Experimental Biology Laboratory: Prof. Shimizu Laboratory

Damianos Melidis
dmelidis@student.ethz.ch

4th July, 2013

Abstract

During this laboratory rotation, I learned how to identify the size of specific gene using both genomic DNA and by reverse transcription of RNA to cDNA. We amplified these genes using PCR and reverse PCR respectively. I have also become familiar with the Next Generation Sequencing data. We try to apply this knowledge to contribute in the improvement of reconstruction of two closely related subspecies of Arabidopsis thaliana

Part1: Experimental Biology

A very familiar way to amplify a gene of interest among whole genome, is the PCR (Polymerase Chain Reaction) [Saiki, R.; Gelfand, D.; Stoffel, S.; Scharf, S.; Higuchi, R.; Horn, G.; Mullis, K.; Erlich, H.]. In this method we exploit the DNA polymerase's ability to form a new DNA strand complementary to the given *template* strand. The specificity of gene production is accomplished by the two smaller DNA oligonucleotides, the forward and reverse primer, as they guide DNA polymerase where to start and end DNA synthesis.
In more details firstly we heat up the solution (94-98℃) in order to break the hydrogen bonds of the double-stranded DNA resulting to single-stranded DNA (denaturation step). The next step is the annealing step where the temperature is decreased to 50-65℃, to facilitate the binding of primares on the single-stranded DNA. Afterwards the polymerase synthensizes the complementary of each single-stranded DNA, only if it is binded with the primers. This step is known as extension/elongation and its temperature depends on the polymerase's type (usually 72℃). To remove polymerase from single-stranded DNAs we need to increase the temperature to 94℃. After this step we create the template strand and its complementary one (2 DNA strands in total). If we cycle through the previous steps, the resulted DNA strands are 4. The third time it will produce 8 DNA strands. We explpoit this exponential growth of DNA strands production to get sufficient quantity of our desired gene. After cycling these steps 35 times, we let the solution on 72 for 5 minutes to help the potential remaining synetheses to be completed (final elongation step). The last step is to cool or freeze our solution on 4 or -20 respectively.
We need this method to produce sufficient amount of the investigated gene to be further analyzed, as we know/suspect only the gene's length and not all its base pairs. In the figure 2 we can see the PCR work flow(`http://en.wikipedia.org/wiki/File:PCR.svg`):

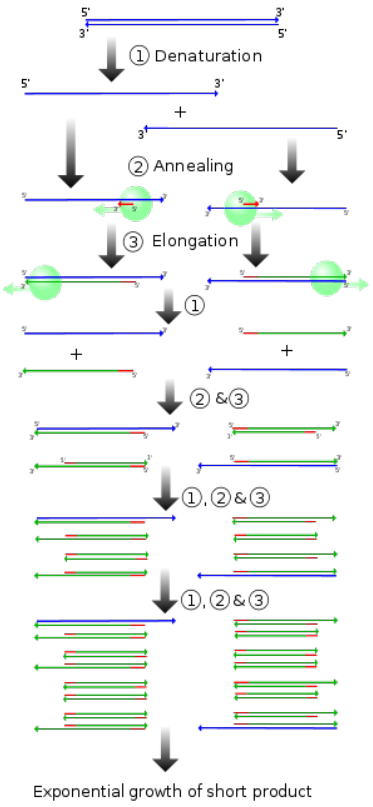

Figure 1: PCR steps in theory

In more details the PCR reaction is composed by the next steps:

1. Denaturation: Denaturate your DNA template at 94℃, to separate the double strand DNA form.
2. Annealing: Facilitate the primers attachment to desired genome region by cooling to 55℃.
3. Extension/elongation: Facilitate DNA synthesis by the DNA polymerase heating up to 72℃.
4. Second Denaturation: Heat up your solution to remove DNA polymerase and primers from the DNA strands also to inhibit polymerase to continue sunthesizing.
5. Go to step 1, until you get sufficient amount of replicated desired gene (usually 35 cylces needed).
6. Final elongation: Heat to 72℃, giving the chance to all remaining replication to be completed
7. Final hold: Cool the PCR production at 4℃, or freeze it at -20℃

In the laboratory assuming that we would like to create X samples for PCR, we design the experiment like we need X+0.5(or +1.0) samples to overcome human and pipette error. For X+0.5 samples we calculate the appropriate portion of desired polymerase (Taq), of the specifical forward and reverse primers for the wanted gene and of the nucleotides to be used by polymerase. We also use the corresponding portion of water and a buffer solution (to help attachment of the enzyme and small DNA fragments to the genome). We add all these substances to a 'master mix' tube. After dividing this tube to X smaller tubes, we add the specific DNA or cDNA *template* on them. As final step we place the X tubes to the PCR machine. When we used RNA to create cDNA we had to clean our desk properly and also save the RNA in ice or on the freezer.
If we need to understand how the DNA is expressed in different tissues of a organism we can extract RNA and use the reverse transcriptase to synthensize the corresponding DNA (with only

its exons). Then with the resulted DNA, often called cDNA, we can run PCR and examine the difference amount of desired gene expression in the tissues.
After the PCR reaction, we need to perform electrophoresis [1] to visualize our amplified PCR product. Because DNA is negatively charged we can run the fragments through an agarose gel where larger products move slower toward to a positive electrical charge. As first step we create the agorase gel with salted water and agorase, abjusting its density accordingly to the mean expected genes size. After letting the gel be a homogeneous liquid we place it to the tray using comb to create holes in it. Then we put the gel in electrophoresis machine, supplying with salted water until it covers the gel. We load the PCR result to the holes and run the electrophoresis. In this way we separate DNA fragment by their size. The last step is to place the gel in a UV fluorescent dye and after some time we use UV light to observe the separation of DNA fragments by size.
We tried to apply this technique in order to verify the sequence length of *HMA4* gene of the species Aradidopsis halleri and then using Next Generation Sequencing to learn its whole sequence. This gene is shown to give the ability to transfer the metal components (like Zinc) from its root to its leaves, helping this plant to survive in soils with metal hyperaccumulation [Ina N. Talke, Marc Hanikenne and Ute Krmer] and
[Hanikenne M, Talke IN, Haydon MJ, Lanz C, Nolte A, Motte P, Kroymann J, Weigel D, Krmer U.].

Part 2: Mapping Next Generation Sequencing Reads

A new way of getting know the whole sequence of an investigated gene is the *Next Generation Sequencing* (NGS) method. In this method (cyclic-array methods), common adaptors are ligated to fragmented genomic DNA. Then these fragments subjected to one of several protocols, result in an array of millions of spatially immobilized PCR colonies. Each PCR colony consists of many copies of a single shotgun library fragment. As all colonies are binded to a planar array, all array's features can be manipulated in paarllel. Afterwards imaging-based detection of fluorescent labels incorporated with each extension can be used to acquire sequencing data on all features in parallel. Successive iterations of enzymatic interrogation and imaging are used to build up a contiguous sequencing read for each array feature. Each contiguous sequencing result is called *read*.
We can see that this method is advantageous compared to Sanger sequencing technique [Sanger F, Coulson AR] on some parts. Firstly the in-vitro construction of the sequencing library and the in-vitro clonal amplification to generate sequencing features helps to the parallelism of sequencing. Also in NGS hundreds of millions reads can potentially be obtained in parallel by conventional imaging techniques. Another advantage is that the array features can be enzymatically manipulated by a single reagent volume. The last advantage is the lower price per base. The disadvantages of this generation of sequencing are the small read length (limited to 500 Kbps) and the raw error introduced by the platforms. The previous introduction to NGS is based on [Jay Shendure and Hanlee Ji].
The application of NGS lies on getting know the DNA or the RNA (after splicing step) of one or more individual in one or more population. Then we would like to search for structural variations in their whole genomes. In this scope, we need to *map* our reads to an approximate genome reference and then compare the reads of each individual. Alternatively we can create the draft of whole genome for a species *assembled* by these small reads. There are many algorithmic challenges for short-read mapping, the most popular and efficient solution is the application Bowtie 2 [Langmead B, Salzberg SL.]. This method tries to map each read to the reference using sequence similarity, in/exact matching, more specifically uses the Burrows-Wheeler transformation and FM-index technique to search each read on the reference on sublinear time and linear space, allowing inexact matching. Also it tries to find the alignment significance by computing the posterior probability of a match given the read and the reference, this value is also known as the *mapping quality*.
The resulted sequence alignments are saved to *.sam* and *.bam* files. The former contains the header section, if any, where we can find the version of SAM file and information of alignments sorting (if there are sorted). In the second section, alignment section, each alignment has 11 mandatory

[1] `http://www.nature.com/scitable/definition/gel-electrophoresis-286`

fields containing basic alignment information. The most important information is the read id, the reference part, scaffold or contig, where the read is mapped, the position of reference where the mapping starts, the mapping quality and the read quality of each base, etc.. The BAM file contains the same read alignment information but in a compressed form, binary format and it is accompanied with an index for more efficient search time. Many times a visualization of read mapping to the reference helps us to understand qualitatively the performed mapping and also we may observe test cases which our proposed pipeline does not support. A convenient application is the *IGV* (Intergrative Genome Viewer), where the researcher can load the genome reference and the BAM files from a NGS platform and visualize the read mapping acquiring information for variation features efficiently for a standard desktop pc.

After being introduced to NGS area. We tried to contribute in the current laboratory project where laboratory researchers try to assemble the draft genome references of two diploid species (*Arabidopsis halleri* and *Arabidopsis thaliana*) using NGS data from individuals a polyploid species (*Arabidopsis kamchatica*). The first two species are hybridized to produce the last one, which contains the whole unreduced genomes of A. thaliana and A. halleri [Shimizu-Inatsugi R, et al.]. The proposed workflow has seven steps. The first one is to prepare the A.halleri and A.lyrata *reference* genomes. Then a wild type of A.kamchatica genome *reads* is mapped to both references independently, using Bowtie 2. In the next step we classify the mapped reads into A.halleri or A.lyrata derived. Afterwards we concatenate the reads classified commonly to both species with the reads that correspond to either A.halleri or A.lyrata. Now we are able to check the number of SNPs (single nucleotide polymorphism) detected and calculate the mutation rate. *Currently,* the working method is not concatenating the classified reads by only replace the SNPs and INDELs. By the last two steps we acquire more information for some parts of the species reference so we are able to replace the mutation point in species references. We cycle through this procedure of mapping, grouping together and replacing until no mutation points are detected. By this cyclic method we try to construct a more qualitative reference genome for A.halleri and A.lyrata, based on short reads of A.kamchatica genome. The figure 2 shows the mentioned workflow and it was originally created by professor Jun Sese (Tokyo Institute of Technology), contributing as laboratory collaborator.

In more details the step of read classification to the two references follows the intuition that low number of mismatches of read with the one reference compared to the other indicates higher similarity of the read with the former. Consequently in this step the script file (`read_classify.py`) counts the number of mismatches of a given read with the A.halleri and A.lyrata references. If the number of mismatches is equal for the two cases we classify the read as *common* read, but for example if the read has lower number of mismatches for the first reference then we classify it as *origin* for the first genome in the opposite case we should classify it as *other*, because it is more similar to the second reference genome. During the laboratory rotation with collaboration with my supervisor Dr. Masaomi Hatakeyama we improve the classification procedure by accounting the alignment significance of each read to the two references measured by the *mapping quality* value. As described before this variable measures the probability of the read being aligned to a reference genome, indicating the significance of the alignment. Given the previous workflow this variable is measured separately for the A.halleri and A.lyrata reference genomes, thus accessing its value does not need additional time. In more details after in each comparison of number of mismatches in `read_classify.py`, we compare also the significance of the alignment to the two distinct references before assign the read to either *common*, *origin* or *other*.

To evaluate the proposed method, scripts were written to test the performance of the previous and current classification over 8 cycles of mapping assembly workflow. In the next figures (3-6) we can assess the difference between the number of detected INDELs, SNPs and the mean quality and read coverage for the two approaches. We can observe that the new method falls short to decrease the number of detected INDELs and SNPs over each iteration compared to the old one but it produces alignments with higher mean quality (lower only for A. halleri) but with better read coverage (for both references). This indicates that the old classification can reduce the number of INDELs and SNPs more than the new method with a constant degree. However following the

intuition of mere number of mismatches, to classify a read, cannot yield to more qualitative alignments also resulted alignment lack coverage compared to the enhanced method. Interestingly the runtime of the enhanced method is slightly higher than the corresponding of old method. A visual inspection using IGV should facilitate a more detailed evaluation of the new method, as the expert can have a more empirical evaluation of the references construction quality. Finally an enchanhed method for replacing SNPs and INDEL may contribute to a better assembly of A. halleri and A. lyrata genomes. All method scripts, including of the evaluation, can be distributed upon request.

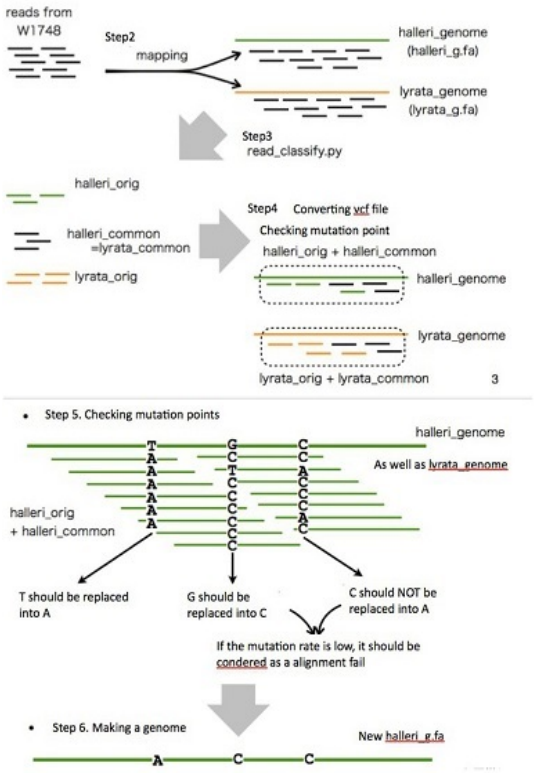

Figure 2: Mapping-Assembly Workflow

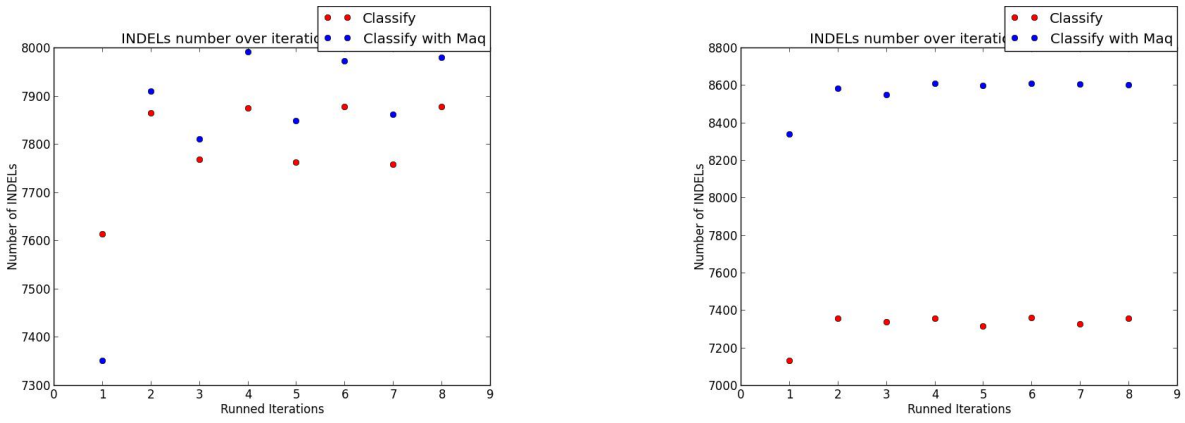

Figure 3: Number of detected INDELs over test iterations for A. halleri (left) and A. lyrata (right)

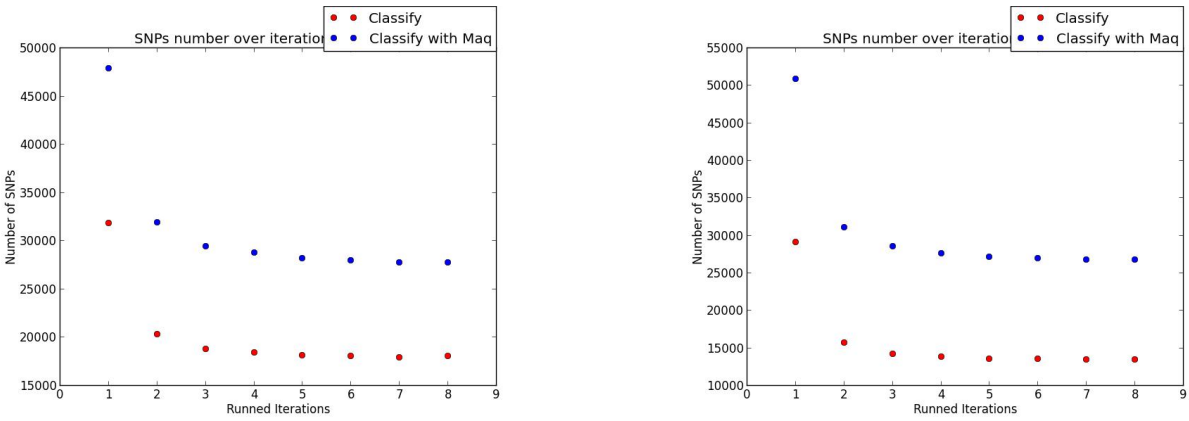

Figure 4: Number of detected SNPs over test iterations for A. halleri (left) and A. lyrata (right)

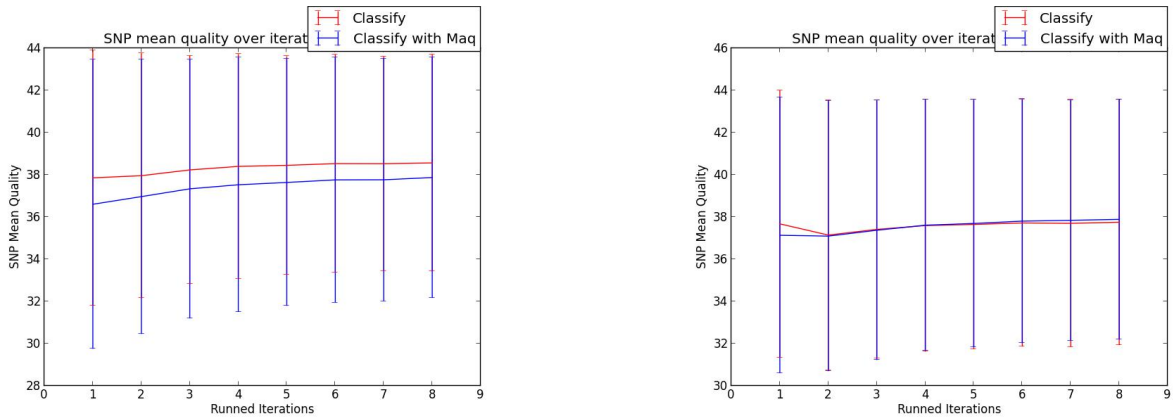

Figure 5: Mean quality of the alignments to A. halleri (left) and A. lyrata (right) over test iterations

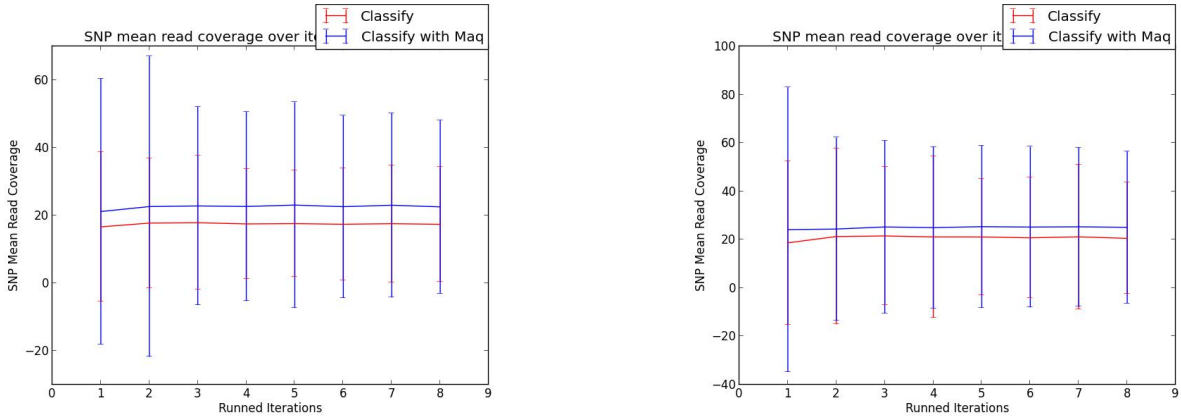

Figure 6: Mean read coverage of the alignments to A. halleri (left) and A. lyrata (right) over test iterations

Conclusion

In this laboratory rotation I learn basic but important biological methods for investigating an organism's genome. Also I was introduced to exciting area of NGS and under the supervision of Timothy Paape and Masaomi Hatakeyama I tried to contribute to their current research interest. I would like to thank from my heart all researchers in professor Kentaro Shimizu laboratory for the respect that show me and for their politeness in every day situations, which encouraged me in a great manner during the laboratory rotation.

References

[Saiki, R.; Gelfand, D.; Stoffel, S.; Scharf, S.; Higuchi, R.; Horn, G.; Mullis, K.; Erlich, H.] Saiki, R.; Gelfand, D.; Stoffel, S.; Scharf, S.; Higuchi, R.; Horn, G.; Mullis, K.; Erlich, H. (1988) Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase. Science, Vol. 239, pp. 487491, 1988.

[Ina N. Talke, Marc Hanikenne and Ute Krmer] Ina N. Talke, Marc Hanikenne and Ute Krmer (2006) Zinc-Dependent Global Transcriptional Control, Transcriptional Deregulation, and Higher Gene Copy Number for Genes in Metal Homeostasis of the Hyperaccumulator Arabidopsis halleri. Plant Physiology, Vol. 142, pp. 148-167, September 2006.

[Hanikenne M, Talke IN, Haydon MJ, Lanz C, Nolte A, Motte P, Kroymann J, Weigel D, Krmer U.] Hanikenne M, Talke IN, Haydon MJ, Lanz C, Nolte A, Motte P, Kroymann J, Weigel D, Krmer U. (2008) Evolution of metal hyperaccumulation required cis-regulatory changes and triplication of HMA4. Nature. Vol., 453, May 2008.

[Sanger F, Coulson AR] Sanger F, Coulson AR (May 1975) A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase J. Mol. Biol., Vol. 94, pp. 4418, May 1975.

[Jay Shendure and Hanlee Ji] Jay Shendure and Hanlee Ji (2008) Next generation DNA sequencing. Nature Biotechnology, Vol. 26, pp. 1135 - 1145, 2008.

[Langmead B, Salzberg SL.] Langmead B, Salzberg SL. (2012) Fast gapped-read alignment with Bowtie 2. Nat Methods., Vol. 4, pp. 357-9, Mar 2012.

[Shimizu-Inatsugi R, et al.] Shimizu-Inatsugi R, et al. (2009) The allopolyploid Arabidopsis kamchatica originated from multiple individuals of Arabidopsis lyrata and Arabidopsis halleri. Mol Ecol., 18:4024-48, Oct 2009.