Improving Your Scientific Writing

Frederic D. Bushman

Contents

1. Forward

Your job as a scientist involves more writing with each promotion. By the time you become a lab director, you spend most of your day writing papers, grant applications, recommendation letters, texts and emails. Despite the importance, many of us received little formal training, and write ineffective prose.

Few recognize how much work it is to write well.

In grant proposals, it is common to see sentences underlined or highlighted in bold letters. The Gates Foundation even requires underlining to mark the hypothesis of the proposal. This is only necessary because typical scientific prose is so wandering and wordy that it is difficult to extract the meaning. Underlining is a desperate last effort to communicate through the clutter. Millions of dollars are on the line with large grant proposals, but inept writing creates needless obstacles for many applicants.

We scientists need to create interest in our work. In 2017, according to one measure, the United States spent $500 billion on science. The public has a right to know where their money is going, and a right to be grumpy if scientists can't justify the expense.

Scientists are uniquely qualified to educate the public on the most important issues of our day—think of global warming, human population growth, and global pandemics. To be successful, this requires effective communication.

Here I present suggestions for improving your scientific writing. Over the years I have given the same advice to young scientists again and again, and some have told me it was useful. Write in short sentences. Cut out every unnecessary word. Start paragraphs with strong topic sentences. One idea per paragraph. Simplify wherever possible. Let the facts carry the story.

My training came from writing classes, tough critiques from early mentors, firm guidance from professional editors, and feedback from readers. Much of the best advice I received parallels three classic works on expository writing: "Politics and the English Language" by George Orwell, "The Elements of Style" by Strunk and White, and "On Writing Well" by William Zinsser. Each of these is well worth reading today, though none are specific to scientific writing. There are guides to scientific writing (several are listed at the end), but I haven't found them as useful as the three classics. Furthermore, scientific writing has been changing, for example with the new focus on bioinformatics and Big Data, resulting in new challenges that are not covered well in published guides.

Here I update the three classics and apply their advice to contemporary scientific writing. In places this guide is tough going, working through examples of weak prose or muddy figures and how to fix them. I've tried to make the text more inviting by mixing in examples from really great scientific writing. In a few places I've also added extreme or even outlandish examples from other sources to amplify the main points and add interest--Chapter 2 features a run-on sentence of 126 words that is actually good; Chapter 8 presents what may be the worst explanatory diagram ever made; and Chapter 8 further features several DNA tattoos that are unfortunately the mirror image of the correct structure.

This booklet starts with the elements of editing, emphasizing removing clutter to highlight your content (Chapter 2). Chapters 3-6 discuss the specifics of writing research papers, grant applications, graduate preliminary exams, and emails. Chapter 7 reviews usage of words and phrases common in scientific writing. Chapter 8 deals with the visual display of quantitative data--here the aesthetic is the same—removing clutter emphasizes the main points and allows addition of more content. Chapter 9 presents a few points on writing and thinking. Chapter 10, newly added in January 2023, discuss the challenges presented by ChatGPT, an artificial intelligence program that can write clear prose and edit effectively. Additional material includes suggested reading (Chapter 11), editing exercises (Chapter 12), and samples of letters important in managing scientific publication that may be unfamiliar to young scientists (Chapter 13).

2. Editing

The simpler the better.

Simplify. Every dispensable word you remove highlights your content. In Zinsser's words:

"Few people realize how badly they write. Nobody has shown them how much excess or murkiness has crept into their style and how it obstructs what they are trying to say. If you give me an eight page article and I tell you to cut it to four pages, you'll howl and say it can't be done. Then you'll go home and do it, and it will be much better. After that comes the hard part: cutting it to three".

William Zinsser, in "On Writing Well".

Even for experienced writers, it is remarkable how much of a first draft can be cut out with hard work, and how much the shortening improves the final product.

In the next sections of this chapter we first go over writing effective sentences, then merging sentences into paragraphs. The chapter ends with some general points on editing.

WRITING AND EDITING SENTENCES

Write in short sentences

Keep sentences short. Short sentences are easier to read than long sentences, and they help keep your own thoughts in order. Wandering muddy sentences reflect wandering muddy thinking. All the great scientists I've known wrote in short declarative sentences.

For example, here is the first sentence of Crick and Watson's paper on the double helical structure of DNA.

"We wish to suggest a structure for the salt of deoxyribose nucleic acid (DNA)".

Crick and Watson[1].

DNA is a polyanion, so a cation is commonly added to neutralize the charge in water, thus "salt" is the precise description. They also use the first sentence to define the abbreviation "DNA". Just 14 words suffice to introduce the advance in the paper and address two technical points needed in what follows.

It is possible to write well in run-on sentences, but it's rare. David Foster Wallace was famous for run-on sentences. In the below, "Ennet House" is a halfway house for recovering addicts; "AA" is "Alcoholics Anonymous".

"Gately's biggest asset as an Ennet House live-in Staffer–besides the size thing, which is not to be discounted when order has to be maintained in a place where guys come in fresh from detox still in Withdrawal with their eyes rolling like palsied cattle and an earring in their eyelid and a tattoo that says BORN TO BE UNPLEASANT–besides the fact that his upper arms are the size of cuts of beef you rarely see off hooks, his big plus is he has this ability to convey his own experience about at first hating AA to new House residents who hate AA and resent being forced to go and sit up in nose-pore-range and listen to such limply improbably cliched drivel night after night".

David Foster Wallace "Infinite Jest".

Run on sentences can make for intriguing post-modern fiction, but are usually confusing in scientific writing. If you are just getting started, use short sentences only. As you become more experienced, it can add interest to vary the length of your sentences. David Foster Wallace's prose, for example, often involved short sentences—he just cut lose once in a while with a really long one. Sometimes you can make a point in one longish sentence instead of two shorter ones, and use fewer words in the process.

Variety can add interest, but mostly keep sentences short.

As a last example, consider the start of Martin Luther King's famous speech from the March on Washington in 1963:

"I have a dream."

Martin Luther King, from the steps of the Lincoln Memorial.

Just four words, each of a single syllable, were enough for a riveting start.

Weak intensifiers always hurt you.

Avoid using "very, interestingly, strikingly, new, novel, excitingly…" Only the content itself can be interesting, striking, or novel. Editorializing—proclaiming your opinion that something is interesting or whatever—only invites skepticism. Many scientists go their whole careers without catching on to this. The only route forward is to provide interesting content, and let readers conclude for themselves that it is interesting.

Annoying intensifiers can also have an emotional coloration, as in "I deeply believe in the importance of cancer research". Imposing your emotions on others in a professional context is manipulative, and in me elicits the opposite of the hoped-for effect—quit jerking me around and explain why cancer research is important or I'll find something else to read.

Some words are always dispensable

Here is a sentence from a paper on the growth of carbon nanotubes.

"These results suggest that it would be fundamentally difficult to achieve a fast growth with a long lifetime."

Here is the sentence without "fundamentally"?

"These results suggest that it would be difficult to achieve a fast growth with a long lifetime."

There is no difference in meaning between "difficult" and "fundamentally difficult". The two sentences differ in that the first contains a useless word of five syllables. The sentence also has other problems—the authors should have written "fast growth rate" instead of "a fast growth", or still better something more specific.

Always delete "fundamentally" from your writing. Similarly, delete "certainly" and "basically". "Basic" is fine when it means high pH, but not when interchangeable with "fundamentally". Scrutinize your prose for additional words that add nothing and can be deleted.

Verb tense

Be careful to keep verb tense consistent within sections of a paper or written piece. For example, the Results section of a paper is usually in the past tense, because the experiments have already been done. General principals disclosed by experimentation can be described in the present tense, since the conclusion is ongoing. Be consistent.

Don't start sentences with long modifying clauses.

Here is a painful example:

"Using phosphorescence imaging as a form of biological oximetry, we confirm the oxygen poor environment of the gut lumen and demonstrate the existence of a dynamic equilibrium with an established gradient whereby the mammalian gut releases oxygen into the gut lumen".

Anonymous, early paper draft

A reader will likely need to read the sentence several times to get the meaning.

The much shorter revision below, which lacks the modifying clause, captures most of the content:

“We used phosphorescence imaging to characterize oxygen gradients in the gut lumen and found higher levels near the gut wall”.

Here is a simple declarative sentence from a published paper:

“Escherichia coli IHF protein is a prominent component of bacteriophage lambda integration and excision that binds specifically to DNA”.

Goodman, Nicholson, and Nash[5]

The Nash sentence introduces several points but is clear in one reading.

Starting with long modifying clauses will usually require the reader to go back and reread the sentence once they know from the second half what the first half was modifying.

Get to the subject and verb early

Consider the following difficult sentence:

“Here, a study of microbial communities inhabiting mangrove sediments across southeastern China, spanning mangroves in six nature reserves, was conducted.”

Rearranging to place the subject and verb early in the sentence improves clarity:

“We conducted a study of microbial communities inhabiting mangrove sediments in six nature reserves in southeastern China.”

Rephrase for brevity

Editing is hard work. Below are three before-and-after examples. The first is a wordy paragraph I wrote in a 1999 review article on retroviral integration[6]. “PIC” stands for “pre-integration complex”; “integrase” is an enzyme encoded by retroviruses.

1) Original: “Much interest has centered on the question of whether host proteins are important for the function of PICs in vivo. This article will first review proposals for important proteins arising from studies of PICs, then review studies employing reactions with purified integrase. Proteins thought to influence integration by binding target DNA will be considered in a following section”. (58 words)

Here is a version rephrased for brevity that is also more accurate.

“Are host proteins important for the function of PICs? Below I review proposals derived from in vitro studies of 1) PICs, 2) purified integrase, and 3) purified target DNA binding proteins”. (31 words)

The next two examples are contrived for this work with the goal of illustrating specific editing steps.

2) Original: “A wide variety of factors influence the success of treatment of multiple human cancers.” (14 words)

Rephrased: “The success of cancer therapy is affected by multiple factors.” (10 words)

Phrases like “A wide variety of…” can usually be replaced by reorganizing a sentence. The thick phrase “…influence the success of the treatment of…” is clumsy and again invites rewording and shortening. The rephrased declarative version is four words shorter and the meaning clearer.

3) Original: “Based on data presented here and the published literature (21-23), we propose a model in which HIV can exploit binding to multiple cell surface proteins to enter cells efficiently.” (29 words).

Rephrased: “Evidently HIV can bind multiple cell surface proteins to facilitate entry (this work and 21-23)”. (15 words).

The phrase “we propose a model in which” is a careful statement of the scientific process, keeping distinct the data and ideas about what they mean, but the phrase is also wordy. Consider “evidently” as a one-word summary for “based on evidence”. The long initial modifying clause (“Based…”) is difficult and offers an opportunity for rephrasing for brevity. Clarifying that others have made the same point as in your paper is delicate, but the parenthetical clause is shorter and adequately respectful.

Chapter 11 presents three examples that you can try to edit, then compare your edited text to revised versions that are presented on following pages.

Pompous opinionating is particularly inviting to cut. Just get rid of all of it and let the facts carry the story. Mark Twain, traveling in Europe, became heartily sick the vague language sophisticates used to discuss famous paintings. Commenting on one masterpiece, he wrote:

“The colours are fresh and rich, the ‘expression’, I am told, is fine, the ‘feeling’ is lively, the ‘tone’ is good, the ‘depth’ is profound, and the width is about four and a half feet, I should judge.”
“The Innocents Abroad”, Mark Twain, 1869.

In technical writing, we would cut the above to “The painting is ~4.5 feet wide”.

Minimize novel abbreviations

Inexperienced writers seem to find it exhilarating to define novel abbreviations. I’m saving words! Maybe my new abbreviation will be the next IBM!

The trouble is that each time you encounter a novel abbreviation, you need to make the effort of remembering the new coinage. This may be OK for one, maybe two new abbreviations. Beyond that readers rebel, continuing to read without remembering the abbreviation, progressively losing the thread. It doesn’t take long to until they give up.

Instead make it easy—minimize new abbreviations, or eliminate them altogether.

When to spell out numbers

Most scientific papers will include numbers. Write out all numbers less than 10 (i. e. “nine” not “9”). Write out any number at the start of a sentence. For sentences starting with long numbers, it is usually best to rearrange:

“Four hundred and sixty one subjects were analyzed” can be changed to “We analyzed 461 subjects”.

If you are going to reuse tired phrases, at least learn what they mean.

How often have lazy scientists written that “A is the hallmark of B”? Did they know what a hallmark actually is? Do you? In Great Britain, in the Renaissance, metal workers banded together into guilds that worked out of guildhalls. They would stamp a mark specific to their hall onto completed gold and silver pieces—the hallmark. In saying that “A is the hallmark of B”, how often do writers really mean “stamped on logo”? In my experience, not often.

Another is “paradigm shift”. I once heard an NIH grant review administrator go on at length on how high-scoring grants must represent paradigm shifts. She had no idea how much baggage the term carried.

“Paradigm shift” was introduced in 1969 by Thomas Kuhn in “Structure of Scientific Revolutions”, in which he argued that science is not cumulative. His view was that some revolutions were so profound that they falsified everything that went before (think of gravity before and after Einstein). To make this work, he had to separate technology (which clearly is cumulative) from science, which seems to me a bit forced. The NIH administrator had no idea she was demanding that reviewers only support research that falsified large fields—what she dimly intended to support was really high impact science.

Compare the vision of a government clerk talking about “paradigm shifts” to David Foster Wallace’s description of guys in withdrawal with “eyes rolling like palsied cattle”. My recommendation is to avoid using “hallmark”, “paradigm shift”, and all similar tired metaphors and phrases. Because of blurry overuse, different people will interpret these differently, causing confusion. Just say what you mean simply and precisely, or find a new image (“palsied cattle”) that is particularly apt.

WRITING AND EDITING PARAGRAPHS

Start paragraphs with punchy topic sentences.

A topic sentence should introduce and summarize what follows in the paragraph. You can’t compress the whole paragraph into the first sentence, but you can indicate what is to follow and create interest. Think of the hook in the first paragraph of a newspaper article. Ideally, reading through the topic sentences alone overviews the whole piece.

Here is an example of a poor topic sentence:

“The bacterial microbes that inhabit the intestinal tract, together with their genes and the environment collectively known as the gut microbiome, is a densely populated and complex community dominated by obligate anaerobic organisms from both the *Firmicutes* and *Bacteroidetes* Phyla.”

Anonymous, *early paper draft.*

A reader groans—slogging through such lengthy and tortuous sentences for a whole paper will be an ordeal.

The next example, in contrast, is simple and to the point:

“The repressor of bacteriophage lambda is a protein containing two domains of approximately equal size.”

Mark Ptashne and coworkers[2]

After reading this sentence you expect another short sentence that expands on the function of lambda repressor and begins to develop the direction of the paper.

End paragraphs with sentences that collect what was important and set up what follows.

Consider the last sentence of the abstract of Howard Nash's classic bend-swap paper:

"In recent years the capacity of proteins to bend DNA by binding to specific sites has become a widely appreciated phenomenon. In many cases, the protein-DNA interaction is known to be functionally significant because destruction of the DNA site or the protein itself results in an altered phenotype. An important question to be answered in these cases is whether bending of DNA is important per se or is merely a consequence of the way a particular protein binds to DNA. Here we report direct evidence from the bacteriophage lambda integration system that a bend introduced by a protein is intrinsically important. We find that a binding site for a specific recombination protein known to bend DNA can be successfully replaced by two other modules that also bend DNA; related modules that fail to bend DNA are ineffective".

Goodman and Nash[3]

The final sentence both presents the main data and serves to wrap up the story. Nash had the guts to end his abstract describing a control, confident that the simple presentation of the idea and experiment made further comment unnecessary. How many less secure writers would have gone on to add "Thus we conclude that the data supports a hypothesis in which…"? Nash's last sentence leaves a reader eager to continue on to the main text.

One idea per paragraph

Help your readers by presenting a single idea in each paragraph. When editing, it is often possible to improve your prose by breaking a lengthy complicated paragraph into two or more shorter paragraphs with one idea each. It is fine to write paragraphs with only three sentences, or even two or one.

To avoid the underlining mentioned in the introduction, consider creating a short paragraph presenting each idea you wish to emphasize. That way the prominence of the topic sentence adds emphasis while allowing the prose to read more smoothly. To be fair, opinions do vary among good scientific writers on the virtues of underlining—more on this in the chapter on grants.

Avoid starting with lengthy generalizations.

Mark Ptashne tells a story of his experience writing a review article for editor Al Hershey (Nobel laureate). Hershey was a leader in the lambda field, and Mark the rising star. In Mark's words[4]:

"I wrote a 20 page paper for him and got it back with most lines crossed out and the occasional phrase circled and marked 'Good'. So I rewrote and rewrote and it came back with not a mark on the first page! Not a mark on the second! Then the third page: a line through the middle, a penciled-in 'START HERE', and then most lines thereafter crossed out."

Inexperienced writers often begin with lengthy generalizations, and only start in on specifics part way in. It is usually best to get to the facts as early as possible. Be confident that the general points will be implicit in the specifics.

Cutting deadwood makes possible more cutting

When editing is going well, you sometimes find upon rereading that you can dispense with a lot more text. As the meaning becomes clearer, you don't need to keep reminding readers of stuff that is already fixed in their minds--you can just cut out the unneeded reminders.

FURTHER CONSIDERATIONS

Scientific writing and gender.

Women comprise half of the population but are under-represented in top positions in science. The Hopkins Report disclosed that from 1985 to 1997, the MIT faculty was comprised of less than 10% women. Despite the report, by 2011 the proportion had risen to only 19% women.

In scientific writing, it is common to see the masculine "he" or "him" used when both women and men are intended. The sexist use of "he" for both genders is grating like fingernails on a blackboard--inaccurate writing that also highlights gender inequity. Of course, there are cases where gender-specific pronouns are correct and necessary, as in a medical case report on a female subject. However, "he" appears to be overused in the scientific literature. On June 27, 2015, I carried out a PubMed search using "he" as a keyword, and obtained 132,253 hits. A search on "she" yielded only 87,810 hits.

So what to do? There is no single answer. Substituting "he" with "he or she" is one solution, though wordy. Using instead "she or he" can be gracious (ladies first), but can sound like the writer is showcasing their political correctness. "He/she" is slightly more compact, but also a distraction—think of reading "he/she" out loud and hesitating on how to pronounce it.

Often it is possible to rephrase a sentence to avoid sexist language. Zinsser was eloquent on this point—he recognized that early versions of "On Writing Well" contained sexist sections, and in later editions he described a variety of remedies. "Where a certain occupation has both a masculine and feminine form, look for a generic substitute. Actors and actresses can become performers". My daughter tells me that “you guys” can include girls too, but why not substitute “you folks”?

Consider the following gender-biased sentence:

“Every student should decide what he thinks is best for his own education in biology”.

What to do? One approach is to use the gender-neutral third person plural:

“Every student should decide what they think is best for their own education in biology”.

In some cases switching to the second person can also side-step sexist usage:

“You should decide what is best for your education in biology”.

In this writing guide I’ve used the second person frequently, in part because it lends a direct and familiar tone, but also to avoid sexist language.

In summary, gender bias is part of the history of science and is with us today. Learn to recognize sexist language and rephrase to avoid it.

Avoid over-condensing your writing.

It is possible for writing to be over-condensed. You do need to anticipate questions that a reader may have and write in a way that answers them. Give your readers what they need to follow your points, and do so in short simple sentences. It is fine to use an occasional sentence as a road map, telling the reader what follows and why. Be a generous guide, while keeping your prose spare and effective.

In math and chess, it is common to see phrases like “the rest of the proof is obvious” (which it often isn’t), or in chess “the win is now a matter of technique”. This is arrogant grandstanding designed to highlight the intelligence of the writer. It is also cowardly—if you wrote out how to win the chess game, you would be exposing yourself to the possibility that another player could find a hole in your analysis. Far better to briefly spell out the specifics.

In “In Defense of Food”, Michal Pollan condensed all the advice in his book into three short sentences: “Eat food. Not too much. Mostly plants.” To unpack a little, by “Eat food” he meant avoid rebuilt chemical confections from the food industry (Twinkies etc.). “Not too much” and “Mostly plants” are self-evident. All of an outstanding book condensed into three short sentences. Is this over-condensed? The whole book is well worth reading, but I think the seven word summary is useful, and even brilliant.

Recognize and enjoy outstanding expository writing

Great expository writing is great art. Here is the entire introduction to Hershey's paper on the discovery of circularization of phage lambda DNA.

"Aggregation of DNA is often suspected but seldom studied. In phage lambda we found a DNA that can form characteristic and stable complexes. A first account of them is given here".

Al Hershey[7]

A perfectly appropriate introduction section in three short sentences.

Also from the lambda field, here is an outstanding short abstract from Mark Ptashne.

"The lambda phage repressor is both a positive and a negative regulator of gene transcription. We describe a mutant lambda phage repressor that has specifically lost its activator function. The mutant binds to the lambda phage operator sites and represses the lambda phage promoters P_R and P_L. However, it fails to stimulate transcription from the promoter P_{RM}. The mutation lies in that portion of repressor--namely, the amino-terminal domain--that has been shown to mediate stimulation of P_{RM}. We suggest that the mutation has altered that region of repressor which, in the wild-type, contacts RNA polymerase to activate transcription from P_{RM}".

Guarente, Ptashne and coworkers[8]

Orwell's Rules

George Orwell ends "Politics and the English Language" with six rules for writing clearly, which are as pertinent today as in 1946. Orwell's rules make an appropriate finish here.

1. Never use a metaphor, simile, or other figure of speech which you are used to seeing in print.

2. Never use a long word where a short one will do.

3. If it is possible to cut a word out, always cut it out.

4. Never use the passive where you can use the active.

5. Never use a foreign phrase, a scientific word, or a jargon word if you can think of an everyday English equivalent.

6. Break any of these rules sooner than say anything outright barbarous.

3. Writing scientific papers

This section presents specific advice on writing a scientific paper. There are many ways to do so, and approaches vary among experience authors. I recommend the recipe below for those just starting out.

Generating a draft

Begin by writing an outline. Use separate headings for Introduction, Materials and Methods, Results, Discussion, and Figure Legends. List the main points for each section under the appropriate heading. Discuss the outline with mentors and colleagues.

Writing the text requires a clear idea of the overall direction and the specific data to be included. What is the main story? Writing the outline focuses attention on your most important points.

The next step is to work up relatively final versions of the figures and figure legends. Show the figure prototypes to coauthors and coworkers. Edit based on common reactions from experienced commentators.

Next write the Materials and Methods. After completing the Figures and Materials and Methods, the experimental content of the paper should be fairly clear.

Then write the Results. The text proceeds with a sequential discussion of each Figure. End the section on each Figure with a brief statement of the conclusion, but leave detailed interpretation for the Discussion section.

Next write the Discussion. The first paragraph is typically a summary of the main findings of the paper. Additional paragraphs expand on the interpretation and relationship to previous work. Don’t just repeat the Results section—instead focus on questions like “what can we do now that we have this new data” or “what gap in previous knowledge is now filled” or “what surprises did we encounter”.

Then go back and write the Introduction. Keep it relatively brief--just enough to get things started. Explain why this study addresses an important question.

At this point, show the draft to coworkers. Go through cycles of editing until the draft becomes easily readable and the main points plain and obvious.

Around this stage the references can be put in. I use Endnote. Make a separate database for each paper (big databases can lead to big problems).

Once the manuscript draft is in near-final form, circulate it to lab members and colleagues for comments. Then carry out another round of editing based on comments.

Finishing a paper is a lot of work. Everything needs to be consistent—it is amateurish to have “Fig.”, “Fig”, and “Figure” in the same paper. Reviewers notice. Italicize Linnean names according to standard conventions (check Wikipedia or PubMed on usage if uncertain). Search on in-text markers (I use XXX) to get out all marked comments. Make sure all spelling is correct. If using Microsoft Word, remove all the trash Microsoft adds to documents (comments, marked edits etc). Check each figure call out. Check that the references are consistent—for example, database screw-ups often result in duplicating references in the final list. Check that all figures are of high quality after uploading to the journal and downloading the final PDF.

It is important to check and adhere to the Author Instructions for the journal selected for submission. Check the order of elements and reference style of the journal to which the paper will be submitted--if the wrong journal format is used, the editors might think paper was already rejected by another journal. You need to adjust each of these items for each journal submission

Take responsibility for producing a clean submission-ready document.

Plagiarism

It is not rare for prose in one scientific paper to be similar or identical to prose in another. Sophisticated software for detecting plagiarism is available online (e. g. http://plagiarism.bloomfieldmedia.com/wordpress/software/wcopyfind/) and regularly run over the scientific literature. Findings are reported. If a researcher is found guilty plagiarism, severe punishment is likely.

Don’t take chances. Never copy paste prose from others into your work. Check prose sections contributed to your papers by collaborators if you have any suspicions. Don’t even copy paste from your own papers—rephrase sections despite the fact that they may be saying the same thing.

Writing about statistics

For your results to be convincing, it is important to carry out and document statistical analysis of your data. All measurements are a mixture of signal and noise. It is usually necessary to carry out replicates of experimental and control measurements, and assess the outcomes statistically by comparing variation within each condition to variation between conditions. Construct your prose around how you reject the null hypothesis of no difference between groups.

There are various ways of presenting statistical analysis. I suggest a detailed presentation in the Results section of the main text. After all, you are trying to convince a reader of the soundness of your conclusions, and it is the statistics that do the job. There are cases where analysis may be better placed in the Figures, Methods, or Supplemental sections, but I favor the Results where possible.

As an example of good style, here is a sentence from a paper by Jeff Gordon and coworkers on immaturity in the microbiota of malnourished children [9].

“Family membership explained 29% of the total variance in relative microbiota maturity measurements (log-likelihood ratio**=**102.1, $P<0.0001$; linear mixed model).”

The paper set up an index to quantify maturity of the gut microbiota, then applied it to their samples. As they say, one of the most important determinants turned out to be family membership. They report on the effect size (29% of the variance), the log-likelihood ratio, the statistical significance as the P value, and the test used. Insertion of the parenthetical details does disrupt the text slightly, but it answers the question “why should I believe this”, which to me outweighs the downside of the interruption.

Unpacking the above a bit more—p values conflates sample size and effect size. It is possible to have highly significant differences that are tiny effects. This is part of the basis of Mark Twain’s grumbling about “lies, damn lies, and statistics.” Gordon documents the effect size by specifying the amount of the variance explained. There are many types of statistical tests, and often more than one can be applied to a particular data type. Thus it is important to specify the test used as well.

Note that writing the statistics out carefully allows economies in other areas. There is no need to say “Family membership significantly…”. The P value not only establishes that the result is significant, but quantifies how significant. With a clear explanation in the Results, the presentation of statistical approach in the Methods or Supplemental section can be truncated.

Responding to reviewers’ comments.

Research papers are typically submitted for peer review, then comments come back to the authors. The paper may be rejected outright or accepted subject to revision and re-review as specified in the reviewers’ comments. If the paper is rejected, you need to resubmit to another journal. Don’t despair! Many famous papers have been rejected as early submissions. An important component of success in science is the ability to withstand rejection and keep moving forward. There is always another journal, and the quality of the editorial staff is famously uneven.

If the editors indicate interest in a resubmission, then the text is modified based on the reviewers' comments. Reviewers may ask for more experiments and data, more explanation of the results, or clarification of specific points. In responding, be careful to address to every comment. Remember that the paper will likely be reread by the first round reviewers, and that this is burdensome for them. Reviewing is a duty to the scientific community, but it takes time away from other activities. Write responses to every comment as a gesture of respect. Say clearly how you changed the paper in response to feedback. If you don't, your reviewers may well respond unfavorably.

Slogging through a response to reviewers' comments can itself be slow and annoying. You recruit support by keeping your responses clipped and short while doing a thorough job of addressing the reviewers' comments.

Add data

Responses to reviewers' comments are always strengthened by saying you added more data. Find something to add, and mention it in the first paragraph of the response letter. It doesn't need to be a major new finding. Additional data does need to be meaningful and is best packaged as a response to reviewers' comments, but this can take many forms. Create a favorable first impression.

4. Writing grant applications

There's a lot on the line when scientists write grants, often millions of dollars. Effective grant writing can have a huge effect on your career. If you write effective grants you are expanding and pushing frontiers. Struggle for funding and you are dragged into a grinding battle to survive.

The effect of good writing on grant success is hard to overstate.

Imagine a grant review committee. Reviewing grants is pretty much torture. The writing of the typical applicant is so outlandishly bad that making it through is like climbing Mt. Everest. The rare well-written grant, in contrast, can be an enjoyable opportunity to learn about advanced ideas in an unfamiliar area. Grant reviewers respond strongly to well written grants, often without fully realizing that they are doing so.

You can greatly improve your chances of success by writing strong prose. Below are a few tips.

Write simple prose

The recommendations for simplicity in earlier chapters apply with particular force to grant applications. Write in simple short sentences. Edit out every unnecessary word. Write simple short paragraphs with one idea per paragraph. Let the facts carry the story.

Write readable prose and you are way ahead of the competition.

Follow the instructions

Read the instructions carefully, and talk with grant administrators at the program to which you are applying. It is their job to work with you, so don't be shy about cold-calling them. Most are well-meaning and glad to help. Hopeless applications are no fun for them either.

Work hard to figure out what the funding agency is seeking to support. Explain in clear simple prose why your proposal is aligned with the agency's goals. Work with grant administrators to craft a proposal that matches what they want to fund.

Write to recruit support for your proposal.

Study sections are tough. Imagine a room full of mid-career scientists who have been going over poorly written prose for many hours. People are tired and grumpy. Maybe the same two panelists have been bumping heads all morning. Today NIH doesn't even provide coffee, making things even worse.

There are far more good grants than there is funding. This is well known to the study section members, adding to the gloom.

There is no hope of getting your grant application funded unless it earns the support of an advocate on the review panel. Someone on the panel needs to read your grant and be genuinely excited about it, so that they step up and support your grant before the group. If something new and exciting comes along, it lightens the mood, relieves the depression, and recruits the support of all involved. To be funded, yours needs to be that grant.

So write with your advocate in mind. Most scientific projects are well conceived and have some clever technology involved. Write in a simple and clear way to describe the goals of the project and why they are important. Explain new technology in detail, so that anyone could understand it. Explain in an honest way why you are excited about it. Hit the main points as early as possible.

Remember that your advocate on the panel needs something relatively simple to relay to the study section, most of whom have not read the grant. Write out a simple factual pitch for why your idea is a major advance. Elsewhere, in the more technical sections, explain to specialists exactly what you are going to do and why they should believe your goals are achievable.

This may sound daunting, but the competition makes you look good. If you can describe an exciting project in simple effective prose, you have a strong chance of obtaining funding.

To underline or not to underline?

I try to avoid highlighting text in grants by underlining or bold lettering, but I'm in the minority. Underlining seems to me unnecessary if the prose is well written. Instead I use paragraph structure to highlight important sections. By writing short paragraphs, each with a single idea, you can use the topic sentence to highlight your point. So why disrupt your text with cheesy underlining?

However, good grant writers have argued this with me, and I think they have a point. Think of a tired grant reviewer trying to remember what they liked about a grant among the dozen they read. It may be easier to glance over the highlighted sentences to review the main points, then relay these points to the committee. Given the burden on grant reviewers, the argument goes, anything to make it easier is useful.

You can make your own decision.

Avoid inverted pyramids

When a funding agency supports a project, they want there to be a return on their investment. If a project might fail completely, then it is unlikely to be funded. As a result, the review process is quite conservative.

A common source of problems is the “inverted pyramid”, where a key experiment needs to work for the downstream investigation to be warranted. What if the key step doesn’t work? The whole research program is finished, and the grant money wasted. If your proposal is judged an inverted pyramid it will usually result in a poor score.

There are various solutions for this. The best is to work through the pivotal experiment in advance of funding, then explain in your proposal how your preliminary data makes possible the downstream steps. Another is to propose multiple routes to the same goal, so that you are not dependent on any one experiment working. Ideally, you can use your preliminary data to bolster the idea that the program is feasible.

Inverted pyramids are a common pitfall for new grant writers. After you write a proposal, get away from it for a bit, then reread to check for inverted pyramids. If necessary, rewrite or generate new data to strengthen the case.

Get way back and get way in—minimize the middle ground

It is common for inexperienced grant writers to write much of their application at a middle level of detail. Writers jump right into the problem, and explain mechanism in vague conceptual diagrams. Experimental details are presented in a general sense only. This can be dull and ineffective.

Far better is to get way back from the data and describe why yours is an important question, then get way in and be specific about the engineering involved, particularly for the most novel parts.

Some of the best lectures I’ve heard were by Matt Meselson in the 1980s, and the same aesthetic holds for scientific writing. He used relatively few slides. At the start, he walked forward from the podium, sat on the edge of the stage, and described at some length why he began the research projects he planned to present. He discussed his thinking leading up to the study, conversations with other scientists, and how he ultimately began experimentation on the topic. This was followed by a small number of slides describing key pieces of new data. In presenting each slide, he described the x and y axes, and went over the distribution of data in each graph in an unrushed fashion. This was followed by a sophisticated discussion of the relationship of data to ideas, models for causality, and a realistic assessment of the importance.

It is painfully common today to see scientists in lectures waving at ultra-complicated slides and summarizing the conclusions, usually in a rushed tone.

This is then followed by another slide with many complicated graphs, and no orderly discussion of the blizzard of details.

Grants applications are often the equivalent. Over-compressed in both the conceptual and technical parts, boring and baffling at the same time.

Say in a careful way why you care about the problem, then describe the engineering in depth. Get way back, and get way in. Use short simple sentences that follow one another in an orderly fashion. With a little practice, your grant can be the one that brightens up a study section.

5. Writing preliminary exam proposals

Students in PhD programs will typically take a preliminary examination at the end of their second year. The exam can take many forms. At the University of Pennsylvania, in the Microbiology, Virology, and Parasitology program, preliminary exams take the form of a grant proposal describing planned thesis work. The proposal is refined in consultation with faculty and students, then defended in front of a faculty panel.

A well-written prelim proposal will be a good grant proposal, and the advice on grant writing applies to prelims as well. Keep sentences short. Start paragraphs with strong topic sentences. Present one idea per paragraph. Avoid inverted pyramids. Get way back, and get way in. There are, however, a few features that are more emphasized in Prelim proposals.

You are documenting your scholarship, so be careful to include all the main citations in your field. Be prepared to answer questions on background. Similarly, on the engineering side, explain in professional terms the engineering steps required, and be prepared to answer questions. Committee members will likely keep quizzing you until they reach the limits of your knowledge. Don’t be afraid to say you don’t know, but this sounds much better if you have described a bunch of important factual information before getting there. Typically the most important facts are written into the proposal up front.

Don’t be boring. Students presenting in front of faculty often take a very conservative approach, in the hope that they will be less exposed to criticism. This can erode support, because over-conservative proposals are usually dull. Be clear on why your work will result in high impact papers. If it will not, find a better project.

The committee will want to see some doable sections to the prelim, so that they believe the student can complete the PhD, but it is also important to put in ambitious studies that might not work. Just be clear that you are aware that some of the proposed experiments are hard, and that you will re-prioritize if things don’t go well. Many huge advances were only possible because a talented fanatic confronted a gigantic challenge. Think of Barbara McClintock discovering transposons, gene control, and epigenetics all at the same time. It is OK to say that you are taking on a daring challenge--just make sure your committee is also convinced that you can complete a doctoral degree.

6. Writing emails

Write simple emails.

Email is murder. I receive hundreds of emails a day. It’s gotten to the point where I think of dealing with the onslaught as “killing” emails as I process each one.

When you write an email, think of the recipient who is dealing with this deluge. The goal of the email recipient, as they open your message, is to carry out what needs to be done as quickly and simply as possible, then get on the next damn email.

For this reason, professional email needs to be distilled and simplified. Indicate what the point is in the subject line. Aim to write the minimum number of words that achieves the job that needs to be done. Write in complete sentences to avoid confusion about who is doing what. Provide any needed context and background at the start, so that the business to be done is easily grasped on first reading. Jaunty jokes and blurry personal references can obstruct communication in professional email.

For example, consider this email announcing a seminar on limb transplantation:

“Given the number of times you said that you’d give an arm or a leg for something, compared to the number of limb transplants actually performed, I’d have to say, really? And yet, there is a chance to redeem yourself, or at least to calibrate your bargaining position. Today's speaker does limb transplants and more,…”

It is just annoying to wade through someone’s free association to get to the point. Even if you succeed you often are only 80% sure you understood the intended meaning. In the above email, you need to read through two needless sentences to get the point. He thinks he is being funny. I’m pissed off.

I suggest using the following framework for business emails. Begin by writing "Dear Dr. Smith (or whoevever)". Often there are multiple people cc'ed, and it can be unclear for whom the email is intended. Then write a first sentence that overviews the purpose of the email and the business to be done--"I'm writing to explain the reason for delays in completing this year's budget". Once the email recipient understands the purpose of the email it is fine to go into the details, but be careful not to add more detail than is necessary to get the job done. End with a sentence summarizing action items "Thus I may need your support in obtaining the needed information from the Cancer Biology Department".

Write clear simple prose that gets the job done in the fewest possible words.

Be careful to distinguish between professional and personal email.

It is fine to write in a more personal tone in emails or texts between friends, where the primary goal is not completing some professional task. Just be careful with this distinction. Try to indicate at the start of an email whether the intent is personal or professional.

Email is not private.

Professional emails are not private. When writing an email, imagine a hostile lawyer waving the text at you in court. I've been close to multiple cases of people being fired for the content of their emails.

Keep a professional tone in email correspondence, using short clear sentences and appropriate content.

7. Notes on usage

Below are some troublesome words and phrases. The list comes from years of wrestling with definitions in papers, grant applications, and student theses. The list is intended to be read from start to finish, not consulted like a dictionary.

Basic. A word that is fine when describing pH, but which has no meaning when used synonymously with "fundamental". Always delete in the second case.

Briefly. Also "In brief". Never write this. Just be brief.

Certainly. Just invites skepticism. Always delete this.

Fundamentally. See "basically" and "certainly". Always delete these useless words.

Gene. A widely used word for the unit of genetic function, which has a surprisingly vague meaning. Gene regulatory regions can extend for long regions along DNA, making the edges of genes hard to define. Genes can overlap. In flies, there are even effects on regulation by sequences on sister chromosomes (synapsis-dependent complementation, termed "transvection"). Where possible, favor more precise words, like "transcription unit", indicating just the part that's transcribed, or "locus", meaning just a linear region of a chromosome.

Gender. Actually not a synonym for "sex". "Gender" implies both biology and the cultural context that comes with it. Today some people would prefer to describe their gender as nonbinary. "Sex" is the strictly biological attribution based on morphology and function. "Gender" is the preferred usage when the focus is human culture, "sex" is preferred when the topic is animal phenotype.

If. "If" is often used incorrectly for "whether". David Foster Wallace explains: "They are not synonyms—*if* is used to express a conditional, *whether* to introduce alternative possibilities…in this case there's a wonderfully simple test you can use: If you can coherently insert an "[or not]" after either the conjunction or the clause it introduces, you need *whether*. Examples: "He didn't know whether [or not] it would rain"; "She asked me straight out whether I was a fetishist [or not]"; "We told him to call if [or not, no] he needed a ride [or not? no]" (from "Twenty-Four Word Notes", in "Both Flesh and Not").

Impact. When used as a verb, as in "the intervention impacted health", the word is a needless neologism. Favor "influence".

Influence. A good verb that should be favored over "impact".

In vitro. In vitro means “in glass”, as in a test tube or culture dish. This means different things to different people. In mechanistic biochemistry, “in vitro” usually means reactions in test tubes using purified components. In virology, “in vitro” may mean studies of viral replication in culture dishes. The phrase “in vitro” can be useful, but consider whether more specific phrases can be substituted (e. g. “reconstituted reactions”, “studies of HIV replication in SupT1 cells”, etc.).

In vivo. This means “in a living organism”, but usually a more specific phrase can be used. Consider instead writing “in Drosophila”, or “in teenage human subjects”, taking the opportunity to remind your readers of the system tested using only a small number of additional syllables.

Life cycle. This is the process of replication from birth through reproduction and death. Generally a fine phrase, but when describing viruses favor “replication cycle”, in order to avoid picking a fight over the unanswerable question of whether or not viruses are alive.

Participant. Individuals undergoing testing in clinical trials are understandably sensitive about they way researchers refer to them. “Participant” is respectful and collegial. “Patient” often isn’t right, because participants may be healthy controls, or participants with a chronic disease may not currently be in treatment. “Subject” can sound dehumanizing, in the direction of “laboratory animal”. Favor “participant”.

Protein. Proteins are linear polymers of amino acids. Avoid blurry mixing of protein and DNA, as in “we mutated alanine 161 to valine”. Mutations happen in DNA. Favor either “we mutated DNA encoding alanine 161 to encode valine”, or “we substituted alanine for valine in the protein”.

Prove. Also proof. Acceptable to use in the specific sense of a mathematical proof. Not appropriate in biomedical science—data never proves a model or idea true or false, but only influences our assessment of the likelihood.

Replication cycle. All biological entities replicate, but whether viruses are alive or not is debatable. The word “life” is so loosely defined that it is not possible to test borderline cases--such as viruses--to determine whether they are alive or not. Nobel Laureate Harold Varmus wisely taught his trainees to favor “replication cycle” for viruses over “life cycle”.

Sex. See “gender”.

Significant. Use in scientific prose only in the sense of “p value <0.05”--that is, statistically significant. Still better, just cite the p value, leaving “significant” implicit.

Subject. See “participant”.

Utilize. On this one David Foster Wallace said it all: “A noxious puff-word. Since it does nothing that good old *use* doesn’t do, its extra letters and syllables don’t make a writer seem smarter; rather, using *utilize* makes you seem either like a pompous twit or like someone so insecure that she’ll use pointlessly big words in an attempt to look sophisticated. The same is true for the noun *utilization*, for *vehicle* as used for *car*, for *residence* as used for *house*, for *presently*, *at present*, *at this time*, and *at the present time* as used for *now*, and so on. What’s worth remembering about puff-words is something that good writing teachers spend a lot of time drumming into undergrads: “formal writing” does not mean gratuitously fancy writing; it means clean, clear, maximally considerate writing.” (from “Twenty-Four Word Notes”, in “Both Flesh and Not”).

Very. The quintessential weak intensifier. Always delete this.

Whether. See “if”.

8. Constructing figures

The same aesthetic applied to scientific prose above applies equally to figures. If you strip away everything unnecessary, you highlight what is important. This then provides the opportunity to add more content in the same space.

Edward Tufte wrote a series of outstanding books on this topic. I strongly recommend his first book “The Visual Display of Quantitative Information”. All his books beautifully present good and bad visual summaries of data. He teaches how to remove “chart junk” to focus attention on the intended point, allowing addition of more layers of information in the same graphic.

Below I go over a few examples, applying Tufte’s technique of pointing out strengths and weaknesses.

The first figure, from a paper of mine, attempts to summarizes the results of reactions in vitro testing the properties of purified HIV integrase[10]. The figure is needlessly difficult and the legend is almost unreadable.

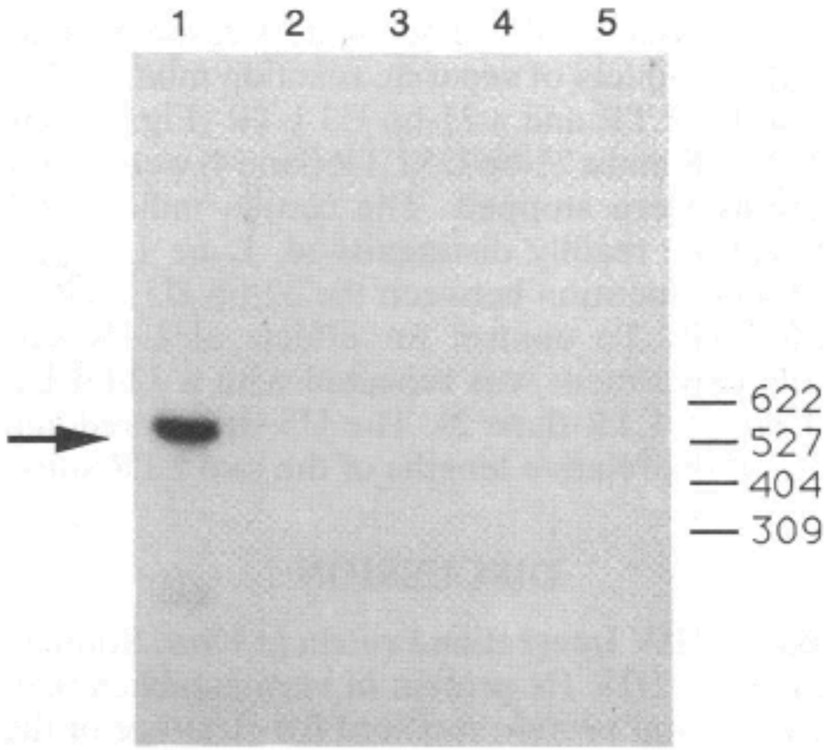


FIG. 4. Insertion of model LTR substrates into a circular DNA target. Standard strand transfer reactions, modified as described, were carried out in the presence of 5 ng of a 174-bp circular DNA. Lane 1, complete reaction mixture. The predominant product is marked with the arrow. Lane 2, 15 mM $MgCl_2$ substituted for 15 mM $MnCl_2$; lane 3, 15 mM EDTA substituted for 15 mM $MnCl_2$; lane 4, target DNA omitted; lane 5, control protein fraction from insect cells infected with a wild-type baculovirus substituted for the IN protein-containing fraction. The dashes beside the gel mark the mobility of size standards (pBR322 digested with *Msp* I) of the indicated length (given in bp). The labeled LTR substrate used in this experiment, LTR J, is identical to LTR A except for an addition of 11 bp to the right side as drawn in Fig. 1. The sequence of this addition was 5′-GGATCCTATCG-3′ and its complement. The use of the 174-bp circular target, synthesized using the *loxP-cre* recombination system of phage P1, and this lengthened LTR substrate facilitated subsequent characterization of reaction products after denaturation (see text).

The original version does not tell a story by itself. One needs to read the wandering figure legend and maybe the rest of the paper to work out what's what.

Compare the revised version of the figure below.

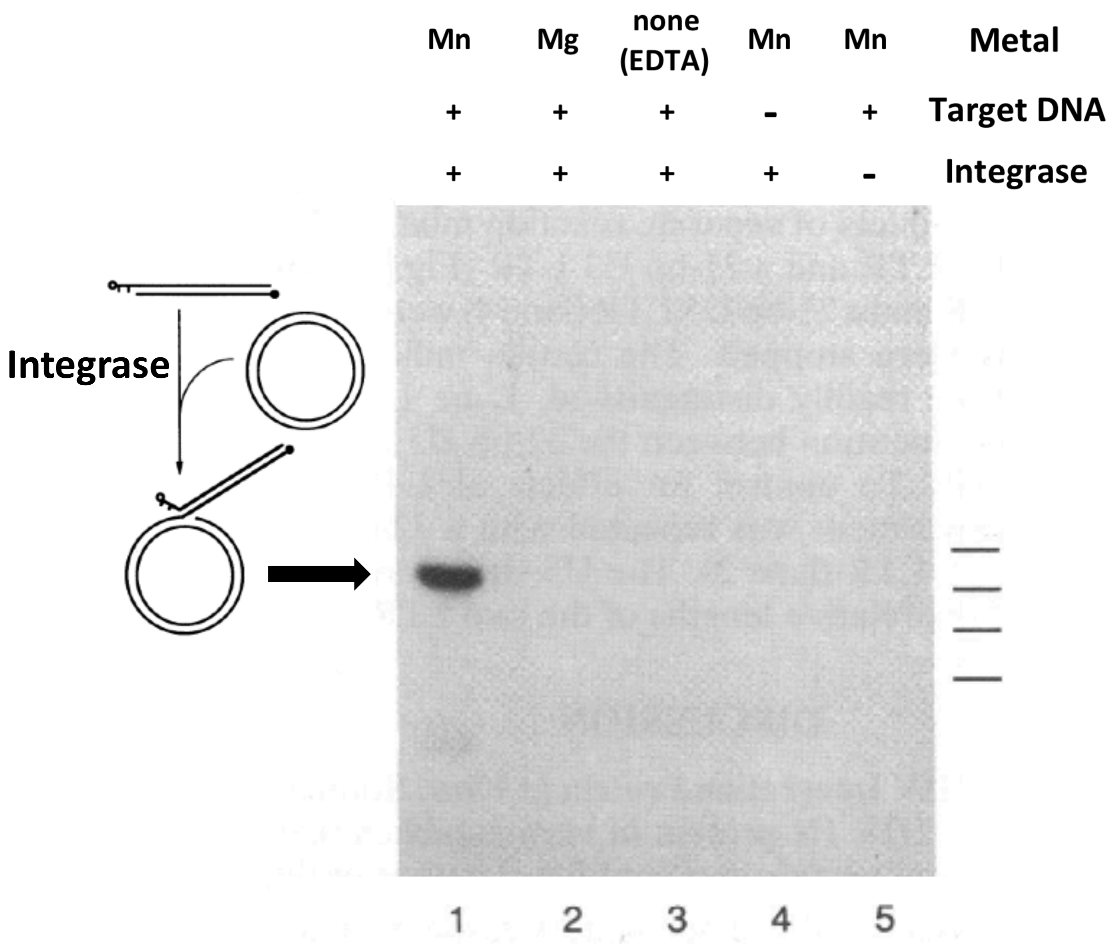


Figure legend. Requirements for HIV DNA integration in vitro. The reaction is diagrammed at the left. A linear DNA mimicking the viral DNA end (top left) becomes integrated into a small circular DNA in the presence of integrase. The viral end oligonucleotide was end labeled, and reaction products separated on a native electrophoresis gel and visualized by autoradiography. Reactions contained: lane 1, complete mixture; lane 2, Mg instead of Mn; lane 3, no added metal (containing only the chelator EDTA); lane 4, no target DNA; and lane 5, a blank protein fraction lacking integrase. Dashes to the right indicate size markers of 622, 527, 404, and 309 bp.

The new version of the figure is much more self-explanatory, diagramming the reaction substrates and products, and specifying the contents of each of the five assays. We learn at a glance that you need integrase, Mn, and target DNA for the reaction to yield product. Improving the figure also allowed simplification of the figure legend.

In the new age of Big Data, it is common to see network diagrams like those in the examples below. These and other Big Data displays are often easy to make, but their value is variable.

In the first example below, a network diagram was generated summarizing co-occurrence among different types of viral genes on a collection of partial viral genome sequences from human gut[11]. The value of the diagram is modest.

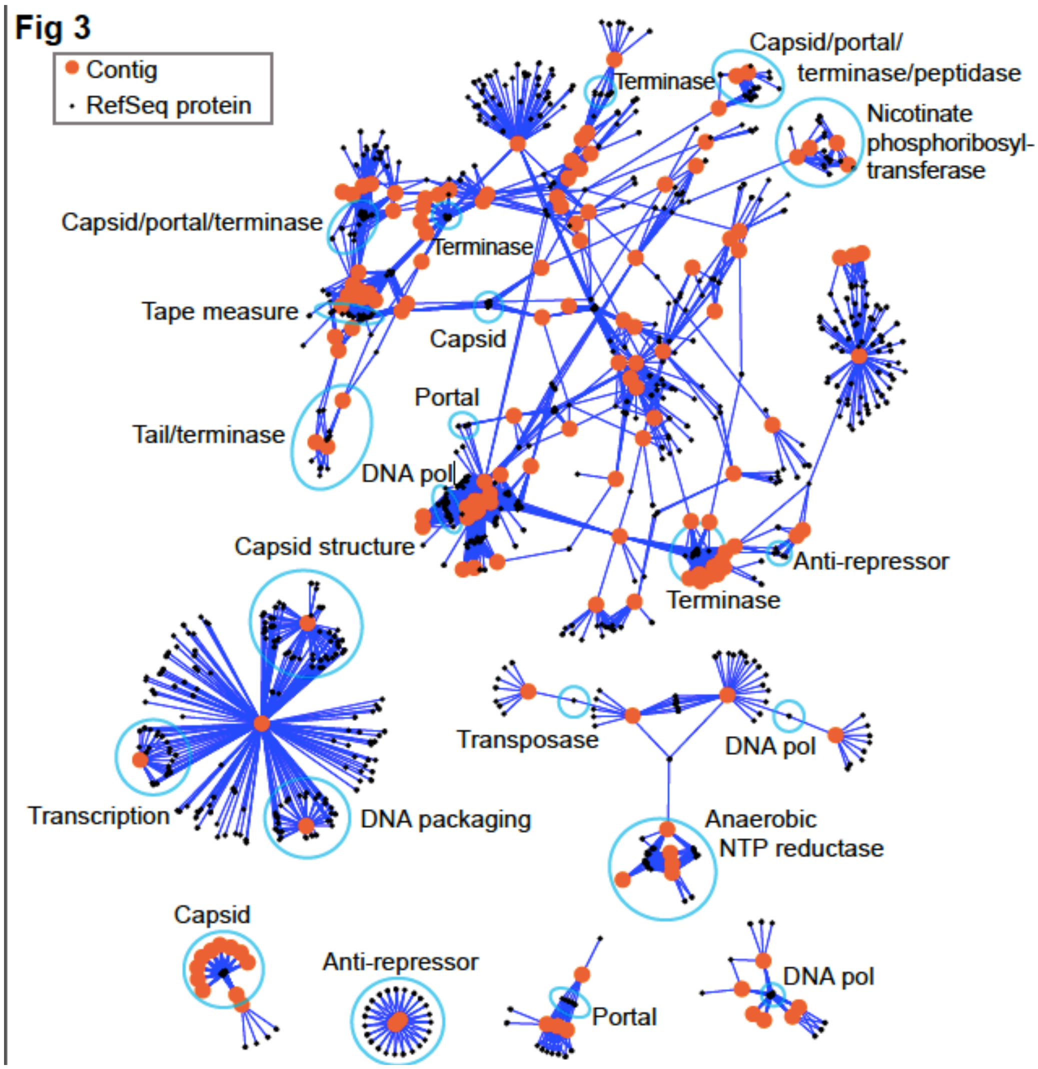


Figure legend. Network based annotation of viral contigs. Orange circles represent viral contigs no shorter than 3 kb. Black circles represent proteins in the RefSeq viral database. RefSeq proteins are connected to viral contigs when an ORF encoded by that contig resembles that protein at $E<10^{-50}$ (blastp). Blue outlines indicate groups of RefSeq proteins and ORFs from contigs that share the function indicated by the adjacent label.

The image shows viral genes linked up by their co-occurrence in DNA sequence populations. The diagram is a fair presentation of the results, but there is not much further you can do with this--if the diagram had come out a lot differently, it wouldn't have made much difference. The picture is purely descriptive and does not support any larger conclusions or follow up. Many of the complicated visual presentations of Big Data today have this quality.

The second example (below) is much more useful. This figure was generated for a training grant application to support HIV research. The nodes indicate researchers participating in the training program. The lines indicate whether any pair of trainers shared a joint publication, and the thickness of the lines indicates the number of joint publications.

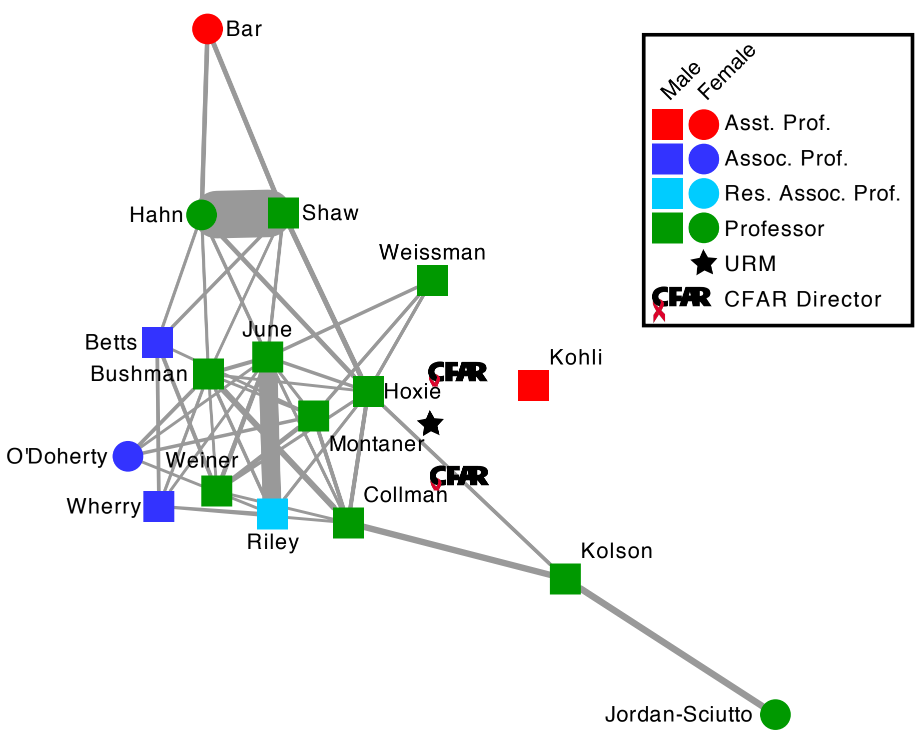
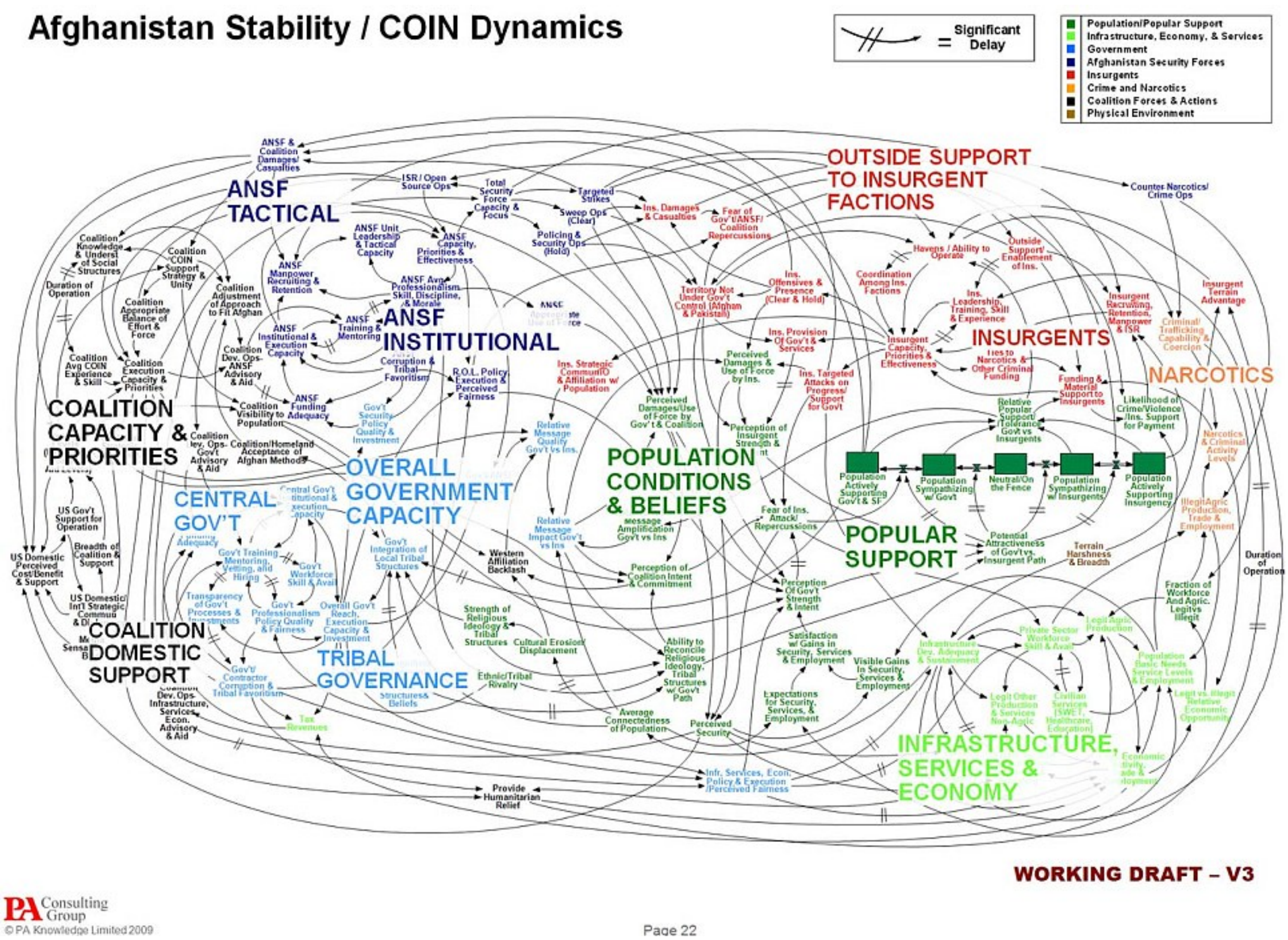


Figure 2. Network analysis of joint publications linking our trainers. The Figure summarizes joint publications as a cytoscape interaction network. The thickness of the gray lines indicates the number of joint publications linking each pair of trainers (e. g. the thin line linking Shaw and Bushman indicates a single publication, the thick line linking Hahn and Shaw indicates 151 publications). The colors show academic rank and the shapes indicate female or male trainers. CFAR directors are marked by the CFAR logo; the star indicates our URM trainer. Analysis and visualization were carried out using R.

Image by Kyle Bittinger.

The density of lines connecting trainers shows that the members of the training program genuinely work together closely. The diagram also summarizes the academic rank and gender, marks the directors of the Center for AIDS Research, and indicates that one of the trainers is from an under-represented minority group. The grant was selected for funding in a tough competition, likely in part because this diagram helped solidify the idea that the trainers worked together effectively.

Here is one of the worst diagrams ever made. Once this just seemed ridiculous, today tragic. The diagram is from the US military describing how we were going to win the war in Afghanistan.

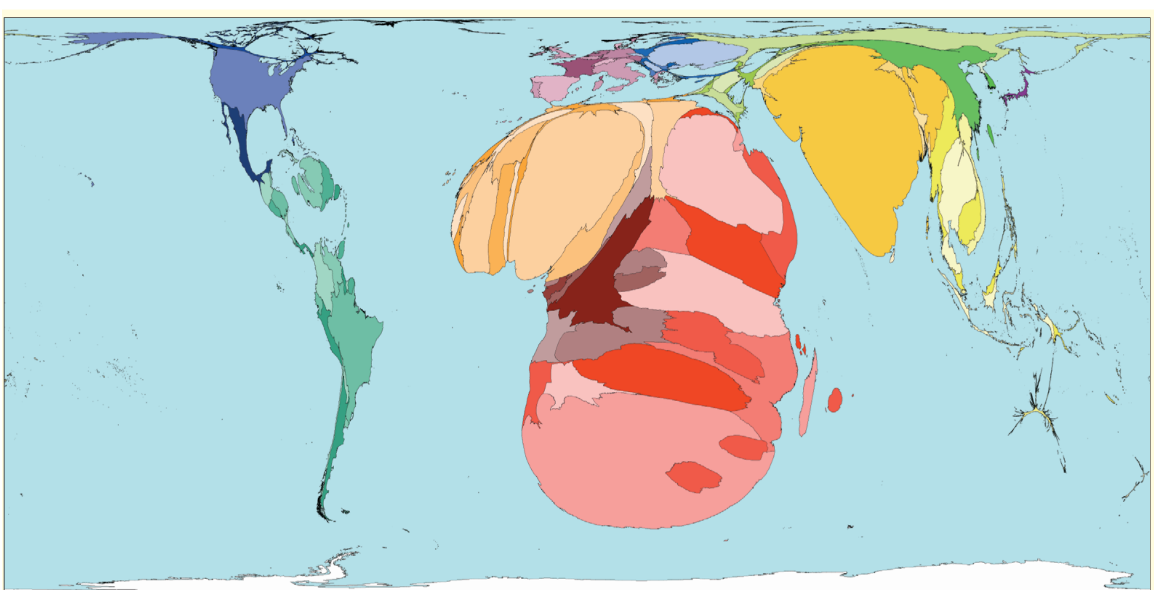


The image looks like a bowl of spaghetti—no one will find anything useful in the tangle of connections. The headings mix different categories. CENTRAL GOV'T, POPULAR SUPPORT, and NARCOTICS are each in capital letters and about the same point size, but one is a political institution, one is a sentiment, and the last is a class of physical objects. If your goal is to confuse your readers, nonparallel lists are a great choice.

Upon seeing this, General McChrystal, leader of US forces in Afghanistan, remarked that “when we understand that slide, we’ll have won the war”. A rebellion against such diagrams followed. General James Mattis commented “Powerpoint makes us stupid”, and some military leaders began banning powerpoint from their staff meetings.

Take advantage of the strengths of data visualization, but prune out the unnecessary. Focus on what you want your readers to get out of your diagram, and edit to highlight your point. In the war diagram above, there was no point to begin with, which the author tried to conceal with extreme complication.

Arresting images often relay your points effectively. A picture isn’t always worth a thousand words, but sometimes it is. The image below was made using “WorldMapper” software, where countries are resized in proportion to their burden of HIV infections.



It is clear at a glance that Africa and India are particularly hard hit. The impact is amplified by the fact that the figure shows a distorted version of a well-known image—we are drawn in by the contrast between the familiar and strange.

Enlarge the lettering

Almost every time a trainee prepares a figure, I end up asking them to make the lettering larger. Among computational biologists, there seems to a religion based on making the lettering as small as possible. Be careful about the type size in labeling each figure--make size consistent and don’t use very large point sizes and very small point sizes in the same figure. Remember that figures are typically reduced in size upon publication, so the lettering needs to be abnormally large at the start to be readable after reduction.

DNA structure

Often the DNA helix is shown in scientific diagrams. It is common in the popular press and even advertising. Few people seem to realize that B-form DNA comes in two mirror images, which are right-handed and left-handed helices, and only one is found biologically.

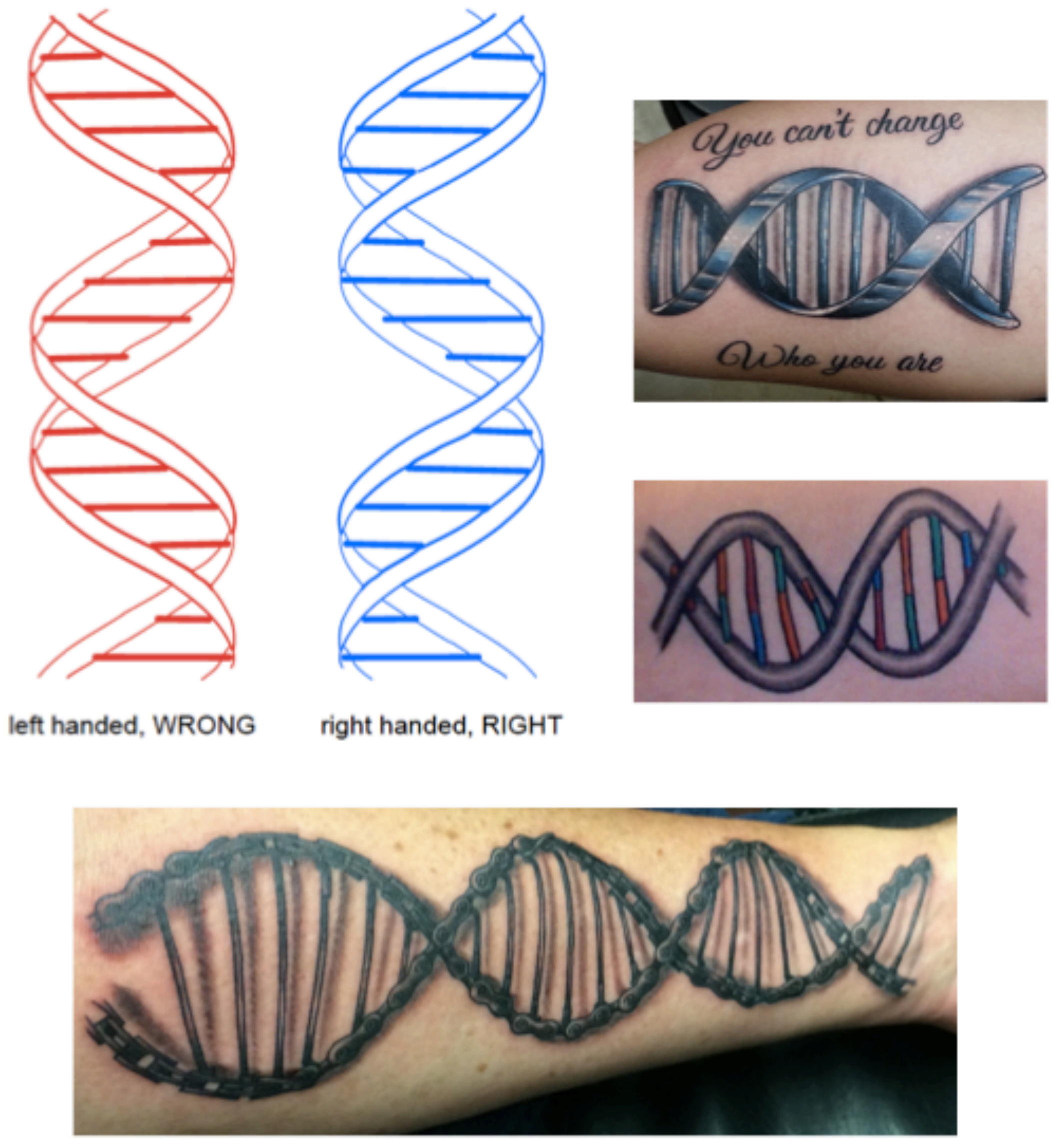


Figure legend: Comparison of lefthanded (wrong) and right handed (correct) B-DNA, and some examples of left handed B-DNA tattoos.

Biologically occurring B-DNA is right-handed only. In the 1980s there was a proposal for left handed B-DNA in *E. coli* [12], but it was later shown to be wrong [13]. Z-DNA is genuinely left handed, but Z-DNA is not a simple B-DNA-like helix and forms only under extreme conditions that are rare or absent inside cells.

In popular culture images are split about 50:50 right-handed versus left-handed B-DNA. Many people who think they are professional biologists make this mistake. I once saw the Chair of a Genetics Department show diagrams with left-handed B-DNA in a lecture, a needless credibility buster.

If it comes up in your work, learn to recognize right and left-handed helices. Get it right in your own work. Don't be like the woman I saw in the New Orleans airport, who had a left-handed B-DNA helix tattooed on her upper arm.

Color

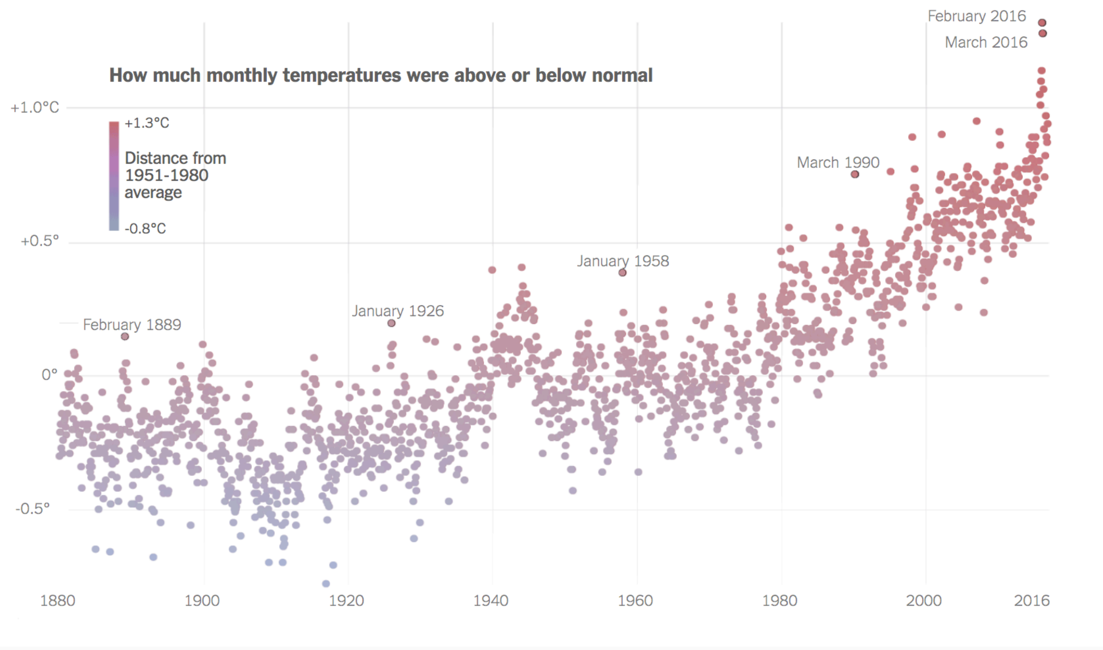
It is possible to wrote precisely about color. Charles Darwin, on the voyage of the Beagle, brought along "Werner's Nomenclature of Colours", which had pictures of over 100 colors, and names for each tint. Darwin held up the book to natural objects, matched the object's color to the entry in the book, and used the name of the closest color in his writing.

Today, we have several digital scales, for example RGB, that allows us to be precise about color. Pantone color scales provide names. Once in a while, color is really important, for example in describing the look of an infected tissue, or the extent of bleaching in a coral. Take advantage of quantitative color scales if helpful.

And avoid sexism in coloring your figures. Early in the days of transcriptional profiling, it was common to see heat maps transitioning from red to green. Unfortunately, 10% of the male population is red-green colorblind. Choose other colors.

Emphasis

The attached diagram, from the January 19, 2017 New York Times, lays out the case for global warming in no uncertain terms.



The simple presentation of time on the x-axis and temperature differential on the y-axis makes the case, but in addition the diagram emphasizes the density of the data by showing each measurement individually. Increasing temperature is indicated twice for emphasis, both as the position on the y-axis and the color of the point.

Not much room for doubt looking at this.

Image manipulation

Today, given the vast tool set for digital image manipulation, it can be hard to be sure what is allowable. Young scientists commonly have excellent skills with computers, often extending to digital art. Photoshopping your lab-mate's head onto the body of a giraffe is standard fun. When managing data, often the first output from an experiment is an image in digital form. So what is fair manipulating the data?

For example, it seems benign to change the contrast on an image, so that a weaker signal is more evident. At the other extreme, making chimeric images, where data features important to the interpretation are added or removed, is clearly wrong. Other manipulations can fall in between. How to decide?

An excellent guide was published in the Journal of Cell Biology, titled "What's in a picture? The temptation of image manipulation" [14]. This is a valuable read for anyone training in the sciences. One message is that many manipulations are OK so long as you do the same thing to the entire image. Differential modification misrepresents the original. Another point is that it is critical to report fully any image manipulation carried out.

If images were generated in separate experiments, don’t try to smooth over boarders between figure panels—make it clear that different pieces are from different experiments.

Follow journal guidelines carefully. Going wild with image manipulation can result in fun art for the lab walls. Just keep it out of your papers.

9. Writing and thinking

Be your own toughest critic.

Garry Kasparov, probably the strongest chess player of all time, was famously hard on his own play. Kasparov:

“I've seen - both in myself and my competitors - how satisfaction can lead to a lack of vigilance, then to mistakes and missed opportunities.”

Kasparov annotated the games from his rise to fame in the 1986 book “The Test of Time”, where he pointed out flaw after flaw in his own play. It got to the point that other top players began weighing in to defend Kasparov’s play against his own attacks. Kasparov won the world chess champion in 1985 at the age of 22.

Like Kasparov, be your own toughest critic. When carrying out experiments, after a lot of hard work you sometimes get an exciting result. Many scientists see their hopes in the data, and not the reality. Easier to dream of glory than confront messy experimental flaws. You are far better off assuming that your result is the most embarrassing possible artifact, and getting to work trying to rule that out. If you fail to falsify your finding, move on to the second worst artifact, and try to falsify that. If, after a lot of hard work, you consistently fail to falsify your result, then maybe you are on to something.

Take the same attitude with your writing. Assume that what you just wrote is weak and look for ways to improve it.

The approach to experimentation described above lends itself to strong writing. Write carefully about the idea and the experiment supporting the idea. Then describe the control experiments that address alternative explanations. This moves the account forward in a natural way and earns the credibility of your readers.

Critique your own motives.

As a new assistant professor I once wrote a collaborative paper with the great chemist Leslie Orgel, who was an outstanding scientific writer. I wrote something like “We predicted that xxx would be the case, and so carried out the following experiments. The data in fact matched our prediction, supporting the idea that…”.

I can still hear Leslie saying “We predicted it, did we--now weren’t we clever…”

The pompous padding was removed from the next draft, shortening and improving the text.

Write to explain something important, not to express how great you are. Write to approach the truth. Readers get this. You advance your own cause much more with clarity and consideration of your readers than with any amount of self-serving chest pounding.

When smart people speak or write confusingly, they often are trying to get away with something.

This is one of the main points of “Politics and the English Language”. Orwell wrote the essay in 1946, having just lived through politics going horribly wrong in the Second World War. His essay spotlighted the deliberate misuses of language that went with it.

Obscurity in writing is often self-serving. Imagine someone saying “it’s not about the money”, then going off on some confusing tangent. It is, of course, about the money. Politics is rife with this kind of nonsense, which was pilloried by Orwell.

It is remarkably common to see scientists writing or speaking in a deliberately obscure way—implying “you can’t understand me, so I must be smart”. Such deliberate obfuscation is an epidemic among computational biologists. Once you take the skeptical Orwellian attitude, these tricks become quite transparent.

In your own writing or public presentations, respect your audience. Make an effort to learn their backgrounds. Start by briefly reviewing stuff they probably already know, to get everyone lined up at the same starting point. Then tell them what you are going to teach them. Follow up and explain your points in a simple step-by-step fashion in terms they can understand. Be realistic. It is extremely common for scientists to lapse into the jargon of their discipline and lose their audience due to laziness or arrogance. Explain the content in a simple and orderly way, be it in public speaking or writing. People notice and appreciate the effort.

Science and proof

It is common to find scientists, often MDs, writing about “proving” a model true. A search of the biomedical literature using PubMed on the keyword “prove” yielded more than 70,000 hits. The problem is that the relationship between idea and experiment is too complex for “prove” to be appropriate. A scientist forms an idea about how the world might work, then uses it to make a prediction for a non-obvious outcome in an experiment. If the experimental result is as predicted by the idea, then the idea is supported. A strong idea may further allow the development of new technology, providing further support.

None of this, however, means that an idea is proven true. It’s just increasingly likely. Never use “prove” or “proof” in scientific writing unless you are referring to the strict mathematical sense of proving theorems, which is fine.

Writing, thinking and public speaking.

Orwell, after blasting deliberately blurry language, pointed to even deeper issues. “If thought corrupts language, language can also corrupt thought”. Habitually vague writing not only reflects vague thinking, but makes the thinking all the more vague. Our thought processes are themselves often in words—think of your internal monolog--so that corrupted writing corrupts your core data processing.

The hard work of cleaning up your writing--stripping out every unneeded word, reorganizing for clarity, honing the key points--cleans up your thinking as well. And it doesn’t stop there. Most scientists will regularly present their research in public lectures. I find that sentences I revised and improved in publications often come out when I describe the data in seminars. Feedback from lectures then refines the description. Questions asked by listeners are particularly valuable--in the next lecture, I commonly answer the best questions from the previous lecture during the seminar. The better-crafted wording is then available for the next paper in the series.

The refined words, and their implications, are in my mind as I stare out the window riding the train to work, chewing them over and searching for additional implications.

Good writing, clear public speaking, and effective thinking all reinforce each other. They are parts of the same whole. Refine them all together to boost your success as a scientist.

Last words

The main points again:

Write in short sentences.
Cut out every unnecessary word.
Start paragraphs with strong topic sentences.
One idea per paragraph.
Simplify wherever possible.
Let the facts carry the story.

The same holds for figures—cut out everything unnecessary to highlight the content.

Strip out the junk so the main points are inescapable.

David Foster Wallace once again:

"The truth is that most US academic prose is appalling—pompous, abstruse, claustral, inflated, euphuistic, pleonastic, solecistic, sesquipidelian, Heliogabaline, occluded, obscure, jargon-ridden, empty: resplendently dead."

David Foster Wallace in "Authority and American usage" 1999.

Many scientists are horrible writers. You can do better.

10. Use of artificial intelligence algorithms in writing and editing

Our world changed in November, 2022, with the introduction ChatGPT, a free artificial intelligence algorithm that can write clearly (https://chat.openai.com/chat). The repercussions are just starting to be felt. Below is an example of a prompt given to ChatGPT, and the output.

Prompt:

Write a five sentence paragraph on the importance of studying SARS-CoV-2 infection in deer.

ChatGPT output:

Studying SARS-CoV-2 infection in deer is important because it can provide insight into how the virus spreads and how it affects different species. It can also help scientists understand the role that wildlife plays in the overall transmission of the virus. Additionally, studying the infection in deer can help identify potential new treatments or vaccines that can be used to prevent the spread of the virus. Furthermore, it is important to study SARS-CoV-2 in deer because it can provide insight into the virus's evolution and how it may change over time. Understanding the virus in different species is key for preventing further outbreaks and controlling the spread of the disease.

Perhaps the text is a little dry and repetitive, but accurate and readable. The scientific literature contains much worse.

Furthermore, ChatGPT can edit effectively. Here is a paragraph from an early draft of a paper I coauthored:

Interest in the human microbiome is rapidly expanding due to increased capacity for DNA sequencing and analysis. The recognition that our microbes may be linked not only to acute infections such as *Clostridium difficile* but to chronic diseases such as heart disease, cancer, obesity, autoimmune disorders and multiple sclerosis is prompting substantial interest from both basic and clinical perspectives. Although we are all essentially identical at the level of our human genomes, our microbiomes are far more variable both in terms of which organisms are present, the relative abundance of shared organisms, and the genes they carry. Additionally, although our genome is static throughout life, each of our microbial communities changes profoundly from infancy through adulthood, when they continue to change through ongoing exposures to diet, drugs and environment. Understanding the microbiome may thus be critical for diagnostics and, eventually, interventions based on the microbiome itself.

Here are instructions given to ChatGPT (written by Dr. Joe Fraietta). The output follows below.

I want you to act as an AI writing tutor for academic research grants. I will provide you with text and your task is to use artificial intelligence tools, such as natural language processing, to re-write the text to be as clear and concise. In addition, the text should be easy to read and flow naturally. Furthermore, the text should use the active voice. You should also use your rhetorical knowledge and experience about effective writing techniques in order to improve the text without changing the underlying meaning of the text. You should consider at least 5 different versions and show me only the version that you think is the best one.

Output from ChatGPT:

The human microbiome refers to the diverse community of microorganisms that live in and on our bodies. Recent advances in DNA sequencing have revealed that the microbiome plays a role in a wide range of health conditions, including chronic diseases such as heart disease, cancer, obesity, autoimmune disorders, and multiple sclerosis. The composition and function of the microbiome vary greatly between individuals, and can change over time in response to factors such as diet, drugs, and environment. Understanding the microbiome is an area of active research, with potential applications in diagnostics and interventions based on manipulating the microbiome.

Perhaps the ChatGPT response was a little less detailed, but still shorter and more readable.

Where does this leave scientific writing? It is important to recognize that AIs are sometimes quite wrong, so relying on prose from ChatGPT or others risks embarrassing blunders. However, the prose is often fine.

It is tempting to forbid use of AIs in scientific writing, and some efforts have been made to create algorithms that will identify AI-generated prose. However, this just starts an arms race between AI detection and countermeasures that is in no one's best interest.

So what is an appropriate use of ChatGPT and resources like it? As of January 2023 the community is just starting to wrestle with this, and best practices are sure to evolve. My initial recommendations are as follows:

-Don't use ChatGPT to generate first drafts. It is still being determined whether this constitutes plagiarism—I think it probably does, but opinions vary. It is in your interest to learn to write clearly, so don't use AIs as a crutch.
-It may be OK to use ChatGPT to help edit, but don't use the AI changes verbatim—treat them as suggestions.

-Always acknowledge use of ChatGPT at the end of any text it was used to generate or modify. Be specific regarding any prompts used, and which parts of the text were modified.
-Allowable uses of ChatGPT will increasingly be specified by course directors, scientific journals, etc. Follow the regulations.

11. Suggested Reading

One of the most effective ways of improving your writing is to read examples of the best. Pay attention to how talented writers construct their prose and develop a written piece. Use the most effective of their methods in your own scientific writing.

Below are a few of my favorites.

<u>Books and essays on expository writing:</u>

William Zinsser. "On Writing Well." Harper and Row, New York, 1985.
A gem.

William Strunk Jr. and E. B. White. "The Elements of Style." Third Edition. Macmillan Publishing Co., Inc. New York. 1979.
The early classic.

George Orwell. "Politics and the English Language."
Should be required reading for every citizen of the planet.

<u>Examples of outstanding scientific writing:</u>

Mark Ptashne. "A Genetic Switch." 3rd Edition. Cell and Blackwell Press.
On the growth of phage lambda.

Jonathon Weiner. "The Beak of the Finch". Random House.
Pulitzer Prize winning book on evolution of birds in the Galapagos.

Jonathon Weiner. "Time, Love, Memory". Random House.
On genes and behavior, focusing on the career of Seymour Benzer.

Ed Yong. “An Immense World”. Random House.
How animal senses differ from ours, and the implications for their perceptions of the world.

The classic research papers cited in the text above are also well worth reading.

<u>Books on the visual display of quantitative information:</u>

Edward Tufte. "The Visual Display of Quantitative Information".
Outstanding book on editing visual displays. Called "a visual Strunk and White" by the Boston Globe.

Edward Tufte. "Envisioning Information".
The successor to the above, also excellent.

Books and essays on scientific writing

Mimi Zeiger. "Essentials of Writing Biomedical Research Papers", Second Edition.
Excellent in many ways, but lengthy and a published before the big data era.

Robert A. Day, Barbara Gastel. "How to Write and Publish a Scientific Paper", 7th Edition.
A detailed discussion of the elements of writing scientific papers.

"Writing Successfully in Science", by Maeve O'Connor

"Am I Making Myself Clear?: A Scientist's Guide to Talking to the Public", by Cornelia Dean

"The Chicago Manual of Style" 16th Edition Sixteenth Edition, by University of Chicago Press

"CBE (Counsil of Biological Editors) Style Manual, 8th Edition". Counsil of Biological Editors, Inc. Bethesda MD.

12. Editing exercises

Below are three examples of scientific writing that can be improved. Try editing them yourself. Following each example is an edited version you can compare to your own efforts.

Example 1. A paragraph from a student thesis (125 words). "T/F" means "transmitted/founder" Hepatitis C virus genomes.

"In another study by Mitchell and colleagues, significant differences in the potency and nature of the innate responses to RNAs generated from T/F molecular clones were detected in cultured hepatocyes and immortalized cell lines. Additionally, these were found to correlate with respect to the genotypes of the T/F genomes with genotype 3 RNAs stimulating an enhanced pro-inflammatory profile as compared to that of genotype 1 and 4 T/F RNAs. The cell intrinsic response to genotype 3 RNAs included enhanced expression of RIG-1, STAT1, and TLR3. Intriguingly, these findings may provide a mechanistic explanation for the unique clinical characteristics of genotype 3 infections including a higher rate of spontaneous clearance, and a strong association with accelerated cirrhosis and hepatocellular carcinoma".

Example 1, revised.

Here is a version edited for brevity (86 words).

"Mitchell and colleagues found differences in innate responses after transfection of T/F molecular clones of different HCV genotypes into cultured hepatocyes and immortalized cell lines. Transfection with genotype 3 clones, but not 1 and 4, resulted in a pro-inflammatory cellular response including enhanced expression of RIG-1, STAT1, and TLR3. In patients, genotype 3 HCV infections show a higher rate of spontaneous clearance, and a strong association with accelerated cirrhosis and hepatocellular carcinoma, potentially reflecting the proinflammatory properties seen in cell culture".

After editing, the length is reduced to 86 words, for a savings of 39 words. Note how the use of "…, but not 1 and 4,…" allowed deleting a longer clause. Deleting the weak linkers "Additionally" and "Intriguingly" allowed the paragraph to read more smoothly. Rephrasing the sentences for simplicity allowed considerable further shortening with gain of clarity.

Example 2. An abstract from a published paper, a useful study of the gut microbiome and its possible roles in cardiovascular disease.

Intestinal microbiota composition modulates choline bioavailability from diet and accumulation of the proatherogenic metabolite trimethylamine-N-oxide.

Romano KA, Vivas EI, Amador-Noguez D, Rey FE

Choline is a water-soluble nutrient essential for human life. Gut microbial metabolism of choline results in the production of trimethylamine (TMA), which upon absorption by the host is converted in the liver to trimethylamine-N-oxide (TMAO). Recent studies revealed that TMAO exacerbates atherosclerosis in mice and positively correlates with the severity of this disease in humans. However, which microbes contribute to TMA production in the human gut, the extent to which host factors (e.g., genotype) and diet affect TMA production and colonization of these microbes, and the effects TMA-producing microbes have on the bioavailability of dietary choline remain largely unknown. We screened a collection of 79 sequenced human intestinal isolates encompassing the major phyla found in the human gut and identified nine strains capable of producing TMA from choline in vitro. Gnotobiotic mouse studies showed that TMAO accumulates in the serum of animals colonized with TMA-producing species, but not in the serum of animals colonized with intestinal isolates that do not generate TMA from choline in vitro. Remarkably, low levels of colonization by TMA-producing bacteria significantly reduced choline levels available to the host. This effect was more pronounced as the abundance of TMA-producing bacteria increased. Our findings provide a framework for designing strategies aimed at changing the representation or activity of TMA-producing bacteria in the human gut and suggest that the TMA-producing status of the gut microbiota should be considered when making recommendations about choline intake requirements for humans.

238 words

Example 2, revised. Here is a shortened version of this abstract.

“Choline is a water-soluble nutrient essential for human life. Gut microbial metabolism of choline results in the production of trimethylamine (TMA), which upon absorption by the host is converted in the liver to trimethylamine-N-oxide (TMAO). TMAO is reported to exacerbate atherosclerosis in mice and is positively correlated with the severity of atherosclerosis in humans. Which microbes contribute to TMA production in the human gut remains largely unknown. We screened a collection of 79 sequenced human intestinal bacterial strains from XXX phyla and identified nine strains capable of producing TMA from choline in vitro. Gnotobiotic mouse studies showed that TMAO accumulates in the serum of animals colonized with TMA-producing species, but not in the serum of animals colonized with strains incapable of generating TMA. Even low levels of colonization by TMA-producing bacteria significantly reduced choline levels available to the host—more efficient colonization reduced levels further. Our findings suggest approaches to controlling TMA production in human gut and optimizing recommendations for choline ingestion based individual microbiota composition.”

166 words

In this case the original abstract wasn’t too bad, and the research presented significant. Nevertheless, editing shortened the abstract by 72 words with gain of clarity. Phrases like “Recent studies revealed that” in the original can usually be shortened with thoughtful rephrasing. The sentence starting “However,…” was a 44 word run on, providing an opportunity for shortening with improved readability.

Example 3. Recently I wanted a challenging abstract as an editing exercise for a class. I thus searched PubMed on “basically”, and found some that were crying out for help. One is below.

Attention deficit hyperactivity disorder (ADHD) is a complex disorder, which can be seen as a disorder of life time, developing in preschool years and manifesting symptoms (full and/or partial) throughout the adulthood; therefore, it is not surprising that there are no simple solutions. The aim of this paper is to provide a short and concise review which can be used to inform affected children and adults; family members of affected children and adults, and other medical, paramedical, non-medical, and educational professionals about the disorder. This paper has also tried to look into the process of how ADHD develops; what are the associated problems; and how many other children and adults are affected by such problems all over the world basically to understand ADHD more precisely in order to develop a better medical and or non-medical multimodal intervention plan. If preschool teachers and clinicians are aware of what the research tells us about ADHD, the varying theories of its cause, and which areas need further research, the knowledge will assist them in supporting the families of children with ADHD. By including information in this review about the connection between biological behavior, it is hoped that preschool teachers and clinicians at all levels will feel more confident about explaining to parents of ADHD children, and older ADHD children themselves about the probable causes of ADHD.

223 words

Example 3, revised.

Attention deficit hyperactivity disorder (ADHD) is complex and lifelong. ADHD can develop in the preschool years and manifest symptoms (full or partial) throughout adulthood. There are no simple solutions. This paper provides a concise review of ADHD to inform medical and educational professionals, those with the disorder, and their family members. This paper also investigates how ADHD develops and associated problems, with the goal of improving interventions. Better information on ADHD will assist teachers and clinicians in supporting families with affected members.

83 words.

13. Sample submission and resubmission letters

Young scientists just starting out may never have seen the letters surrounding publication—their mentors usually handled it. Below are attached two letters from the publication process as models for correspondence of your own.

Cover letters are needed for initial submission of papers for publication. Once these were mailed along with paper manuscripts. Today they are uploaded into web sites, but the letters have not changed much. Reviews then come back with a cover letter from the Editor and several anonymous reviews. Once the paper is modified to address the reviewers' comments, it is resubmitted along with detailed responses.

Two letters from publications of ours are below. The first letter is a submission letter, the second a rebuttal letter.

For rebuttal letters, it seems that editors today want a letter with each reviewer comment repeated, then your response afterwards. There was a time when condensing reviewer comments and answering them collectively was allowable and in my opinion more readable and efficient, but journals have mostly gone in a different direction. Thus unfortunately I recommend the structure in the letter below. The letter involves a paper in a journal, mBio, in which the reviews are not confidential.

Perelman School of Medicine
University of Pennsylvania

Department of Microbiology Frederic Bushman, Ph.D.
3610 Hamilton Walk Professor
426A Johnson Pavilion
Philadelphia, PA 19104-6076
Tel 215.573.8732 Fax 215.573.4856
bushman@mail.med.upenn.edu
www.med.upenn.edu

March 14, 2013

Dear Madam or Sir,

Attached please find a draft paper, "Fungi of the murine gut: episodic variation and proliferation during antibiotic treatment", by Dollive et al., that we would like to submit for publication in PLoS One. Many clinical papers have suggested that fungi may grow out when patients are treated with antibiotics to suppress bacterial growth, but analysis in humans can be complicated by the complexities of the underlying condition and use of therapies in addition to antibiotics. To assess the effects of antibiotics on fungi in isolation, we treated mice with a cocktail of antibiotics, then used metagenomic methods to monitor fungal and bacterial growth.

We found that fungi indeed grew out prominently in the mouse gut, and that cessation of treatment allowed the community to return to a state that was similar but not identical to the starting state. Notably, *Candida* persisted at a higher level at the last time point tested. These data suggest that treatment with antibiotics can result in the outgrowth of a medically relevant fungus, and that antibiotic effects can persist for long periods after cessation of treatment. These data are also important because antibiotic cocktails are often used to deplete bacteria in immunological studies in mice, but the fact that they promote fungal growth may need to be considered as well.

In addition, we found to our surprise that fungal populations were highly variable even in control mice, and that variations were specific to each cage of mice studied, disclosing a new and likely important variable in microbiome research.

Reviewers qualified to comment on our paper include:

(Attach here names, addresses and email addresses for five reviewers. Aim for a good mix of geographic location, academic ranks, and genders.)

None of this work is submitted elsewhere for publication. All authors have viewed and approved the manuscript. Thank you very much for considering this submission.

Best regards,

Frederic Bushman

Department of Microbiology Frederic Bushman, Ph.D.
3610 Hamilton Walk Professor and Chair

426A Johnson Pavilion
Philadelphia, PA 19104-6076
Tel 215.573.8732 Fax 215.573.4856
bushman@mail.med.upenn.edu
www.med.upenn.edu

November 22, 2017

Response to reviews of “Allometry and ecology of the bilaterian gut microbiome”

Dear Madam or Sir,

Attached are our responses to reviews from Dr. XXX and Dr. XXX of our paper “Allometry and ecology of the bilaterian gut microbiome” by Sherrill-Mix et al. We are grateful to the reviewers for their helpful comments. We have revised the manuscript extensively in response, and feel it is considerably improved as a result. Specific responses are as follows. In each case we copied the reviewers’ points into the letter, then added our responses below.

Comments from Dr. XXX

We are gratified that Dr. XXX felt that “Overall the paper is very well done and an interesting read.”

1) The authors note in the introduction that previous studies were mixed in whether species-area relationships exist for the gut microbiota, with ref 4 not showing a relationship and ref 5 reporting a relationship, which was again recovered here. It would be a nice addition to the discussion for the authors to reflect on why these inconsistencies between different studies may have occurred.

In the revised draft we have added more discussion on the possible origins of the discrepancy.

2) The samples from the larger animals were from feces and from smaller animals (insects) were generally from dissected intestines. How might these different methods affect the degree to which the entire intestinal microbiota is observed and thus species-area relationships? E.g. might we recover bacteria from more available niches in whole intestines versus feces? It would be good to add some discussion of this. Maybe this is why slope is not as large as may be expected?

We now mention this point in the Discussion as suggested.

3) At the bottom of page 7/top of page 8 the average sequence reads per specimen is described and the range in the number of OTUs across samples (e.g. 21 in box elder). Are these OTU estimates calculated on rarefied data and if so what level was the data rarefied at?

We mostly used data without rarefication in order to maximize the amount of sequence information available. In cases where the amount of sequence analyzed could have affected the outcome, as in the species-area analysis, we did use rarefied data. We have expanded on this in the revised draft to improve clarity.

4) On the bottom of page 8 it is explained that "12 percent of the sequences remained unassigned using the greengenes classifier". Does this mean that the classifier could not even characterize these as some type of bacteria? Also, what is the "greengenes classifier"? I think that greengenes is just a database. E.g. QIIME will use the RDP-classifier to assign sequences taxonomically using the greengenes database. It would be good if these results regarding unclassified reads were included in Figure 1.

We have clarified the description of the use of the Greengenes database, and added new analysis on the reads that could not be assigned using Greengenes (new Figure Axxx).

5) On the bottom of page 9 and Figure 1D, the proportion of OTUs unique to each species is reported and it is concluded that this did not correlate with phylogenetic placement of the host. How does it relate to how densely the different parts of the host tree is sampled? For instance the low amount of unique diversity in any given monkey species may be more related to that part of the host phylogeny being sampled pretty deeply.

We have removed the Proportion of unique OTUs from Figure 1 and supplementary figures in favor of using PD as suggested in 6 below, which we agree is better.

6) In addition to the # of unique OTUs per species, it might be interesting to look at the amount of unique branch length in the 16S rRNA phylogeny per species. This concept was described by Dan Faith as “G” for “Gain in phylogenetic diversity” in the same paper that he introduced the PD (Phylogenetic Diversity) concept (it is implemented in QIIME). G might be more interesting than number of unique OTUs because it also would give more information on whether those OTUs are on deep branching phylogenetic lineages.

We have added PD to Figure 1, and we explore deep branching lineages in more detail as described below.

7) For the unclassified sequences, it might be interesting to look at where they fall in the 16S rRNA phylogenetic tree to get a sense of whether there are any deep unclassified lineages that are widespread across animals. (e.g. do many of the unclassified 16S rRNA all cluster together?). David Relman’s group recently described novel deep branching lineages from the dolphin’s mouth. Might be a good paper to reference in terms of an example of a host-associated community revealing much novel diversity.

We have carried out an analysis of deep branching lineages, and added a figure on this to the supplemental information. We also now reference the Relman dolphin study.

8) On Page 14 it is described that “Almost all gut samples had a long tail of rare species with many OTUs..” The authors discuss how this tail of rare species is not likely an artifact driven by Chimeras since Chimeras were rare. In terms of artifacts that boost our measurement of the “rare biosphere” I actually worry more about sequencing error than Chimeras. There are a couple of new tools out there that work quite well for denoising. One is dada2, which I have used a lot and it really seems to clean things up nicely. The Knight lab also recently released a similar tool called “deblur”, which I have used less but I hear works better on really large datasets like this one than dada2. It would be interesting to see what happens to this long tail if one of these denoising tools was applied.

We have used deblur to denoise our data set and compared results. We still found the long tails of rare species in rank-abundance curves. This is now mentioned in the revised draft.

9) On Page 14, it is determined that a fit based on a Power Series or Poisson lognormal curve provides the closest match to the data. Does this tell us anything more about the ecological drivers of community structure beyond that it is not neutral?

We have been very interested in this point. In the revised draft we have added results for another version of the neutral model which incorporates vertical acquisition of microbes, but this too does not fit our observations. It would be exciting to devise generative models that produce the data we observe, but available models mostly don't make unique predictions. We now mention this and cite a thorough study.

10) On the top of page 21, the version of QIIME used should be noted.

The version of QIIME is now specified.

11) On line 6 of page 21 it is described that bacterial cell growth information was determined from Bergey's manual. Is this growth rate? Where is this information used? For both this and the oxygen information, what was done for species/OTUs that were not defined at the species level and/or not found in these textbooks?

Here, we use "aerobic" or "anaerobic" to indicate the designations from published literature, which capture information on whether bacteria can grow in aerobic or anaerobic environments. We have adjusted the text to clarify this point. In the methods section, we now describe our approach for assignment of OTUs that are not defined at the species level or not found in our references. We added an additional table in supplemental information (new Table A4) to provide details on the aerobic or anaerobic status of each taxon, and how the status was determined.

12) In the tree on Figure 1, it would be helpful if the class and order names that are used in Figure 2 were labeled on the tree. This might be too much info/messy, but it would help someone understand how these are related to each other when looking at Figure 2.

We wrestled with this, but could not find a way to make Figure 1 readable with added labeling. In the revised text we have directed readers to Figure A1, which has full information on phylogeny.

13) There are many supplemental figures that apply all of the analyses in Figure 1 to publicly available datasets. This is interesting/useful except that there is very, very little discussion of this plethora of results in the text. A little more interpretation on the degree to which these analysis of public data provide unique results and or results that are consistent across studies would benefit the paper.

We have added more discussion of the supplemental data to the main text as suggested.

14) Figure A2, The legend should have more info on what “dissection” “extraction: “NegControl” and “posControl” are showing exactly.

We have explained the nature of these controls in the figure legend as requested.

15) Typos
  a. Middle of Page 8, Bacteroides is misspelled as “Bacteriodes”

Corrected.

  b. Page 13 – line 5 “carried our a species-area analysis of over 1100”… -insert the word “of”

Corrected.

Comments from Dr. XXX

We are gratified that Dr. XXX felt “This will no doubt be of interest to many readers of mBio.”

As suggested, we have added more on toothed versus filter-feeding whales, thus bringing in more discussion of the supplementary material.

p.5: We added “rRNA” as suggested.

p. 5: We added “other processes” as suggested.

p.9: Reworded as suggested.

p.9: Reworded to make consistent.

p.11: The misspelling is corrected.

p.12: We have reworded to clarify the distinction between filter-feeding and toothed (carnivorous) whales.

p.14: The spelling mistake has been corrected.

p.16: We reworded to reduce redundancy as suggested.

p.16: We clarified that the indicated sentence is new information.

p.19: The question about the DNA purification protocol has been addressed in the revised text.

We again thank the reviewers for their thoughtful comments on our study, and hope the paper is now suitable for publication in mBio.

Sincerely,

Rick Bushman

Acknowledgements

Thanks to Arwa Abbas, Thomas D. Bushman, Sara Cherry, Eric Clarke, Yvonne Patterson and Scott Sherrill-Mix for helpful comments on this manuscript.

14. Selected References

1 Watson, J. D. & Crick, F. H. Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. *Nature* **171**, 737-738 (1953).

2 Sauer, R. T., Pabo, C. O., Meyer, B. J., Ptashne, M. & Backman, K. C. Regulatory functions of the lambda repressor reside in the amino-terminal domain. *Nature* **279**, 396-400, doi:10.1038/279396a0 (1979).

3 Goodman, S. D. & Nash, H. A. Functional replacement of a protein-induced bend in a DNA recombination site. *Nature* **341**, 251-254, doi:10.1038/341251a0 (1989).

4 Ptashne, M. On learning to write. *Current biology : CB* **17**, R394-395, doi:10.1016/j.cub.2007.04.013 (2007).

5 Goodman, S. D., Nicholson, S. C. & Nash, H. A. Deformation of DNA during site-specific recombination of bacteriophage lambda: replacement of IHF protein by HU protein or sequence-directed bends. *Proceedings of the National Academy of Sciences of the United States of America* **89**, 11910-11914 (1992).

6 Bushman, F. D. Host proteins in retroviral cDNA integration. *Advances in virus research* **52**, 301-317 (1999).

7 Hershey, A. D., Burgi, E. & Ingraham, L. Cohesion of DNA Molecules Isolated from Phage Lambda. *Proceedings of the National Academy of Sciences of the United States of America* **49**, 748-755 (1963).

8 Guarente, L., Nye, J. S., Hochschild, A. & Ptashne, M. Mutant lambda phage repressor with a specific defect in its positive control function. *Proceedings of the National Academy of Sciences of the United States of America* **79**, 2236-2239 (1982).

9 Subramanian, S. *et al.* Persistent gut microbiota immaturity in malnourished Bangladeshi children. *Nature* **510**, 417-421, doi:10.1038/nature13421 (2014).

10 Bushman, F. D. & Craigie, R. Activities of human immunodeficiency virus (HIV) integration protein in vitro: specific cleavage and integration of HIV DNA. *Proceedings of the National Academy of Sciences of the United States of America* **88**, 1339-1343 (1991).

11 Minot, S., Wu, G. D., Lewis, J. D. & Bushman, F. D. Conservation of gene cassettes among diverse viruses of the human gut. *PloS one* **7**, e42342, doi:10.1371/journal.pone.0042342 (2012).

12 McKay, D. B. & Steitz, T. A. Structure of catabolite gene activator protein at 2.9 A resolution suggests binding to left-handed B-DNA. *Nature* **290**, 744-749 (1981).

13 Schultz, S. C., Shields, G. C. & Steitz, T. A. Crystal structure of a CAP-DNA complex: the DNA is bent by 90 degrees. *Science* **253**, 1001-1007 (1991).

14 Rossner, M. & Yamada, K. M. What's in a picture? The temptation of image manipulation. *J Cell Biol* **166**, 11-15, doi:10.1083/jcb.200406019 (2004).