Sequencing *Medicago truncatula* expressed sequenced tags using 454 Life Sciences Technology

Cheung F et al (2006) BMC Genomics 7:272

This article uses similar techniques as the *Gunnera* cDNA sequencing project (though, as we'll see, towards a very different goal), and it will be instructive to see how they have been used. Like any research article, this one is confusing to those outside the field, and so I'll accompany you as you read it. I like to ask questions of myself and of the article as I go along, and I hope you'll do the same. However, since I'm on tape, you'll be able to hear my questions, but I won't be able to hear yours. Very irritating! If you ask but find no answer, please write down the question and send it to me.

Abstract

I generally give the abstract about 30 seconds to say something interesting, and if nothing happens within that period, I toss it and move on. Understand that authors are under severe space limitations in writing abstracts, and they usually want to cram in lots of details so Google and similar will have something to work with. As a result, abstracts are often incomprehensible. Well, that's about 30 seconds. I'm gone.

Background

My first goal is to grasp the big question the authors seek to address. I must say, they don't make it easy on us, presuming a good deal of prior knowledge. Let's examine a few of the terms that look to be of critical importance:

- *Medicago*… what's that? You could easily find out, but I'll save you the trip. It's alfalfa, an important crop plant. Humans don't eat a lot of alfalfa, but it's a major component of animal feed.
- Expressed Sequence Tags (ESTs)… Never mind the historical reasons for this term. In brief ESTs are genome sequences that are transcribed from genes to mRNA. This is generally a small fraction of the genome, but it is the fraction that people are most interested in.
- cDNA… (introduced in paragraph 3) DNA made in the lab that is ***c***omplementary to mRNA present in a cell. It's easier to work with DNA than RNA, so the isolated mRNA is reverse-transcribed into DNA. ESTs are sequences obtained from cDNAs. Full-length cDNAs, extending from the beginning of the mRNA transcript to the poly-A tail marking its end, range from several hundred to several thousand nucleotides in length.
- Library normalization… cDNA libraries (collections of cDNAs) generally have lots of copies of cDNAs from genes that are highly expressed and few copies from genes that are expressed at a low level. Some very important genes are expressed at a low level – many regulatory proteins don't need to be expressed to a high level in order to be effective – and the authors have taken the trouble to reduce the fraction of the library taken up by high copy cDNAs.
- Deep sequencing… Through normalization and brute force, sequencing to so great an extent that a significant number of rare sequences are sampled.

- Spliced alignment / gene structure… Plant genes often have introns – extra DNA within genes – that are spliced out during the process of mRNA maturation. cDNA (derived from mRNA) therefore lack introns. A comparison of cDNA and genome sequences can be used to identify the extent of introns, which they call the gene structure. It is often not easy to predict from a genomic sequence where an exon stops and an intron begins without the help of cDNA.
- SSR… Simple sequence repeat. This term is unaccountably not defined until well into the **Results** section. AGAGAG… is an example of a simple sequence repeat.

SQ1. What is the big goal into which this work fits? (Note: the authors don't state a big goal, presuming (probably correctly) that most readers of BMC Genomics won't need a reminder)

SQ2. What did the authors hope to learn from their work?

My second goal is to get a sense of what the authors actually did. This is often found in the last paragraph of the **Introduction** (sometimes called **Background**) section, so I fast forward there. I learn that the authors want to know whether the protocol they describe can lead to the discovery of predicted genes not found in an existing database. For this question to make any sense, we have to know what the existing database is about. I ***presume*** that the 226,923 ESTs are the result of conventional (Sanger) sequencing, with read lengths in the several hundred nucleotides range. It is disappointing that even after going to references 13 and 14 I don't know for sure.

Never mind. I came to this article primarily to learn what they did, with less interest in why they did it.

Results and Discussion

cDNA library production

Most of this section concerns itself with a characterization of the cDNA library using conventional sequencing. I'm going to blip over all this. There are two points worth making, however. The construction of the cDNA library (detailed in the **Methods** section) is important to understand what kinds of things can go wrong, but I'll defer that topic for face-to-face discussion. The other issue is:

SQ3. I note that the authors used pooled RNA. Given their purposes, why did they choose to do this? Why is the Gunnera project *not* using pooled RNA but rather getting it from a single source?

454 sequencing

SQ4. The authors give the number of reads, their average length, and the total length. Do these numbers hang together?

Most of this section concerns itself with the presence in the library of 5' and 3' ends of the original cDNA sequences. We'll discuss face-to-face the exact nature of directional adapters, but even without deep insight into this, perhaps you can see why most 454 reads don't contain them.

SQ5. Why is it that some 454 reads contain adapters specific for the 5' end or 3' end of the cDNA but most contain no adapters?

454 cDNA sequence assemblies

The first step after sequencing – no matter what method is used – is to get rid of adapters and other unwanted sequences. Only then is it possible to assemble the reads into larger entities. Assembly is achieved by comparing reads, looking for regions of overlap that would indicate that the reads came from the same cDNA. Assembling many overlapping reads may allow the reconstruction of the original cDNA.

SQ6. Why are adapter sequences unwanted? How would they interfere with the assembly process?

SQ7. Make holistic sense out of the numbers given for the number of high quality reads, the number of reads incorporated into assemblies, the number of assemblies, and the number of singletons. If one of these numbers were changed, what other numbers would need to be changed?

SQ8. The authors refer to Table 2 to justify the statement that most assemblies are short. Does the evidence convince you? Why is it that the total under Reads does not equal the total under Post-Assembly?

That total number of 29 million is a lot of nucleotides! But is it enough? Much of this section addresses the question. The ***coverage*** of a set of target sequences (in this case the transcriptome, i.e. the set of all transcripts) is the total number of nucleotides sequenced divided by the total number of nucleotides in the target set.

SQ9. Calculate the coverage yourself. The authors give the answer as 0.28. Are they right?

SQ10. No special knowledge of statistics is required to predict the number of reads that are singleton, but most would find this calculation challenging. Why not give it a try? Hints: (1) Presume that all reads are the same length as the average. (2) Ask yourself how likely it is that one such random read overlaps with another such random read.

And the rest…

The rest of the article takes us beyond the aspects I think we will be able to cover, so I'll stop here.