Project 4977 – ConGenOmics Short Visit Grant Report

Quantifying epigenetic influence on phenotype and epigenetic profile changes under stress

Purpose of Visit

During my 6-day visit to the Netherlands Institute for Ecology, Wageningen (14th-20th January 2013), I collaborated with Thomas van Gurp. We made excellent progress towards our aim of developing a workflow to quantify DNA methylation differences between individuals using next-generation sequencing. In brief, we started with paired-end sequence reads in the same individual digested with two isoschizomer enzymes, and moved towards quantifying differences in fragment presence/absence that correspond to differences in methylation state. I learnt a great deal from Thomas about bioinformatic workflow design, software, and UNIX and Python programming, and I introduced him to the flexibility and utility of R as a programming language and graphics tool. I also gave a seminar to the Institute about my current work on evolutionary shifts to selfing and phylogeographic history of *Arabidopsis lyrata*, and met with Dr. Koen Verhoeven and other researchers in ecology and ecological genetics fields.

Work Carried Out

I brought with me a dataset of approx. 600 million Illumina sequencing reads from a Restricted Representation Library sequenced using paired-end sequencing in two flow cell lanes. This included 13 barcoded individuals that had been digested in parallel with MspI and HpaII, isoschizomeric restriction enzymes that have the same recognition site but differ in methylation sensitivity. Thomas and I examined the raw data and the barcode-deconvoluted data to check quality. Using command-line tools, we replaced read numbers with informative individual + enzyme names, and mapped all reads to the *A. lyrata* reference genome using BWA.

We then wrote a Python script to perform the following steps:

1. select only uniquely-mapped mate pairs

2. check for restriction site at the start of the forward read and the end of the reverse read and exclude mate pairs without it
3. classify each mate pair by start and end position, and length
4. determine the location of internal restriction sites in the reference sequence for each fragment
5. output the information as a table (Table 1)

This entire workflow was uploaded into Galaxy, for future replicability and to allow ongoing data and workflow sharing.

Table 1: Example of output table from Python script. Read counts are shown in italics for each sample at each genome-mapped fragment detected in the entire dataset

Fragment No.	Chromosome	Start position	End position	Length (bp)	Internal CCGG	Sample1_MspI	Sample1_HpaII	Sample 2_MspI	Sample2_HpaII
1	1	1049	1700	651		*9*	*38*	*10*	*0*
2	1	5890	6151	261	192	*0*	*0*	*0*	*50*
3	1	14088	14523	435		*35*	*23*	*19*	*12*
4	1	14524	14699	175		*5*	*20*	*1*	*25*
5	1	30001	30331	330	237, 300	*1*	*1*	2	*1*
6	1	33221	33612	391		*0*	*73*	*38*	*37*
7	1	62754	62897	143		*15*	*34*	*28*	*43*

We then began further analysis of this output table in R (The R Foundation for Statistical Computing 2012).

Main Results Obtained, Projected Publications and Future Collaboration

We looked at the depth of coverage across the mapped fragments, identifying several highly-sequenced chromosome regions (e.g. the rRNA multi-copy region) that will probably be analysed separately. Coverage per individual was useful for our purposes and also for SNP calling, though despite experimental attempts to maintain equal template amounts per individual, we still detected up to 10-fold variation in mapped read counts, probably simply due to currently-unavoidable stochastic factors and measurement inaccuracies. Another important experimental finding was as follows. Although the RRL library preparation included a size-selection step that targeted 300-500 bp fragments only, a high proportion of fragments obtained were smaller than 300 bp. We suggest this is due to smaller fragments forming sequencing clusters on the Illumina chip more readily. This knowledge is useful for ongoing library preparation in Yvonne Willi's group (Evolutionary Botany Lab, Université de Neuchâtel), as it will affect the overall coverage predictions and so affect the optimum number of individuals pooled in a single library.

By examining the fragment count output together with the reference-mapped fragment display, we found several issues that required troubleshooting at the data processing level. We discovered that adapter overhangs in short fragments were inflating the fragment count, since reads with different end points were counted as different fragments. Near the end of the week, we also discovered some discrepancies that seemed to be due to incorrect barode deconvolution. To verify this, and to perform better adapter trimming, we are currently reprocessing the entire raw dataset with an alternative barcode deconvolution / renaming method and including adapter trimming and quality control.

To proceed with our aim of looking at methylation differences, we selected all fragments with no (or very few) reads with one enzyme, but many reads in the other, for the same individual. By looking at these fragments and their close neighbours (if present), we could distinguish probable methylation at some sites. One important initial observation is that heterozygosity in methylation state seems quite common. This can be detected when a long HpaII fragment with an internal CCGG site is found (implying the internal CCGG site is methylated) *together* with internal fragments (implying the internal CCGG is unmethylated); whereas the long fragment is not produced using MspI. Methylation-sensitive AFLP (MS-AFLP) studies would interpret this pattern (band presence with HpaII digestion but absence with MspI digestion) as evidence of a methylated internal C in the restriction site (e.g. Salmon *et al.* 2008; Richards *et al.* 2012), which would obviously be incorrect in many cases according to our results. We see this as an important and highly publishable result that will be the main focus of our first publication on this topic.

We also discussed and planned the next steps for this analysis. Namely:

1. including in the Python script a check in the read (where possible) the presence of internal restriction sites found in the reference sequence
2. re-running the script with the newly processed data for optimum accuracy
3. deciding how to treat fragments < 300 bp (or whether the coverage will be biased for this small size range)
4. quantifying a threshold of certainty for assigning read presence/absence, taking into account the mean and variance in overall coverage for that individual/enzyme combination

In summary, this visit was an extremely valuable contribution to our ongoing collaboration. We planned and carried out initial steps in the analysis of next-generation sequencing data. Once the workflow rerun has finished, we will be able progress further in quantifying the utility of this method to conservation genetics studies in comparison to current alternatives like methylation-sensitive AFLP. We expect to have a manuscript ready for submission on this topic in late 2013, so our collaboration will continue at least until this point. In addition, I am currently applying for further funding with the aim of applying this sequencing analysis method to experiments I have performed that investigate heritability of DNA methylation and its inducibility in plants under drought stress.

References

Richards CL, Schrey AW, Pigliucci M (2012) Invasion of diverse habitats by few Japanese knotweed genotypes is correlated with epigenetic differentiation (M Vellend, Ed,). *Ecology Letters*, **15**, 1016–1025.

Salmon A, Clotault J, Jenczewski E, Chable V, Manzanares-Dauleux MJ (2008) Brassica oleracea displays a high level of DNA methylation polymorphism. *Plant Science*, **174**, 61–70.

The R Foundation for Statistical Computing (2012) R version 2.15.2.