Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Elsevier hereby grants permission to make all its COVID-19-related research that is available on the COVID-19 resource centre - including this research content - immediately available in PubMed Central and other publicly funded repositories, such as the WHO COVID database with rights for unrestricted research re-use and analyses in any form or by any means with acknowledgement of the original source. These permissions are granted for free by Elsevier for as long as the COVID-19 resource centre remains active.
Genomic signatures

This chapter addresses some of the many issues associated with the identification of signatures based on genomic DNA/RNA, which can be used to identify and characterize pathogens for diverse goals such as medical diagnostics, biodefense, and microbial forensic. For the purposes of this chapter, we define a “signature” as one or more strings of contiguous genomic DNA or RNA bases sufficient to identify a pathogenic target of interest at the desired resolution and that could be instantiated with particular detection chemistry on a particular platform. The target resolution may be taxonomic identification of a whole organism, an individual functional mechanism (e.g., a toxin gene), or simply a nucleic acid region indicative of the organism. The desired resolution will vary with each program’s goals but could easily range from family to genus to species to strain to isolate. Resolution need not be taxonomically based but could be pan-mechanistic in nature: detecting virulence or antibiotic-resistance genes shared by multiple microbes. Entire industries exist around different detection chemistries and instrument platforms for identification of pathogens, and we only briefly mention a few of the techniques that have been used at Lawrence Livermore National Laboratory (LLNL) to support our biosecurity-related work since 2000. Most nucleic acid–based detection chemistries involve the ability to isolate and amplify the signature target region(s), combined with a technique to detect that amplification. Signatures may be employed for detection and/or characterization of known organisms, by focusing on unique genomic differences, or as ways of discovering new ones by focusing on highly conserved genomic regions.

Genomic signature–based identification techniques have the advantage of being precise, highly sensitive, and relatively fast in comparison with biochemical typing methods and protein signatures. Classic biochemical typing methods were developed long before knowledge of DNA and resulted in dozens of tests (Gram’s stain, differential growth characteristics media, etc.) that could be used to roughly characterize the major known pathogens (of course, some are uncultivable). These tests could take many days to complete and precise resolution of
species and strains is not always possible. In contrast, protein recognition signatures composed of antibodies or synthetic high-affinity ligands offer extremely fast results but require a large quantity of the target to be present. False positives/negatives are also a problem with some protein-based techniques (home pregnancy kits use this basic approach).

### Different types and resolutions of genomic signatures

Genomic signatures can be intended for many different purposes and applied at multiple different resolutions. At LLNL, we have been working on genomic signatures that can be broken out into several categories: (i) organism signatures, (ii) mechanism signatures, and (iii) genetic engineering-method signatures (or method signatures).

Organism signatures are intended to uniquely identify the organism(s) involved. Mechanism signatures can be best thought of as identifying sets of one or more genes that result in functional properties such as virulence, antibiotic resistance, or host range. The primary reason to identify mechanisms, independent of organisms, is to detect potential genetic engineering. A secondary reason is that nature has shared many important mechanisms over evolutionary time, and thus they may not be sufficiently unique to identify specific “chassis” organisms. Knowledge of whether a particular isolate has the full virulence kit or possesses unusual antibiotic-resistance properties and whether it is human transmissible is important for biodefense and public health responses. Method signatures present yet another dimension of analyzing pathogens: evidence of potential bacterial genetic engineering may be seen in a genome by checking for traces of the bacterial vector(s) that may have been used to insert one or more foreign genes and related components (promoters, etc.) into the genome being modified. In the future, host range signatures might indicate that an otherwise uncharacterized pathogen was potentially capable of evading or defeating the immune system of a particular host organism.

It is also possible to think in terms of detection, diagnosis, and characterization as different classes of activities that may have diverse types of signature needs. Using anthrax as an example, a detection signature might be trying to answer the question “Is there *Bacillus anthracis* in this air or soil or surface swipe sample?” One or more signatures might be used for environmental sampling to indicate the possible presence of that pathogen. In contrast, a diagnostic signature would be attempting to answer, “Does this person have *Bacillus anthracis* in their nasal cavity now?” Such a signature would need to be embodied in an assay that is approved by the US FDA (Food and Drug Administration; we are not aware of any licensed anthrax diagnostic assays at this time), even if the genomic signature is the same as in the environmental example (where regulation is not required). Characterization signatures would be trying to answer questions such as “Which known strain(s) are closest to this sample and what differences are observed?” or “How similar is this attack sample to another sample from a suspect’s home laboratory?”; both of which might be answered by whole-genome sequencing or a large set of single-nucleotide polymorphisms (SNPs) or other genomic markers via PCR or an array. Thus, signatures involve an intended use (detection, diagnostic, characterization), a specific required resolution (genus, species, strain, gene, SNP), a desired low error rate, plus cost, and time constraints specific to the use case.

### Potential target organisms

Genetic signatures can be used to identify any living organisms that contain intact DNA or RNA. Focusing on biosecurity, we are interested
primarily in identifying bacteria, viruses, and fungi that could potentially be used to threaten human, animal, or plant life, to disrupt our economy, or to disturb our social order. Note that there is a wide range of genome sizes involved. RNA viruses are generally small (foot and mouth disease virus is about 8 kbp, SARS coronavirus is about 30 kbp), whereas the variola virus (causative agent of smallpox) is a large DNA virus of about 200 kbp. High-threat bacterial pathogens tend to be in the 2–5-Mbp size range (Yersinia pestis, causative agent of plague, is about 4 Mbp while Bacillus anthracis is about 5 Mbp.). Fungi can range from 10 Mbp to over 700 Mbp. As costs scale roughly with genome size, the sequencing databases have many more viral genomes than bacterial and many more bacterial genomes than fungal. In comparison, the human genome is about 3 Gbp and wheat is about 16 Gbp.

**Signature resolution**

Organism detection signatures must be conserved sequences, reliable, and able to detect all variations of the target organism to minimize false negatives. The signatures should be unique to the target organisms and not detecting nontarget organisms to minimize false positives. Organism detection signatures can be at different taxonomic resolution, typically genus, species, or strain.

In biosecurity applications, high-resolution signatures are needed to precisely identify isolates or strains. In past years, a large distinction was drawn between detection signatures and forensic signatures, where forensic signatures were typically thought of as at the strain level or below (typically thought of as substrain or isolate specific). When microbial sequencing was quite expensive, techniques such as MLVÅ (multiple locus variable number of tandem repeat Analysis) were employed for forensic characterization (Keim et al., 2000). PCR amplicons covering regions containing variable numbers of tandem repeats were measured to provide patterns which corresponded to evolutionary distance. More recently, the distinction between detection and forensic signatures has become blurred because historic taxonomic distinctions have become less certain and because new signature techniques provide increased resolution levels. Using either whole-genome sequencing or current commercially available microarray technologies that allow a million or more signatures to be designed on each chip, one can simultaneously interrogate the entire resolution range (genus, species, strain, and isolate) for desired pathogen targets, providing both detection and forensic resolution. Signature design today is a combination of the desired signature purpose, our current understanding of the diversity of the organism being targeted, and the specific mission constraints that may dictate the detection chemistry and platform to be used for either biodefense or public health. Given the drastically lowered cost of whole-genome sequencing of microbes, we anticipate that the use of all techniques with lower resolution for genomic characterization will diminish in the coming years.

**Genomic sequence data: what to use and where to get it**

There is no single resource for all microbial genomic sequence data pertinent to signature design. The most comprehensive public source for genomic sequence data is the National Center for Biotechnology Information (NCBI) website (http://www.ncbi.nlm.nih.gov/). The NCBI has reciprocal data exchange agreements with the European Molecular Biology Laboratory in the United Kingdom and the DNA Data Bank of Japan, which are equivalent databases used heavily in those parts of the world. Most authors of published sequence data usually submit a final version of their sequence datasets to
GenBank. However, numerous sequence data-bases exist that have organism-specific data that may not be found in GenBank during the interim period of data generation and manuscript preparation and those sites would need to be probed directly to obtain the most recent and up-to-date sequence data. Some examples of these publicly available resources are the Integrated Microbial Genomics project at the Joint Genome Institute (http://img.jgi.doe.gov), the Genomes OnLine Database at the Joint Genome Institute (https://gold.jgi.doe.gov/), the Broad Institute (https://www.broadinstitute.org/data-software-and-tools), and the Sanger Institute in the United Kingdom (https://www.sanger.ac.uk/resources/downloads/).

There are also numerous specialty genome databases at gene and/or protein resolution. Examples of these include UniProt, CARD (Canadian Antibiotic Resistance Database, https://card.mcmaster.ca/), and VFDB (Virulence Factor Database, http://www.mgc.ac.cn/VFs/). One issue is that gene databases typically focus on protein sequences and may or may not contain a good representation of the DNA sequence variants that code for functionally equivalent proteins.

Sequence data most useful for signature design fall into two major categories: finished and draft data of isolated organisms. A third category, raw sequence reads from an isolate, may be encountered but can be readily assembled into a draft genome. Draft genomes are composed of multiple sets of overlapping reads, called “contigs,” potentially with little or no information about the order or orientation of the contigs relative to the original genome. Draft sequence is often described by a depth factor, which is a numeric statement about the average redundancy of coverage at any base position, and thus confidence. A 30X draft sequence would have, on average, at least 30 overlapping reads that contain each base in the genome being sequenced; this is a common minimal average depth for modern draft sequencing, but microbial projects with far greater read depth are also common.

Finished whole-genome microbial sequences have undergone an iterative process to assemble contigs and then use a variety of techniques to order and orient them and close any gaps. This often lengthy and costly process, when and if completed, produces a single string of high-quality bases from the individual and scrambled contigs of the draft sequence. Obviously, finished genomes are superior to drafts when it comes to performing annotation of gene content or other features, as well as for performing multiple-sequence alignments (MSAs) to compare two or more genomes. In our early Sanger sequence experience at LLNL, a 10X draft microbial genome provides sufficient information for DNA signature design purposes (Gardner et al., 2005); modern isolate sequencing of microbes typically yields at least 100X coverage (meaning on average at least 100 separate reads cover each base in the final genome). When you consider that finished microbial genomes can be 10 times as expensive as draft, due to the large amounts of skilled labor required to close draft sequencing gaps, it is not surprising that many microbial genomes may never be finished. Increasingly, short-read sequences are being mapped to reference genomes in lieu of a de novo assembly, and we expect finished genomes to become increasingly rare as the cost of draft sequencing continues to decrease.

Another increasingly important category of data is metagenomic sequence, where no attempt has been made to isolate individual organisms for sequencing. Sometimes this is because of the lack of method to isolate and culture the particular organism(s) of interest. Only a tiny fraction of organisms can be cultured in vitro, and our knowledge base remains greatly skewed toward those that can. At other times, it is because what is desired is a sampling of an entire community of organisms. Although numerous metagenomic samples have been sequenced, it is exceedingly rare for complete assemblies of sequences from multiple organisms to result. One exception is a very small symbiotic bacterial community found living in an extremely harsh acidic environment in a

III. Methodology
mine (Allen et al., 2007). Metagenomic data are not currently of much utility for genomic signature development. A paper on the acid mine microbial community is providing clues about the evolution of viral resistance (Banfield and Andersson, 2008), which illustrates the vital role metagenomic sequencing will play in expanding our systems biology knowledge at both the organism and the ecosystem level.

Searching for sequence data based on free-text queries can be problematic. For example, GenBank does not enforce consistency with sequence designation. Not all complete genomes have “complete genome” in the title, and some that do are not actually complete genomes. We have encountered complete genomes that were labeled “complete cds” (coding sequence), “complete gene,” or otherwise unlabeled as a complete genome. Curation is required to validate any sequence data obtained from a public resource, and periodic in-house testing against benchmark data is necessary to maintain a database of high fidelity. A related problem is distinguishing when a newly finished genome should replace a prior draft, as strain name, authors, or institutions may have changed and the linkage between the two forms of the same genome may be missing.

Genome database quality is another important consideration when acquiring sequence data for signature design. In addition to the sequence designation issues mentioned above, genome databases may contain sequencing and/or assembly errors. Contamination errors may include sequence from other organisms present beside the desired isolate that assembled into your target sample genome, or contigs from other organisms present that were not properly screened out before database submission. Physical contamination of sequencing instruments or sequencing reagents have also caused problems with database quality, as have errors in the DNA barcodes used to multiplex samples on a single sequencing run. We have encountered situations where some human sequence present in one obscure draft microbial genome led that organism to show hundreds of thousands of reads in every human sample analyzed, until the offending genome was removed from the microbial reference database. Beware that the other direction of contamination can also occur (microbial sequence incorrectly present in human genomes.) Ground truth is a slippery concept when it comes to genomic databases of any ilk due to all the possible sources of contamination and errors. Deliberate introduction of judicious genome data errors with malicious intent to confound detection of subsequent genetic engineering is a possibility that should be kept in mind.

### Identifying conserved sequence among targets

Finding regions of conservation across all target genomes can be done with “alignment-based” methods and with “alignment-free” methods. The difference between the two methods revolves around a trade-off between time and quality.

The first issue to be faced when searching for conservation with an MSA is the amount of sequence (breadth) that an alignment method can handle. Alignments sometimes fail when input sequences are very long or when there is a large number of sequences to be aligned (depth), even if the sequences are not particularly long. Failure happens because an MSA takes impractically long to finish due to the intractable algorithmic complexity involved or due to a lack of memory on the machine being used. These limitations mean that the optimal alignment approach may vary depending on the breadth and depth of sequences used as input. The recent explosion of genome sequence data has resulted in a lack of MSA algorithms that can scale appropriately. Clustering is one common way to attempt to reduce input data size. This works well when the goal is to analyze for regions in common to design signatures; however, it works less well when one is trying to detect SNPs that provide maximal differentiation. In recent years, with the massive numbers

III. Methodology
of whole genomes available for many species of interest, MSA has become an increasingly poor solution choice.

Another topic of concern when identifying conserved sequence regions is whether an approach can incorporate incomplete and/or draft sequences. Incomplete sequences do not cover the complete genome of the organism. Draft sequences may cover the complete genome and may be of lower quality, particularly near the ends of contigs. Increasingly, the number of genomes being finished to completion is significantly fewer than the number of genomes that will remain incomplete and in draft form. MUMmer (Kurtz et al., 2004) is a notable MSA program in this respect because it can align draft and complete genomes. Note that any use of incomplete genomes carries an inherent risk because regions not present in the incomplete genome(s) will not appear to be conserved and thus may not be considered for signature mining.

Alignment-free methods for finding consensus are now required to handle the full range of available microbial genomes for signature design in practical amounts of time. PriMux (Hysom et al., 2012) is one example of a non-MSA signature design approach that scales to handle thousands of input genomes, including draft genomes.

Finally, viruses are often highly divergent at the nucleotide level. This high degree of divergence, common among many RNA viruses, can cause even alignment-free methods that rely on a pairwise sequence search to fail at finding all shared genetic regions. Some nonviral organisms have also been observed with enough divergence to make using alignment-free methods prone to error. To help overcome the hurdles of divergent targets, we have developed a novel method of signature generation, “minimal set clustering” (MSC), described later.

### Mining for signatures

After pathogen target regions of sufficient length from conserved and unique regions are found, they are mined for detection signatures. Signatures are found by searching for oligonucleotides with appropriate length, melting temperature, and GC ratio and by searching for oligonucleotide combinations with appropriate overall amplicon size and minimal interoligonucleotide hybridization potential. Programs such as Primer3 (Rozen and Skaletsky, 2000) can perform some or all the signature selection work from a given target sequence input. Primer3 can be integrated into any signature development pipeline, unlike other primer design packages that typically only offer a manual graphic interface.
RNA viruses present additional challenges

High rates of mutation and lack of genome repair mechanisms in many viruses generate increased levels of intraspecies diversity and result in quasispecies, particularly for many single-stranded RNA viruses. Consequently, PCR-based signatures for viral detection often require high levels of degeneracy or multiplexing to detect all variants robustly. Large amounts of sequence data are often required to represent the range of target diversity, sometimes hundreds to thousands of genomes. As noted previously, building MSAs with many diverse genomes taxes the capabilities of most available software. Once an alignment is built, it may reveal insufficient consensus for even a single primer, much less a pair, to detect all members of some species (e.g., human immunodeficiency virus 1 or influenza A).

One solution is to subdivide the targets into smaller or more closely related subgroups, such as clade, serotype, or phenotype, of interest (examples of phenotypes could include virulent vs. nonvirulent, domestic vs. foreign), and attempt to find signatures separately for each subgroup. This approach implies that multiple signatures will be required for species-level detection of all subgroups. One must make an assessment in advance of signature design of how best to subdivide the target sequences.

A second approach is to allow degenerate or inosine bases so that a single signature will detect the diverse genomes within a target species. Specificity may suffer if some combinations of degenerate bases also pick up nontarget species. Sensitivity may decline, as the specific priming sequence for a given target is diluted in the degenerate mix. Several tools that require a MSA as input are available for degenerate primer design (e.g., SCPrimer Jabado et al., 2006 PrimaClade http://primacleade.org/cgi-bin/primacleade.cgi, Amplicon http://amplicon.sourceforge.net/, and HYDEN http://acgt.cs.tau.ac.il/hyden/).

A third approach employed is called MSC (Gardner et al., 2003). Because it avoids the need for MSA or a priori subgrouping of target sequences, this method can be run blindly without expert knowledge of the target species. It begins by removing nonunique regions from consideration as primers or probes from each of the target sequences relative to a database of nontarget sequences. The remaining unique regions of each target sequence are mined for candidate signatures, without regard for conservation among other targets, yet satisfying user specifications for primer and probe length, $T_m$, GC%, amplicon length, and so on. All candidate signatures are compared to all targets and clustered by the subset of targets they are predicted to detect. To predict detection, we may require that a signature’s primers and probe have a perfect match to target in the correct orientation and proximity, or we may relax the match requirements to allow a limited number of mismatches, as long as $T_m$ remains above a specified threshold or those mismatches do not occur too close to the 3’ end of a primer. Signatures within a given cluster are equivalent in that they are predicted to detect the same subset of targets, so by clustering we reduce the redundancy and size of the problem to finding a small set of signatures that detect all targets. Nevertheless, computing the optimal solution of the fewest clusters to detect all targets is an “NP complete” or intractable problem, so for large datasets we use a greedy algorithm to find a small number of clusters that together should pick up all targets. LLNL has used this method to design signature sets for numerous RNA viruses, including influenza A HA serotypes, foot and mouth disease, Norwalk, Crimean–Congo hemorrhagic fever, ebola, and other divergent viruses (Gardner et al., 2014). Fig. 20.1 shows the result of an MSC computation for Crimean–Congo hemorrhagic fever performed in 2005, with the resulting signatures displayed against a whole-genome phylogenetic tree of all the sequences available at that time.

A fourth approach is to forego sequence alignment altogether and to look for sets of primer-length oligomers of length k, or “k-mers,” present in many targets and unique relative to
nontarget sequences. Using combinatoric or greedy algorithms and allowing degenerate bases, PriMUX builds a signature set of k-mers such that each target contains at least two k-mers to function as forward and reverse primers. This approach demands large computing memory to store all candidate k-mers for large or many genomes, especially as k increases above 20, and may require suffix trees or other techniques for data compression (Gardner and Hall, 2013).

**Signatures of potential bacterial genetic engineering**

Detecting evidence for genetic engineering in bacteria is challenging when the target modification is not known and the effects of an outbreak on human health are not well understood. We may, for example, anticipate a biological outbreak that employs a bacterial host containing a foreign toxin, but the observed effects of the toxin may not implicate a known gene. Even in cases where the gene is known, it may be difficult to rule out a natural origin for the outbreak. In such cases, it may be useful to search for more direct evidence of the genetic engineering tools used to insert and express foreign genes in a bacterial host. Among the most widely used and readily available tools for genetic engineering in bacteria are artificial vector DNA molecules. Genetic engineering with artificial vectors began with efforts to improve on early work using natural plasmids for gene cloning. Natural plasmids are extrachromosomal replicons (self-replicating molecules) that come in both circular and linear form and are generally nonessential genetic material for the bacterial host but can confer important phenotypes such as virulence and drug resistance. These plasmids are mobile

FIGURE 20.1 Result of minimal subset clustering signatures for Crimean–Congo hemorrhagic fever virus (CCHFV) displayed against a whole-genome phylogenetic tree of available target genomes. Note that signatures 45 and 51 cover a wide range of isolates from one geographical location, whereas signatures 28, 39, and 50 cover isolates found in Eastern Europe. Signatures 1, 5, 17, and 22 are required to detect some historical isolates that are not likely to be in current circulation.

III. Methodology
We use the term “artificial vector” to refer to replicons created through human intervention to explicitly distinguish them from their natural plasmid precursors. Sequence features designed to support genetic manipulation form the basis for methods used to distinguish artificial vector sequence from natural plasmids. The most common artificial vector-specific feature is the multiple cloning site region, which is a sequence insert containing clusters of restriction enzyme sites used to facilitate insertion of the foreign gene elements. Selection marker genes also play an important role in selecting bacteria, which maintain the artificial vector. The gene transcription control unit, which includes a promoter sequence and transcription terminator sequence for the foreign gene elements, is also an important feature, along with the origin of replication site required for maintenance of the artificial vector in the bacterial colony (Solar et al., 1998). Detecting an artificial vector sequence in a mixed bacterial sample is best accomplished via metagenomic sequence analysis. We note that this task can be very difficult if any E. coli is present in the sample, as most artificial vectors are derived from an E. coli chassis.

Recent advances in synthetic biology and genetic editing techniques such as Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR) (Adli, 2018) make deliberate genetic engineering harder to detect than classical vector-mediated genomic insertion. The newer technologies make it easier to synthesize entire genomes or make arbitrary edits without leaving behind any “vector scars” that aid identification of nonnatural changes (Noyce et al., 2018). A recent National Academy of Sciences report (Biodefense) highlights some of the increasing concerns about advances in synthetic biology and the risks of deliberate or accidental harm. The implication for microbial forensics is that current technologies make it possible to make nearly arbitrary genomic constructs or changes without leaving obvious genetic signs that the construct is unnatural or connecting it to a specific individual or group.

Viral and bacterial detection array signatures

Numerous microarrays have been designed for viral discovery, detection, and resequencing (Wang et al., 2002; Palacios et al., 2007; Lin et al., 2006; Jabado et al., 2008). Resequencing arrays can provide sequence information for viruses closely related (0.90% similarity) to sequences from which the array was designed. Discovery arrays to detect more diverse and more distantly related organisms have been built using techniques for selecting probes from regions of known conservation based on BLAST nucleotide sequence similarity (Wang et al., 2003) or profile HMM and motif indications of amino acid sequence conservation (Jabado et al., 2008). Array design to span an entire kingdom on a single microarray demands substantial investment in probe selection algorithms. Beginning in 2007, LLNL designed a microarray to detect all bacteria, plasmids, and viruses based on all available whole-genome, whole-segment, and whole-plasmid sequences. We attempted to find probes that are unique to each viral and bacterial family and favor probes conserved within a family. We used probes 50–65 bases long, enabling sensitive detection of targets with some sequence variation relative to the probe. We used a greedy minimal set coverage algorithm to ensure that all database sequences (genomes or chromosomes or viral segments) have at least 50 (for viruses) or 15 (for bacteria and plasmids) probes per sequence. We allowed some mismatches between probe and target, based on previous mismatch experiments in which we determined that probes
with a contiguous match at least 29 bases long and with 85% sequence similarity between probe and target still gave a strong signal intensity (Gardner et al., 2010). We developed a novel statistical method that is based on likelihood maximization within a Bayesian network, incorporating a sophisticated probabilistic model of probe-target hybridization developed and validated with experimental data from hundreds of thousands of probe intensity measurements (McLoughlin, 2011). The method is designed to enable quantifiable predictions of likelihood for the presence of each of multiple organisms in a complex, mixed sample, which is especially important in an environmental sample or one with chimeric organisms.

Glass-slide platforms containing up to 400,000 total probes and capable of running 1, 2, 4, or 8 samples at a time were used for a very wide range of studies on human, animal, environmental, and product samples. This includes finding vaccine contaminants (Victoria et al., 2010), viral association with bladder cancer (Paradžik et al., 2013), identifying ancient pathogen DNA in archaeological samples (Devault et al., 2014), combat wound infection analysis (Be et al., 2014), and finding emerging viruses in clinical samples (Rosenstierne et al., 2014).

Recently, our detection array has migrated to a new high-throughput platform, the Applied Biosystems Axiom Microbiome Array (Thermofisher), which can process 24 or 96 samples at a time, each sample being exposed to 1.4M DNA probes about 35bp in length. Over 12,000 unique bacterial, viral, fungal, archaeal, and protozoan species are represented. With reagent costs as low as $40/sample, this array is best suited for screening large numbers of samples to determine which samples clearly have pathogens of interest present and which samples may benefit from further expense of metagenomic or isolate sequencing. Recent advances in targeted sequencing, which can be viewed as “liquid arrays with sequence readout,” may indicate that arrays are headed for niche roles in the future. One such development are the semiconductor arrays developed by DNA Electronics (Dnae), which can do targeted sequencing with direct electronic readout. While initial products are still under development (Genomeweb), the inherent cost scalability of semiconductor devices means that point-of-need targeted sequencing arrays are likely to become available in the relatively near future to compete with benchtop sequencing for detection of known organisms.

The future of genomic signatures

Issues related to scaling, taxonomy, and technology advances continue to be main drivers for the future of genomic signatures.

Scaling problems all stem from the exponential rate at which genomic sequence data are growing. In recent years, the PulseNet Project (https://www.cdc.gov/pulsenet/participants/international/wgs-vision.html) (Pulsenet) has deposited many thousands of complete bacterial foodborne pathogen genomes into GenBank, far exceeding the capacity of most previously used bioinformatics tools to handle them. At the time of this writing in August 2018, NCBI holds 12,000+ E. coli, 9000+ Salmonella, 2500+ Campylobacter, 2500+ Listeria, and 1800+ Shigella genomes. This represents nearly three orders of magnitude increase over the genomes available when the first edition of this book was prepared in early 2003.

Both analytical tools and techniques have evolved in recent years to deal with this flood of sequence information, much of which is in draft genome form. Comparative genomic tools involving MSA algorithms used in the early part of this century could not scale to thousands of genomes and have been replaced by alignment-free tools such as KRAKEN (Wood and Salzberg, 2014), LMAT (Ames et al., 2013), etc., which utilize “k-mers” (short strings of length k) to more efficiently locate regions of
similarity and difference. One benefit of this class of tools is that they work well whether the input sequence is raw reads from a pure microbial isolate or a complex clinical or environmental sample. Another technique referred to as “read mapping” compares reads from an unknown sample against a high-quality RefSeq genome to reduce complexity (Ncbi). While utilizing a small high-quality reference database provides faster analysis speed, good coverage of strain variation can only be achieved by utilizing draft genomes as well.

Signature design tools have been similarly challenged by the flood of available genomic sequence data. Digesting the complete set of available microbial finished and draft genomes to design the Applied Biosystems Axiom Microbiome Array described above required over 5 CPU years of computation time, using software that had not been revised to handle the explosive growth. An additional factor in scaling comparative genomic codes, including signature design codes, is that large-memory computation nodes can make a big difference in program execution time. A comparison of over 9000 full human genomes against all microbial genomes (Ames et al., 2015), looking for contamination in each, ran in 6 days on a cluster with over 800 Gb of memory on each node; running the same comparison on a standard cluster with 128 Gb of memory per node would have taken several months because accessing data over a network is one to two orders of magnitude slower than accessing local memory. The relationship between analysis algorithms and computer architecture needs to be considered to achieve optimal genomic analysis performance.

Earlier we mentioned difficulties with the evolving taxonomy of pathogenic organisms, as classification schemes originally developed based on phenomenology are faced now with genomic inconsistencies. The current flood of metagenomic data is presenting us with an even larger problem: what exactly do concepts such as “species” and “strains” mean if it turns out that microbial life is a broad spectrum with few well-defined transitions? It is now common to refer to a “core genome” and additional distinct gene content variations (“pan genome”) that presumably is responsible for different phenotypes (Nelson et al., 2004). It is possible that new concepts and terminology will be needed to map existing taxonomic categories into the genomic reality of the 21st century. Similarly, compressed genome data storage techniques, including just mapping differences relative to reference genomes of one or more species, may be leveraged to reduce data storage, transfer, and computation bottlenecks (Hosseini et al., 2016).

The rate of advancement in sequencing technology exceeds that even of computers, fueled by the promise of personalized medicine if individual drug and disease reactions can be determined and if individual genetic variation can be determined efficiently via low-cost sequencing. The field of pathogen diagnostics is riding this technology wave, too small a market to have any direct influence. Note that the read lengths of some new sequencing technologies may be too short to provide confident pathogen identification based on a single read, meaning that direct metagenomic identification of human pathogens from complex clinical or environmental samples contains some degree of uncertainty. Microarrays will have to ride their own faster/less expensive/more information-per-chip curve if they are not to become obsolete within a few years. Alternatively, one could argue that future advances in protein detection technology could lead to breakthroughs in fast dipstick assays (similar to current home pregnancy test kits) that could provide fast, accurate, and inexpensive results for pathogen detection. In all likelihoods, all these techniques will continue to compete as they evolve asynchronously.

Another technological advance is seen in the recent breakthroughs in gene and genome synthesis (Gibson et al., 2008) and editing
techniques such as CRISPR/Cas9 (Adli, 2018). Not only do we need to deal with emerging natural viruses from every remote corner of the planet, but now we also need to deal with the fact that for rapidly decreasing amounts of money, it is possible to synthesize combinatorial versions of any DNA one might wish to (re) create. This potential ability to create a new class of supercharged pathogens, as well as the possibility of synthesized pathogens that do not exist in nature, puts a new urgency into ensuring that we have adequate tools to deal with these evolving biothreats.

What all this means for genomic signature design is that we will have to exist in a combination of a data avalanche, new analysis tools, and rapidly evolving new technologies. Against this background of change, we will have to deal with new missions and new challenges from adversaries equipped with the latest technologies. Fittingly for biodefense, it is indeed a very Darwinian challenge that faces us.

**Protein signatures: a new forensic approach?**

Recent work at LLNL has established that amino acid changes in human hair, teeth, bone, and skin samples can provide discriminating power in cases where no DNA is available for standard human forensics based on short tandem repeats. Called GVPs (genetically variable proteins), they are the protein analogue of SNPs in DNA (Parker et al., 2016; Mason et al., 2018).

The initial emphasis of this work has been on hair which, due to its high protein content, makes an ideal forensic specimen for proteomic forensic analysis. Hair also has the added benefit of high stability over long time frames and under extreme conditions. Comprised primarily of keratins and keratin-associated proteins, hair exhibits high durability that contributes to its persistence. Packed into coiled-coils within the hair shaft, hair keratins are stabilized via cross-linking by cystine disulfide bonds as well as by isopeptide bonds between proteins (Zhang et al., 2015; Wolfram, 2003). Finally, evidence comprised of hair, bone, and other major tissue types of interest are easily parsed, largely eliminating complications associated with mixed or multicontributor samples that can be limiting for DNA analysis.

Although research continues to develop optimal sets of GVPs and to determine the maximum resolution possible, it has already been demonstrated that GVPs can provide objective and statistically valid identity discrimination. It has also been shown that it is possible to determine the sex from skeletal remains, specifically teeth, of children under 15 years of age, a range where classical morphological techniques are not reliable. In addition, different ethnicities have been shown to carry common amino acid mutations. These markers have utility in providing biogeographic background information, even from archaeological specimens. Once the procedures are validated and GVP reference libraries are established, GVPs will open up an entirely new field of forensics.

**Acknowledgments**

This chapter is an update of the previous edition, authored by Tom Slezak, Shea Gardner, Jonathan Allen, Elizabeth Vitalis, Marisa Torres, Clinton Torres, and Crystal Jaing from Lawrence Livermore National Lab. We honor the groundbreaking work in this field performed by the late Shea Gardner.

**Disclaimer**

This chapter was prepared as an account of work sponsored by an agency of the US government. Neither the US government nor Lawrence Livermore National Security, LLC, nor any of their employees makes any warranty, expressed or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed or represents that its use would not infringe on privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement,
recommendation, or favoring by the US government or Lawrence Livermore National Security, LLC. The views and opinions of the authors expressed herein do not necessarily state or reflect those of the US government or Lawrence Livermore National Security, LLC, and shall not be used for advertising or product endorsement purposes. This work was performed under the auspices of the US Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344.

References

Adli, M., 2018. The CRISPR tool kit for genome editing and beyond. Nat. Commun. 9, 1911.

Allen, E.E., Tyson, G.W., Whitaker, R.J., Detter, J.C., Richardson, P.M., Banfield, J.F., 2007. Genome dynamics in a natural archaeal population. Proc. Natl. Acad. Sci. U.S.A. 104 (6), 1883–1888.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J., 1990. Basic local alignment search tool. J. Mol. Biol. 215, 403-410.

Ames, S.K., Hysom, D.A., Gardner, S.N., Lloyd, G.S., Gokhale, M.B., Allen, J.E., September 15, 2013. Scalable metagenomic taxonomy classification using a reference genome database. Bioinformatics 29 (18), 2253–2260.

Ames, S., Gardner, S., Marti, J., Slezak, T., Gokhale, M., Allen, J., 2015. Using populations of human and microbial genomes for organism detection in metagenomics. Genome Res. (25), 1–13.

Banfield, J.F., Andersson, A., 2008. Virus population dynamics and acquired virus resistance in natural microbial communities. Science 230, 1047–1050.

Be, N., Allen, J., Brown, T., Chromy, B., Eldridge, A., Luciw, P., Forsberg, J., Kirkup, B., Elster, E., Jaing, C., 2014. Molecular profiling of combat wound infection through microbial detection microarray and next-generation sequencing. J. Clin. Microbiol. 52, 2583–2594. https://www.nap.edu/catalog/24890/biodefense-in-the-age-of-synthetic-biology.

Boratyn, G.M., Camacho, C., Cooper, P.S., Coulouris, G., Fong, A., Ma, N., Madden, T.L., Matten, W.T., McGinnis, S.D., Merezhuk, Y., Raytselis, Y., Sayers, E.W., Tao, T., Ye, J., Zaretskaya, I., 2013. BLAST: a more efficient report with usability improvements. Nucleic Acids Res. 41, W29-W33.

Devault, A.M., Jaing, C., Gardner, S., Porter, T.M., Enk, J., Thissen, J., Allen, J., Borucki, M., DeWitte, S.N., Dhody, A.N., Poinar, H.N., 2014. Ancient pathogen DNA in archaeological samples detected with a microbial detection array. Sci. Rep. 4, 4245. https://www.dnae.com/.

Frost, L.S., Leplae, R., Summers, A.O., Toussaint, A., 2005. Mobile genetic elements: the agents of open source evolution. Nat. Rev. Microbiol. 3, 722–732.

Gardner, S.N., Hall, B.G., 2013. When whole-genome alignments just won’t work: kSNP v2 software for alignment-free SNP discovery and phylogenetics of hundreds of microbial genomes. PLoS One 8 (12), e81760.

Gardner, S.N., Kuczmarek, T.A., Vitalis, E.A., T.R., 2003. Slezak, Limitations of TaqMan PCR for detecting divergent viral pathogens illustrated by hepatitis A, B, C, and E viruses and human immunodeficiency virus. J. Clin. Microbiol. 41, 2417–2427.

Gardner, S.N., Lam, M.W., Smith, J.R., Torres, C.L., Slezak, T.R., 2005. Draft versus finished sequence data for DNA and protein diagnostic signature development. Nucleic Acids Res. 33 (18), 5838–5850.

Gardner, S.N., Jaing, C., McLoughlin, K., et al., 2010. A microbial detection array (MDA) for viral and bacterial detection. BMC Genom. 11, 668. https://doi.org/10.1186/1471-2164-11-668 published Nov 25, 2010.

Gardner, S.N., Jaing, C.J., Elsheikh, M.M., Pena, J., Hysom, D.A., Borucki, M.K., 2014. Multiplex degenerate primer design for targeted whole genome amplification of many viral genomes. Adv. Bioinf. 2014, 8. https://www.genomeweb.com/sequencing/barda-awards-519m-dnae-develop-rapid-diagnostics.

Gibson, D.G., Benders, G.A., Andrews-Pfannkoch, C., Denisova, E.A., Baden-Tillson, H., Zaveri, J., et al., 2008. Complete chemical synthesis, assembly, and cloning of a Mycoplasma genitalium genome. Science 319, 1215–1220.

Hosseini, M., Pratas, D., Pinho, A.J., 2016. A survey on data compression methods for biological sequences. Information 7 (4), 56.

Hysom, D.A., Naraghi Arani, P., Elsheikh, M., Carrillo, A.C., Williams, P.L., Gardner, S.N., 2012. Skip the alignment: degenerate, multiplex primer and probe design using k-mer matching instead of alignments. PLoS One 7 (4). Article ID e34560.

Jabado, O.J., Palacios, G., Kapoor, V., Hui, J., Renwick, N., Zhai, J., Briese, T., 2006. Lipkin WI: greene SCPrimer: a rapid comprehensive tool for designing degenerate primers from multiple sequence alignments. Nucleic Acids Res. 34 (22), 6605–6611. https://doi.org/10.1093/nar/gkl966.

Jabado, O.J., Liu, Y., Conlan, S., Quan, P.L., Heygi, H., Lussier, Y., et al., 2008. Comprehensive viral oligonucleotide probe design using conserved protein regions. Nucleic Acids Res. 36 (1), e3.

Keim, P., Price, L.B., Kleevtska, A.M., Smith, K.L., Schupp, J.M., Okinaka, R., Jackson, P.J., Hugh-Jones, M.E., 2000. Multiple-locus variable-number tandem repeat analysis reveals genetic relationships within Bacillus anthracis. J. Bacteriol. 182, 2928–2936.

III. Methodology
Kurtz, S., Phillippy, A., Delcher, A.L., Smoot, M., Shumway, M., Antonescu, C., et al., 2004. Versatile and open software for comparing large genomes. Genome Biol. 5 (2), R12.

Lin, B., Wang, Z., Vora, G.J., Thornton, J.A., Schnur, J.M., Thach, D.C., et al., 2006. Broad-spectrum respiratory tract pathogen identification using resequencing DNA microarrays. Genome Res. 16 (4), 527–535.

Mason, K.E., et al., 2018. Protein-based forensic identification using genetically variant peptides in human bone. Forensic Sci. Int. 288.

McLoughlin, K., 2011. Microarrays for pathogen detection and analysis. Briefings Funct. Genomics 10 (6), 342–353.

https://www.ncbi.nlm.nih.gov/refseq/.

Nelson, K.E., Fouts, D.E., Mongodin, E.F., Ravel, J., DeBoy, R.T., Kolonay, J.F., et al., 2004. Whole genome comparisons of serotype 4b and 1/2a strains of the food-borne pathogen Listeria monocytogenes reveal new insights into the core genome components of this species. Nucleic Acids Res. 32 (8), 2386–2395.

Noyce, R.S., Lederman, S., Evans, D.H., 2018. Construction of an infectious horsepox virus vaccine from chemically synthesized DNA fragments. PLoS One 13, e0188453.

Palacios, G., Quan, P.L., Jabado, O.J., Conlan, S., Hirschberg, D.L., Liu, Y., et al., 2007. Panmicrobial oligonucleotide array for diagnosis of infectious diseases. Emerg. Infect. Dis. 13 (1), 73–81.

Paradzík, M., Bućević-Popović, V., Šitum, M., Jaing, C., Degoricija, M., McLoughlin, K., Ismail, S., Pundapolić, V., Terzić, J., 2013. Association of Kaposi’s sarcoma-associated herpesvirus (KSHV) with bladder cancer in Croatian patients. Tumor Biol. 1–6.

Parker, G.J., et al., 2016. Demonstration of protein-based human identification using the hair shaft proteome. PLoS One 11, e0160653.

https://www.cdc.gov/pulsenet/participants/international/wgs-vision.html.

Rosenstierne, M.W., McLoughlin, K.S., Olesen, M.L., Papa, A., Gardner, S.N., Engler, O., Plomet, S., Mirazimi, A., Weidmann, M., Niedrig, M., Fomsgaard, A., Erlandsson, L., 2014. The microbial detection array for detection of emerging viruses in clinical samples—a useful panmicrobial diagnostic tool. PLoS One 9, e100813.

Rozen, S., Skaltsky, H., 2000. Primer3 on the WWW for general users and for biologist programmers. Methods Mol. Biol. 132, 365–386.

Solar, G.D., Giraldo, R., Ruiz-Echevarria, M.J., Espinosa, M., Díaz-Orejas, R., 1998. Replication and control of circular bacterial plasmids. Microbiol. Mol. Biol. Rev. 62 (2), 434–464.

https://www.thermofisher.com/order/catalog/product/902904.

Victoria, J.G., Wang, C., Jones, M.S., Jaing, C., McLoughlin, K., Gardner, S., Delwart, E.L., 2010. Viral nucleic acids in live-attenuated vaccines: detection of minority variants and an adventitious virus. J. Virol. 84, 6033–6040.

Wang, D., Coscoy, L., Zylberberg, M., Avila, P.C., Boushey, H.A., Ganem, D., et al., 2002. Microarray-based detection and genotyping of viral pathogens. Proc. Natl. Acad. Sci. U.S.A. 99 (24), 15687–15692.

Wang, D., Urisman, A., Liu, Y.-T., Springer, M., Ksiazek, T., Erdman, D.D., et al., 2003. Viral discovery and sequence recovery using DNA microarrays. PLoS Biol. 1 (2), e2.

Wolfram, L.J., 2003. Human hair: a unique physicochemical composite. J. Am. Acad. Dermatol. 48, S106–S114.

Wood, D.E., Salzberg, S.L., 2014. Kraken: ultrafast metagenomic sequence classification using exact alignments. Genome Biol. 15, R46.

Zhang, Y., Alsop, R.J., Soomro, A., Yang, F.-C., Rheinstädter, M.C., 2015. Effect of shampoo, conditioner and permanent waving on the molecular structure of human hair. PeerJ 3, e1296.

III. Methodology