community and its potential for activity rather than an actual measure of real-time gene expression levels or the corresponding enzymatic activity. Ongoing methods development to directly measure actual gene expression levels will likely enhance the resolution and accuracy of studying microbial contributions or response to key environmental functions and activities. Although our data link key gene abundance data with measured greenhouse gas potential, changes in gene expression levels could transiently change a given soil or treatment from a greenhouse gas source to a sink. Thus, making more direct molecular measurements of flux in microbial community gene expression patterns (for example, through environmental transcriptomics) is indeed highly desirable for future work.

This study provides the first quantitative assessment of the effect of land management practice on multiple microbial community constituents at both the functional and phylogenetic level. We showed that microbial assemblages do not readily return to a native or baseline community state following agricultural disturbance, consistent with previous findings that soil nutrient levels require decades or more to recover after agriculture (Robertson *et al.*, 1988, 1993; Drinkwater *et al.*, 1998; Knops and Tilman, 2000). We also present the first quantitative study illustrating interactions between different bacterial activities and their role in controlling nitrogen flux as a response to ecosystem changes.

In conclusion, we note that this initial analysis linking bacterial gene abundance data to process-level greenhouse gas emission rates represents an early step in integrating key bacterial activities to larger-scale biogeochemical cycles. Additional research in this area will extend such capabilities and allow us to assess microbial contributions and responses to ecosystem, and even to global-scale ecological phenomena such as climate change. This is particularly important as it is widely acknowledged that microorganisms govern or at least contribute to global biogeochemical cycles, yet their roles and activities are generally not even considered in current large-scale models for climate change and other global phenomena.

Acknowledgements

This project was funded by the US Department of Agriculture–National Research Initiative (USDA–CSREES Grant No. 2004–03501). Soil samples for this project were graciously provided by the Kellogg biological station long-term ecological research project. We gratefully acknowledge the anonymous reviewers and Dr RM Callaway for helpful comments leading to and during revision of this manuscript.

References

Blackwood CB, Paul EA. (2003). Eubacterial community structure and population size within the soil light fraction, rhizosphere, and heavy fraction of several agricultural systems. *Soil Biol Biochem* **35**: 1245–1255.

Bothe H, Jost G, Schloter M, Ward BB, Witzel K. (2000). Molecular analysis of ammonia oxidation and denitrification in natural environments. *FEMS Microbiol Rev* **24**: 673.

Bremer C, Braker G, Matthies D, Reuter A, Engels C, Conrad R. (2007). Impact of plant functional group, plant species, and sampling time on the composition of nirK-type denitrifier communities in soil. *Appl Environ Microbiol* **73**: 6876–6884.

Broughton LC, Gross KL. (2000). Patterns of diversity in plant and soil microbial communities along a productivity gradient in a Michigan old-field. *Oecologia* **125**: 420–427.

Bruns MA, Fries MR, Tiedje JM, Paul EA. (1998). Functional gene hybridization patterns of terrestrial ammonia-oxidizing bacteria. *Microb Ecol* **36**: 293–302.

Bruns MA, Stephen JR, Kowalchuk GA, Prosser JI, Paul EA. (1999). Comparative diversity of ammonia-oxidizer 16S rRNA gene sequences in native, tilled, and successional soils. *Appl Environ Microbiol* **65**: 2994–3000.

Buckley DH, Schmidt TM. (2001). The structure of microbial communities in soil and the lasting impact of cultivation. *Microb Ecol* **42**: 11–21.

Buckley DH, Schmidt TM. (2003). Diversity and dynamics of microbial communities in soils from agro-ecosystems. *Environ Microbiol* **5**: 441–452.

Cavigelli MA, Robertson GP. (2001). Role of denitrifier diversity in rates of nitrous oxide consumption in a terrestrial ecosystem. *Soil Biol Biochem* **33**: 297–310.

Chim Chan O, Casper P, Sha LQ, Feng ZL, Fu Y, Yang XD *et al.* (2008). Vegetation cover of forest, shrub and pasture strongly influences soil bacterial community structure as revealed by 16S rRNA gene T-RFLP analysis. *FEMS Microbiol Ecol* **64**: 449–458.

Culhane AC, Perrière G, Higgins DG. (2003). Cross-platform comparison and visualisation of gene expression data using co-inertia analysis. *BMC Bioinformatics* **4**: 59.

Culhane AC, Thioulouse J, Perriere G, Higgins DG. (2005). MADE4: an R package for multivariate analysis of gene expression data. *Bioinformatics* **21**: 2789–2790.

Dray S, Chessel D, Thioulouse J. (2003). Co-inertia analysis and the linking of ecological data tables. *Ecology* **84**: 3078–3089.

Drinkwater LE, Wagoner P, Sarrantonio M. (1998). Legume-based cropping systems have reduced carbon and nitrogen losses. *Nature* **396**: 262–265.

EIA (2008). Emissions of Greenhouse Gases in the United States 2007 report number DOE/EIA-0573 (2007). In Corti J and Sweetnam GE (eds).

Feris KP, Ramsey PW, Gibbons SM, Frazar C, Rillig MC, Moore JN *et al.* (2009). Hyporheic microbial community development is a sensitive indicator of metal contamination. *Environ Sci Technol* **43**: 6158–6163.

Henry S, Bru D, Stres B, Hallet S, Philippot L. (2006). Quantitative detection of the nosZ gene, encoding nitrous oxide reductase, and comparison of the abundances of 16 s rRNA, narG, nirK, and nosZ genes in soils. *Appl Environ Microbiol* **72**: 5181–5189.

Henry S, Texier S, Hallet S, Bru D, Dambreville C, Cheneby D *et al.* (2008). Disentangling the rhizosphere effect on nitrate reducers and denitrifiers: insight into the role of root exudates. *Environ Microbiol* **10**: 3082–3092.

Houlton BZ, Wang YP, Vitousek PM, Field CB. (2008). A unifying framework for dinitrogen fixation in the terrestrial biosphere. *Nature* **454**: 327–330.

Iniguez AL, Dong Y, Triplett EW. (2004). Nitrogen fixation in wheat provided by Klebsiella pneumoniae 342. *Mol Plant Microbe Interact* **17**: 1078–1085.

Izumi H, Anderson IC, Alexander IJ, Killham K, Moore ERB. (2006). Diversity and expression of nitrogenase genes (nifH) from ectomycorrhizas of Corsican pine (*Pinus nigra*). *Environ Microbiol* **8**: 2224–2230.

Knops JMH, Tilman D. (2000). Dynamics of soil nitrogen and carbon accumulation for 61 years after agricultural abandonment. *Ecology* **81**: 88–98.

Leininger S, Urich T, Schloter M, Schwark L, Qi J, Nicol GW *et al.* (2006). Archaea predominate among ammonia-oxidizing prokaryotes in soils. *Nature* **442**: 806–809.

Minamisawa K, Nishioka K, Miyaki T, Ye B, Miyamoto T, You M *et al.* (2004). Anaerobic nitrogen-fixing consortia consisting of Clostridia isolated from gramineous plants. *Appl Environ Microbiol* **70**: 3096–3102.

Morales SE, Cosart TF, Johnson JV, Holben WE. (2009). Extensive phylogenetic analysis of a soil bacterial community illustrates extreme taxon evenness and the effects of amplicon length, degree of coverage and DNA fractionation on classification and ecological parameters. *Appl Environ Microbiol* **75**: 668–675.

Morales SE, Holben WE. (2009). Empirical testing of 16S rRNA gene PCR primer pairs reveals variance in target specificity and efficacy not suggested by *in silico* analysis. *Appl Environ Microbiol* **75**: 2677–2683.

Nandasena KG, O'Hara GW, Tiwari RP, Sezmis E, Howieson JG. (2007). *In situ* lateral transfer of symbiosis islands results in rapid evolution of diverse competitive strains of mesorhizobia suboptimal in symbiotic nitrogen fixation on the pasture legume Biserrula pelecinus L. *Environ Microbiol* **9**: 2496–2511.

Phillips CJ, Harris D, Dollhopf SL, Gross KL, Prosser JI, Paul EA. (2000a). Effects of agronomic treatments on the structure and function of ammonia oxidizing communities. *Appl Environ Microbiol* **66**: 5410–5418.

Phillips CJ, Paul EA, Prosser JI. (2000b). Quantitative analysis of ammonia oxidising bacteria using competitive PCR. FEMS Microbial Ecology. *FEMS Microbiol Ecol* **32**: 167–175.

Ridgway KP, Marland LA, Harrison AF, Wright J, Young JPW, Fitter AH. (2004). Molecular diversity of Frankia in root nodules of Alnus incana grown with inoculum from polluted urban soils. *FEMS Microbiol Ecol* **50**: 255–263.

Robertson G, Grace P. (2004). Greenhouse gas fluxes in tropical and temperate agriculture: the need for a full-cost accounting of global warming potentials. *Environ Dev Sustain* **6**: 51–63.

Robertson GP, Crum JR, Ellis BG. (1993). The spatial variability of soil resources following long-term disturbance. *Oecologia* **96**: 451–456.

Robertson GP, Huston MA, Evans FC, Tiedje JM. (1988). Spatial variability in a successional plant community: patterns of nitrogen availability. *Ecology* **69**: 1517–1524.

Robertson GP, Paul EA, Harwood RR. (2000). Greenhouse gases in intensive agriculture: contributions of individual gases to the radiative forcing of the atmosphere. *Science* **289**: 1922–1925.

Coelho MR, de Vos M, Carneiro NP, Marriel IE, Paiva E, Seldin L. (2008). Diversity of *nif*H gene pools in the rhizosphere of two cultivars of sorghum (*Sorghum bicolor*) treated with contrasting levels of nitrogen fertilizer. *FEMS Microbiol Lett* **279**: 15–22.

Rösch C, Bothe H. (2005). Improved assessment of denitrifying, N2-fixing, and total-community bacteria by terminal restriction fragment length polymorphism analysis using multiple restriction enzymes. *Appl Environm Microbiol* **71**: 2026–2035.

Staley JT, Reysenbach AL. (2002). *Biodiversity of Microbial Life: Foundation of Earth's Biosphere*. John Wiley & Sons, Inc.: New York.

Stres B, Mahne I, Avgustin G, Tiedje JM. (2004). Nitrous oxide reductase (nos Z) gene fragments differ between native and cultivated Michigan soils. *Appl Environ Microbiol* **70**: 301–309.

Suwanwaree P, Robertson GP. (2005). Methane oxidation in forest, successional, and no-till agricultural ecosystems effects of nitrogen and soil disturbance. *Soil Sci Soc Am J* **69**: 1722–1729.

Throback IN, Enwall K, Jarvis A, Hallin S. (2004). Reassessing PCR primers targeting nirS, nirK and nosZ genes for community surveys of denitrifying bacteria with DGGE. *FEMS Microbiol Ecol* **49**: 401–417.

Tiedje JM. (1988). Ecology of denitrification and dissimilatory nitrate reduction to ammonium. *Biol Anaerob Microorg* **179**: 244.

Tilman D, Cassman KG, Matson PA, Naylor R, Polasky S. (2002). Agricultural sustainability and intensive production practices. *Nature* **418**: 671–677.

van der Heijden MGA, Bakker R, Verwaal J, Scheublin TR, Rutten M, van Logtestijn R *et al.* (2006). Symbiotic bacteria as a determinant of plant community structure and plant productivity in dune grassland. *FEMS Microbiol Ecol* **56**: 178–187.

Wakelin SA, Macdonald LM, Rogers SL, Gregg AL, Bolger TP, Baldock JA. (2008). Habitat selective factors influencing the structural composition and functional capacity of microbial communities in agricultural soils. *Soil Biol Biochem* **40**: 803–813.

Yergeau E, Kang S, He Z, Zhou J, Kowalchuk GA. (2007). Functional microarray analysis of nitrogen and carbon cycling genes across an Antarctic latitudinal transect. *ISME J* **1**: 163–179.

Young JPW, Haukka KE. (1996). Diversity and phylogeny of Rhizobia. *New Phytol* **133**: 87–94.

Zumft WG. (1997). Cell biology and molecular basis of denitrification. *Microbiol Mol Biol Rev* **61**: 533–616.

Supplementary Information accompanies the paper on The ISME Journal website (http://www.nature.com/ismej)