In 2001, microbial forensics faced one of its greatest challenges during the anthrax letters microbial investigation. Briefly, an unidentified individual sent letters filled with Bacillus anthracis to a Congressional office. Exposure to B. anthracis, a level 1 pathogen, can cause serve lung infections that are potentially fatal. This organism was chosen as a biological weapon because of its ability to produce easily inhaled spores, thus simplifying transportation. The clonal nature of B. anthracis means every individual is an almost exact copy of its parent. While DNA fingerprinting is reliant on individual degrees of genetic variability, a clonal organism, such as B. anthracis, makes this an almost impossible application. This is analogous to DNA fingerprinting twins. Their DNA is, for all intents and purposes, identical. Thus their DNA fingerprints are similar. Investigators utilized known B. anthracis GENETIC MARKERS TO DIFFERENTIATE STRAINS. Through this typing method, investigators deciphered the letters contained in B. anthracis AMES, a widespread research laboratory strain. At the time of the investigation, scientists typed 20 known polymorphic sights spanning the AMES genome in hopes of discovering the lab of origin. Unfortunately, the results were inconclusive, and all available molecular techniques were exhausted; the best means of identifying the source of the strain was through visual differentiation of colony morphology. This technique is both subjective and difficult to quantify, thus emphasizing the need for a highly sensitive methodology capable of distinguishing clonal bacteria. Current knowledge of the DNA replication mechanism provided scientists with an opportunity to study rare variance among B. anthracis strains; each time DNA is copied low-level mistakes are made at an error rate of 0.001. While this is too low to be detected through DNA fingerprinting, the modern era of next-generation sequencing provides the ability to identify these rare variants.

Next generation sequencers can rapidly sequence a whole bacterial genome and produce billions of reads, giving investigators an advantage over DNA fingerprinting. While DNA fingerprinting is both cheap and reliable, it only utilizes small, targeted regions of DNA. Whole genome sequencing allows for the identification of discovering SNPs genome-wide; however, in the rare cases like the 2001 anthrax letters, a broader approach is necessary to find variation. The high throughput method sequences the whole genome and identifies SNPs throughout the genome. Along with whole-genome sequencing, new algorithms are needed to be developed to scrutinize the genome for rare variants. This requires lots of sequence data and computing power to complete the analysis. The level of detection needed to identify rare variants with a frequency of 0.1% in the population would require sequencing the sample to a minimum of 39000x coverage. This is done so at least 25x coverage is obtained on the rare variants sites. This equates to ~ 4 billion reads generates for one B. anthracis genome. This level of sequencing costs about $15000 with today’s technologies, but in 2001 this was an unfathomable feat. With the arrival of current technologies that allow for high sequence coverage, new sets of issues arise. Errors are inevitably accumulated from the sequencer. Adding to the challenge of rare variant detection is the sheer size of the dataset as well as the implementation of the alignment algorithm. Aligners insert more errors into the investigation through misalignment and false positives.

Along with cost, an inherent issue with high throughput sequencing is the incorporation of errors into the sequence reads. The error rate for an Illumina HiSeq is 0.928%, a frequency dangerously close to the rare variant frequency scientists are trying to detect. This error rate is for single end sequencing. In single-end sequencing, the read is sequenced from only the forward end. This is much cheaper but is associated with a higher error rate. One possible way to reduce the error rate from sequencing is to do an alternative method of sequencing called paired-end sequencing. An advantage of the paired-end mode is the reduction of the error rate by order of magnitude. However, if the reverse and forward reads do not align, the reads will be discarded. Usually, about 20% of the reads are removed because of this misalignment. Also, the cost is significantly higher than single end sequencing. Through trial and error, it was discovered that mistakes are not randomly distributed across the genome. Some patterns are associated with sequence errors. For example, the Illumina HiSeq has significantly more errors at the end of the reads than the beginning (0.3% vs. 3.8%).

Along with a distribution pattern, there are sequence patterns that are at high risk of error. An example is whole polymer errors, which occur in strings of nearly homologous bases, like “CCAC.” The sequencer has a propensity to replace the ‘A’ with a ‘C’ making the sequence ‘CCCC.’ The knowledge of the existence of such error patterns is beneficial when distinguishing true SNPs during the alignment phase.

It is computationally intensive to align ~ 4 billion bases against a reference genome. There are two different methods of whole genome mapping sequence aligners: heuristic (Burrows-Wheeler Transform (BWT)) and exhaustive(hash-tables).The heuristic methods utilize a very efficient data indexing method that transforms the genome into a FM-index and searches the reads against the index. There is a marked performance increase but there is an increase in alignment errors and a decrease in the percentage of the mapped genome.The exhaustive method utilizes hash tables, which creates an index of overlapping k-mer (fragment of DNA size k) of the reference genome. After the index is created, the reads are broken into smaller sub-strings and matched against the genome; when all substring locations are found they combine to make the final alignment. The exhaustive method, as Dr,Slava put it succinctly.

The reference genome, due to its topography, can have instances where SNPs are silenced and alignments will fail at a higher rate. The issue arises from recent gene duplication events, meaning there is a high degree of similarity. These regions are difficult for any of the mapping algorithms to distinguish if a SNP happens in a duplicated it will be silenced because the all the reads will align to one region. A further issue is the level of similarity between regions, that are not idential. In this case, if a heuristic approach is utilized, the first alignment found determines whether the read aligns or not. The alignments start by using a seeding sequence. If the seed is found in multiplr regions there is a significantly higher chance that many reads will be aligned to the wrong region and discarded. One method developed to identify these silencing and high-level misalignment regions is a brute force method; the reference genome is broken into k-mers. The k-mers are modified to incorporate all possible SNPs as well as insertions and deletions(indels). This allows the reference genome to be annotated in order to identify regions that could produce erroneous or low confidence SNPs.