WTH is “ILR”?

By Alex Washburne & Jamie Morton

Several recent papers we’ve been involved in have utilized the isometric log-ratio (ILR) transform to analyze microbiome datasets. The papers and their software packages range from a phylogenetic transform (PhILR), a phylogenetic version of factor analysis (phylofactor), and using balance trees for hierarchical clustering (gneiss). In this post, we will demystify the ILR transform to help readers disentangle the literature that uses this transform in different ways to perform different analyses.

The elevator speech is that the ILR transform is a convenient way of measuring the difference between two groups of species, R and S, and the three methods above, which all use the ILR transform to measure differences, differ in which groups of species they measure the difference between.

The “isometric log-ratio transform” sounds scary. Look, below is its scary formula:

$$y = \sqrt{\frac{rs}{r+s}}\log\frac{g(x_R)}{g(x_S)}.$$

AHH! Don’t worry, we’re mathematicians (very rarely do we get to say that).

To demystify this scary formula, we will conveniently break up the “isometric log-ratio transform” up into four parts: “ratio”, “log”, “isometric”, and “transform”. We’ll leave the explanation of the hyphen, “-“, to a linguist. To lubricate the mind for the next sections, now is the time to pick up your soothing tea, triple-shot of espresso, or sipping scotch (AW prefers the latter). <sips scotch>.

Ratios – It’s all relative

A fundamental premise/assumption in our analyses of microbiome datasets is that sequence-count data are “compositional”. By “compositional”, we mean that these data provide only information on relative – not absolute – abundances. When we assume that the data are compositional, the benefit of using ratios becomes clear. The compositionality of sequence-count data has been motivated in the primary literature, such as here and here, but we’ll benefit from building our own, simple examples to motivate the compositional assumption for sequence-count data <sips scotch>.

Two thought experiments motivate the compositional assumption. Consider a bacterial community composed of only two types of bacteria, Bacteroides (B’s) and Firmicutes (F’s). Suppose we have one sample in which there are 100 sequence-counts of Bacteroides and 200 counts of Firmicutes.

For the first thought experiment, imagine we double the sequencing depth of the exact same community, yielding 200 B’s and 400 F’s – if we looked at the differences in counts, we’d say that B’s increased by 100 and F’s increased by 200! Typically, this is why people rarify data, but it isn’t necessary to rarify our counts if we just look at ratios: 200/100 = 2 is the same as 400/200 = 2. Using ratios correctly indicates that the abundances of B’s relative to F’s did not change.

For the second thought experiment, assume abundances DO change, but the sequencing depth does not. Imagine that we go from 100 B’s and 200 F’s to 200 B’s and 100 F’s. What can we infer about the underlying bacterial community from these data? How did they change? Illumina sequence-counts are not qPCR – they can't tell us about changes in DNA concentrations in the real community. These data can't tell us whether the changes we observe are due to B's increasing in abundance, F's decreasing in abundance, or some combination of the two. The only thing we can say is that B's change *relative to* F. How do we quantify this change? By ratios of ratios <sips scotch>.

Suppose we're using ratios of F's to B's. In our original sample, we had 200/100 = 2. In our second sample, we have 100/200 = 0.5. How "different" are these, i.e. by how much did they change? Not only can we use ratios to measure the difference between B's and F's within a sample, but we can and often should use ratios to measure the difference across samples. In particular, 2/0.5 = 4 is a natural measure of difference in how much the community changed. In the real community, our two bacterial populations changed by a combined factor of 4 to produce our new samples.

Try out a couple of examples of this: Suppose that the true community has 1 million B's and 2 million F's. Possible scenarios for our second sample are: 4 million B's and 2 million F's (multiply B's by 4), 1 million B's and 0.5 million F's (divide F's by 4), 2 million B's and 1 million F's (multiply B's by 2 and divide F's by 2), or any number of fractional combinations a*1 million B's and b*2 million F's yielding a combined change of a/b=4.

The train of thought, then, is: (1) sequence-count data appear compositional, (2) compositional data only allow us to infer relative abundances, and (3) differences in relative abundance are measured via ratios, which allow us to make inferences that stay within the confines of the data's limitations. On a deeper note, when we're measuring differences with ratios, ratios are acting like subtraction, and logarithms can return us to the familiar world of using subtraction to measure differences. <sips scotch>.

Logarithms – Changing Multiplication to Addition Since 1614

When we use ratios to measure differences, and ratios are like subtraction, what should be addition? Addition is the inverse of subtraction, so we want (a+b)-b=a, from which it's clear that multiplication should be our compositional "addition" since (a*b)/b=a. If multiplication is the compositional data analyst's "addition", what is the compositional data analyst's "division" & "multiplication"?? Well, layman's multiplication is defined as repeated addition, i.e. a+a+a = 3*a, so the compositional data analyst's "multiplication" will be repeated "addition", i.e. a*a*a=a^3, or layman's exponentiation.

That we can redefine "addition" and "subtraction" for data on weird spaces is mind-bending and deeply satisfying if you're like us. However, most people like to steal lunch money from people like us <sips scotch and cries a little>. So, before you erupt into a fit of rage and steal lunch money from the nearest mathematician, or curl into a ball and cry (different people have different responses), let's simplify all of this. Take your count data and take the logarithm (assume there are no zeros, or replace zeros with a positive number less than 1 such as 0.5 – we'll talk about this later). Now, the logarithm of a product is the sum of logarithms… the logarithm of a ratio is the difference of logarithms… the logarithm of a variable raised to a power is the logarithm multiplied by the power. The funky operations for the compositional data analyst's "addition" and "multiplication" are reduced to their layman's analogues by simply taking logarithms.

Let’s try this out. Say we have a vector of counts, $\boldsymbol{x}$, from n species. What is the mean count across species? An appropriate measure of central tendency would use the “addition” and “division” operations defined above, giving us the geometric mean,

$$g(\boldsymbol{x}) = \prod_{i=1}^{n} x_i^{1/n}.$$

However, if we look at log-count data, we get

$$\log g(\boldsymbol{x}) = \sum_{i=1}^{n} \frac{\log x_i}{n},$$

which is the familiar, arithmetic mean of the logarithm. Logarithms turn annoying products and ratios into more friendly addition and subtraction.

Now that we’ve discussed logarithms as ways of making compositional operations less annoying, we can understand the “log-ratio” portion of the “isometric log-ratio transform”,

$$\log \frac{g(\boldsymbol{x}_R)}{g(\boldsymbol{x}_S)},$$

where $\boldsymbol{x}_R$ is the vector of counts in group R and $\boldsymbol{x}_S$ the vector of counts in group S. The ILR transform is a measure of difference between two groups, R and S, best intuited by noting that the log-ratio of geometric means is the difference of arithmetic means of logarithms. In even simpler words, the ILR is a difference of means between two groups.

Suppose our Firmicutes in our original sample above are composed of 25 OTUs with 8 counts each (200 counts), and suppose our Bacteroides are composed of 50 OTUs with 2 counts each (100 counts). Letting R be our Firmicutes and S be our Bacteroides, $g(\boldsymbol{x}_R) = 8$ is the geometric mean of counts of Firmicutes and $g(\boldsymbol{x}_S) = 2$ is that of Bacteroides. The log-ratio part of our ILR transform will be $\log(4)$, which is greater than 0, indicating that Firmicutes OTUs are, on average across OTUs, more abundant than Bacteroides OTUs.

Isometric - … <sips scotch>

The “isometric” part of the isometric log-ratio transform refers to a preservation of distances. To *really* understand how the ILR transform is preserving distances, we’d have to introduce, motivate and write-out the Aitchison distance and, within seconds of writing down the Aitchison distance, we’re worried readers would grab torches & pitchforks and all of our remaining lunch money would be stolen. However, we can provide some intuition as to why having an “isometric” transform is important <sips scotch>.

The first step towards understanding “preserving distances” is to do some mind-yoga and understand distances. Distances are well-behaved measures of the difference between two data points. If we’re using ratios like layman’s subtraction to measure differences, we should probably use ratios to measure distances. A guy named Aitchison fleshed this out and defined a consistent measure of distance between two compositional data points. Much like someone noted that the distance between two points on the Earth is most appropriately measured as the shortest path along the surface of a sphere, Aitchison had to note that compositional data are constrained to a surface and we need distances that measure the length of the shortest path between two points, where “long” and “short” is measured by taking ratios or subtracting logarithms. Thinking about spheres provides intuition that you can work with: compositional data are constrained to a weird surface (the simplex) and require a weird distance (Aitchison distance).

So then why "preserve" distances? Transforms warp the data – taking transforms, especially non-linear transforms like logarithms and ratios, can bend data points closer together and farther away from one-another. If we were being careless and transforming latitude-longitude data to log-latitudes and regular-longitudes (defining 0 latitude as the south pole), the south pole would shoot to minus infinity and the Euclidean distance between two people holding hands near the south pole would seem greater than the distance between Chile and Canada. If we transform the data, we want to ensure we don't warp distances in ways that give us silly results. The way to do that is to define both a transform and a distance such that the distance between transformed data is equal to our original distance between our original, un-transformed data. The isometric log-ratio transform does just that, and in a manner that is very convenient: the Euclidean distance between two ILR-transformed data points is equal to the Aitchison distance between our original, compositional data.

The mathemagicians who conceived the ILR transform noted that the isometry is accomplished through a combination of the use of geometric means to measure central-tendency in a group, as opposed to summing relative abundances (due to the annoying fact that log(a+b) is impossible to simplify in terms of log(a) and log(b)), and the inclusion of a mysterious constant in front of our log-ratio,

$$\sqrt{\frac{rs}{r+s}}$$

where $r = 25$ is the number of OTUs in group R and $s = 50$ the number of OTUs in group S. Don't be afraid of this mysterious constant – this constant is our friend. Even though we've had enough scotch to start calling constants "friends" (an indication of the scotch we've had, or our complete lack of friends) and you've had enough math, it's worth taking a second to thinker a bit and understand why that constant is our friend <sips scotch>.

In mathspeak, the ILR transform is a change of basis from the CLR transform, and the constant ensures that the basis being used is orthonormal <hands over remaining lunch money>. More practically, that constant is our friend because it ensures we don't warp the data too much, like our log-latitudes did. Warped data can cause us some coordinates (e.g. our log-latitudes) to have warped variances, and our constant friend stabilizes the variance of log-ratios of geometric means, which is super important since almost all of statistics involves explaining variance (regression & model fitting), analysis of variance (ANOVA) and finding axes along which the data tend to co-vary (PCA). There would be no statistics without variance <sips scotch> – every data point would be equal to the mean - and so what happens to the variance under our ILR transform is worth paying special attention to.

Let's thinker our way through this. The logarithm of the ratio of geometric means is a difference of arithmetic means:

$$\log\frac{g(\boldsymbol{x}_R)}{g(\boldsymbol{x}_S)} = \overline{\log \boldsymbol{x}_R} - \overline{\log \boldsymbol{x}_S}.$$

However, the variance of the arithmetic mean depends on the sample size. That's why the sample mean is awesome – it's an unbiased estimate of the true mean and its variance shrinks to zero as our sample size gets large. Suppose the log-count data for each OTU, $\log x_i$, were all independent and all had some true variance, σ^2. The variance of the arithmetic mean of r OTUs, $\sum \log x_i / r$, will be $\frac{\sigma^2}{r}$. Consequently, the variance of our un-scaled log-ratio above is

$$\frac{\sigma^2}{r} + \frac{\sigma^2}{s} = \sigma^2 \left(\frac{r+s}{rs}\right).$$

The variance of our log-ratio of geometric means depends on the sizes of the groups, R and S. If we have a bunch of ILR coordinates, the ones corresponding to smaller groups would be "warped" like our log-latitude data and have higher variances. Consequently, if we used PCA to skewer our hot-dog cloud of data points, on a null dataset of a bunch of ILR variables corresponding to a bunch of groups without using this constant, we would expect the loadings (the skewers) to be heavily weighted by ILR coordinates corresponding to small groups that have very high variance. However, we could correct for this by multiplying our log-ratio by our friend,

$$\sqrt{\frac{rs}{r+s}}.$$

Now, you can perform PCA without being fooled by small groups, you can compare the variances of different ILR coordinates without being obviously biased towards/against small groups, and do pretty much whatever you would normally do with your hot-dog cloud of data points scattered about in space. People can hold hands at our compositional South Pole without feeling too distant.

Transform – Changing variables with a tree <sips scotch>

Now that we know all about the isometric log-ratios,

$$y = \sqrt{\frac{rs}{r+s}} \log \frac{g(\boldsymbol{x}_R)}{g(\boldsymbol{x}_S)},$$

we can finish this rant by discussing why it's a transform. Basically, we can take every coordinate – the counts of each OTU – and change them into new variables that correspond to log-ratios of groups. We are all familiar with changing coordinates from x-y-z Cartesian coordinates to spherical coordinates. We refer to our locations on Earth in terms of latitude and longitude because it's a "more natural" way of viewing data (locations) given their constraints (on surface of a giant ball). Changing coordinates helps us calculate distances more easily over directions we are likely to travel given the constraints. Instead of using Aitchison distances of compositional data, we can use Euclidean distances (which everyone knows) between ILR-transformed data.

Let's understand how we go from one isometric log-ration transform, like the one listed above, to a whole coordinate system. We're all comfortable with changing to spherical coordinates using formulas such as:

$$radius = \sqrt{x^2 + y^2 + z^2}$$

and

$$latitude = \arcsin \frac{z}{radius}$$

defining two of three "coordinates" to help us locate points on Earth. We can pinpoint any location on Earth provided all three coordinates carry distinct information.

The ILR transform can define a set of coordinates carrying distinct, non-overlapping information, through a sequential binary partition (think: a strictly bifurcating tree or dendogram). Since the ILR transform is a difference between two groups, each coordinate can be interpreted as a split in a tree separating one group into two. For instance, the coordinate

$$y = \sqrt{\frac{rs}{r+s}} \log \frac{g(\boldsymbol{x}_R)}{g(\boldsymbol{x}_S)}$$

Separates the group containing both R and S into two groups. This coordinate carries information about the relative abundance of group R to group S. Our first coordinate carries no information about differences in relative abundance within R , and so another unique coordinate that doesn't overlap with our first one could be defined easily: an ILR transform splitting R into two groups measuring the differences in relative abundances of these two groups within R. Likewise, another unique coordinate can split S into two groups. This process can be repeated until we have a full coordinate system in the form of a bunch of isometric log-ratios that sequentially partition our OTUs into smaller and smaller groups. The groups at the end of this sequential binary partitioning procedure will be much smaller than the groups in the beginning, but thankfully the ILR transform is isometric and thus ensures that, in null datasets, all coordinates will have the same variance <sips scotch>.

Now that we understand how this is a coordinate system, we need to pick a particular set of coordinates. Which groups, R and S, should we choose? Given a dataset with n OTUs, there are approximately a kajillion different trees (precisely ($2n$-3)!! possible trees) we could make by defining different partitions and consequently there are approximately a kajillion different ILR coordinates. While the Earth spins about its axis, giving us a universally agreed natural choice for a reference axis to measure latitude, and longitude is fixed by convention at the Prime Meridian, microbiome datasets don't immediately offer a choice of partition. Which one of the kajillion coordinates should we use? The three papers we've been involved in, PhILR, phylofactor and gneiss, all went different directions on how to pick a set of ILR coordinates for analyzing & interpreting microbiome data. Each has their interpretation and justification, and the utility of one or the other depends on your what kinds of effects you think you might observe in your data.

The good news, however, is that since the ILR coordinates are basically "new axes" for the CLR transformed data, the hot-dog cloud of data points remains the same regardless which partition you use. Much like your position in Earth remains the same regardless what trickery mathematicians are doing, changing coordinates as they see fit <glares at mathemagicians>, the position of your data remains fixed regardless which coordinates you use. The only thing that changes is the ease of calculation of different kinds of changes in the data and the biological interpretation of inferences on individual coordinates.

<u>So, WTH is ILR?</u>

The ILR transform is a change of variables from relative abundances to a set of log-ratios that preserve Aitchison distances and have stabilized variances under null data. We use ratios because we can only infer relative abundances, we use logarithms because ratios are annoying, and we make things "isometric" because we want to look at meaningful notions of distance in our constrained data and use standard statistical tools that partition variance.

The ILR transform is a mathematical tool. The science comes in deciding how we use it for biology. The biggest open challenges with using the ILR transform are, in our opinion, (1) choosing the sequential binary partition for a biologically meaningful interpretation, (2) deciding how to incorporate sequencing depth into the certainty of our inferences, and (3) dealing with zeros, because logarithms don't like zeros (sneak-peak: all methods use pseudo-counts, some more justified than others, and the choice of pseudo-counts defines the "distance" between a 0-count and a 1-count).

The ILR transform can be intuited as an average difference between two groups of species, and different methods, such as PhILR, phylofactor and gneiss, differ in open-challenge (1): which two groups to differentiate and how to interpret the coordinates. PhILR and phylofactor both use the phylogenetic tree as a scaffolding for coordinates. PhILR differentiates sister clades, and so there is only one PhILR transform for a given tree, there can be no polytomies in the tree, and coordinates correspond to differences between sister clades weighted by the branch length separating the sister clades. Phylofactor differentiates clades along edges in the tree according to which edge is "coolest", so there are many phylofactorizations for a given tree, depending on the data and how you define "cool", and coordinates are interpreted as inferences on edges along which an important trait may have arisen. Gneiss differentiates groups of OTUs in more general hierarchical clustering schemes to investigate partitions that cannot be explained by phylogeny (and, for a fluent user, the machinery in Gneiss could be used to perform phylogenetically-informed hierarchical clustering – stay tuned). Each method has its virtues and faults, and we'll discuss these three methods in more detail later, but, by knowing WTF the ILR is, you should now be able to understand the heart of these methods, navigate the waters yourself and perhaps take your own stab at defining a phylogenetic "Prime Meridian".

<sips whiskey>.