# Microbial and Plant Genomics

Jonathan Eisen
The Institute for Genomic Research (TIGR)

The purpose of this presentation is to discuss the genome projects underway at the Institute for Genomic Research (TIGR), where the first complete genome was determined in **(19XX)**. However, it is important to note that there are numerous genome projects taking place worldwide. Many of these projects involve understanding micro-organisms of plants of various eukaryotic pathogens, as well as the genome sequences of humans and other organisms. This presentation begins with background on microbial genomes. I will then describe three applications of genomics, including 1) using genomics to improve vaccine development; 2) applications of genomics in bioremediation; and 3) the role of genomics in improving our understanding of the evolution of species.

## Background: The Role of Micro-Organisms

It is important to note that gene transfer occurs naturally in the field. Microbiology had it's beginnings near Delft where Van Leeuwenhoek was the first to document the existence of microbes. Today it is recognized that microbes dominate this planet. In many instances they are the major cause of pathogenic diseases, and, still the major cause of death around the world.

In terms of evolutionary diversity, an evolutionary tree of life exists **(refers to slide)** and highlighted in **Table X** are all of the micro-organisms. Only at the tip of the eukaryotic evolutionary branch do we see large collections of larger organisms. The remaining microorganisms are bacteria (called the 'arkea' or 'arke bacteria'). Even the majority of the eukaryots are micro-organisms.

Micro-organisms play major a major role in global energy and nutrient cycles, as well as in pathogenesis. They are also responsible for many features of non-micro/macro-organisms through symbiosis with those organisms. Micro-organisms count for a large portion of the bio-mass on the planet and for the majority of biochemical and physiological diversity.

In 1995, prior to my tenure at TIGR, the first complete genome was determined of the bacteria "hymofelous" influenza. Since that genome was completed and published just 5 years ago there has been an accelerated rate of publication and release of genome sequences. In addition, much research is underway in private corporations and entities that are not really releasing them. At present, there are literally hundreds of bacterial genome projects going on around the world.

**[Dr. Eisen- the next paragraph describes a color slide which will not work well in the text...can we omit?] Table X** illustrates a time line to reveal the different genomes: in **red** are the ones that have been done at TIGR and **yellow** are the ones done at other institutions. This chart also reveals the different types of organisms, including the bacterial genomes that continue to be sequenced. The ones in **pink** are arke-bacterial genomes whereas the ones in green are eukaryotic genomes or chromosomes. The determination of complete genome sequences is rapidly increasing and I anticipate an exponential increase in the next few years.

*MISSING BIO*

## Rationale for Complete Genome Sequencing

Why do we want the complete genome sequence, rather than just most of the genome sequence of an organism? When the complete genome sequence is obtained it is possible to determine not just the presence of particular genes that might difficult to locate through other methods; however, it is also important to understand the absence of particular genes in certain genes within the organism. For example, some pathogenic organisms (such as pathogenic bacteria), they are missing key genes that regulate the mutation rates. Many pathogens have high mutation rates which, in some cases, is due to the absence of genes that limit the mutation rates in other organisms. This can be determine from a complete genome sequence. In addition, having the complete genome sequence enables scientists to obtain a firm grip on numerous *genome features* (such as large genome duplications) which cannot be done with incomplete genomes. Finally, the *order* of genes can be important for understanding genetic function of certain genes, and, it can provide data on where within the genome might replication might originate.

Even when there is a small amount of genome sequence missing, the implications can be significant. For example, TIGR just finished sequencing chromosome II of the plant "Arabadopsis Thaliana." Through a sort of "curculian" effort of sequencing technologies we went into sequencing some of the hard to sequence regions in the centrum area of the genome where we discovered there were some very interesting features. In many cases people are avoiding sequencing some of these hard to sequence regions in genomes, but it turns out that there will be data of interest and it is worth pursuing.

Completed genomes provide an enormous resource to scientists as well as to people applying the genomes to particular studies. For example, when all the genes in the genome are identified, we run a variety of prediction programs to try and guess what the function of those genes are. Through this process we can identify novel metabolic pathways that may not have been expected in certain organisms and species. In addition, we are able to better understand novel regulatory elements and to improve our general characterization of regulatory elements.

It is also possible to identify genes which are relatively unique to that organism or which have no predictable function which may be worth pursuing as novel candidates for interesting biochemistry or physiology. Further, completed genome sequences enable the identification of pathogenesis features, "envirolence" features, which can be used to identify vaccine candidates. In addition, it is possible to utilize the complete genome sequences to foster a better understanding of both species diversity through "comparative genomics." Finally, through characterization of unculturable organisms by "hybridization" it is possible to improve the understanding of an array of genomes from soil samples enabling detection of species in the soil and there unique properties.

The major reason and need for genome projects is that they provide information and position us at the beginning of a new world of biological research. The ability to understand the molecular basis of the genome enables us to conduct research that may have been limited before by not having sequence information. Most of what is done from the complete genome sequence requires further experiments and follow- up to test the guesses and predictions that are made. As such, genomes provide a starting point for characterization of particular species.

As stated previously, there have been a number of genomes sequenced at TIGR; in addition, I have conducted a search of Medline databases and have plotted the number of publications about that particular organism over time. As is apparent in **Table X**, as soon as the genome of "mycoplasm genetalium" was first published, the number of other publications about this species increased exponentially and continues to increase even further now. The same things happens with other microbial genomes. Genome sequencing stimulates significant biomedical research.

## Applications to Vaccine Development

A major benefit of having a complete genome sequence is that it allows a targeted approach to vaccine development. For example, by having sequence comparisons, it is possible to identify genes in the genome of interest that are similar to genes that are known to be antigens in other species. With

this information, it is possible to use these as targets for vaccine development for that particular species. Through predictions of the structure of the genes in a genome, one can also identify candidates that might be on the surface of the organism that might, therefore, be antigenic peptides. Based on this understanding, it is now possible to use the sequence to make the DNA-based vaccines, which are being developed worldwide.

Further, in many organisms, the genes that are antigenic are also genes that are under strong selective pressures in that species to vary. Many organisms have developed means to create this variation. One of those means is by having phase-variable genes, such that the mutation rate in these genes in that organism is very, very high. This allows the organism to evolve responses to the immune system rapidly. This enables identification of candidate variable genes from the genome sequence, based on the sequence features.

There are numerous methods that are being used, based on some of the sequence data emerging from the malaria organism, "plasmodium phosyperum". There is a global effort underway, of which TIGR is part, to sequence the genome of this organism. Our approach is to take that genome and to make whatever predictions we can, based on the gene sequence and to then identify candidate antigens. These antigens can then be fed into a variety of systems to test vaccines. This process requires specific methods of experimentation in that particular species; again, the genome sequence is the starting point for identifying any of these candidate antigens.

At TIGR, we have also completed the sequence of the bacteria "Aneasyria Menengititus", which is one of the causative agents of meningitis. From this genome, it's known from previous biological work that phase variation is very important in determining the pathogenesity of this organism. Prior to having the genome sequence, around fifteen to sixteen phase variables genes were known in this organism. From the genome sequence, we can now identify at least fifty new phase variable genes (or likely phase variable genes from this species). Therefore, this research increases the possibility of developing better vaccines, based on these genes and our understanding of the molecular basis of the organism and the species.

## Environmental Applications

Numerous applications of genome sequencing exist. One major area of interest is in improving our understanding of organisms that survive in extreme environments and the potential applications of that capacity. In this case, TIGR has just finished sequencing the genome of the bacteria "Dynacoccus Radiodurands", which is most radiation-resistant bacteria known. It can survive doses of gamma radiation that will start to melt the test-tubes they are grown in and it appears to be resilient to high doses of radiation. These are doses a thousand times greater than the bacteria E-coli can survive. This organism is also extremely resistant to desiccation, to mutagenesis, and to ultraviolet (UV radiation).

**Table X** provides a map of the genome sequence. From the genome sequence, we went through and tried to identify likely genes involved in some of these resistant processes. In particular, we focused on looking at genes that might be involved in protecting the organism and the genome from this damage (for example, we studied scavenge oxygen radicals). There are various mechanisms that will allow an organism to tolerate the damage, even when it occurs. In particular, what we were most interested in is in identifying the genes that were involved in actually repairing the damage. In particular, repairing the damage to the DNA of the genome.

**Table X** provides a listing of the DNA repair genes (putative DNA repair genes in the genome that we have identified) based on comparison to other species. This enables scientists to take that list of DNA repair genes and either to target disruptions in those genes to see if they really do account for the extreme resistance of this organism, or, to transfer those genes into other organisms to see if they make those other organisms equally radiation resistant. So far, that has not been the case for the genes that have been transferred into other organisms.

Increasingly, Dynacoccus is actually being used to take pathways from organisms that can degrade toxins and transfer them into Dynacoccus Radiodurands. Because Dynacoccus can survive extremely high doses of radiation and for long periods of time, the hope is that in mixed contamination waste sites it will be possible to introduce toxin

degrading pathways into Dynacoccus, which will result in the release Dynacoccus in that environment where it will be able to survive the radiation and at the same time degrade the other toxins in the environment. The only way that scientists have been able to isolate it, is by irradiating the sample until it's the only thing that's left. However, prior to irradiation, it's very difficult to isolate it. It is present in very small levels and it doesn't out-compete the other organisms that are there. This application has potential as a bio-remediation device, however, much remains unknown with regard to its efficiency at such a process.

## Contributions of Genomes Sequence Data in the Study of Evolution

Another major application of complete genome sequence data is in the study of evolution. It is now possible to learn about gene transfer among organisms through evolutionary comparisons. The general model for studying the evolution of organisms that existed for many decades was the model of "vertical inheritance". That is, one begins with some individual organism or a small population of organisms. These are split into different lineage's and those lineage's evolve separately into different species at the end. It has been known that among populations there is a reasonable amount of gene flow within a species, and that there is some gene flow between closely related species (hybridization in nature).

However, until recently, there was little information on gene transfer between organisms. There were specific cases that were well documented (for example, exchange of plasmids). In this case, for example, there was just a model of some green species of bacteria and a red species of bacteria next door and there are many well documented cases where small pieces of circular DNA (called plasmids) could be exchanged from one organism to the other. It was shown that if the green feature is due to the genes on this plasmid, then this organism would essentially become green. To date there have been many examples of gene transfer in the environment (involving viruses, transposable elements, and the agro-bacterium plasmid). This bacteria transfers some of its genes to plants, in order to cause the plants to produce compounds that are useful to that bacteria. The general amount of this gene transfer is considered relatively minimal compared to the total size of genomes.

Specific cases of gene transference involving large amounts of DNA have also been well documented with transfer genes from organelles to the nucleus, and, from the mitochondrion chloroplast to the nucleus of organisms. Mitochondrion chloroplasts were originally living bacteria. They are now a symbiosis with eukaryots and have transferred many of their genes to the nucleus. From "Arabadopsis" chromosome two sequence, we have found many more cases of this organelle to nuclear gene transfer then we previously expected, including a 275 kilobase section of the mitochondrial genome inserted, very recently, into the centromere of Arabadopsis chromosome.

From all the genome sequences that have been identified, we have determined that gene transfer appears to be more of a rule than an exception. Studies of the deep-branching, early-evolving bacteria "Thermatogo Maritima" clearly show that there have been enormous amounts of gene transfer between thermophillic bacteria Anarchia (as mentioned previously in Arabadopsis Thaliana). In Dynacoccus, we believe that some of the smaller chromosomal elements will actually be referred to as the chromosome two and the mega-plasmid may actually have come from other sources. Again, this appears to be the rule in organisms rather than the exception - enormous amounts of transfer involving huge chunks of the genome either at one time or over long periods of time.

As such, this is what the tree of life actually really looks like. It's not a tree of life, it's a web of life or a network of life where gene transfers are occurring constantly between organisms. What this means is not just that the tree of life is really a web of life, but, related to this conference, that genetic engineering is actually going on in microbes quite a bit. They're transferring genes and sampling diversity of processes by soaking up DNA from other organisms and thereby possible evolving new pathways or new resistances to things. So this gene transferring is occurring all the time. It can actually be extremely precise in nature, in that single genes inserted into very particular regions of the genome through regulated processes, very similar to targeted genetic mod-

ifications being done in labs now. However, what this also means - it is a double edge sword - it also means that it is likely that when we make genetically modified organisms some of those genes could be transferred to other organisms and should be considered a potential process either through viruses or bacteria or recombination or hybridization. All sorts of mechanisms allow this. What appears to be the most important thing, at least from the evolutionary analysis to determine or not whether these gene transfers occur. Not the possibilities of the genes being exchanged. That appears to be very easy. But whether or not, the gene that is exchanged can be maintained in the new organism. First you have to have a strong selection to keep it there. In addition it has to have the right features to fit into that organism's replication and genome processes.

So just to end...The genome sequences are the starting point for a lot of more detailed research. They have provided a lot of insight into things like lateral gene transfer. They also allow people to do genome-wide experimental studies, which we will probably hear more about later, such as genome micro arrays and things like that. Things that have not yet been done, but we'll continue to do in the future are comparisons of closely related species, such as human and mouse or very closely related bacteria. And that provides a very different type of information than the comparison of distantly related organisms. JB&B