TMTI Task 1.6 Genetic Engineering Methods and Detection

Thomas R. Slezak

Ray Lenhoff, Jonathan Allen, Monica Borucki, Elizabeth Vitalis, and Shea Gardner
Lawrence Livermore National Laboratory
December 7, 2009

Task 1.6 Genetic Engineering Methods and Detection

The Lawrence Livermore National Laboratory (LLNL) Bioinformatics group has recently taken on a role in DTRA's Transformation Medical Technologies Initiative (TMTI). The high-level goal of TMTI is to accelerate the development of broad-spectrum countermeasures. To achieve those goals, TMTI has a near term need to assess methods of genetic engineering (GE) and their ease of execution as well as the difficulty of detection of these methods in light of available detection approaches.

Executive Summary

A large number of GE techniques can be adapted from other microorganisms to biothreat bacteria and viruses. Detection of GE in a microorganism increases in difficulty as the size of the genetic change decreases. In addition to the size of the engineered change, the consensus genomic sequence of the microorganism can impact the difficulty of detecting an engineered change in genomes that are highly variable from strain to strain. This problem will require comprehensive databases of whole genome sequences for more genetically variable biothreat bacteria and viruses. Preliminary work with microarrays for detecting synthetic elements or virulence genes and analytic bioinformatic approaches for whole genome sequence comparison to detect genetic engineering show promise for attacking this difficult problem but a large amount of future work remains.

Introduction

Genetic engineering is defined in this document as scientific alteration of the structure of genetic material in a living organism. It involves the production and use of recombinant DNA in order to produce a new trait in the engineered organism. Both bacteria and viruses can be engineered however the level of difficulty is usually considered higher for the engineering of viruses as their genomes are smaller than those of biothreat bacteria and therefore more easily disrupted by genetic manipulation.

Methods of Genetic Engineering

Numerous techniques have been developed for genetic manipulation of bacteria and viruses. They fall into two categories:

- Extrachromosomal genetic engineering (EGE) involves the transfer of an autonomously replicating DNA (plasmid, cosmid, bacterial artificial chromosome) to the engineered organisms. This DNA can have a range of sizes of genetic material from 2 to 200 thousand base pairs.
- Intrachromosomal genetic engineering (IGE) involves the transfer or modification of chromosomal DNA from the organism through an in vivo or in vitro process.

A number of approaches can be used for IGE. The common mechanism for most approaches is recombination between the genome of the engineered organism and a designed external fragment of DNA. Most of the methods differ in the makeup of this external DNA fragment. It can be a plasmid, a bacterial artificial chromosome, a transposable element, a virus or viral fragment, or even a short sequence of synthetic DNA. Table 1 summarizes many of the techniques used for genetic engineering.

Genetic Change	Base pairs (bp) or Kilobase pairs (kbp) changed	Mechanism of Change	Reference
Plasmid or λ, Cre, FLP recombination	2-20kbp	Extrachromosomal or Recombination	Schweizer H. 2008 Horzempa J. 2009
Bacterial or Phage Artificial Chromosome BAC or PAC	20-200kbp	Extrachromosomal or Recombination	Ioannou P.A. 1994
Transposable element	20-60kbp	Recombination	Fu J. 2008
Engineered virus	2-60kbp	Recombination	Jacobs B.L. 2009
Multiple SNPS	20-100 base pairs	Recombination	Wang H.H. 2009
Single SNP	Single base pairs	Recombination	Heerman R. 2008 Datsenko K.A. 2000
Synthetic genome	2-?kbp	Synthesis	Gibson, D. G. 2008 Cello, J. 2002

Table 1. Techniques used for genetic engineering of microorganisms

Because recombination is generally an inefficient process, the external DNA fragment usually contains a selectable marker to allow differentiation of the genetically modified organism from the high background of unmodified organisms. In many cases this selectable marker can leave a unique non native nucleic acid sequence (such as a gene for antibiotic resistance) in the genome of the modified organism that is later detectable as a marker of genetic modification. Genetically modified food crops are identified thorough the detection of such markers. In some cases a second step may remove most of the marker gene but still leave genetic remnants of the alteration (Datsenko and Wanner, 2000). While not all genetic engineering leaves behind a non native sequence (Sun et al., 2008), genetic engineering must by its nature result in an altered genome and by some means this alteration from the natural antecedent must be detectable. The level of difficulty of detection however varies depending on the laboratory process used to create the mutation and the availability of the antecedent strain.

Genetic Engineering detection overview

A microorganism with an engineered genome could have a genetic alteration of thousands of base pairs or a single base pair and the difficulty of detection will depend on the amount of genetic difference between the engineered and non engineered organism. This relationship between size of genetic manipulation and ease of detection is summarized in Fig 1 below.
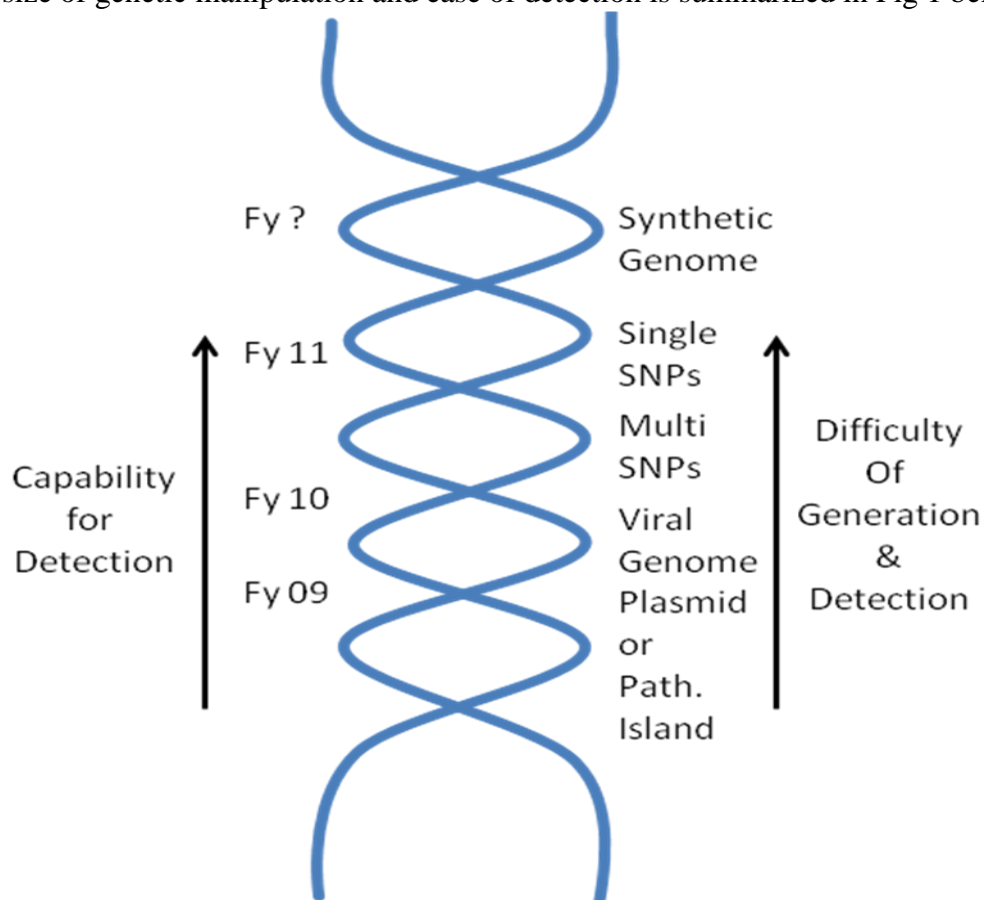
Figure 1. Hierarchy of genetic manipulation verses capability for detection

In addition to the impact of the size of the genetic change on the ability of detection, variation in the consensus genome sequence of the engineered organism can also impact the ability to detect genetic alteration. For example, when the complete genome sequences of many strains of *Bacillus anthracis* are compared there is very little genetic difference in the genome organization and sequence of these genomes. However, *Burkholderia pseudomallei* strains can have tens to hundreds of thousands of base pairs of sequence difference between individual isolates (Tuyanyok A et al., 2008). This wide latitude in core conserved sequence between different biothreat bacteria necessitates the assembly of a large database of sequence information for each of the bacterial agents. Only with a fairly comprehensive database of genomic sequence can one hope to discriminate a naturally occurring variation between individual isolates from an engineered change. Recommendations for additional genome sequencing for each of the

biothreat bacteria and hemorrhagic fever viruses has been covered in a previous TMTI report on sequencing priorities submitted to DTRA in October of 2009.

Genetic Engineering detection

Work by the LLNL bioinformatics department has led us to define two distinct but complimentary approaches to detecting genetic engineering in microbes when using DNA extracted from a microbial sample as evidence. The first approach includes methods that look for direct evidence of genetic engineering. This applies to efforts to look for specific genetic markers associated with artificial genome manipulation. Examples of this include vector mediated engineering, and the various forms of site specific integration and homologous recombination where "scars" are left in the resulting genome from the vectors or primer sequences used to facilitate a genome manipulation event. The second category of genetic engineering detection approaches covers the methods that look for indirect evidence of genetic engineering by predicting that the presence of certain genetic sequences or sequence patterns are unlikely to occur from natural evolutionary processes.

Previous work detecting genetic engineering with microarrays

Progress has been made in detection of both forms of genetic engineering with the development of several microarrays. As an initial attempt to develop bioassays that find direct evidence for genetic engineering, a microarray was designed that hybridizes query samples to DNA probes specific for artificial plasmid vectors associated with genetic engineering in bacteria (Allen et al., 2008). Experiments have shown the ability to identify engineered Bacillus subtilus and Bacillus thuringiensis samples as well as to correctly not call natural Bacillus strains as being genetically engineered. A related effort demonstrated the potential to detect specific types of genetic engineering in plants using microarrays (Tengs et al., 2007). In this effort a tiling microarray was used to detect a small number of plasmid vectors used for genetic engineering in Arabidopsis thaliana and Oryza sativa. It may be the case, however, that the genetic engineering method leaves no evidence of an altered genome, other than the desired manipulation. For example, a site specific insertion event, where a gene is inserted but no signature sequence is left at the flanking sequence. In such cases, we are left to consider a second approach to detecting genetic engineering, which identifies the differences in a microbial sample that distinguish it from the collection of known microbial samples and determines whether the differences are due to natural evolution or human intervention. One approach, which has been undertaken at LLNL, is to focus on the rapid and efficient detection of genetic differences in the microbial sample that are predicted to be of greatest interest for medical diagnostics and countermeasures. In particular, extensive testing of microarrays has been undertaken to detect genetic elements associated with virulence factors and antibiotic resistance (Jaing et al. 2008). For example, if a normally benign strain of Bacillus subtilus displays an unexpected phenotype, the query microbial sample could be interrogated with a microarray to determine whether genes associated with virulence and antibiotic resistance are detected even when previously sequenced strains lack these genetic elements, and as a consequence these findings could indicate a genetic engineering event.

Motivation for whole genome sequencing to detect genetic engineering

Of particular relevance to TMTI is determining what additional information might be gained from sequencing a microbial isolate to identify potential genetic engineering. There are two relevant approaches to this type of sequencing, metagenomic and whole genome. Metagenomic sequencing involves directly sequencing all of the DNA isolated from a single sample. Metagenomic sequence is considerably more difficult to analyze since only parts of the genome of interest may be captured along with partial genome sequences of other microbes. Unfortunately, this may be the only form of sequencing possible if the microbe cannot be isolated and cultured for whole genome sequencing. This report focuses on the challenge of using whole genome sequence for genetic engineering detection with the understanding that additional future challenges must be addressed for dealing with metagenomic samples.

There are three challenges to using whole genome sequence for detecting genetic engineering: 1) detecting the differences between the newly sequenced isolate and near neighbor reference strains, 2) functional characterization of the genome (in particular the differences with nearest neighbors), and 3) estimating the likelihood that genetic differences are due to a genetic engineering process.

Identifying the differences themselves requires strong comparative genomics expertise, and the ability to build a comprehensive catalogue of genetic differences depends on how closely related the query sequence is to reference nearest neighbors and the degree of genome structure variability within the bacterial family. For example, since *Bacillus anthracis* strains exhibit a more stable genome the likelihood increases that any new query *Bacillus anthracis* genome will map in its entirety to near neighbors. In contrast, it may be more difficult to map a novel *Francisella tularensis* sequence to its near neighbors, further increasing the number of reported genetic differences and thus making it more difficult to pinpoint the functionally important changes.

A sequence annotation pipeline applied to the genome sequence can give a functional characterization of the genome yielding results similar to microarrays but with the potential for greater precision. Thus, it may be possible to increase the sensitivity of detection of functionally relevant genetic elements, such as the presence of antibiotic resistance genes or other virulence markers that may be missed due to the limits on the number of virulence probes available for detection on the microarray. This stage of analysis could include comparing the newly sequenced genome to a database of genetic markers associated with known genetic engineering methods.

The first two steps to whole genome analysis will be of principle importance for developing broad range diagnostics and countermeasures, irrespective of whether genetic engineering is a factor. Nevertheless, it is important to recognize that simply having access to the complete microbial genome and the first two steps of the analysis (identification of genetic differences to near neighbors, and functional characterization of the differences) does not automatically give the ability to determine whether the differences are due to genetic engineering when known genetic engineering scars from a predetermined set are absent. Knowing whether changes are

due to genetic engineering may in fact be of secondary importance, however, having the ability to make this determination could provide supplemental information in prioritizing which differences with nearest neighbors should be analyzed more closely. For example, if the manipulation is the insertion of a motif or mutation, the genetic alteration may get lost in a sea of differences. Similarly, the significance of finding the presence or absence of certain genes may go unnoticed if their functional role remains unknown. Having additional methods to prioritize the analysis of a potentially large number of genetic differences could speed up the time it takes to identify and design effective countermeasures.
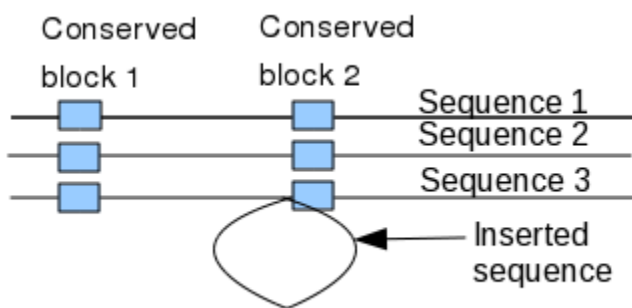

Figure 2. Schematic of a detected DNA sequence insertion event

Novel approaches to detecting genetic engineering with whole genome data

To begin to address these problems we have begun to explore a specific genetic engineering detection test case, to better understand the scope of the problem and determine if bioinformatics methods can translate TMTI's sequencing capabilities into fast and accurate functional characterization of microbial genetic engineering events. To focus initial efforts, the question of site specific insertion and deletion events was considered. A simplified schematic is shown in Figure 2. The goal is to start with the (relatively) more simple detection problems, which can serve as a basis for developing additional detection approaches for other forms of genome manipulation such as gene swapping events, which more closely resemble naturally occurring lateral gene transfer events and nucleotide polymorphisms.

The broad question posed is, given the comparison of a newly sequenced isolate to nearest neighbor reference sequences; can genetic differences due to some process other than natural evolution be identified? To address this question, we first define a model for the natural process of evolution. In our initial examination, the definition of a "functional element" was left intentionally very broad to include detecting insertion or deletion of sequences that may contain promoters, genes, transcription terminators or parts of genes. There is likely no one model of evolution for the gain and loss of such sequence. It is important to note here that we are explicitly excluding in this definition mobile genetic elements. Two data sets were examined to generate preliminary results: 19 *Bacillus anthracis* genomes, and 12 *Yersinia pestis* and *Yersinia pseudotuberculosis* genomes and preliminary analysis on the Yersinia data set is given here.
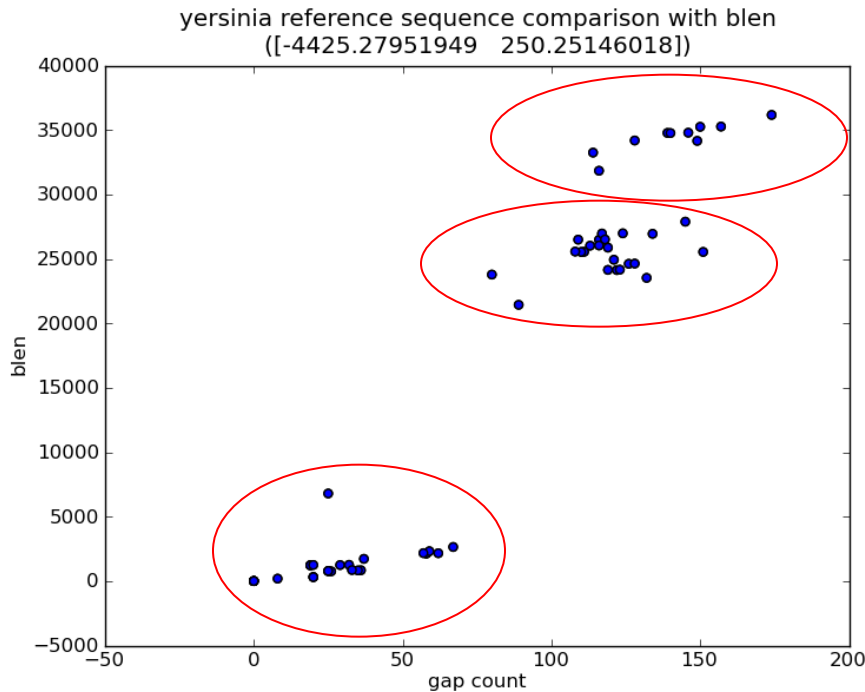
Figure 3. Pair wise nucleotide differences between Yersinia genomes (Y-axis) versus number of pairwise distinct insertion/deletion events

Figure 3 shows a comparison of pairwise genetic distances in the form of phylogenetic tree branch lengths using neighbor joining tree inference. The y-axis is marked "blen" for the branch length and the x-axis marked "gap count" is the number of observed insertion/deletion sequences (length > 30 nucleotides) between two Yersinia sequences. A linear fit of the data is used to estimate a gap insertion/deletion "rate" relative to the genetic distance observed between two sequences. The rate is estimated to be 1 insertion/deletion per 250 branch length units. However, it is clear from the figure that there are three clusters, with potentially distinct insertion/deletion rates. Figure 4 and Figure 5 show the breakdown of the two different clusters, which correspond to *Yersinia pestis* (Figure 4) and *Yersinia pseudotuberculosis* (Figure 5). The third cluster represents distances between the two subgroups. The *Yersinia pestis* rate is estimated to be higher with 1 insertion/deletion event per 37 branch length units. In contrast, the *Yersinia pseudotuberculosis* genomes appear to have a much lower site specific insertion/deletion rate at 1 per 274 branch length units. Looking at two of the more closely related *Yersinia pestis* strains, KIM and Nepal 516 as an example, predicts that we should see 5 distinct insertion/deletion gaps (the actual observed number is 8) given the number of nucleotide substitutions inferred between the two sequences. If accurate upper bounds on the rate parameter can be determined, this gives a powerful tool for determining the level of maximal sequence divergence for which a query strain can be compared to a reference strain and still easily identify an event attributable to genetic engineering. For example, if the maximal upper bound rate is a branch length 50 for each insertion/deletion event, the relative genetic distance between the KIM and Nepal 516, would expect to see 4 natural insertion/deletions between the two sequences. Therefore, engineering events that led to repeated insertions above 4 would stand out, but

7

gauging the significance of 4 or less insertion/deletions would require comparing the sequence to a closer near neighbor to determine whether natural evolutionary divergence is the cause of the differences. This information could prioritize, whether new *Yersinia pestis* strains are need be sequenced to fill in gaps that would facilitate genetic engineering detection. Although a lower insertion/deletion rate is estimated for *Yersinia pestis*, the minimal branch length between *Yersinia pseudotuberculosis* strains is 6794, implying that at least 25 insertion/deletions are observed among pairwise comparisons of the existing set of sequenced strains. Thus, this would indicate the need to consider filling in specific sequence gaps for *Yersinia pseudotuberculosis* in the tree.
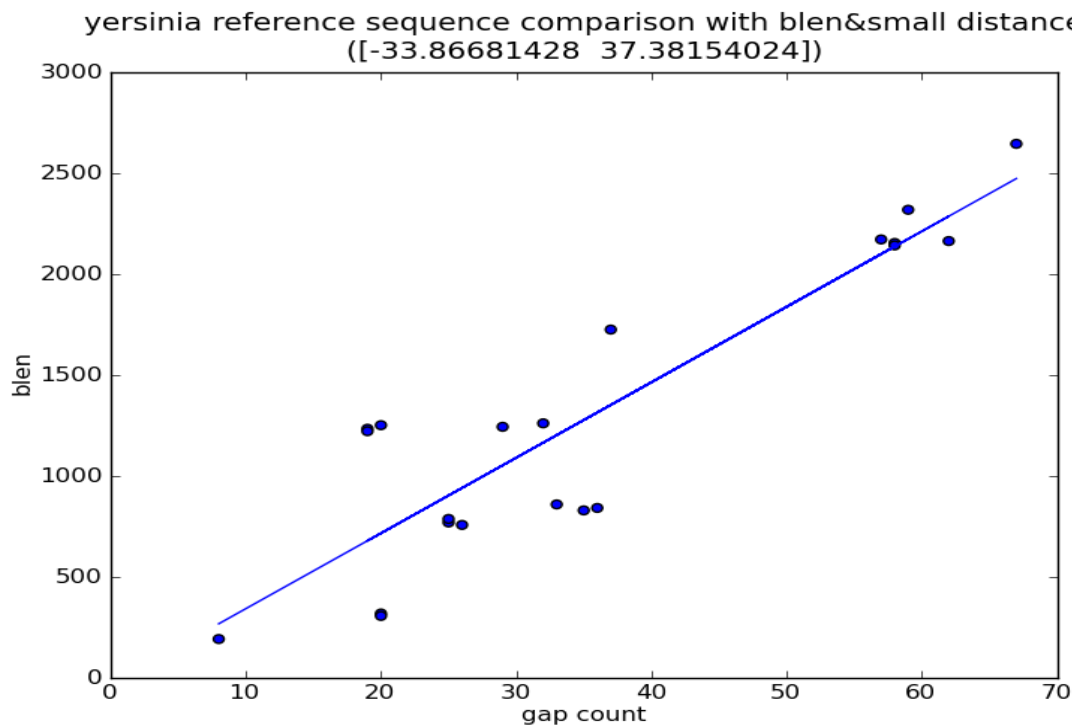
Figure 4. *Yersinia pestis* **insertion/deletion rate**

8
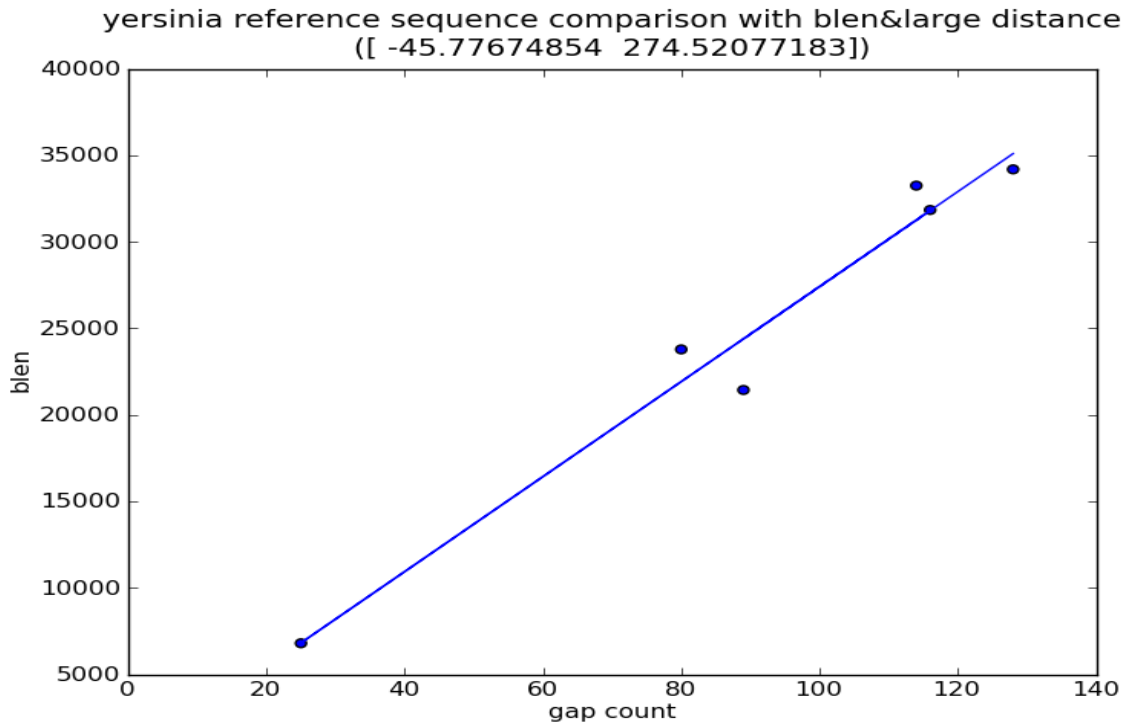
Figure 5. *Yersinia pseudotuberculosis* **insertion/deletion rate**

The results for the Yersinia data set presented here are only preliminary and require follow up verification. However, they illustrate the potential utility of applying bioinformatics analysis to support prioritizing sequencing of near neighbors for the purpose of rapid identification of genetic engineering. These preliminary results suggest the need may arise in the future for a larger sequencing effort to achieve the goal of a rapid accurate and comprehensive detection system where more near neighbors are sequenced. Absent follow on sequencing to support improved detection, bioinformatics analysis will help quantify the confidence levels for asserting whether differences in a novel sample are attributable to genetic engineering or are the product of natural evolution.

The initial work reported here points to important areas for developing genetic engineering detection methods using whole genome comparisons which are listed below:

- Evaluation of more bacterial strains using existing sequence is needed. Initial work has started on *Bacillus anthracis*, but many more bacterial genomes should be analyzed. There may be additional difficulties encountered for bacterial genomes with a high level of genome plasticity.
- The accuracy of the insertion/deletion rate must be improved to ensure that a tight upper bound is achieved.
- Automated detection of different insertion/deletion rates within a bacterial group may be needed.

- Methods can be extended to consider sequence replacement events, rather than insertion/deletion events, which would be similar to other forms of lateral gene transfer.
- Develop distinct rates for specific functional events. For example, distinguish gene insertion/deletion rates, from intergenic insertion/deletion and partial gene insertion/deletion.
- Experiment with different similarity cutoffs to decide what constitutes a non-homologous (laterally transferred or engineered) sequence.
- Evaluate the insertion/deletion patterns of multiple sequences simultaneously to more precisely determine, which insertion/deletion events are more likely to be due to genetic engineering when multiple candidates are available.

Other methods for detecting genetic engineering

The comparative genomics method discussed above provides a powerful tool for detecting various forms of laterally transferred genes, which may be naturally occurring or due to genetic engineering. The utility of this method will decrease, however, as the distance to the nearest neighbors increase beyond a certain threshold where the genome alignments become error prone and incomplete. More work is needed to better characterize what these distance thresholds are and the values will depend on the particular organism family in question. Our initial work with *Bacillus anthracis*, *Yersinia pseudotuberculosis* and *Yersinia pestis* shows the ability to generate alignments that cover most of the sequences, but the quality of the alignments may not hold for other organisms with greater levels of divergence.

A second approach to get around this problem is to consider a gene comparison, or more generally a genome segment comparison, where smaller sub regions are considered for their potential for lateral gene transfer independent of surrounding levels of conservation. This could give another method for targeting regions of interest absent a comprehensive functional characterization of the region and could provide evidence for genetic engineering. One approach is to build a consensus phylogenetic tree for a bacterial family of interest from a concatenation of orthologous genes, and then build individual gene trees to search for regions of the tree where disagreement in tree topology gives evidence for lateral gene transfer (Beiko et al. 2005). A detailed method for implementing an analysis pipeline is given in Beiko and Ragan, 2008. This is an approach that is closely related to methods for detecting genetic shift in the form of Phylogenetic networks (Huson et al. 2006).

One source (Ragan 2001) gives a good overview of different methods for detecting lateral gene transfer including the use of unexpected codon patterns. Codon usage patterns have been recently used to aid functional characterization of genes in term of their predicted expression levels (Bailly-Bechet et al 2006), and codon usage patterns could provide an additional source of evidence for detecting genetic engineering.

Future directions and needs

Given the myriad of approaches that can be used for genetic engineering and the widely varying forms of genome alteration (from single SNP to whole genome synthesis), it is suggested that at a minimum a database should be developed and maintained that would track the various forms of genetic engineering along with the corresponding genomics based detection tools (where such tools exist). This will allow TMTI to maintain a dynamic resource of risks and responses to potential genetic engineering that can be rapidly accessed as well as identify where the capability gaps are expected to persist when characterizing newly identified samples. As a first step LLNL has already assembled a collection of the commercially available plasmid vectors and designed vector specific DNA signatures, which can be used on a microarray or *in silico* to screen bacterial genomes for evidence of genetic engineering. A logical next step would be to expand this database to include diagnostic genetic signatures for other forms of genetic engineering that leave behind specific alterations in the genome. The challenge will be to maintain a comprehensive survey of genetic engineering techniques in light of the expanding toolbox of genetic engineering techniques under development in laboratories around the world. To guard against missing unpublished methods or methods that leave no distinct genetic signature, it is worth including sequence analysis tools that look for genetic engineering methods by detecting (potentially unexpected) differences to near neighbor sequences. Here too, considerable work remains to determine what genomic consensus set of sequences are needed for each biothreat microorganism in order to precisely identify a change in the genome sequence as a product of genetic engineering. The comparative genomics tools that are developed and applied specifically to the genetic engineering detection task will help guide what sequence set of bacterial genomes should be used or developed, and will determine the limits of future rapid detection and characterization capabilities for novel genetically engineered biological threats.

References

Allen JE, Gardner SN, and Slezak TR. DNA signatures for detecting genetic engineering in bacteria. Genome Biol 9:R56, 2008

Bailly-Bechet M, Danchin A, Iqbal M, Marsili M, and Vergassola M. Codon usage domains over Bacterial Chromosomes. 2:4, 2006

Beiko RG, Harlow TJ, Ragan MA. Highways of gene sharing in prokaryotes. PNAS 102(40), 14332-14337, 2005

Beiko RG and Ragan MA. Detecting lateral genetic transfer a phylogenetic approach. Bioinformatics, Volume 1: Data, Sequence Analysis, and Evolution, vol 452, 2008

Cello J, Paul AV, and Wimmer, E. Chemical Synthesis of Poliovirus cDNA: Generation of infectious Virus in the Absence of Natural Template. Science 297, 1016-1018, 2002

Datsenko KA, and Wanner BL. One-step inactivation of chromosomal genes in Escherichia coli K-12 using PCR products. PNAS 97(12), 6640-6645, 2000

Fu J, Wenzel SC, Perlova O, Wang J, Gross F, Tang Z, Yin Y, Stewart AF, Muller R, and Zhang Y. Efficient transfer of two large secondary metabolite pathway gene clusters into heterologous hosts by transposition. Nucleic Acids Research 36(17), e113, 2008

Gibson DG, Benders GA, Andrews-Pfannkoch C, Denisova EA, Baden-Tillson H, Zaveri J, Stockwell TB, Brownley A, Thomas DW, Algire MA et. al. Complete Chemical Synthesis, Assembly, and Cloning of a Mycoplasma genitalium Genome. Science 319, 1215-1220, 2008

Heermann R, Zeppenfeld T, and Jung K. Simple generation of site-directed point mutations in the Escherichia coli chromosome using Red/ET Recombination. Microbial Cell Factories 7(14), 2008

Horzempa J, Shanks RMQ, Brown MJ, and Russo BC. Utilization of an unstable plasmid and the I-SceI endonuclease to generate routine markerless deletion mutants in Francisella tularensis. Journal of Microbiological Methods (in press), 2009

Huson, DH and Bryant D. Application of Phylogenetic Networks in Evolutionary Studies. Mol Bio and Evol, 23:254-267, 2006

Ioannou PA, Amemiya CT, Garnes J, Kroisel PM, Shizuya H, Chen C, Batzer MA, and de Jong PJ. A new bacteriophage P1-derived vector for the propagation of large human DNA fragments. Nature 6, 84-89, 1994

Jacobs BL, Langland JO, Kibler KV, Denzler KL, White SD, Holechek SA, Wong S, Huynh T, and Baskin CR. Vaccinia virus vaccines: Past, present and future. Antiviral Research 84, 1-13, 2009

Jaing C, Gardner S, McLoughlin K, Mulakken N, Alegria-Hartman M, Banda P, Williams P, Gu P, Wagner M, Manohar C, and Slezak T. A functional gene array for detection of bacterial virulence elements. PLoS ONE 2008

Ragan MA. Detection of lateral gene transfer among microbial genomes. Current Opinion in Genetics & Development 11:620-626, 2001

Schweizer HP. Bacterial genetics: past achievements, present state of the field, and future challenges. BioTechniques 44, 633-641, 2008

Sun W, Wang S, and Curtiss III R. Highly efficient method for introducing successive multiple scarless gene deletions and markerless gene insertions into the Yersinia pestis chromosome. Appl Environ Microbio, 74(13), 4241-4245, 2008

Tengs T, Kristoffersen A, Berdal K, Thorstensen T, Butenko M, Nesvold H, and Holst-Jensen A. Microarray-based method for detection of unknown genetic modifications. BMC Biotechnology, 7(1):91, 2007

Tuyanyok A, Leadem BR, Auerbach RK, Beckstrom-Sternberg SM, Beckstrom-Sternberg JS, Mayo M, Wethiekanun V, Brettin TS, Neirman WC, Peacock SJ, et. al. Genomic islands from five strains of Burkholderia pseudomallei. BMC Genomics, 9:566, 2008

Wang HH, Isaacs FJ, Carr PA, Sun ZZ, Xu G, Forest CR, Church GM. Programming cells by multiplex genome engineering and accelerated evolution. Nature 460, 894-899, 2009