frontiers in
PLANT SCIENCE

OPINION ARTICLE
published: 03 June 2014
doi: 10.3389/fpls.2014.00250

A biologist, a statistician, and a bioinformatician walk into a conference room. . . and walk out with a great metagenomics project plan

Ann E. Stapleton*

Biology and Marine Biology, University of North Carolina Wilmington, Wilmington, NC, USA
**Correspondence: stapletona@uncw.edu*

Edited by:
Corné M. J. Pieterse, Utrecht University, Netherlands

Reviewed by:
Roeland Lucas Berendsen, Utrecht University, Netherlands

Keywords: computational resources, metagenomes, metagenomic experimental design, metadata, statistical analysis, optimal experimental design, multivariate data analysis

Reviews of metagenomics analysis often emphasize the interdisciplinary and technical aspects of data analysis (Knief, 2014; Sharpton, 2014). How might these recommendations be implemented for future projects? In this opinion, I provide some areas to consider, especially in experimental design and full-data-cycle planning, and in expertise areas of value to metagenomics projects. This opinion is structured as a hypothetical conversation that reviews state-of-the art in these areas and brings out the various aspects of metagenomics project design.

Let's consider an example project in plant-microbe metagenomics—analysis of microbial metagenome functional genes that predict plant yield differences in a horticultural crop species. I've framed a discussion of key points as a conversation, perhaps at the third or fourth meeting, after participants have described their range of expertise.

Biologist: We are here today because we're all interested in doing a great research project in the rapidly growing area of metagenomics. We've heard about this in human biology and health (Human Microbiome Project Consortium, 2012; Morgan et al., 2013), now we'd like to be sure we think through the research aspects for crop biological systems. Let's consider some biological characteristics, such as homeostasis—resilience to disturbance—and adaptation, as general background. Homeostasis, or robustness, is the ability to respond transiently, and then go back to something that functions like the original measured state. In biology, we usually talk about this in the simplest examples using an X-Y line graph with a peak (Calabrese and Blain, 2005; Paine et al., 2012). For example, responses to plant hormones often show a peak at a certain concentration (Taiz and Zeiger, 2006). For communities of organisms, this is often described as ecological resilience and may be measured at multiple levels of organization. We'd like to understand if resilience is happening and if it is important.

Statistician: There are some interesting statistical implications for defining your important questions as curves. Let's relate this to recent "design-of-experiment" research, which is about how to create the most efficient experiment. For curves, you will need to think about how few points can be used to fit such curves (you will need several amounts from your X and Y axes), and how the replicates should be arranged. . . for example, should there be more replicates on the steep sections of a curve or at the tails. This is an area of research called response surface design. Current approaches in this field include low-dimensional Bayesian (Ryan et al., 2014) and Gaussian models (Harari and Steinberg, 2013).

Biologist: Another biological aspect to consider is adaption, the ability to detect a stimulus after the system stabilized, which is usually graphically illustrated as a step-shaped X-Y plot, with the adaptive process happening in the "step" phase, with the response in the "riser" sections (Lim et al., 2013). So, in my particular plant-microbe research area/model system, I am interested in analyzing metagenome changes that can capture these patterns and determine if they are different in low-yield and high-yield plants.

Statistician: Another applied statistical topic is the effect of assumptions behind various analysis methods, from more classical assumptions of normality to choosing a specific possible distribution as a Bayesian prior. This is especially important to consider as a metagenomic sample is highly multivariate (there are many gene sequences within each sample), and underlying assumptions about distributions will constrain what you can reliably detect. There can be useful information for understanding your biological system in the higher-order correlation and autocorrelation (Gallagher et al., 2014) within the samples, so it is worth spending time thinking about how to incorporate what you already know about your system into your analysis choices.

Bioinformatician: It does no good to have data that you can't analyze in a reasonable time frame! We will need to plan for storage of the raw data and feeding of the raw data into the quality control programs (Knight et al., 2012). How much data and how complex is the analysis going to be?

Statistician: It's quite a balancing act to determine the number of samples. We will need to ensure that we have the resources to do a careful walk-through and thorough testing of the data analysis, with the same seriousness we would use for pilot tests of lab procedures, for various options. For

example, we should locate any existing known-truth data and develop software code to produce known-truth data—this is where we embed a known pattern, such as a particular gene present in large amounts, in a background of other genes. This known-truth generation process is usually called simulation in statistics. We would also want to use the most similar already-available real data for testing of our analysis methods. We want to know as much as possible about accuracy and precision before beginning the experimental data collection.

Biologist: I am hearing that we need to focus the question or we will have a huge number of samples. What kinds of pilot tests can we do that would help us keep the sample numbers low but have the maximum power to make predictions?

Bioinformatician: For processing raw reads before doing statistical tests, we will need to test the options for quality control processing (the parameters); it's important to understand how these work before selecting ranges to test, to avoid wasting time testing things that don't affect the output much and to define how some parameter choices depend on other parameter choices (Zhou and Rokas, 2014). This is a place where the known-truth simulations that were mentioned can be helpful. We will also want to track the current best practice in the field using listservs and web resources (Li et al., 2012), as optimal methods can be updated very quickly.

Statistician: We would want to leverage the multivariate aspects of the data for statistical comparison. Typically I would use R packages for this and I'd like your opinion on the computational feasibility. I would also like more details on preprocessing—how extensive is the data cleaning?

Bioinformatician: We would want to assemble sequences from the reads that come from the sequencing machine to reduce the error and increase the information in each "sequence unit," but there is no single best assembly method; using combinations of methods will increase the computational demand substantially. Another important computational consideration is minimizing the trafficking across the network and doing data transfers efficiently. With large numbers of large samples, we will need to use efficient code for data analysis. If the analysis code is written in R, we need to ensure that certain key parts are in C, determine if high performance computing resources are needed and how most easily access those resources. Running statistical R code on a computing cluster or the national XSEDE resource does not guarantee speedup, so we would need to figure out how to optimize the analysis enough to finish it in an acceptable length of time. Another consideration is how to determine how many times the analysis will be tested/re-run, to decide how to organize the code for re-use.

Biologist: Another aspect of metagenomic sequence data is that it can be considered at multiple levels, with annotations of function that come from sources ranging from ontologies (Ashburner et al., 2000) to literature citations (Raychaudhuri et al., 2009), and can be placed in groups ranging from one annotation per sequence to one annotation category that includes thousands of sub-category sequences.

Statistician: Multilevel, or hierarchical, models can be used to handle data labels that have subgroups like the GO annotations, but they can be computationally challenging to fit. We will need to consider these ways of labeling groups and the resulting constraints on comparing samples as we test different analysis methods, in order to choose models that can handle these types of graphs. Different levels of nesting, correlation and comparisons of sets from different parts of an acyclic graph present challenges, for example (Tryputsen et al., 2014).

Biologist: Let me summarize what I see as the dimensions of data analysis we are considering…experimental design tradeoffs, quality control, model fit, assumptions, and their interactions and dependencies. This certainly requires true collaboration, and we should think about formalizing what we've discussed in high-level systems modeling tools http://insightmaker.com/, (North et al., 2013) to explore the costs and benefits and thus optimize our experimental plan.

Statistician: This kind of high-level modeling is sometimes called decision support, and it certainly could help us convince ourselves and our reviewers and colleagues that we have the best possible experimental plan. We do seem to have a good start on synthesis across our different fields from this conversion and these suggestions.

Bioinformatician: We also need to consider metadata, storage, and classroom or citizen use—it's not just the publication, it's the impact, the reuse as well as citations (Piwowar and Vision, 2013; Roche et al., 2014). In fact, there are people who specialize in this—let's add an information science librarian to the mix to advise us on curation (Whyte and Allard, 2014). Now that all the pieces of a great project are on the table, the whiteboard, and the shared computer files, we can think more about the details for our next project meeting, and we have an excellent background to do superb metagenomic science.

This conversation highlights current recommendations and considerations for efficient metagenomics data collection and data analysis. I recommend that project teams consider these general topic areas and involve experts in all these areas when they next develop project plans.

ACKNOWLEDGMENT

Editorial advice from Dr. Patrick Erwin is much appreciated.

REFERENCES

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., et al. (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* 25, 25–29. doi: 10.1038/75556

Calabrese, E. J., and Blain, R. (2005). The occurrence of hormetic dose responses in the toxicological literature, the hormesis database: an overview. *Toxicol. Appl. Pharmacol.* 202, 289–301. doi: 10.1016/j.taap.2004.06.023

Gallagher, C. M., Fisher, T. J., and Shen, J. (2014). A Cauchy estimator test for autocorrelation. *J. Stat. Comput. Simul.* 1–13. doi: 10.1080/00949655.2013.874424. (in press).

Harari, O., and Steinberg, D. M. (2013). Optimal designs for Gaussian process models via spectral decomposition. *J. Stat. Plan. Inference* doi: 10.1016/j.jspi.2013.11.013. (in press).

Human Microbiome Project Consortium. (2012). Structure, function and diversity of the healthy

human microbiome. *Nature* 486, 207–214. doi: 10.1038/nature11234

Knief, C. (2014). Analysis of plant microbe interactions in the era of next generation sequencing technologies. *Front. Plant Sci.* 5:216. doi: 10.3389/fpls.2014.00216

Knight, R., Jansson, J., Field, D., Fierer, N., Desai, N., Fuhrman, J. A., et al. (2012). Unlocking the potential of metagenomics through replicated experimental design. *Nat. Biotechnol.* 30, 513–520. doi: 10.1038/nbt.2235

Li, J.-W., Schmieder, R., Ward, R. M., Delenick, J., Olivares, E. C., and Mittelman, D. (2012). SEQanswers: an open access community for collaboratively decoding genomes. *Bioinformatics* 28, 1272–1273. doi: 10.1093/bioinformatics/bts128

Lim, W. A., Lee, C. M., and Tang, C. (2013). Design principles of regulatory networks: searching for the molecular algorithms of the cell. *Mol. Cell* 49, 202–212. doi: 10.1016/j.molcel.2012.12.020

Morgan, X. C., Segata, N., and Huttenhower, C. (2013). Biodiversity and functional genomics in the human microbiome. *Trends Genet.* 29, 51–58. doi: 10.1016/j.tig.2012.09.005

North, M. J., Collier, N. T., Ozik, J., Tatara, E. R., Macal, C. M., Bragen, M., et al. (2013). Complex adaptive systems modeling with repast simphony. *Complex Adapt. Syst. Model.* 1, 3. doi: 10.1186/2194-3206-1-3

Paine, C. E. T., Marthews, T. R., Vogt, D. R., Purves, D., Rees, M., Hector, A., et al. (2012). How to fit nonlinear plant growth models and calculate growth rates: an update for ecologists. *Methods Ecol. Evol.* 3, 245–256. doi: 10.1111/j.2041-210X.2011.00155.x

Piwowar, H. A., and Vision, T. J. (2013). Data reuse and the open data citation advantage. *Peer J.* 1, e175. doi: 10.7717/peerj.175

Raychaudhuri, S., Plenge, R. M., Rossin, E. J., Ng, A. C. Y., Purcell, S. M., Sklar, P., et al. (2009). Identifying relationships among genomic disease regions: predicting genes at pathogenic SNP associations and rare deletions. *PLoS Genet* 5:e1000534. doi: 10.1371/journal.pgen.1000534

Roche, D. G., Lanfear, R., Binning, S. A., Haff, T. M., Schwanz, L. E., Cain, K. E., et al. (2014). Troubleshooting public data archiving: suggestions to increase participation. *PLoS Biol* 12:e1001779. doi: 10.1371/journal.pbio.1001779

Ryan, E. G., Drovandi, C. C., Thompson, M. H., and Pettitt, A. N. (2014). Towards Bayesian experimental design for nonlinear models that require a large number of sampling times. *Comput. Stat. Data Anal.* 70, 45–60. doi: 10.1016/j.csda.2013.08.017

Sharpton, T. J. (2014). An introduction to the analysis of shotgun metagenomic data. *Front. Plant Sci.* 5:209. doi: 10.3389/fpls.2014.00209

Taiz, L., and Zeiger, E. (2006). *Plant Physiology*. Sunderland, MA: Sinauer Associates.

Tryputsen, V., Cabrera, J., de Bondt, A., and Amaratunga, D. (2014). Using Fisher's method to identify enriched gene sets. *Stat. Biopharm. Res.* doi: 10.1080/19466315.2014.888013. (in press).

Whyte, A., and Allard, S. (2014). '*How to Discover Research Data Management Service Requirements,' DCC How-to Guides*. Edinburgh: Digital Curation Centre. Available online at: http://www.dcc.ac.uk/resources/how-guides. See more at: http://www.dcc.ac.uk/how-discover-requirements#sthash.yhj8EobY.dpuf

Zhou, X., and Rokas, A. (2014). Prevention, diagnosis and treatment of high-throughput sequencing data pathologies. *Mol. Ecol.* 23, 1679–1700. doi: 10.1111/mec.12680

Conflict of Interest Statement: The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 15 April 2014; accepted: 15 May 2014; published online: 03 June 2014.

Citation: Stapleton AE (2014) A biologist, a statistician, and a bioinformatician walk into a conference room… and walk out with a great metagenomics project plan. Front. Plant Sci. ***5****:250. doi: 10.3389/fpls.2014.00250*

This article was submitted to Plant Genetics and Genomics, a section of the journal Frontiers in Plant Science.