What's all the hype about HybSeq?

A brief history and introduction to target enrichment in Compositae

Jennifer R. Mandel[1]

[1] Department of Biological Sciences, University of Memphis, Memphis, TN 38152, *jmandel@memphis.edu*

Keywords: Target enrichment, high-throughput sequencing, phylogenomics, phylogenetics.

INTRODUCTION

The year was 2010, and the place was the Compositae White Paper meeting in Davis, California, USA. Members of the National Science Foundation (NSF-USA) funded Compositae Genome Project (CGP; Figure 1) had convened in northern California, in the land of grapevines, to present about their progress in developing tools and studying the genomes of comps (crops). It is still unclear how Dr. Vicki Ann Funk caught wind of this meeting at Davis, but she was intent on attending and talking with (educating) this growing group of genomicists studying sunflowers, lettuce, and their relatives, and convincing the attendees that they needed to look beyond crops of the family. Vicki brought along with her a six foot plus (*ca.* 2 m)poster with the Compositae metatree of Funk et al. (2009) (Figure 2). She posted it outside in the lobby where those in attendance could not help but see the vast diversity and beautiful possibilities of expanding their genomic knowledge and tools to study the evolution of Compositae.

During that trip, Vicki met with Dr. Loren Rieseberg (University of British Columbia) and Dr. John Burke (University of Georgia) who already had extensive experience studying evolution of members of the family. At that 6AM breakfast, I (a postdoctoral researcher with Burke at the time) was asked to come along. Vicki had an idea: apply this "next-gen" thing to studying the systematics and evolution of Compositae. Studies using ITS/plastid markers left many gaps in our understanding of the evolutionary relationships in the family: it was time for more data and a new approach. The three of them, Loren, John, and Vicki, hatched a pilot project and I would be the hands at the bench and the keyboard. The rest is sort of history (Mandel et al. 2014; 2015; 2017; 2019), well, after a lot of lab trouble-shooting and banging of head on keyboard for the bioinformatics (even with the help of the wiz Dr. Michael McKain, University of Alabama).

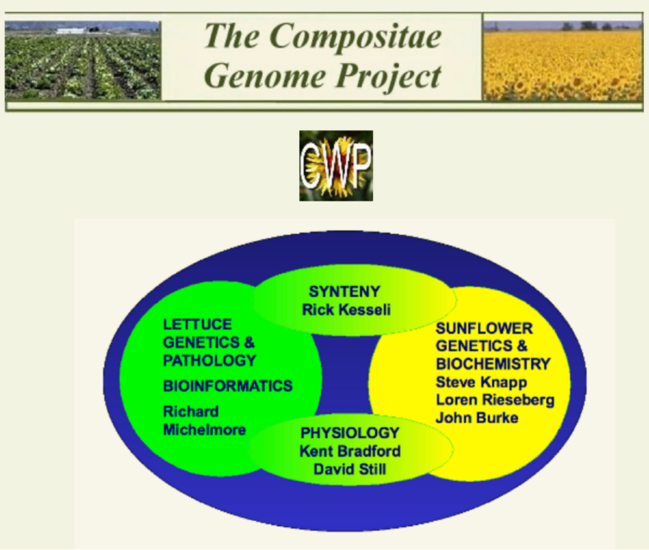

Figure 1. Research facets on the Compositae Genome Project webpage and its Principal Investigators from circa 2000s.

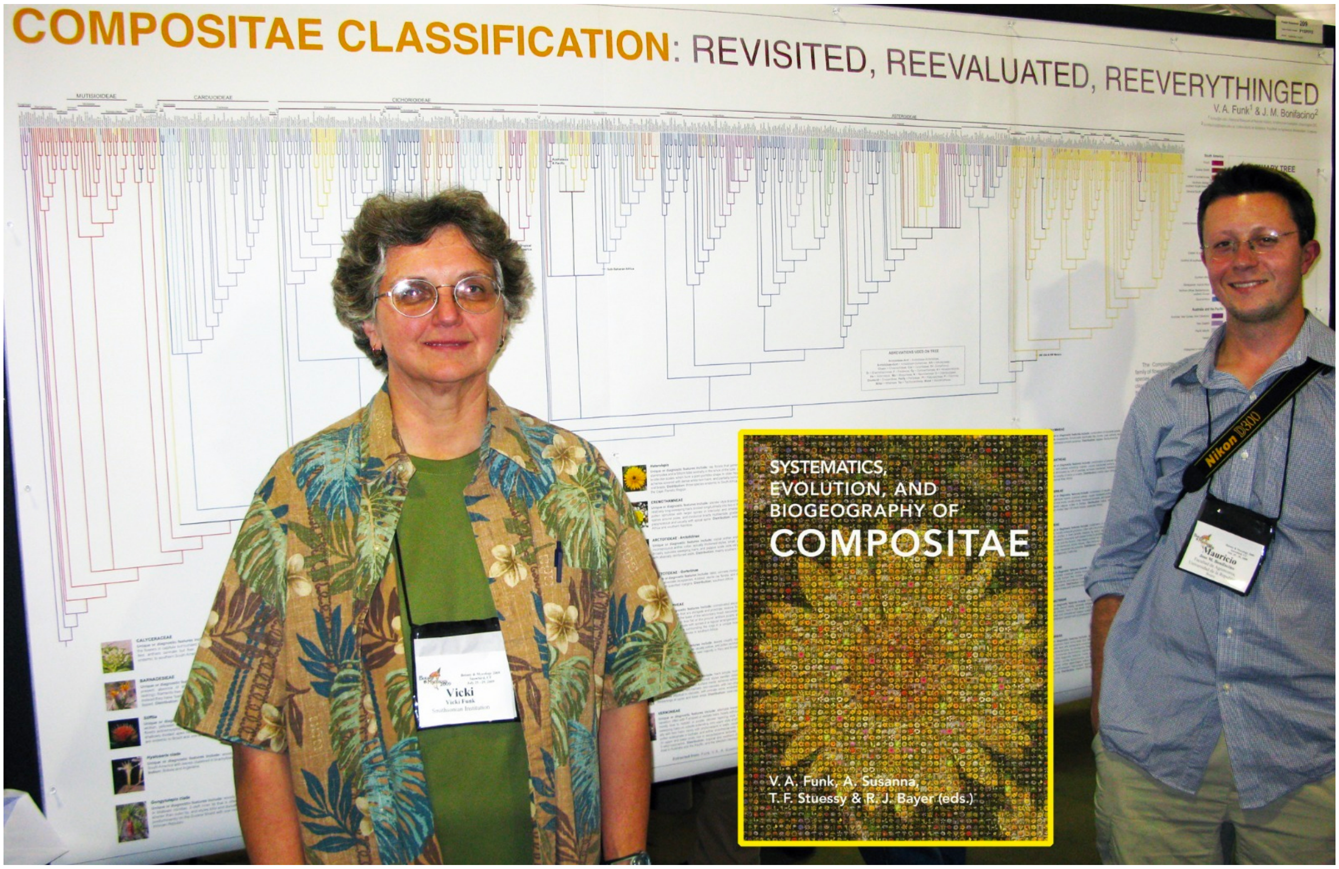
COMPOSITAE CLASSIFICATION: REVISITED, REEVALUATED, REEVERYTHINGED
V. A. Funk & J. M. Bonifacino
SYSTEMATICS, EVOLUTION, AND BIOGEOGRAPHY OF
COMPOSITAE
V. A. Funk, A. Susanna, T. F. Stuessy & R. J. Bayer (eds.)
Vicki
Mauricio
Nikon

Figure 2. Dr. Vicki Funk with Dr. Mauricio Bonifacino and the huge Compositae metatree.

During this time, Vicki introduced Jennifer to Dr. Rebecca Dikow (Research Data Scientist and leader of the Smithsonian Institution Data Science Lab) and the three had a wonderful collaboration through this project and Jennifer and Rebecca continue to collaborate on genomics projects in Compositae.

The great promise of employing big data through HybSeq (aka: target enrichment, sequence capture; Weitemier et al. 2014) to questions of evolutionary relationships are still being seen. This summer, a special issue in APPS (Applications in Plant Sciences) is devoted to "Exploring the Potential of Angiosperms353, a Universal Toolkit for Flowering Plant Phylogenomics" and will highlight the possibilities of phylogenomics with this toolkit for flowering plants. Along with Dr. Carolina Siniscalchi and members of Royal Botanic Gardens at Kew, we present a comparison of the broad toolkit with the Compositae specific kit in, "Lineage-specific vs. universal: comparison of the Compositae1061 and Angiosperms353 enrichment panels in the sunflower family" (Siniscalchi et al. 2021).

The objective of this article is to break down the HybSeq process for the novice, wishing someone had done this for us when we started!

Into the details of HybSeq

Let's start with the Seq: In 2010, it was clear that massively parallel or so-called next-generation sequencing (NGS) methods were going to radically change the way evolutionary biologists (well everyone in science) would be able to approach research questions. The concept of NGS is that high-throughput sequencing technology can be used to determine the base-pair sequence of DNA/RNA molecules at much larger quantities than previous end-termination based sequencing techniques (e.g. Sanger sequencing). See Ekblom and Wolf's (2014) fantastic "A field guide to whole-genome sequencing, assembly and annotation" for more details on methods, approaches, and definitions. Briefly, NGS involves sequencing millions of small fragments of DNA (or RNA converted to cDNA) in parallel. While there are several platforms

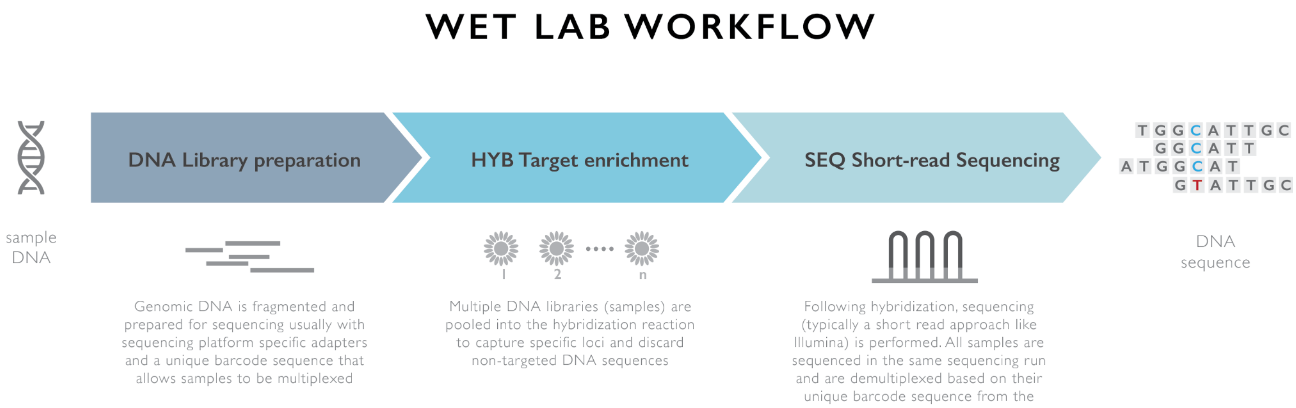
WET LAB WORKFLOW
DNA Library preparation
HYB Target enrichment
SEQ Short-read Sequencing
TGGCATTGC
GGCATT
ATGGCAT
GTATTGC
sample DNA
1
2
n
DNA sequence
Genomic DNA is fragmented and prepared for sequencing usually with sequencing platform specific adapters and a unique barcode sequence that allows samples to be multiplexed
Multiple DNA libraries (samples) are pooled into the hybridization reaction to capture specific loci and discard non-targeted DNA sequences
Following hybridization, sequencing (typically a short read approach like Illumina) is performed. All samples are sequenced in the same sequencing run and are demultiplexed based on their unique barcode sequence from the

Figure 3. Overview workflow of the HybSeq workflow inlcluding DNA library preparation, hybridization of targeted sequences, and high-throughput sequence of captured, or enriched, DNA.

that can accomplish this, the most widely used in studies of phylogenetics and evolution is the Illumina platform. Illumina is ever expanding its sequencing technologies and capabilities, but a few staples are the MiSeq, HiSeq, and more recently the NovaSeq. Sequencing reactions using the Illumina technology take place on the surface of a glass semiconductor (just bigger than a microscope slide) etched with patterned nanowells where the DNA sequencing reactions will occur, called a flow cell. After DNA is extracted, or isolated, from a tissue, the DNA is prepared for NGS sequencing through a library preparation step that involves randomly shearing the DNA and incorporating specific sequences that allow DNA to adhere to the sequencing flow cell and the sample to be identified. And while fresh tissue usually ends up providing the best library, silica dried, frozen, and herbarium material can be used for successful library preparation and sequencing. Library preparation methods vary in how much time they take to complete, but usually can be performed in 4-6 hours for a single sample and scale nicely as with some training, 96 samples (in plates) could easily be done in a full day. The resulting libraries are quality checked, quantified, and sequenced on the machine. The great part about the library prep step is that you can add a unique barcode, or index, of 6 or more nucleotides to label your

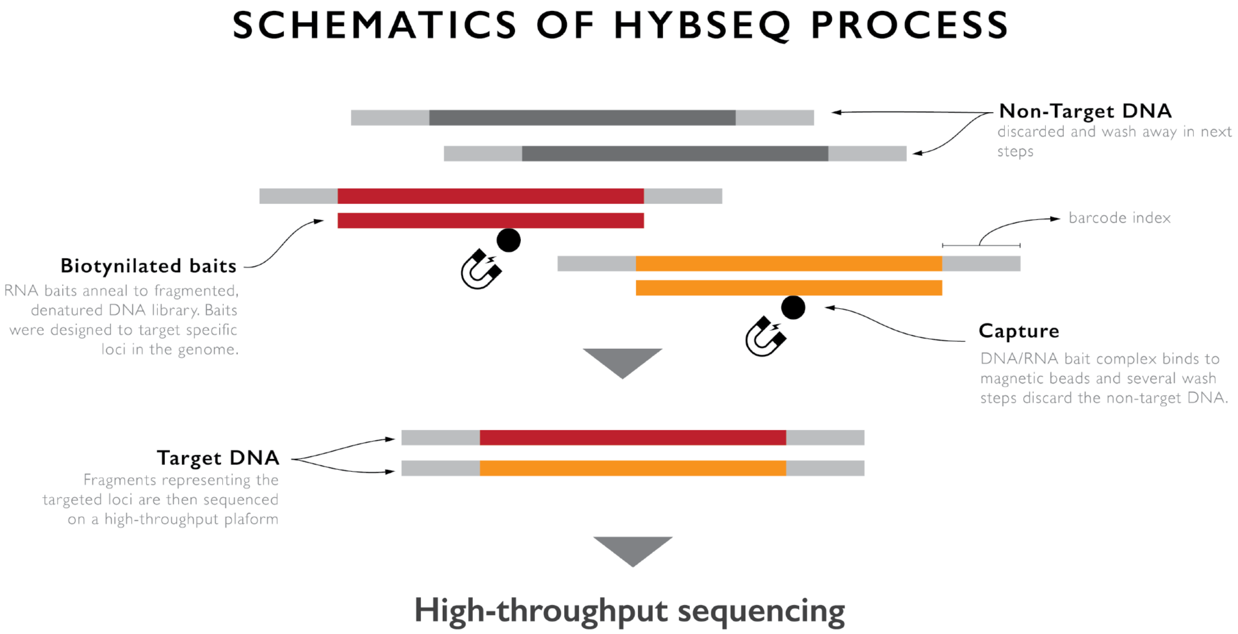
SCHEMATICS OF HYBSEQ PROCESS
Non-Target DNA
discarded and wash away in next steps
barcode index
Biotynilated baits
RNA baits anneal to fragmented, denatured DNA library. Baits were designed to target specific loci in the genome.
Capture
DNA/RNA bait complex binds to magnetic beads and several wash steps discard the non-target DNA.
Target DNA
Fragments representing the targeted loci are then sequenced on a high-throughput plaform
High-throughput sequencing

Figure 4. Overview of the hybridization process where targeted genomic sequences are captured using RNA biotynilated baits designed by the user. Non-targeted DNA is discarded during wash steps and the resulting product is a reduced representation of the total genome for sequencing.

individual DNA sample, e.g., *Helianthus annuus* L. might get GACAATTC and *H. verticillatus* Small could get CAACCGAG. The barcode is added near the ends of every single fragment you will sequence. This allows you to pool many samples together on the same flow cell, called multiplexing. Then, the barcode is sequenced along with the rest of the fragment, allowing you to determine which sequenced fragments came from which sample. When the sequencing run finishes, software from the instrument sorts each sequenced fragment into its own directory/folder based upon that barcode, so all *H. annuus* fragments are together in your files.

So, that's the Seq part of HybSeq, what about the Hyb part? The premise here is that instead of sequencing every fragment of DNA from your sample, we target only specific genes. Here, the DNA regions of interest (e.g., 1061 loci for phylogenetics from the Comp1061 MyBaits Kit) are captured or enriched in the sample using long oligonucleotide baits, or probes. The baits are designed to have complementarity to your genes of interest and can be generated de novo (custom captures) or using a pre-designed kit (e.g., Angiosperm353, Johson et al. 2019; Comp1061). The baits are biotinylated, so they bind to streptavidin-coated magnetic beads; the hybridization is performed in liquid solution in a microcentrifuge or smaller tube. Essentially your DNA (the library you made) sticks to baits, DNA/baits complex sticks to magnetic beads, then you put the tube with the DNA/baits/bead in a magnetic rack stand, wash away all of the DNA fragments that aren't bound to the beads, and elute off the beads, and you've captured your targeted DNA! Since the DNA was randomly sheared during the library prep step, capture fragments overlap and are unique. Most approaches PCR this targeted sample (because it is a very small amount) and then send it for sequencing. That's it! Thanks for reading!

ACKNOWLEDGMENTS

First and foremost, I thank the late Dr. Vicki Ann Funk for her willingness to mentor and work with me for so many years. Her generous and inclusive spirit meant that I was always part of the conversation. It was her idea for using big data to address the evolution of Compositae and I am grateful for her taking me along on the ride. I also want to thank countless others over the years who have helped us out with the lab work, the bioinformatics, the writing, and the interpretation. Team Compositae is the best!

LITERATURE CITED

Andermann, T., Torres Jiménez, M.F., Matos-Maraví, P., Batista, R., Blanco-Pastor, J.L., Gustafsson, A.L.S., Kistler, L., Liberal, I.M., Oxelman, B., Bacon, C.D. & Antonelli, A. 2020. A guide to carrying out a phylogenomic target sequence capture project. *Front. genet.* 10: 1407.

Ekblom, R. & Wolf, J.B. 2014. A field guide to whole-genome sequencing, assembly and annotation. *Evol. App*. 7: 1026–1042.

Johnson, M.G., Pokorny, L., Dodsworth, S., Botigue, L.R., Cowan, R.S., Devault, A., Eiserhardt, W.L., Epitawalage, N., Forest, F., Kim, J.T. & Leebens-Mack, J.H. 2019. A universal probe set for targeted sequencing of 353 nuclear genes from any flowering plant designed using k-medoids clustering. *Syst. Biol.* 68: 594–606.

Mandel, J.R., Dikow, R.B., Funk, V.A., Masalia, R.R., Staton, S.E., Kozik, A., Michelmore, R.W., Rieseberg, L.H. & Burke, J.M. 2014. A target enrichment method for gathering phylogenetic information from hundreds of loci: an example from the Compositae. *App. Plant Sci.* 2(2): 1300085.

Mandel, J.R., Dikow, R.B., & Funk, V.A. 2015. Using phylogenomics to resolve mega-families: an example from Compositae. *J. Syst. Evol.* 53: 391–402.

Mandel, J.R., Barker, M.S., Bayer, R.J., Dikow, R.B., Jones, K., Keeley, S., Siniscalchi, C.M., Susanna, A., Thapa, R., Watson, L.E., & Funk, V.A. 2017. The Compositae Tree of Life in the age of phylogenomics. *J. Syst. Evol.* 55: 405–410.

Mandel, J.R., Dikow, R.B., Siniscalchi, C.M., Thapa, R., Watson, LE., & Funk, V.A. 2019 A fully resolved backbone phylogeny reveals numerous dispersals and explosive diversifications throughout the history of Asteraceae. *Proc. Nat. Acad. Sci. USA* 116: 14083–14088.

Weitemier, K., Straub, S.C., Cronn, R.C., Fishbein, M., Schmickl, R., McDonnell, A. & Liston, A. 2014. Hyb-Seq: Combining target enrichment and genome skimming for plant phylogenomics. *App. Plant Sci.* 2: 1400042.