특/별/기/고

Genome analysis with bioinformatics and microarray technology

Heenam Kim
The Institute for Genomic Research
Hkim@tigr.org

We are living in the golden age of genomics. More than 200 microbial genomes have been sequenced, and the number is increasing nearly exponentially in the coming years. Many eukaryotes including yeasts, *C. elegans*, *Arabidopsis*, rice, Drosophila, Fugu, mosquito, *Plasmodium*, and other eukaryotic models are finished or well advanced. A "working draft" of the human and mouse genome are now available. These vast amounts of genome sequences provide the step stones to the future revolution of life sciences. Large scale production of the genome sequences became possible with the development of high-throughput longer-reading sequencing machines, fast computing powers, and the revolutionary shotgun sequencing strategy developed at The Institute for Genomic Research (TIGR). Soon much faster and economical technologies are expected to arrive and revolutionize genome sequencing once more.

However, we are taking far less than a full advantage of the genome data. There are two main reasons for this. First, we still have very little knowledge in genomic/molecular-level life processes, and thus we are simply blind to much of the information written in the genomes. This problem will become less severe as more and more sequences are available. Comparative analysis of closely related genomes will reveal lots of life's secrets. Secondly, we have only limited experimental means to conduct genomic-level research. Since reading information directly off the genomes currently has problems, we need to deduce the information by conducting appropriate experiments. However, we have relatively good tools to study transcriptomes, but very elementary level tools for the biological processes involving proteomes. To date, the key to best genome analysis mainly lie in how efficiently we can use bioinformatic tools and DNA microarray technology.

Bioinformatics can be best described as the application of computational tools and techniques to the management and analysis of biological data. It is needed in all aspects of the genomic studies from the beginning (i.e. sequencing) to the end (i.e. genome analysis). For this reason, to be a good genome biologist can be quite challenging. This new breed of biologists not only should have solid biological background but also need to be excellent computing tool users with some ability to write short programs/scripts (i.e. Perl, SQL) for specific customized problems. Work efficiency can be greatly increased using perl scripts, and thus perl scripting has become the major part of bioinformatics these days. As an experimental means, DNA microarray technology has been established as the standard of high throughput genome analysis tool. The major use of the technology is to obtain transcriptome profiles relevant to various biological processes. The second use of DNA microarrays is to conduct Comparative Genome Hybridization (CGH) analysis. Comparative genomic analysis of closely related species or within the same species with various deletions often can reveal valuable biological information of the organism. The best comparative genomic analysis can be done with fully sequenced genomes at the sequence level. However, low resolution but still worthwhile comparative analysis conducted using DNA microarrays is often useful.

There are a few different kinds of DNA microarrays. They are cDNA arrays, genomic DNA amplicon arrays, and DNA oligomer arrays. DNA oligomer arrays again can be distinguished as spot arrays and the arrays on which

oligomers are synthesized. Each type of arrays have their own good and bad things, and thus scientists should carefully choose arrays based on the quality of genome annotations and their major scientific interests. In our group at TIGR, we have chosen genomic DNA amplicon strategies for our research involving two bacterial and three eukaryotic microarray projects. Figure 1 shows the strategy that we developed designing the primers for PCR amplification for each gene in the genome of *Arabidopsis*. We applied the same strategy to two fungal genomes, *Aspergillus fumigatus and Aspergillus flavus*. For bacterial projects, we simply took the region inside of the Open Reading Frames (ORFs) and designed primers based on them.

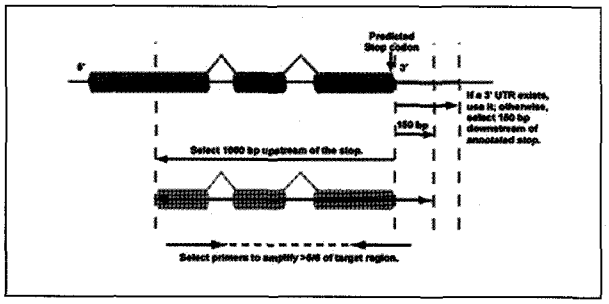

Figure 1. Primer selection strategy for *Arabidopsis* genome.

After designing primers and confirm their uniqueness using blastn and e-PCR, we produced PCR amplicons. Then the amplicons were printed on amino silane coated slides using IAS printing robots (Figure 2).

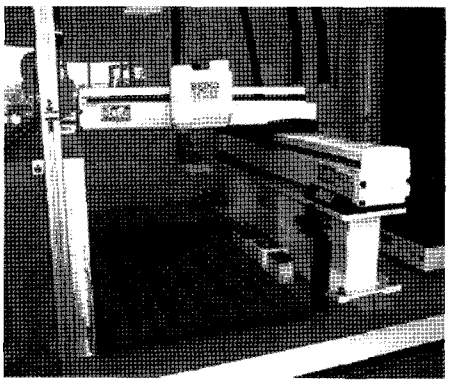

Figure 2. IAS microarray printing robot from Intelligent Automation Systems.

We use the software tools developed in TIGR for microarray data analysis (Figure 3). Current technology uses two samples labeled with different fluorescent dyes (i.e. Cy3 and Cy5) are co-hybridized to arrays. Then the images are taken using a scanner. The TIFF images stored from the scanning process can be converted into intensity numbers of spots by TIGR Spotfinder. These raw spot intensity data are normalized between the two channels and statistically analyzed using TIGR MIDAS. And then, these final data are organized in clusters based on expression patterns using TIGR MEV. TIGR MEV has all the important algorithms implemented for clustering analysis. All of these software tools are freely downloadable from our website < http://www.tigr.org/software >.

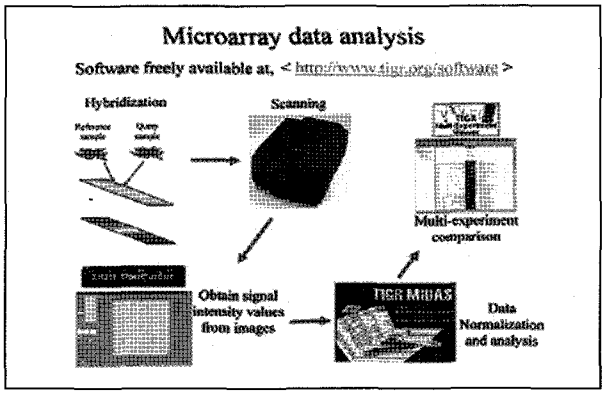

Figure 3. Microarray data analysis scheme in TIGR.

Now, I will present some projects we have done using bioinformatics and DNA microarrays. We have used *Arabidopsis thaliana* microarrays for three main goals, 1) facilitate and confirm genome annotation, 2) transcriptome studies, and 3) evolution studies. For annotation purpose, we have estimated the expression of the chromosome 2 genes and found that about 84% of the annotated genes are expressed under at least one condition out of 40 tested. This analysis was helpful to confirm our *Arabidopsis* annotation because eukaryotic genome annotation is still very inaccurate (even in *Arabidopsis*) and thus detecting expression of gene models provides confirmation of their presence. This is especially valuable for hypothetical genes.

특/별/기/고

For the transcriptome studies, we have surveyed response to a number of biotic and abiotic stresses. These studies will shed light on understanding plant stress biology at the genomic level.

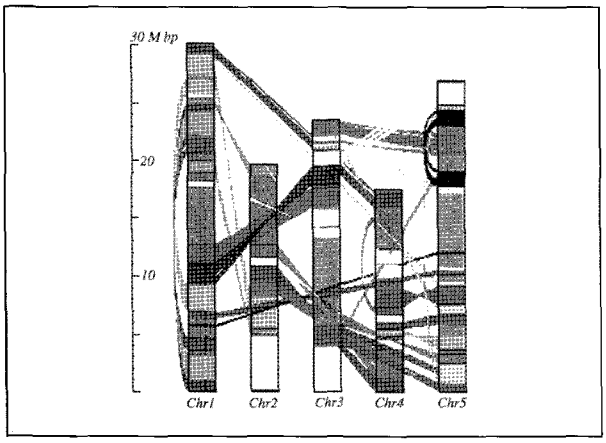

Figure 4. Segmental duplication and shuffling in Arabidopsis genome identified with MUMer (TIGR).

For the evolution studies, we examined the recent segmental duplication in Arabidopsis genome which occurred 24-40 million years ago. Still 1/3 of the genes in the duplicated blocks conserve high homology. So, the question we asked was how many of the duplicate genes still conserve their original expression patterns. We examined this in the context of oxidative stress response, since it comprises the central stress responsive system and thus has to evolve continuously in the ever changing environment. The result was that among the duplicate gene pairs at least one member was significantly responded to oxidative stress, only a fraction of them showed significantly regulated expression in both pairs. This indicates that most duplicate genes have diverged expressions. Even among the genes that both genes are regulated, the expression patterns varied from similar to very dissimilar. Different expression patterns may suggest divergent evolution of the genes, although experimental confirmation remains to be done.

We also have bacterial genome projects. *Burkholderia mallei and Burkholderia pseudomallei* were used during world wars I and II by Germany and Japan, respectively. They were weaponized by former Soviet union, and are US CDC category B bio-warfare agent. Despite the importance, details of biology and pathogenicity are largely unknown. *B. thailandensis* is closely related to the two strains, but it is non-pathogenic. We have sequenced *B. mallei* and *B. thailandensis*, and Wellcome Trust Sanger Institute in England sequenced *B. pseudomallei*. They all have two chromosomes. *B. mallei* and *B. pseudomallei* have different host range, and while *B. pseudomallei* can live in the environment, *B. mallei* has never been isolated from soil and is believed to be almost obligate pathogen. So, we have an excellent system for comparative genomics analysis, two pathogen with different host range, and a non-pathogen. We first examined which regions match in different strains (Figure 4). In this example, the lines link matching regions between *B. mallei* and *B. pseudomallei*. You can see that there is a great deal of shuffling and *B. mallei* has many deletions.

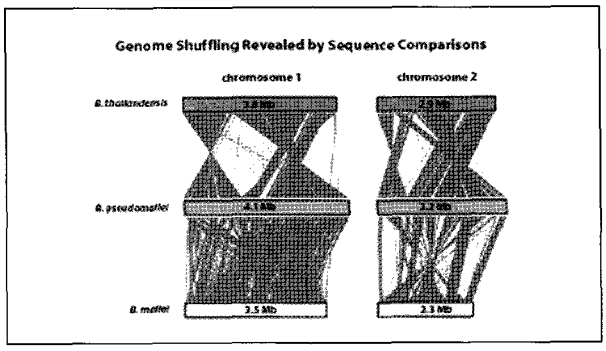

Figure 5. Comparative genomic analysis among three *Burkholderia* genomes.

Besides extensive sequence level analysis, we also did comparative genomic analysis using whole genome *B. mallei* DNA microarray. In one analysis, we took a number of virulent and avirulent isolates of *B. mallei, B. pusedomallei*, and *B. thailandensis*. This study revealed possible virulence genes in the *B. mallei* genome.

Genomics is blooming. Craig Venter says, "Well, there never is a post genomics. We are in the genome era and we'll be in the genome era for the rest of human history..."