The Cow Diversity Project: introducing undergraduates to molecular population genetics.
R. Kliman, A. Walther, D. Walther

Notes for Instructors

I. Equipment and consumables
II. Considerations
III. Things to watch out for
IV. Sample data with divergence and diversity calculations

I. Equipment and consumables

The following lists assume lab sections of 16 students. The listing of particular kits does not imply endorsement over other products. There are many alternatives. However, we do use these kits, and they are explicitly referenced in the handout provided to students.

A. Equipment required:
- Balance with resolution to 1 mg (week 1)
- Micropipettors
  - P10 (week 3)
  - P20, P200, P1000 (weeks 1, 2, 3)
- Heat blocks or water baths with wells that can accommodate 1.5 mL microcentrifuge tubes (week 1)
  - 55°C for overnight Proteinase K digests
  - 70°C for DNA purification
  - 37°C for λ-*Hin*dII digest
  - 75°C for λ-*Hin*dII digest
- High-speed microcentrifuges that can accommodate 1.5 mL microcentrifuge tubes for Qiagen DNeasy and QIAquick kits (weeks 1, 3)
- Vortex mixer (week 1)
- Thermal cycler with sufficient capacity to support 2-4 samples per student (week 2; week 3 if DNA sequencing is being performed in house)
- Equipment for agarose gel electrophoresis (weeks 2, 3)
  - allow for 2 wells per student per section in week 2
  - allow for 2 wells per student (plus 2 wells/gel for size standards) per section in week 3
- Gel photodocumentation system (weeks 2, 3; printouts needed for week 4)
- UV Spectrophotometer capable of small volume measurements (*e.g.*, Nanodrop™)

B. Recommended additional durable materials
- freezer boxes for samples, stocks, aliquoted consumables
- racks for 1.5 mL microcentrifuge tubes (weeks 1, 2, 3)
- racks for 0.2 mL PCR tubes (weeks 2, 3)
- beakers for liquid waste (weeks 1, 3)
- disposal for used tubes and pipet tips (no biohazards)

• sharps disposal for razor blades (week 1)

C. Consumables:
• Pipet tips for all pipettors
• 1.5 mL microcentrifuge tubes (weeks 1, 2, 3)
per student: 1 for sample
4 for week 1

additional tubes needed for PCR and DNA sequencing master mixes
• 0.2 mL thin-walled PCR tubes (weeks 2, 3)
per student: 2 for week 2 (PCR)
per student: 4 for week 3 (DNA sequencing)
note: we sequence 8 PCR products for each gene in each lab section; thus, we need 32 0.2 mL tubes per lab section in week 3
• Aluminum foil (week 1)
• Weigh paper or small weigh boats (week 1)
• Razor blades, 1 per student (week 1)
• Pellet pestles for 1.5 mL tubes, 1 per student (week 1)
• Qiagen DNeasy™ kits sufficient for each student to make one genomic prep (week 1)
• Proteinase K, 20 mg/mL, 20 μL per student (week 1)
note: a sufficient quantity of Proteinase K may come with the DNA purification kit; a spearate purchase may not be necessary
• 95% EtOH suitable for molecular biology, 200 μL per student (week 1)
• 6× Stop/Load dye, ~10 μL per student (weeks 1, 2, 3)
• "ultrapure" H_2O, ~ 5 mL per section (weeks 1, 2, 3)
• phage λ DNA*, 500 μg/mL, 1.5 μL per student (week 1)
• *Hin*dIII*, 20,000 U/mL (plus associated rx buffer), 2 μL per student (week 1)
• *Taq* polymerase*, 5,000 U/mL (and associated rx buffer), 8 μL per section (week 2)
• dNTPs*, stock solution of 2.5 mM each dNTP, 160 μL per section (week 2)

* In recent years we have purchased these reagents from New England Biolabs. We are not advocating these products over others. However, the protocols in our handout assume the concentrations associated with these products. This is *especially* important for *Taq* polymerase, as NEB OneTaq uses a 2× buffer.

• PCR and sequencing primers (weeks 2, 3)

ND3 forward	5' GGTGCCTGATACTGACATTTCGTAG
ND3 reverse	5' GCATAGAAGGGAGGATATTAGGTGG
D-loop forward	5' CTAACATAACACGCCCATACACAGAC
D-loop reverse	5' GCATCCCCCAAAATAAAAAAATACC

We typically resuspend primers into 500 μM stocks in TE buffer for long-term storage at -20°C. Alternatively, primers can be resuspended into 100 μM

stocks in TE or a buffer with only 0.1 mM EDTA (a "primer dilution buffer"); In our experience, it is fine to use standard TE.

For PCR, we prepare fresh 10 μM PCR primer stocks each year by diluting 100 μM stocks with ultrapure H_2O
- 40 μL of each primer required for each section

For in-house DNA sequencing, we prepare fresh 2.5 μM Sequencing stocks each year by diluting 10 μM PCR stocks with ultrapure H_2O
- 9 μL of each primer required for each section

Note: we use the Beckman DTCS Kit with a Beckman CEQ capillary sequencer. All quantities and concentrations reflect the DTCS chemistry.

- Agarose, quantity varies depending on electrophoresis units (weeks 2, 3)
- TAE buffer, quantity varies depending on electrophoresis units (weeks 2, 3)
- Ethidium bromide for gel staining
  *Note: alternatives such as SYBR® Safe (Invitrogen) may work, but sensitivity should be tested, especially if a λ-Hin*d*III digest is used as a size standard; the 560 bp fragment may be too faint.*
- Qiagen QIAQuick kits sufficient for 2 PCRs per student (week 3)
- Kits for in-house DNA sequencing, sufficient for 32 reactions per section (week 3)
  Note: this is based on our protocol. Each PCR to be sequenced will require 2 sequencing reactions for bidirectional sequencing.
- Additional consumables for in-house DNA sequencing

II. Considerations

A. Samples.

Students should be given 1.5 mL microcentrifuge tubes to store their beef samples. Only 25 mg of tissue is needed for the prep. We explain to students that meat and water have similar density, and that a 1.5 mL tube could hold ~1500 mg of meat; the conical part of the tube would hold ~500 mg. Students always overestimate the amount of meat needed for 15 mg.

After adding the sample, the tube can be filled with isopropyl ("rubbing") alcohol obtained from home. Although ideally it would be kept cold, and covered with alcohol, the procedure appears to be robust to fairly poor sample handling practices. Nevertheless, it is worth reminding students that poor sampling will compromise the study.

In our experience, if students do not obtain samples from a broad geographic distribution, DNA sequence diversity is reduced – sometimes radically. This is especially likely to occur if students obtain their samples from the same source (*e.g.*, campus dining services, a nearby supermarket) on the same day. Our strategy to avoid under-sampling diversity is to

ask to students to obtain students while they are away on a major break that precedes the start of the exercise: either winter break between fall and spring semesters, or a mid-semester break.

B. Overnight Proteinase K digests in week 1.

We have found that most students can follow the instructions for setting up the overnight digests. We set aside a work station for this, with instructions printed in large font and taped to the wall at the work station. For students with less confidence or experience, it is probably wise to arrange for someone to be available at a predetermined time to assist them. The preparation requires the use of micropipettors, so this skill needs to have been covered in in prerequisite course or earlier in the current course. If it has not yet been covered that semester, it is helpful to quickly review micropipettor use the week before starting the lab exercise.

C. Week 1 time management.

We start with the λ-*Hin*dIII digest, completing the DNA purification during the 37° incubation. [*Note: the length of the λ-*Hin*dIII digest can be extended up to 2 hours if needed to accommodate the lab day schedule.*] This should leave time (in a 3-hour lab session) to discuss the overall aims of the project and to lay out the 4-week plan.

Because it may be hard to keep students synchronized, it is worth having at least 2 microcentrifuges available.

D. Week 2 time management and gel pouring.

The bench work in week 2 is fairly minimal and can be used for discussion or exercises that reinforce concepts such as restriction digestions, PCR, and gel electrophoresis. We have also used part of this lab session for other activities (*e.g.*, oral presentations for a different exercise).

If students prepare the gels (Step B.1), it is important to gauge their experience. If a student has not done this before, **it is critical to ensure that the student is not burned when testing whether the gel is "sufficiently cool" to pour the gel.** A rule-of-thumb is that the gel can be poured if the flask can be touched to the inner forearm without feeling too hot. That means that the instructor has to ensure that a student does not perform this test if the flask is too hot.

It is worth noting that even DNA preps that seem to have failed often yield fine PCRs. DNA that is highly sheared and at a low concentration may be hard to see on a gel, but still provide plenty of mtDNA template for the PCRs to work. Thus, we always have every student do the PCRs, regardless of the appearance of their DNA prep.

E. Week 3 time management.

This is a busy week and requires good time management. We recommend having gels poured ahead of the start time for lab, in order to immediately begin electrophoresis of the PCR samples. Once electrophoresis is under way, the PCR cleanups can be performed for all students, although it may turn out that some of them had unsuccessful PCRs. As soon as the cleanups are done, the cleaned PCRs can be assessed by UV spectrophotometry.

For the Beckman DTCS kit (half reactions), we use 0.5 μL of cleaned PCR that has been diluted to 10-20 ng/μL. The sequencing reactions are fairly sensitive to template concentration. As we perform the spectrophotometry, we dilute samples as necessary to land in this range. This can be a little time-consuming, and it helps to have someone keep the students on task. By the end of the session, cycle sequencing reactions must be in the thermal cycler.

F. Selecting PCRs for sequencing in week 3.

In our experience, if an undiluted PCR product is below 10 ng/μL, it does not yield good DNA sequences. The PCRs generally require diluting, so a PCR that falls below the 10-20 ng/μL range is probably a low-quality PCR.

When sequencing is done in-house, we generally sequence 8 PCRs per locus per section, which often means having to decide whose PCRs to use. However, when sending samples out for commercial sequencing, more samples can be sequenced, contingent on budget constraints. We consider the following criteria:

- If possible, every student should have at least one PCR sequenced, even if this means setting aside some superior PCRs.
- A PCR is excluded if more than one band appeared on the gel, the band is absent, or the band indicates a product of incorrect size. [This is rare.]
- Cleaned PCRs in the 10-20 ng/μL range are favored.
- Cleaned PCRs with 260/280 ratios of ~1.8 are favored.
- Cleaned PCRs with 260/230 ratios of ~1.8 are favored (although 260/280 ratios take precedence).

G. In-house sequencing vs. outsourced sequencing.

If a Sanger sequencer is available in-house, then students can set up their sequencing reactions during lab and the samples can be prepared for analysis on the instrument by the instructor outside of lab time. Depending on the type of sequencer, there may be considerable preparation required before samples can be loaded; using our Beckman sequencer, the cleanups for three lab sections requires about four hours of work.

Alternatively, sequencing can be done by a core facility or off site through a commercial vendor. In such circumstances, the cleaned-up PCR samples are combined with the appropriate amounts of primer (check with sequencing facility for specific requirements) and then shipped for sequencing. In the case of outsourced sequencing, the lab may need

to include a "gap" week between the preparation of sequencing samples and week dedicated to the analysis of variation, to allow for the return of the sequencing results. This gap week can be used for the Linear regression exercise, reinforcement of concepts from the lab, or primary literature discussion; or it can be dedicated to a different lab exercise.

H. Curating the data set.

IMPORTANT NOTE: This may take some time. There are essentially 4 steps, which may (or may not) be accomplished concurrently (depending on software):

1. Assessing sequence quality by examining trace (chromatogram) files.
2. Comparing complementary sequences of each sample, to ensure accuracy.
3. Creating multiple sequence alignments.
4. Exporting multiple sequence alignments in an interleaved format that allows students to analyze variation. [See examples at the end of this document.]

As a rule, we exclude any PCR that has not been successfully sequenced on both strands. For *ND3*, we analyze every coding base (345 bp). For D-loop, we analyze the ~370 bp stretch that begins with `AGTACATTAAATTAT` and ends with `AGCCCATGCTCACACATAA`. Allowing PCRs that weren't cleanly sequenced on both strands could upwardly bias estimates of diversity and divergence.

We have been using Sequencher (GeneCodes, Inc.) to edit base calls and manually align sequences. There are alternatives (*e.g.* Tom Hall's BioEdit software allows a user to view and edit standard chromatogram files, build multiple sequence libraries, and align sequences manually or using algorithsm such as ClustalW; www.mbio.ncsu.edu/BioEdit/bioedit.html). Kumar et al's MEGA software (www.megasoftware.net) also provides numerous tools for aligning and editing sequences, and should be able to open ABI trace files. MEGA allows the user to export the alignment in an interleaved format.

I. Linear regression in week 4.

We use this lab as an opportunity to teach students how to perform linear regression. For a 0.6% agarose gel, most of the λ-*Hin*dIII fragments between 500 and 10,000 bp (*i.e.,* the six fragments aside from the largest and smallest) run on the gel at a rate approximately inversely proportion to the log of their length. There is a slight curve, but this does not have much impact on the results. ***Note: the exact sizes of these fragments are provided in the worksheet (see Handout) that students will use to perform linear regression calculations. It is imperative that the top fragment (~23 kb) be skipped, and that only the six remaining visible fragment be used. The 560 bp fragment may be faint, but it should be visible.***

Opinions will differ regarding which variable ($\log_{10}$ of fragment length vs. distance traveled) to place on the *x*-axis and which to place on the *y*-axis. While ultimately we would be using distance traveled to predict PCR product size, distance traveled clearly depends on

the size of the molecule and not *vice versa*. Thus, with second-year students, we reiterate the traditional placement of the explanatory variable (size) on the *x*-axis and the response variable (distance traveled) on the *y*-axis. However, reversing these is defensible and, strictly speaking, perhaps more appropriate from a statistical perspective. [*Note: we use "explanatory" and "response" to clearly denote cause and effect; these correspond to the traditional usage of "independent" and "dependent," respectively.*]

Regardless, we first have students graph the six points on graph paper and draw an eye-fitted line through the points. They then calculate the quantities necessary to calculate the slope of the regression. With *x* being the $\log_{10}$ of length in base pairs of the 6 suitable λ-*Hin*dIII digest fragments, and *y* being the distance traveled on the gel (actually, as measured on a photograph), they calculate the mean of *x*, the mean of *y*, the variance of *x*, and the covariance of *x* and *y*. Once they have the slope ($b = s^2_x$ / COV_{xy}), they can calculate the *y*-intercept, as the regression line runs through the point $(\bar{x}, \bar{y})$.

Once they have the formula for the line ($y = bx + c$), we have them confirm it by calculating *y* values for a pair of arbitrary *x* values, and drawing the line that connects these two points. If they have done their arithmetic correctly, the regression line should be pretty close to their eye-fitted line. If it is not, they have probably made an arithmetic error.

Once they have confirmed that they have a good formula for the line, they can rearrange the formula to $x = (y - c)/b$. They can then measure the distances traveled by their PCR products on the same gel, and estimate the $\log_{10}$ of their lengths from the formula. From here, they take the antilogarithm (*i.e.*, 10^x) of the estimate. [Unless $r^2 = 1.0$ for the line, the answer obtained this way would be different from the answer obtained by reversing the regression. However, the points generally fall on a good line, and the answers would not differ by much.] Students should find that the estimated sizes of the PCR products are fairly close to those predicted by determining the PCR product size based on the placement of the primers in the cow reference sequence (provided in their handout). The actual PCR sizes should be 654 bp for the *ND3* PCR and 506 bp for the D-loop PCR.

J. Analysis of diversity and divergence in week 4.

Generally, we provide printouts of multiple sequence alignments in <u>interleaved</u> format, highlighting variable sites:

`Non-interleaved format:`
`Seq1 AAGACATTAGCCTGA...`
`Seq2 AAGACATTAGCCTGA...`
`Seq3 AAGACATTAGCCTGA...`
	`Interleaved format`
	`Seq1 AAGACATTAG`
	`Seq2 AAGACATTAG`
	`Seq3 AAGACATTAG`
	
	`Seq1 CCTGA...`
	`Seq2 CCTGA...`
	`Seq3 CCTGA...`

Examples of two useful formats are provided at the end of this document. For *ND3*, we include a line that indicates the reading frame. Remember to use the vertebrate mitochondrial genetic code. **Again, it is important to note that producing the handouts of multiple sequence alignments can be time-consuming if you don't have software that exports files in a suitable format. This is definitely not a last-minute task!**

For both diversity and divergence, we use measures that are scaled by sequence length. We have yet to observe length variation in *ND3*. We do not observe length variation in D-loop among cows (although this is certainly possible with sufficient sampling), but there are a handful of alignment gaps for bison vs. cow. While everyone is entitled to their preferred alignment, we use the following D-loop alignment (using *Bison bison* GenBank accession U12864, positions 202-561):

```
bison   AGTACATTAAATTATATGCCCCATGCATATAAGCAAGTACTTATCCTC
cow     AGTACATTAAATTATATGCCCCATGCATATAAGCAAGTACATGACCTC

bison   TATTGACAGTACATAGTACATAAAGTTATTAATTGTACATAGCACATT
cow     TATAG-CAGTACATAATACATATAATTATTGACTGTACATAGTACATT

bison   ATGTCAAATCTACCCTTG---GCA-A-C-A-TGCATACCCCTT-CCAT
cow     ATGTCAAATTCATTCTTGATAGTATATCTATTATATATTCCTTACCAT

bison   TAGATCACGAGCTTAATTACCATGCCGCGTGAAACCAGCAACCCGCTA
cow     TAGATCACGAGCTTAATTACCATGCCGCGTGAAACCAGCAACCCGCTA

bison   GGCAGAGGATCCCTCTTCTCGCTCCGGGCCCATGAACCGTGGGGGTCG
cow     GGCAG-GGATCCCTCTTCTCGCTCCGGGCCCATAAACCGTGGGGGTCG

bison   CTATTTAATGAACTTTATCAGACATCTGGTTCTTTCTTCGGGGCCATC
cow     CTATCCAATGAATTTTACCAGGCATCTGGTTCTTTCTTCAGGGCCATC

bison   TCACCTAAAATCGCCCATTCTTTCCTCTTAAATAAGACATCTCGATGG
cow     TCATCTAAAACAGTCCATTCTTTCCTCTTAAATAAGACATCTCGATGG

bison   ACTAATGGCTAATCAGCC-ATGCTCACACATAA
cow     ACTAATGGCTAATCAGCCCATGCTCACACATAA
```

This alignment is 369 bp in length. Gaps in bison to align with cow are shown in red (total of 9 bp); gaps in cow to align with bison are shown in blue (total of 2 bp). Thus, when calculating divergence, we use only the 358 bp shared by bison and cow. When calculating diversity, we use only the 367 bp shared by all cows. [We have not yet observed length polymorphism among cows.]

The simplest way to calculate divergence of cows from bison is as follows:
Step 1: Identify every site where at least one cow differs from bison.

Step 2: At each of these sites, count the number of cows that differ from bison. Sum these across all sites; this is the numerator of the divergence calculation.
Step 3: If k cows were sequenced, there have been $358k$ pairwise contrasts of a cow base to a bison base. This is the denominator of the divergence calculation. For example, if $k = 12$, this would be 4,296.
Step 4: Divide the numerator by the denominator. This gives the average proportion of pairwise contrasts that differ when comparing a cow to a bison.

The simplest way to calculate diversity among cows is as follows:
Step 1: Identify every site where the cows are not all the same (ignoring bison).
Step 2: At each site, add up the number of individuals having each base. [Usually, there will be only two bases segregating at a site, but there could be more.] If there are only two bases segregating, multiply the number with one base by the number with the other base. For example, if in a sample of 12 cows, 9 have A's and 3 have G's, multiply $9 \times 3 = 27$. On the other hand, if there are 8 A's, 3 G's, and 1 C, you have to calculate and sum three products: $8 \times 3 + 8 \times 1 + 3 \times 1 = 24 + 8 + 3 = 35$. In the extremely rare circumstance that all four bases are segregating, there would be six products to sum ($n_A \times n_G + n_A \times n_C + n_A \times n_T + n_G \times n_C + n_G \times n_T + n_C \times n_T$). Sum these values across all sites; this is the numerator of the diversity calculation.
Step 3. Calculate the number of pairwise comparisons for the k cows. Assuming that 367 bases are aligned among cows, this is $367 \times k \times (k\text{-}1) / 2$. This is the denominator of the diversity calculation. For example, if $k = 12$, this would be 24,222.
Step 4. Divide the numerator by the denominator. This gives the average proportion of pairwise contrasts that differ when comparing one cow to another cow.

There are, of course, alternative approaches to estimate diversity and divergence. However, our students have not completed an upper-level population genetics course, and alternative methods – which often rely on the coalescent – would be "black boxes" to the students. Average pairwise variation is intuitive.

Examples of divergence and diversity calculations are provided for the sample data sets provided in Section IV of this document.

III. Things to watch out for

A. Steps that students regularly mess up.
- Students may add stop/load dye to their purified DNA or PCRs. We have found that it is critical to remind students to never do this; they should only use the stop/load dye for the samples that will be run on a gel. Addition of stop/load dye to either the purified DNA or a PCR will ruin it.
- Inexperienced students, who have not developed the ability to visually recognize volumes, often use the wrong micropipettor. We observe this most frequently with P20s and P200s.

- Students sometimes get confused during DNA and PCR cleanups, accidentally throwing out the cleaned product or saving the wrong liquid when using Qiagen spin kits.

B. Concepts that require reinforcement.
- How PCR and dideoxy (Sanger) sequencing work
- Pairwise variation
- Alignments and alignment gaps
- Synonymous vs. nonsynonymous variation

IV. Sample data with divergence and diversity calculations

We have used a variety of formats, some with more success than others, to present students with the aligned data for analysis. Below, we show two formats that have worked pretty well. For *ND3*, we provide the reading frame in a line above the sequence data. In both alignments, bison sequence is listed first; all other sequences are from cows.

Alignments using the identity symbol ('.')

```
bison     AGTACATTAA ATTATATGCC CCATGCATAT AAGCAAGTAC TTATCCTCTA
TAM_BNF   AGTACATTAA ACTATATGCC CCATGCATAT AAGCAAGTAC ATGACCTCTA
TAM_CBG   AGTACATTAA ATTATATGCC CCATGCATAT AAGCAAGTAC ATGACCTCTA
TAM_CML   AGTACATTAA ATTATATGCC CCATGCATAT AAGCAAGTAC ATGACCTCTA
TAM_DMC   AGTACATTAA ATTATATACC CCATGCATAT AAGCAAGTAC ATGACCTCTA
TAM_KMC   AGTACATTAA ATTATATGCC CCATGCATAT AAGCAAGTAC ATGACCTCTA
```

If an individual uses the same base as the first individual (*i.e.*, the "reference"), the identity symbol ('.') is used in place of the letter representing the base:

```
bison     AGTACATTAAATTATATGCCCCATGCATATAAGCAAGTACTTATCCTCTA
TAM_BNF   ...........C.............................A.GA......
TAM_CBG   ....................................... A.GA......
TAM_CML   ....................................... A.GA......
TAM_DMC   ................A.....................A.GA......
TAM_KMC   ....................................... A.GA......
```

Cow Mitochondrial Genetic Code (DNA)

The vertebrate mitochondrial genetic code differs slightly from the "universal" genetic code. Specifically, ATA encodes methionine (rather than isoleucine), TGA encodes tryptophan (rather than a stop codon), and AGA/AGG encode a stop codon (rather than arginine),

```
AAA lys K      AGA stop *     ACA thr T      ATA met/start M
AAG lys K      AGG stop *     ACG thr T      ATG met/start M
AAC asn N      AGC ser S      ACC thr T      ATC ile I
AAT asn N      AGT ser S      ACT thr T      ATT ile I
GAA glu E      GGA gly G      GCA ala A      GTA val V
GAG glu E      GGG gly G      GCG ala A      GTG val V
GAC asp D      GGC gly G      GCC ala A      GTC val V
GAT asp D      GGT gly G      GCT ala A      GTT val V
CAA gln Q      CGA arg R      CCA pro P      CTA leu L
CAG gln Q      CGG arg R      CCG pro P      CTG leu L
CAC his H      CGC arg R      CCC pro P      CTC leu L
CAT his H      CGT arg R      CCT pro P      CTT leu L
TAA stop *     TGA trp W      TCA ser S      TTA leu L
TAG stop *     TGG trp W      TCG ser S      TTG leu L
TAC tyr Y      TGC cys C      TCC ser S      TTC phe F
TAT tyr Y      TGT cys C      TCT ser S      TTT phe F
```

ND3 Alignment

```
            123123123123123123123123123123123123123123123123123123123123
bison_nd3   ATAAATCTAATACTAGCCCTCCTGACCAATTTTACACTAGCTACCCTACTTGTCATTATC 60
TH14JB      ......T.....................................C.........C.....C...
TH14AF      ......T.....................................C.........C.....C...
TH14KMB     ......T.....................................C.........C.....C...
TH14ER      ......T.....................................C.........C.....C...
T811CF      ......T.....................................C.........C.....C...
T811RH      ......T.....................................C.........C.....C...
T811MCL     ......T.....................................C.........C.....C...
T811KR      ......T.....................................C.........C.....C...
T811CL      ......T.....................................C.........C.....C...
T811ME      ......T.....................................C.........C.....C...
T811AE      ......T.....................................C.........C.....C...
TH14KLB     ......T.....................................C.........C.....C...
T14SB       ......T.....................................C.........C.....C...
T14SC       ......T.....................................C.........C.....C...
T811KZ      ......T.....................................C.........C.....C...

            123123123123123123123123123123123123123123123123123123123123
            GCATTCTGACTTCCCCAATTAAATGTATATTCTGAGAAGACAAGCCCATACGAATGTGGA 120
            ...................C..........C.........A......................
            ...................C..........C.........A......................
            ...................C..........C.........A......................
            ...................C..........C.........A......................
            ...................C..........C.........A......................
            ...................C..........C.........A......................
            ...................C..........C.........A......................
            ...................C..........C.........A......................
            ...................C..........C.........A......................
            ...................C..........C.........A......................
            ...................C..........C.........A......................
            ...................C..........C.........A......................
            ...................C..........C.........A......................
            ...................C..........C.........A......................
            ...................C..........C.........A......................

            123123123123123123123123123123123123123123123123123123123123
            TTTGACCCCATAGGATCAGCCCGCCTTCCCTTCTCCATAAAATTCTTTCTAGTAGCCATC 180
            ....................................T..............G.........
            ....................................T..............G.........
            ....................................T..............G.........
            ....................................T..............G.........
            ....................................T..............G.........
            ....................................T..............G.........
            ....................................T..............G.......C.
            ....................................T..............G.........
            ....................................T..............G.........
            ....................................T..............G.........
            ....................................T..............G.........
            ....................................T..............G.........
            ....................................T..............G.........
            ....................................T..............G.........
```

```
123123123123123123123123123123123123123123123123123123123123123
ACATTCCTCTTATTTGACCTAGAAATTGCACTCCTCTTACCACTACCATGAGCCTCACAA 240
.....................................C.......G..............
.....................................C.......G..............
.....................................C.......G..............
.....................................C.......G..............
.....................................C.......G..............
.....................................C.......G..............
.....................................C.......G..............
.....................................C.......G..............
.....................................C.......G..............
.....................................C.......G..............
.....................................C.......G..............
.....................................C.......G..............
.....................................C.......G..............
.....................................C.......G..............
.....................................C.......G..............

123123123123123123123123123123123123123123123123123123123123123
ACAACAAACCTAAACACAATGCTTACCATAGCCCTCTTCCTAATTATCCTGCTAGCTGTA 300
...G....T.........................................C.........
...G....T.........................................C.........
...G....T.........................................C.........
...G....T.........................................C.........
...G....T.........................................C.........
...G....T.........................................C.........
...G....T.........................................C.........
...G....T.........................................C.........
...G....T.........................................C.........
...G....T.........................................C.........
...G....T.........................................C.........
...G....T.........................................C.........
...G....T.........................................C.........
...G....T.........................................C.........


123123123123123123123123123123123123123123123123123
AGTCTAGCCTACGAATGAACTCAAAAAGGACTGGAATGAACCGAA 345
..C........T..G..................A............
..C........T..G..................A............
..C........T..G..................A......G.....
..C........T..G..................A............
..C........T..G..................A............
..C........T..G..................A............
..C........T..G..................A............
..C........T..G..................A............
..C........T..G..................A............
..C........T..G..................A............
..C........T..G..................A............
..C........T..G..................A............
..C........T..G..................A............
..C........T..G..................A............
..C........T..G..................A......G.....
..C........T..G..................A............
```

Calculation of divergence:

There are 20 positions where at least one cow differs from the bison. Of these, 18 are fixed differences (that is, all 15 cows differed from the bison). At position 244, bison use A and cows use G; the codons are, therefore, ACA (thr) and GCA (ala), respectively. This is the only nonsynonymous fixed difference.

Two of the sites are polymorphic in cows: at position 179, one cow differs from the bison; at position 340, two cows differ from the bison. The change at position 179 is nonsynonymous (ile in bison and thr in cows); likewise, the change at position 340 is also nonsynonymous (thr in bison and ala in cows).

The numerator for the calculation of divergence is

$$18 \cdot 15 + 1 \cdot 1 + 1 \cdot 2 = 273.$$

The denominator for the calculation of divergence is

$$345 \cdot 15 = 5175.$$

Average pairwise divergence, therefore, is

$$\frac{273}{5175} = 0.0528.$$

Calculation of diversity:

There are two polymorphic sites in cows, as noted above. At position 179, one cow differs from the 14 others. At position 340, 2 cows differ from the 13 others.

The numerator for the calculation of diversity is

$$1 \cdot 14 + 2 \cdot 13 = 40.$$

The denominator for the calculation of diversity is

$$345 \cdot \left(\frac{15 \cdot 14}{2}\right) = 36225.$$

Average pairwise diversity, therefore, is

$$\frac{40}{36225} = 0.0011.$$

D-loop Alignment

```
bison    AGTACATTAAATTATATGCCCCATGCATATAAGCAAGTACTTATCCTC
T1_CAS   AGTACATTAAATTATATGCCCCATGCATATAAGCAAGTACATGACCTC
T1_KEU   AGTACATTAAATTATATGCCCCATGCATATAAGCAAGTACATGACCTC
T1_LM    AGTACATTAAATTATATGCCCCATGCATATAAGCAAGTACATGACCTC
T1_TPD   AGTACATTAAATTATATACCCCCATGCATATAAGCAAGTACATGACCTC
T1_VMN   AGTACATTAAATTATATGCCCCATGCATATAAGCAAGTACATGACCTC
R1_GD    AGTACATTAAATTATATGCCCCATGCATATAAGCAAGTACATGACCTC
R1_JND   AGTACATTAAATTATATGCCCCATGCATATAAGCAAGTACATGACCTC
R1_KP    AGTACATTAAATTATATGCCCCATGCATATAAGCAAGTACATGACCTC
R1_KRK   AGTACATTAAATTATATGCCCCATGCATATAAGCAAGTACATGACCTC
R1_MBS   AGTACATTAAATTATATACCCCATGCATATAAGCAAGTACATGACTTC
R1_ME    AGTACATTAAATTATATGCCCCATGCATATAAGCAAGTACATGACCTC
R1_RS    AGTACATTAAATTATATGCCCCATGCATATAAGCAAGTACATGACCTC

TATTGACAGTACATAGTACATAAAGTTATTAATTGTACATAGCACATT
TATAG-CAGTACATAATACATATAATTATTGACTGTACATAGTACATT
TATAG-CAGTACATAATACATATAATTATTGACTGTACATAGTACATT
TATAG-CAGTACATAATACATATAATTATTGACTGTACATAGTACATT
TATAG-CAGTACATAATACATATAATTATTGACTGTACATAGTACATT
TATAG-CAGTACATAATACATATAATTATTGACTGTACATAGTACATT
TATAG-CAGTACATAATACATATAATTATTGACTGTACATAGTACATT
TATAG-CAGTACATAATACATATAATTATTGACCGTACATAGTACATT
TATGG-CAGTACATAATACATATAATTATTGACTGTACATAGTACATT
TATAG-CAGTACATAATACATATAATTATTGACTGTACATAGTACATT
TATAG-CAGTACATAATACATATAATTATTGACTGTACATAGTACATT

ATGTCAAATCTACCCTTG--G-CA-A-C-A-TGCATACCCCTT-CCAT
ATGTCAAATTCATTCTTGATAGTATATCTATTATATATTCCTTACCAT
ATGTCAAATTCATTCTTGATAGTATATCTATTATATATTCCTTACCAT
ATGTCAAATTCATTCTTGATAGTATATCTATTATATATTCCTTACCAT
ATGTCAAATTTATTCTTGACAGTATATCTATTATATATTCCTTACCAT
ATGTCAAATTCATTCTTGATAGTATATCTATTATATATTCCTTACCAT
ATGTCAAATTCATTCTTGATAGTATATCTATTATATATTCCTTACCAT
ATGTCAAATTCATTCTTGATAGTATATCTATTATATATTCCTTACCAT
ATGTCAAATTCATTCTTGATAGTATATCTATTATATATTCCTTACCAT
ATGTCAAATTCATCCTTGATAGTATATCTATTATATATTCCTTACCAT
ATGTCAAATTCATTCTTGATAGTATATCTATTATATATTCCTTACCAT
ATGTCAAATTCATTCTTGATAGTATATCTATTATATATTCCTTACCAT
```

```
Divergence numer
2+12+12+12+1 =
39

Diversity numer
2×10 + 1×11 =
31

Divergence numer
12+12+12+12+12+
12+1+12 =
85

Diversity numer
1×11 + 1×11 =
22

Divergence numer
12+11+12+11+12+
12+12+12+12 =
106

Diversity numer
1×11 + 1×11 +
1×11 =
33
```

```
TAGATCACGAGCTTAATTACCATGCCGCGTGAAACCAGCAACCCGCTA
TAGATCACGAGCTTAATTACCATGCCGCGTGAAACCAGCAACCCGCTA
TAGATCACGAGCTTAATTACCATGCCGCGTGAAACCAGCAACCCGCTA
TAGATCACGAGCTTAATTACCATGCCGCGTGAAACCAGCAACCCGCTA
TAGATCACGAGCTTAATTACCATGCCGCGTGAAACCAGCAACCCGCTA
TAGATCACGAGCTTAATTACCATGCCGCGTGAAACCAGCAACCCGCTA
TAGATCACGAGCTTAATTACCATGCCGCGTGAAACCAGCAACCCGCTA
TAGATCACGAGCTTAATTACCATGCCGCGTGAAACCAGCAACCCGCTA
TAGATCACGAGCTTAATTACCATGCCGCGTGAAACCAGCAACCCGCTA
TAGATCACGAGCTTAATTACCATGCCGCGTGAAACCAGCAACCCGCTA
TAGATCACGAGCTTAATTACCATGCCGCGTGAAACCAGCAACCCGCTA

GGCAGAGGATCCCTCTTCTCGCTCCGGGCCCATGAACCGTGGGGGTCG
GGCAG-GGATCCCTCTTCTCGCTCCGGGCCCATAAACCGTGGGGGTCG
GGCAG-GGATCCCTCTTCTCGCTCCGGGCCCATAAACCGTGGGGGTCG
GGCAG-GGATCCCTCTTCTCGCTCCGGGCCCATAAACCGTGGGGGTCG
GGCAG-GGATCCCTCTTCTCGCTCCGGGCCCATAAACCGTGGGGGTCG
GGCAG-GGATCCCTCTTCTCGCTCCGGGCCCATAAACCGTGGGGGTCG
GGCAG-GGATCCCTCTTCTCGCTCCGGGCCCATAAACCGTGGGGGTCG
GGCAG-GGATCCCTCTTCTCGCTCCGGGCCCATAAACCGTGGGGGTCG
GGCAG-GGATCCCTCTTCTCGCTCCGGGCCCATAAACCGTGGGGGTCG
GGCAG-GGATCCCTCTTCTCGCTCCGGGCCCATAAACCGTGGGGGTCG

CTATTTAATGAACTTTATCAGACATCTGGTTCTTTCTTCGGGGCCATC
CTATCCAATGAATTTTACCAGGCATCTGGTTCTTTCTTCAGGGCCATC
CTATCCAATGAATTTTACCAGGCATCTGGTTCTTTCTTCAGGGCCATC
CTATCCAATGAATTTTACCAGGCATCTGGTTCTTTCTTCAGGGCCATC
CTATCCAATGAATTTTACCAGGCATCTGGTTCTTTCTTCAGGGCCATC
CTATCCAATGAATTTTACCAGGCATCTGGTTCTTTCTTCAGGGCCATC
CTATCCAATGAATTTTACCAGGCATCTGGTTCTTTCTTCAGGGCCATC
CTATCCAATGAACTTTACCAGGCATCTGGTTCTTTCTTCAGGGCCATC
CTATCCAATGAATTTTACCAGGCATCTGGTTCTTTCTTCAGGGCCATC
CTATCCAATGAATTTTACCAGGCATCTGGTTCTTTCTTCAGGGCCATC
```

```
Divergence numer
0

Diversity numer
0

Divergence numer
12

Diversity numer
0

Divergence numer
12+12+11+12+12+
12 =
71

Diversity numer
1×11 =
22
```

```
TCACCTAAAATCGCCCATTCTTTCCTCTTAAATAAGACATCTCGATGG            Divergence numer
TCATCTAAAACAGTCCATTCTTTCCTCTTAAATAAGACATCTCGATGG            12+12+12+12+1 =
TCATCTAAAACGGTCCATTCTTTCCTCTTAAATAAGACATCTCGATGG            49
TCATCTAAAACGGTCCATTCTTTCCTCTTAAATAAGACATCTCGATGG
TCATCTAAAACGGTCCATTCTTTCCTCTTAAATAAGACATCTCGATGG            Diversity numer
TCATCTAAAACGGTCCATTCTTTCCTCTTAAATAAGACATCTCGATGG            2×10 + 1×11 =
TCATCTAAAACGGTCCATTCTTTCCTCTTAAATAAGACATCTCGATGG            31
TCATCTAAAACGGTCCATTCTTTCCTCTTAAATAAGACATCTCGATGG
TCATCTAAAACGGTCCATTCTTTCCTCTTAAATAAGACATCTCGATGG
TCATCTAAAACAGTCCATTCTTTCCTCTTAAATAAGACATCTCGATGG
TCATCTAAAACGGTCCATTCTTTCCCCTTAAATAAGACATCTCGATGG
TCATCTAAAACGGTCCATTCTTTCCTCTTAAATAAGACATCTCGATGG
TCATCTAAAACGGTCCATTCTTTCCTCTTAAATAAGACATCTCGATGG

ACTAATGGCTAATCAGCC-ATGCTCACACATAA                            Divergence numer
ACTAATGGCTAATCAGCCCATGCTCACACATAA                            0
ACTAATGGCTAATCAGCCCATGCTCACACATAA
ACTAATGGCTAATCAGCCCATGCTCACACATAA                            Diversity numer
ACTAATGGCTAATCAGCCCATGCTCACACATAA                            0
ACTAATGGCTAATCAGCCCATGCTCACACATAA
ACTAATGGCTAATCAGCCCATGCTCACACATAA
ACTAATGGCTAATCAGCCCATGCTCACACATAA
ACTAATGGCTAATCAGCCCATGCTCACACATAA
ACTAATGGCTAATCAGCCCATGCTCACACATAA
ACTAATGGCTAATCAGCCCATGCTCACACATAA
ACTAATGGCTAATCAGCCCATGCTCACACATAA
```

Calculation of divergence

In the space to the right of the alignments, pairwise differences at variable positions have been recorded. Wherever the bison sequence is highlighted yellow, there is a fixed difference from cows. However, there are two additional positions (highlighted in pink) where all of the cows differ from bison, but the cows are polymorphic; at these positions. Additionally, wherever there is polymorphism in cows, some cows will differ from bison.

The contributions to the numerator for the diversity calculation are shown to the right of each block in the alignment. These sum to

$$39 + 85 + 106 + 0 + 12 + 71 + 49 + 0 = 362.$$

The denominator of the diversity calculation is

$$358 \cdot 12 = 4296.$$

The average pairwise divergence is

$$\frac{362}{4296} = 0.0843.$$

Calculation of diversity

There were 10 polymorphic sites in cows. Of these, eight sites had one cow differ from the others; at two sites, two cows differed from the other eight. Notice that one of these sites is associated with an alignment gap in bison (highlighted in light blue); it counts toward diversity, but is ignored for divergence.

The numerator of the diversity calculation is

$$8 \cdot (1 \cdot 11) + 2 \cdot (2 \cdot 10) = 128.$$

The denominator of the diversity calculation is

$$367 \cdot \left(\frac{12 \cdot 11}{2}\right) = 24222.$$

The average pairwise diversity is

$$\frac{128}{24222} = 0.0053.$$