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.

18 Omics, Bioinformatics, and Infectious Disease Research

*Konrad H. Paszkiewicz and Mark van der Giezen**

Biosciences, College of Life and Environmental Sciences, University of Exeter, Exeter, UK

18.1 The Need for Bioinformatics

Although bioinformatics is generally perceived to be a modern science, the term had been put forward over thirty years ago by Paulien Hogeweg and Ben Hesper for "the study of informatic processes in biotic systems" (Hogeweg, 1978; Hogeweg and Hesper, 1978). It is necessarily nebulous—bioinformatics spans many disciplines and can have many shades of meaning. Indeed it can be argued that it is the collation and analysis of data from different disciplines that has provided some of the greatest insights. In the field of genomics and transcriptomics, bioinformatics is an incredibly diverse field. Evolution, epidemiology, ecology, and the response of an organism to its environment are all fields that require bioinformatics to accurately process and place into context various sources of data. At the heart of genomics and transcriptomics is the generation and analysis of vast quantities of sequence data. DNA sequencing took off in the late 1980s when Applied Biosystems developed the first automated sequencing machine. The subsequent development of more efficient ways to sequence resulted in the phenomenal growth of the number of sequences deposited in GenBank (Figure 18.1). Obviously, with over 100 million sequences deposited in GenBank, it is not feasible to do any serious manual work with such a large dataset. Data obtained from modern second-generation sequencers is on the order of 1000 times greater than capillary-based sequencers. It is now possible to routinely generate many gigabases of sequence data. Bioinformatics is tasked with making sense of it, mining it, storing it, disseminating it, and ensuring valid biological conclusions can be drawn from it. Many of the recent high-throughput functional genomics technologies rely on a bioinformatics component, though bioinformatics is just one part of the process. For example, identification of proteins by mass spectroscopy, quantitative analysis of expression data, phylogenetics, and so on all make use of bioinformatics tools, methods, and databases. Bioinformatics plays a key role at several steps in genomics,

*Email: m.vandergiezen@exeter.ac.uk

Genetics and Evolution of Infectious Diseases. DOI: 10.1016/B978-0-12-384890-1.00018-2

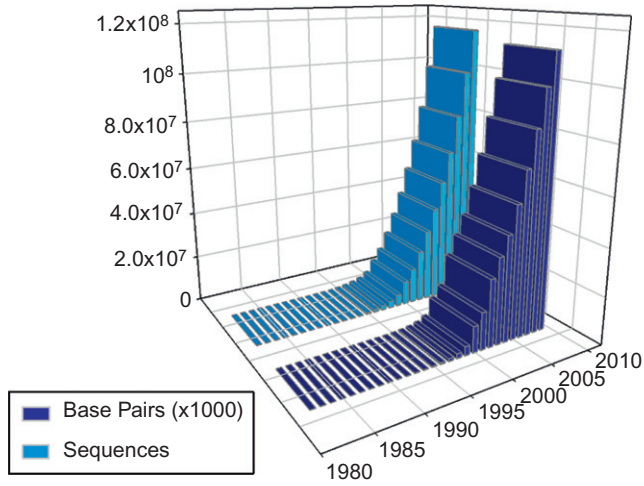

Figure 18.1 The growth of sequences submitted to GenBank. For further info, see http://www.ncbi.nlm.nih.gov/genbank.

comparative genomics, and functional genomics: sequence alignment, assembly, identification of single nucleotide polymorphisms (SNP), gene prediction, quantitative analysis of transcription data, etc. In this chapter, we will discuss the current state of play of bioinformatics related to genomics and transcriptomics and use relevant examples from the field of infectious diseases.

18.2 Metagenomics

The term "metagenomics" was originally used to describe the sequencing of genomes of uncultured microorganisms in order to explore their abilities to produce natural products (Handelsman et al., 1998, Rondon et al., 2000) and subsequently resulted in novel insights into the ecology and evolution of microorganisms on a scale not imagined possible before (see Cardenas and Tiedje, 2008; Hugenholtz and Tyson, 2008 for an overview). However, metagenomics now finds use in infectious disease research as well as the random sequencing of genomes from a variety of organisms from, for example, patient material that could lead to the identification of the cause of disease.

In a quite straightforward metagenomics approach to identify pathogens in sputa from cystic fibrosis patients, standard microbiological culture techniques were compared to molecular methods using 16S rDNA PCR (Bittar et al., 2008). The well-known disadvantage of the microbiological methods is that they normally employ "selective" media that are designed to pick up those bacterial pathogens that are thought to be present. Emerging pathogens will be missed using traditional culture techniques. Indeed, Bittar et al. identified 33 bacteria using cultivation while 53 bacterial species were detected using molecular methods (based on BLAST comparisons; Altschul et al., 1990), interestingly, 30% of the latter were

anaerobes, organisms missed in the routine cultivation methods. Many bacteria identified using the molecular methods are traditionally not thought to be associated with cystic fibrosis. Whether these novel species are associated with the physiopathology of disease remains to be studied. Bittar et al. (2008) also noted that the number of bacteria detected increased with increased numbers of clones sequenced, a well-known phenomenon in environmental sequencing that relates to sample depth (Huber et al., 2007; Huse et al., 2010). However, with the increased use of next-generation sequencing methods in infectious disease research, the lessons learned from environmental studies relating to diversity and relative abundance of different microbes can be put to effective use.

An example of the use of second-generation sequencing in a metagenomics approach of patient material is the study by Nakamura et al. (2009) to identify viruses in nasal and fecal material. In this study, RNA was isolated from patient material obtained during seasonal influenza infections and norovirus outbreaks. This RNA was reverse transcribed into cDNA, which was subsequently subjected to large-scale parallel pyrosequencing resulting in 25,000 reads on average per sample. Although the influenza samples were mainly (>90%) human in origin, it was nonetheless possible to identify the influenza subtypes in each sample (Nakamura et al., 2009). As the fecal samples were cleared of human and bacterial cells, yields were much better and the complete norovirus GII.4 subtype genome was sequenced with an average cover depth of up to 258×. In addition to being able to identify the influenza and noroviruses, two recently identified human viruses were also identified: WU polyomavirus and human coronavirus HKU1 (Nakamura et al., 2009). Major bacterial species normally found in the respiratory tract were also identified. Although Nakamura et al. suggest that the high-throughput sequencing is more sensitive than standard PCR-based analysis and might result in the detection of additional possible pathogens, they also warn that the increased sensitivity might necessitate follow-up work to decide which of the detected pathogens is the actual cause of the disease.

Important results are expected from the Human Microbiome Project (http://www.hmpdacc.org/), which will obtain metagenomic information from various human microenvironments such as the gastrointestinal, nasooral, and urogenital cavities as well as the skin. Understanding the human microbiome is thought to answer questions such as whether changes in the human microbiome are related to human health. However, large-scale metagenomics projects that include eukaryotic genomes have thus far been quite costly and laborious due to the generally large genomes of eukaryotes. The lowering of sequencing costs may alleviate part of the problem, but sequence data are still accumulating at a faster rate than developments in computational analysis (Hugenholtz and Tyson, 2008).

18.3 Comparative Genomics

Organisms that have attracted the attention of genome centers are those that cause disease followed by those from model organisms such as *Saccharomyces cerevisiae* (Goffeau et al., 1996) and *Caenorhabditis elegans* (the C. elegans Sequencing

Consortium, 1998), for example. Indeed, the first bacterial genomes sequenced were those from pathogens (Fleischmann et al., 1995; Fraser et al., 1995; Tomb et al., 1997), and these were preceded by many bacteriophage genomes such as bacteriophage MS2 (Fiers et al., 1976) and φX174 (Sanger et al., 1977) and viral genomes (Fiers et al., 1978). Currently, pathogen genomes represent at least one third of all sequenced genomes.

Obviously, for comparative genomics two genomes are required, and indeed, when the second bacterial pathogen was sequenced (*Mycoplasma genitalium* by Fraser et al., 1995), it was immediately compared with the first one (*Haemophilus influenzae* by Fleischmann et al., 1995). Interestingly, the *H. influenzae* genome was completed using a "bioinformatics" approach. Unlike previous sequencing projects, the used shotgun approach relied on a computational justification that sufficient random sequencing of small fragments would result in a complete coverage of the whole genome. Comparing the *M. genitalium* genome with the *Haemophilus* genome suggested that the percentage of the total genome dedicated to genes is similar albeit that *M. genitalium* has far fewer genes (Fraser et al., 1995). Although the genome of *M. genitalium* is about three times smaller than that of *H. influenzae*, its smaller genome has not resulted in an increase in gene density or decrease in gene size. Detection of several repeats of components of the *Mycoplasma* adhesin, which elicits a strong immune response in humans, suggests that recombination might underlie its ability to evade the human immune response. That this initial genome study was only the tip of the comparative genomics iceberg was already clear from Fleischmann et al. (1995) last sentence: "Knowledge of the complete genomes of pathogenic organisms could lead to new vaccines." A whole-genome effort at identifying vaccine candidates appeared some 5 years later when Pizza et al. (2000) employed bioinformatics to extract putative surface-exposed antigens by genome analysis. Although effective vaccines against *Neisseria meningitidis*, the causative agent of meningococcal meningitis and sepsis, did exist, these vaccines did not cover all pathogenic serogroups. Serogroup B had evaded the development of a good vaccine as its capsular polysaccharide (against which the vaccines of the other serogroups were developed) is identical to a human carbohydrate. In order to identify putative candidates for vaccine development, Pizza et al. decided to sequence the whole genome of a serogroup B strain. All potential open reading frames (ORFs) were analyzed for putative cellular locations using BLASTX. Those ORFs likely to be cytosolic were excluded from further analysis. The remaining ORFs were analyzed to determine whether they encoded proteins that contained transmembrane domains, leader peptides, and outer membrane anchoring motives using a variety of databases such as Pfam (Finn et al., 2010) and ProDom (Servant et al., 2002). This resulted in 570 ORFs encoding putative exposed antigens. These 570 putative genes were cloned in *Escherichia coli* and Pizza et al. successfully expressed 350 ORFs. These 350 recombinant proteins were used to generate antisera that were tested in enzyme-linked immunosorbent assay (ELISA) and fluorescence-activated cell sorter (FACS) analyses to test whether they detected proteins on the outer surface of serogroup B meningococcus strains. In addition, the sera were tested for bactericidal activity. Of the 350

proteins, 85 reacted positively in at least one assay but only 7 were positive in all three assays. These 7 were subsequently tested on a large variety of strains to analyze their efficacy. A total of 5 seemed able to provide protection against 31 *N. meningitidis* strains and in addition, those 5 proteins are 95–99% similar to the homologous *N. gonorrhoeae* proteins, suggesting they might provide successful protection against that pathogen as well (Pizza et al., 2000). Arguably the most striking aspect of this study is that in 18 months the authors identified more vaccine candidates than in the preceding 40 years using a novel genomics/bioinformatics approach (Seib et al., 2009). This study resulted in a vaccine that is currently in Phase III clinical trials (Giuliani et al., 2006).

Protozoan infections are a major burden on developing nations; they take 8 of the 13 diseases targeted by the World Health Organization's Special Program for Research and Training in Tropical Diseases (http://www.who.int/tdr). Over the last 5 years or so, more than 10 parasitic genomes have been sequenced in the hope that their sequences would reveal weak spots to target these pathogens. The trypanosomatids cause serious disease in Africa and South America. *Trypanosoma brucei* causes sleeping sickness in humans and wasting disease in cattle. *Trypanosoma cruzi* is the causative agent of Chagas disease and *Leishmania major* leads to skin lesions. The completion of their genomes (Berriman et al., 2005, El-Sayed et al., 2005a, Ivens et al., 2005) and the comparative analysis of all three genomes (El-Sayed et al., 2005b) may be able to focus efforts toward obtaining vaccines, as current drugs have serious toxicity issues. Although their genomes encode a different number of protein-encoding genes (around 8100 in *T. brucei*; 8300 in *L. major*; 12,000 in *T. cruzi*), comparative analysis resulted in the identification of about 6200 genes that entail the trypanosomatid core proteome. All protein coding genes were compared in a three-way manner using BLASTP (El-Sayed et al., 2005b) and the mutual best hits were grouped as clusters of orthologous genes or COGs (Figure 18.2).

Trypanosomatid specific proteins from these 6200 might be used in a broad-scale vaccine. The remainder of the protein-encoding genes from each parasite (26% of the genes in *T. brucei*; 12% in *L. major*; 32% in *T. cruzi*) consists of species-specific genes. Interestingly, a large proportion of these genes encode surface antigens and this might relate to the different mechanisms these parasites employ to evade the host immune system. In addition, it was noted that many genes

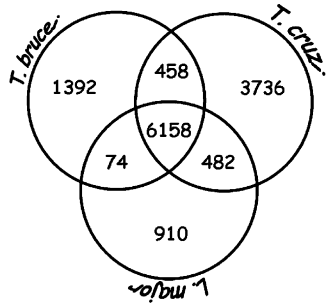

Figure 18.2 Kinetoplastid comparative genomics. A three-way comparison of all protein coding genes from *Trypanosoma cruzi*, *Trypanosoma brucei*, and *Leishmania major* resulted in the discovery of 6200 core proteins that all three kinetoplastids share and various dually shared and unique proteins.
Source: Adapted from El-Sayed et al. (2005b).

encoding surface antigens are found at or near telomeres and that many retroelements seem to be present in these regions as well. This might be related to the enormous antigenic variation observed in both *Trypanosoma* species. The presence of novel genes in these areas might suggest that their products play an unknown role in antigenic variation as well which warrants further studies into these uncharacterized genes (El-Sayed et al., 2005b).

Detailed knowledge of well-studied pathogens might be successfully used to understand the biology of closely related emerging pathogens. This was the driving force for the sequencing of six *Candida* species (Butler et al., 2009). *Candida* species are the most common opportunistic fungal infections in the world and *C. albicans* is the most common of all *Candida* species causing infection. However, *C. albicans* incidence is declining while other species are emerging. Comparison of eight *Candida* species indicated that although genome size was variable, gene content was nearly identical across all species. As the analysis included pathogenic and nonpathogenic species, Butler et al. (2009) specifically studied differences between these two groups. Of the over 9000 gene families analyzed, 21 were significantly enriched in pathogenic species. Many gene families known to be involved in pathogenesis were present in these 21 families (e.g., lipases, oligopeptide transporters, and adhesins). More interestingly, several poorly characterized gene families were also identified, suggesting these might play an unexpected role in pathogenesis as well. This comparative study revealed a wealth of new avenues to explore, which, combined with the large body of work performed on *C. albicans*, will aid understanding the newly emerging pathogenic *Candida* species (Butler et al., 2009).

18.4 Pan-Genomics

Although comparative studies using multiple species can reveal hitherto unknown features as evidenced from the mentioned trypanosomatid and *Candida* studies, they can also reveal something unexpected. Because the definition of a bacterial species has been debated for a long time, Tettelin et al. (2005) set out to address this question by sequencing multiple strains from *Streptococcus agalactiae*, the most common cause of illness or death among newborns. Unexpectedly, despite the presence of a "core-genome" shared between all 8 genomes, mathematical modeling suggested that each additional sequenced genome would add 33 new genes to the "dispensable genome." An additional analysis using *S. pyogenes* also suggested that sequencing additional genomes would continue to add new genes to the pool resulting in a pan-genome that can be defined as the global gene repertoire of a species (Medini et al., 2005). This cannot be extrapolated *ad infinitum*, as a similar analysis of *Bacillus anthracis* indicated that after the fourth genome, no additional genes were identified (Tettelin et al., 2005) in agreement with its known limited genetic diversity (Keim and Smith, 2002). Subsequent analyses have confirmed the presence of pan-genomes for many bacterial species (Hiller et al., 2007; Lefébure and Stanhope, 2007; Rasko et al., 2008; Schoen et al., 2008; Lefébure and Stanhope, 2009) and the ultimate gene repertoire of a bacterial species is much

larger than generally perceived. Whether this would be the case for eukaryotes remains to be shown.

Despite the apparently ever-expanding possibilities of the pan-genome, it has also resulted in a universal vaccine candidate for group B *Streptococcus* (GBS). Because various GBS serotypes exist, current vaccines only offer protection against a limited set of serotypes. Eight genomes from six serotypes were compared resulting in the identification of a core-genome of 1811 genes and a dispensable genome of 765 genes, which were not present in each strain (Maione et al., 2005). Both genomes were analyzed for the presence of putative surface-associated and secreted proteins. Of the 598 identified genes, one third were part of the dispensable genome (193 genes). The authors subsequently produced recombinant tagged proteins in *E. coli* that were used to immunize mice. Ultimately, a combination of four antigens turned out to be highly effective against all major GBS serotypes. Three of these antigens were part of the dispensable genome. In addition, this bioinformatics approach highlights the importance of not dismissing unidentified ORFs on genomes (generally up to 50% of sequenced genomes) as all four antigens had no assigned function. Because of their identification using this method, it became obvious they were part of a pilus-like structure that had never seen before in Group B *Streptococcus* (Lauer et al., 2005). The presence of antigens that provide protection on these pilus-like structures suggest that these might play a role in pathogenicity.

18.5 Transcriptomics

Genomic information is useful as a scaffold. However, in a given environment pathogens and hosts only express a subset of their genes at any one time. The presence of pan-genomes only complicates matters even more. To investigate the response of an organism to an environmental or other stress it is necessary to examine the expression pattern of proteins. At present, this is not possible to accomplish directly on a large scale, but a good approximation can be made by sequencing and counting mRNA molecules. At present the process involves converting the RNA to cDNA, which can introduce biases but nonetheless sequencing has a great many advantages over traditional microarrays (Ledford, 2008). These include high specificity with little or no background noise and one also gains nucleotide level resolution of expression. Despite such drawbacks, microarrays are still extremely powerful tools to understand levels of gene expression, and this is obvious from the study by Toledo-Arana et al., who discovered novel regulatory mechanisms in *Listeria* (Toledo-Arana et al., 2009). *L. monocytogenes* is normally harmless but can lead to serious food-borne infections. Environmental change, from the soil through the stomach to the intestinal lumen and ultimately into the bloodstream, is thought to be responsible for the up- and downregulation of a plethora of genes. Comparative genomics of the nonpathogenic *L. innocua* has resulted in the identification of a virulence locus (Glaser et al., 2001). Using microarrays, transcripts of one strain grown at 37°C in rich medium were compared to three different conditions: stationary phase, hypoxia, and low temperature (30°C). In addition,

knockout mutants in three known regulators of *Listeria* virulence gene expression (PrfA, SigB, and Hfq) were compared to the control strain as well. RNA was also extracted from the intestine of inoculated mice and from blood from healthy human donors that were both infected with three different strains (control and PrfA and sigB knockouts). This analysis resulted in the discovery of massive transcriptional reshaping under the control of SigB when *Listeria* enters the intestines. However, in the bloodstream, gene expression is under control of PrfA. Various noncoding RNAs were uncovered, which show the same expression patters as virulence genes suggesting a potential role in virulence (Toledo-Arana et al., 2009).

Because microarray data are based on a comparative difference in hybridization, high-throughput next-generation sequencing is seen as more quantitative as it based on number of hits for each sequenced transcript (van Vliet, 2010). However, when making cDNA for next-generation sequencing transcriptomics in prokaryotes, there are several difficulties not found in eukaryotes, such as high levels of rRNA and tRNA molecules as well as a lack of poly-A tails, making extraction difficult. Nontheless, it is possible to overcome these by either reducing the amount of rRNA and tRNA using commercially available kits or by bioinformatic removal of such sequences postsequencing (van Vliet, 2010). To date, some 20 RNA-seq style experiments have been performed on prokaryotes. To give an example of the sort of novel insights that can be gleaned using such technology, Passalacqua et al. (2009) sequenced the *Bacillus anthracis* transcriptome using SOLiD and Illumina sequencing and clearly showed the polycistronic nature of many transcripts on a whole genome scale. Although known for individual operons, this had never been shown on a genome-wide scale. They were also able to test the current genome annotations and discovered that 36 loci that were removed as nongenes showed significant transcriptional activity. In addition, 21 nonannotated regions had clear levels of transcription and should therefore be considered as genes (Passalacqua et al., 2009). As internal methionines could have incidentally been identified as start codons, they also checked whether upstream regions were included in the transcribed region. In 11 cases this proved to be the case suggesting the original start codons were incorrectly annotated. Reassuringly, when comparing their data with microarray data, a strong correlation was observed. Interestingly, because of the very high resolution of sequence-based transcriptomics studies, it is possible to identify novel regulatory elements. For example, when comparing expression levels under O_2- and CO_2-rich conditions, the first gene of an eight-gene operon did not show a marked difference in expression level while all the others were significantly upregulated under CO_2 (Passalacqua et al., 2009). Indeed, a bioinformatics approach had suggested the presence of a T-box riboswitch between genes 1 and 2 of this operon (Griffiths-Jones et al., 2005).

A similar approach to study how *Burkholderia cenocepacia*, an opportunistic cystic fibrosis pathogen, responds to environmental changes revealed several new potential virulence factors (Yoder-Himes et al., 2009). As *B. cenocepacia* is routinely isolated from soil, two strains (one isolated from a cystic fibrosis patient and one from soil) were analyzed in their response to changes from growth at synthetic human sputum medium and soil medium. Although their overall nucleotide identity is 99.8%, 179

and 120 homologous genes showed a significant difference in expression between the two strains when grown in synthetic sputum medium and soil medium, respectively. This suggests that despite the high level of relatedness, differential gene expression plays a large role in adaptation to their ecological niche (Yoder-Himes et al., 2009). Interestingly, similar to Passalacqua et al. (2009), several expressed noncoding RNAs were uncovered with different expression levels depending on environmental condition. The significance of this needs to be investigated but highlights the ability of second-generation sequencing to unearth novel findings.

18.6 Proteomics

Despite the fact that a species' genome could well be larger than the actual genome content of one member of that species due to the pan-genome concept, an organism's proteome is by far much more complex. As discussed earlier, transcriptomics will reveal which subset of the genome is expressed under a given condition. However, posttranslational modifications of proteins make the actual proteome far more complex than the transcriptome. This is also the strength of proteomics, as can be seen in a study of the obligate intracellular parasite *Chlamydia pneumonia*. *C. pneumonia* is the third-most-common cause of respiratory infections in the world, which, in part, is made possible due to the unique bi-phasic life cycle of this bacterial pathogen. *Chlamydia* spread via a metabolically inert infectious particle called the elementary body. These elementary bodies enter the host cell where they differentiate into reticulate bodies. As the elementary body is the infectious phase, proteins presented on the outer membrane would be ideal candidates for vaccine development, especially as effective vaccines are lacking and treatment is via antibiotic therapy. A large-scale genomics-proteomics study by Montigiani et al. (2002) systematically assessed putative exposed antigens for possible use in vaccine development. Of the 1073 *C. pneumonia* genes, 636 have assigned functions, 72 of the latter are predicted to be peripherally located and were therefore selected for follow-up studies. In addition, the remaining 437 ORFs were subjected to a series of search algorithms aimed at identifying putative surface-exposed antigens. In total, 141 ORFs were identified as being possibly located on the cell surface. These 141 were subsequently used to produce recombinant proteins in *E. coli*. Because both His-tagged as well as GST-tagged versions were made, a total of 173 recombinant proteins were produced and used for immunizations of mice. All antisera were used in FACS analysis to test if they could bind to the *C. pneumonia* cell surface. This resulted in the identification of 53 putative surface-exposed antigens. Interestingly, apart from well-known antigens, 14 antigens from unidentified ORFs were part of this group of potential vaccine candidates. All 53 candidates were tested on Western blots whether they generated a clean band of the expected size or whether they cross-reacted with other proteins; 33 of the 53 were specific. Finally, Montigiani et al. conducted a proteomic analysis of total protein from the elementary body phase identifying spots using mass spectrometry. Protein sequencing using MALDI-TOF identified 28 putative surface-exposed antigens on the

C. pneumonia 2D gels (Montigiani et al., 2002). A follow-up study by Thorpe et al. (2007) clearly showed that one of the identified candidates, LcrE, induced, amongst others, $CD4^+$ and $CD8^+$ T cell activation and completely cleared infection in a murine model. Interestingly, LcrE is homologous to a protein that is thought be part of the Type III secretion system of *Yersinia*. The exposed nature of LcrE on the *C. pneumonia* cell surface suggests that a Type III secretion system plays a role in *Chlamydia* infection (Montigiani et al., 2002).

The importance of exposed outer membrane proteins as potential vaccine candidates has prompted Berlanda Scorza et al. to assess the complement of outer membrane proteins from an extraintestinal pathogenic *E. coli* strain (Berlanda Scorza et al., 2008). Extraintestinal pathogenic *E. coli* is the leading cause of severe sepsis and current increases in drug resistance warrant the search for novel vaccine targets. In addition, current whole-cell vaccines suffer from undesired cross-reactions to commensal *E. coli* as well. The novel approach by Berland Scorza et al. is based on the observation that some Gram-negative bacteria release outer membrane vesicles (OMV) in the culture media, albeit in minute quantities. A TolR mutant appeared to release much more OMVs than wild-type cells and subsequent large-scale mass spectroscopic analysis of its protein content resulted in the identification of 100 proteins. The majority of these were outer membrane and periplasmic proteins. Intriguingly, three subunits from the cytolethal distending toxin (CDT) were included. This toxin is unusual in that one of its subunits is targeted to the eukaryotic host cell, where it breaks double-stranded DNA resulting in cell death (De Rycke and Oswald, 2001). To check whether the presence of CDT in the OMV was due to the TolR knockout, wild-type extraintestinal pathogenic *E. coli* was tested using Western blotting. Indeed, CDT was detected in wild-type OMV as well (Berlanda Scorza et al., 2008). This suggests that toxin delivery via vesicles might well be the key event in pathogenesis. Interestingly, 18 of the 100 identified proteins were not predicted to be targeted to the periplasm or outer membrane by PSORTb (Gardy et al., 2005). We see here excellent opportunities to train protein targeting algorithms with new wetbench data as these algorithms generally have been trained on a limited set of model organisms that do not reflect the diversity encountered in real life.

18.7 Structural Genomics/Proteomics

Despite the enormous progress in genomics of infectious diseases, the discovery of new drugs has not kept equal pace. For example, no candidate drugs have been identified after 70 high-throughput screens using validated bacterial drug targets (Payne et al., 2007). Although broad-spectrum drugs might be more desirable, there has been a recent trend in targeting specific proteins from specific pathogens using structural biology. Several structural genomics initiatives have been set up to target specific groups of pathogens. For example, the Seattle Structural Genomics Center for Infectious Diseases (http://ssgcid.org) and the Center for Structural Genomics of Infectious Diseases (http://www.csgid.org) work on category A to C agents listed by the National Institute for Allergy and Infectious Diseases (NIAID). Other

centers focus on specific organisms such as *Mycobacterium tuberculosis*. Examples are the Mycobacterium Tuberculosis Structural Proteomics Project (http://xmtb.org) and the Mycobacterium Tuberculosis Structural Proteomics Consortium (http://www.doe-mbi.ucla.edu/TB). The field of structural genomics aims to solve as many protein structures as possible from human pathogens with the aim to come up with new drug targets or vaccines (Van Voorhis et al., 2009). Obviously, correct selection of candidates for structural genomics projects is paramount and various criteria have been put forward (Anderson, 2009; Van Voorhis et al., 2009). If a protein is already a validated drug target obviously aids in selection. The proteins need to be essential for the pathogen and ideally, absent in humans. Proteins involved in the uptake of essential nutrients are another target. Classically, drug design has been focusing on substrate binding sites. More recently, small molecules interfering with subunit binding have started to attract attention. As eukaryotic and prokaryotic inorganic pyrophosphatases differ in composition (the former are homodimers, while the latter are homohexamers), efforts are aimed at compounds that interfere with the oligomeric state of the enzyme. In contrast, the highly conserved active site of inorganic pyrophosphatase would not have been a good target (Van Voorhis et al., 2009). The 2003 SARS outbreak that caught the infectious diseases community (if not the whole world) by surprise is one example where structural genomics has made enormous progress. Despite knowing that coronaviruses caused serious diseases in animals, the fact that they only caused mild disease in humans meant that there was very little knowledge about coronavirus biology. The subsequent effort to understand viral assembly and replication/transcription, for example, has resulted in the elucidation of 12 SARS-CoV solved protein structures. Interestingly, the novel fold-discovery rate was nearly 50%, while it would normally be more close to 6% (Bartlam et al., 2007). In addition, one key protein, the SARS-CoV main protease, has since been at the center of structure-based drug discovery. Because of the nature of the discipline, structural genomics is dependent on various other disciplines such as biochemistry, microbiology, structural biology, computational biology, and bioinformatics and can only foster in a truly interdisciplinary environment (Anderson, 2009).

18.8 A "How-To" of Second-Generation Sequencing

It is now possible to sequence the entire genome of a bacterial pathogen, assemble the raw sequence reads, perform automated annotation, and visualize the results within 3 weeks. At the same time (indeed even on the same sequencer) it is also possible to selectively sequence the transcriptome (RNA-seq) regions of DNA bound to protein (ChIP-Seq) or for relevant species methylated DNA to study epigenetic effects as well as small RNA molecules. It is also possible to perform the very same sequencing on the host organism at the same time.

Bioinformatic algorithms and tools are a crucial tool in analyzing such unprecedented volumes of data. These data volumes have emerged as a result of second-generation sequencers such as the Roche/454, Illumina, and ABI/SoLID systems.

Although useful information can be extracted by single researchers by targeted analysis of the sequencer output, to gain the most information out of such data, it is becoming increasingly common for multiple researchers or research groups with widely differing areas of expertise to collaborate. This collaboration is absolutely crucial if relevant insights are to be gained from large-scale datasets. As a result a vast array of data is generated, which is required to be annotated and curated as well as analyzed for information relevant to any particular experiment. In addition this information needs to be stored, shared, and distributed in a manner that enables reanalysis if and when new hypotheses are generated.

Platforms as produced by the GMOD consortium (http://gmod.org), such as Gbrowse, and underlying databases are excellent web-based tools for visualizing and comparing datasets. However, they currently offer limited scope for collaborative annotation or curation of datasets where relevant expertise can be brought to bear from a variety of different research groups. This problem is magnified with the advent of second-generation sequencers since much smaller groups of researchers tend to be involved, meaning that the expertise that large collaborations can muster (such as the Influenza Research Database [FluDB], http://www.fludb.org/) is much smaller. Thus there is a need for integrated annotation and visualization pipelines to enable individual researchers to perform comparative genomics and transcriptomics.

The Broad Institute offers a number of useful visualization tools to the individual researcher such as ARGO (http://www.broadinstitute.org/annotation/argo/) and the Integrated Genome Viewer (IGV) (http://www.broadinstitute.org/igv/). ARGO offers the ability to manually annotate and visualize a genome as well as provide a good graphical overview for comparative genomics and transcriptomics.

Currently, there is no one standard for bioinformatics pipeline development for next-generation sequencing. Several efforts are underway or can be adapted from Sanger sequencing pipelines. These include the prokaryote annotation pipeline XBase and the ISGA server (Hemmerich et al., 2010). These enable de novo sequenced prokaryote genomes to be annotated automatically and corrected manually at a later date. Alternative Sanger adaptations such as Maker can also be used once an assembly has been generated.

18.9 Alignment or Assembly of Second-generation Sequences

A large array of programs is now available to either align reads to a reference genome or to assemble them de novo (Miller et al., 2010; Paszkiewicz and Studholme, 2010). They will not be listed in detail here as there are many considerations, including sequencing platform used, the read length in use, the expected genome size, length of longest repetitive elements, GC content, and whether paired-end reads are in use.

The proprietary Newbler software from Roche is the most popular method of de novo assembly of 454 reads (typically 400−500bp). Popular assemblers for short reads (i.e., mostly from Illumina or SoLID platforms) are Velvet

(http://www.ebi.ac.uk/~zerbino/velvet) for the assembly of genomic DNA or Oases from the same group dealing with assembly of reads from transcriptomic cDNA (http://www.ebi.ac.uk/~zerbino/oases) (Zerbino and Birney, 2008). Other assemblers such as AbYSS (Simpson et al., 2009), ALLPATHS (Butler et al., 2008) or SOAPdenovo (http://soap.genomics.org.cn/soapdenovo.html) are also popular. AbYSS enables assembly to be parallelized, thus speeding up assembly. ALLPATHS has been shown to offer superior performance when multiple paired-end libraries are used. Independent of read length, it is crucial that paired-end libraries are used when constructing de novo assemblies of any genome. Note that the use of short-read sequences only can lead to significant gaps being left in the final assembly due to repetitive elements. However, for many analyses (especially for prokaryotic organisms) these gaps are generally not considered to be significant. In cases where closure of these gaps is more desirable than the addition of 454, Sanger or long-range PCR data can often help.

Where significant quantities of long- and short-read data are available, then a joint assembly can be attempted. A recommended protocol is to assemble the short and long reads separately using their respective packages and to then merge the two assemblers using programs such as Minimus (Sommer et al., 2007). Another option is to use a template sequence from a related organism to help guide the assembly (note—this is distinct from remapping as described). The amosCMP package is useful for this purpose (Pop et al., 2004). Finally, whatever assembly method is used, it is important to remember that a longer assembly is not necessarily a better one. Examining the reads making up a contig (e.g., using the AMOS package (http://amos.sourceforge.net) or the Tablet viewer (http://bioinf.scri.ac.uk/tablet) and alignment to a core-conserved group of genes should be standard practice to ensure that blatant errors are corrected. Remapping of short reads to a reference genome is also a valid method of comparison. Although software such as BLAT (Kent, 2002) can be used with longer 454 reads, it is not an ideal tool for shorter read technologies where data volumes are much greater. Where such a genome is available, software such as MAQ, its successor, BWA, Bowtie, SOAP, and others offer a wealth of tools to identify indels, SNPs, and other variants which may be of interest. Crucially in these cases it is important to have sufficient depth of coverage to ensure SNP calls are valid. Paired-end data is also valuable to have to highlight the presence of indels. After remapping it is also common practice to assemble unmapped reads using the de novo assembly software to reveal any novel sequence variants, which may be absent in the reference. In the case where pathogens and hosts are sequenced together, if the sequence of at least one is known, then it is relatively straightforward to separate the two using bioinformatic techniques. To deal with transcriptomic data where a reference sequence is available, softwares, such as ERANGE (http://woldlab.caltech.edu/rnaseq/), Tophat (Trapnell et al., 2009), and Cufflinks (http://cufflinks.cbcb.umd.edu/), are extremely useful. The Cufflinks module in particular offers the ability to predict the most likely exon isoform expression pattern using a combination of Bayesian statistics and graph-based algorithms.

18.10 Concluding Remarks

We are aware that our treatment of the use of "omics" and bioinformatics in infectious disease research is not exhaustive. As mentioned in the introduction, what constitutes bioinformatics is not entirely clear and arguably varies depending on who tries to define it. However, we have attempted to show the considerable progress in infectious diseases research that has been made in recent years using various "omics" case studies. In addition, the last section is an attempt to provide a brief overview of the problems and (bioinformatics) solutions that current-day scientists face who embark on second-generation sequencing strategies. This is a fast-moving field, but the provided references and websites should be a good first approach for those who wish to make further strides toward eradicating infectious diseases from our planet.

Acknowledgments

We would like to acknowledge our colleague Dr. David J. Studholme for his suggestions and feedback.

References

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.

Anderson, W.F., 2009. Structural genomics and drug discovery for infectious diseases. Infect. Disord. Drug. Targets 9, 507–517.

Bartlam, M., Xu, Y., Rao, Z., 2007. Structural proteomics of the SARS coronavirus: a model response to emerging infectious diseases. J. Struct. Funct. Genomics 8, 85–97.

Berlanda Scorza, F., Doro, F., Rodríguez-Ortega, M.J., Stella, M., Liberatori, S., Taddei, A.R., et al., 2008. Proteomics characterization of outer membrane vesicles from the extraintestinal pathogenic *Escherichia coli* ΔtolR IHE3034 mutant. Mol. Cell. Proteomics 7, 473–485.

Berriman, M., Ghedin, E., Hertz-Fowler, C., Blandin, G., Renauld, H., Bartholomeu, D.C., et al., 2005. The genome of the African trypanosome *Trypanosoma brucei*. Science 309, 416–422.

Bittar, F., Richet, H., Dubus, J.-C., Reynaud-Gaubert, M., Stremler, N., Sarles, J., et al., 2008. Molecular detection of multiple emerging pathogens in sputa from cystic fibrosis patients. PLoS ONE 3, e2908.

Butler, J., MacCallum, I., Kleber, M., Shlyakhter, I.A., Belmonte, M.K., Lander, E.S., et al., 2008. ALLPATHS: de novo assembly of whole-genome shotgun microreads. Genome Res. 18, 810–820.

Butler, G., Rasmussen, M.D., Lin, M.F., Santos, M.A.S., Sakthikumar, S., Munro, C.A., et al., 2009. Evolution of pathogenicity and sexual reproduction in eight *Candida* genomes. Nature 459, 657–662.

Cardenas, E., Tiedje, J.M., 2008. New tools for discovering and characterizing microbial diversity. Curr. Opin. Biotechnol. 19, 544–549.

De Rycke, J., Oswald, E., 2001. Cytolethal distending toxin (CDT): a bacterial weapon to control host cell proliferation? FEMS Microbiol. Lett. 203, 141–148.

El-Sayed, N.M., Myler, P.J., Bartholomeu, D.C., Nilsson, D., Aggarwal, G., Tran, A.-N., et al., 2005a. The genome sequence of *Trypanosoma cruzi*, etiologic agent of Chagas disease. Science 309, 409–415.

El-Sayed, N.M., Myler, P.J., Blandin, G., Berriman, M., Crabtree, J., Aggarwal, G., et al., 2005b. Comparative genomics of trypanosomatid parasitic protozoa. Science 309, 404–409.

Fiers, W., Contreras, R., Duerinck, F., Haegeman, G., Iserentant, D., Merregaert, J., et al., 1976. Complete nucleotide sequence of bacteriophage MS2 RNA: primary and secondary structure of the replicase gene. Nature 260, 500–507.

Fiers, W., Contreras, R., Haegeman, G., Rogiers, R., Van de Voorde, A., Van Heuverswyn, H., et al., 1978. Complete nucleotide sequence of SV40 DNA. Nature 273, 113–120.

Finn, R.D., Mistry, J., Tate, J., Coggill, P., Heger, A., Pollington, J.E., et al., 2010. The Pfam protein families database. Nucl. Acids Res. 38, D211–222.

Fleischmann, R.D., Adams, M.D., White, O., Clayton, R.A., Kirkness, E.F., Kerlavage, A.R., et al., 1995. Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. Science 269, 496–512.

Fraser, C.M., Gocayne, J.D., White, O., Adams, M.D., Clayton, R.A., Fleischmann, R.D., et al., 1995. The minimal gene complement of *Mycoplasma genitalium*. Science 270, 397–403.

Gardy, J.L., Laird, M.R., Chen, F., Rey, S., Walsh, C.J., Ester, M., et al., 2005. PSORTb v.2.0: expanded prediction of bacterial protein subcellular localization and insights gained from comparative proteome analysis. Bioinformatics 21, 617–623.

Giuliani, M.M., Adu-Bobie, J., Comanducci, M., Aricò, B., Savino, S., Santini, L., et al., 2006. A universal vaccine for serogroup B meningococcus. Proc. Natl. Acad. Sci. 103, 10834–10839.

Glaser, P., Frangeul, L., Buchrieser, C., Rusniok, C., Amend, A., Baquero, F., et al., 2001. Comparative genomics of *Listeria* species. Science 294, 849–852.

Goffeau, A., Barrell, B.G., Bussey, H., Davis, R.W., Dujon, B., Feldmann, H., et al., 1996. Life with 6000 genes. Science 274, 546–567.

Griffiths-Jones, S., Moxon, S., Marshall, M., Khanna, A., Eddy, S.R., Bateman, A., 2005. Rfam: annotating non-coding RNAs in complete genomes. Nucl. Acids Res. 33, D121–124.

Handelsman, J., Rondon, M.R., Brady, S.F., Clardy, J., Goodman, R.M., 1998. Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products. Chem. Biol. 5, R245–R249.

Hemmerich, C., Buechlein, A., Podicheti, R., Revanna, K.V., Dong, Q., 2010. An Ergatis-based prokaryotic genome annotation web server. Bioinformatics 26, 1122–1124.

Hiller, N.L., Janto, B., Hogg, J.S., Boissy, R., Yu, S., Powell, E., et al., 2007. Comparative genomic analyses of seventeen *Streptococcus pneumoniae* strains: insights into the pneumococcal supragenome. J. Bacteriol. 189, 8186–8195.

Hogeweg, P., 1978. Simulating the growth of cellular forms. Simulation 31, 90–96.

Hogeweg, P., Hesper, B., 1978. Interactive instruction on population interactions. Comput. Biol. Med. 8, 319–327.

Huber, J.A., Mark Welch, D.B., Morrison, H.G., Huse, S.M., Neal, P.R., Butterfield, D.A., et al., 2007. Microbial population structures in the deep marine biosphere. Science 318, 97–100.

Hugenholtz, P., Tyson, G.W., 2008. Microbiology: metagenomics. Nature 455, 481–483.

Huse, S.M., Welch, D.M., Morrison, H.G., Sogin, M.L., 2010. Ironing out the wrinkles in the rare biosphere through improved OTU clustering. Environ. Microbiol. 12, 1889–1898.

Ivens, A.C., Peacock, C.S., Worthey, E.A., Murphy, L., Aggarwal, G., Berriman, M., et al., 2005. The genome of the kinetoplastid parasite, *Leishmania major*. Science 309, 436–442.

Keim, P., Smith, K.L., 2002. *Bacillus anthracis* evolution and epidemiology. Curr. Top. Microbiol. Immunol. 271, 21–32.

Kent, W.J., 2002. BLAT—the BLAST-like alignment tool. Genome Res. 12, 656–664.

Lauer, P., Rinaudo, C.D., Soriani, M., Margarit, I., Maione, D., Rosini, R., et al., 2005. Genome analysis reveals pili in group B *Streptococcus*. Science 309, 105.

Ledford, H., 2008. The death of microarrays? Nature 455, 847.

Lefébure, T., Stanhope, M., 2007. Evolution of the core and pan-genome of *Streptococcus*: positive selection, recombination, and genome composition. Genome Biol. 8, R71.

Lefébure, T., Stanhope, M.J., 2009. Pervasive, genome-wide positive selection leading to functional divergence in the bacterial genus *Campylobacter*. Genome Res. 19, 1224–1232.

Maione, D., Margarit, I., Rinaudo, C.D., Masignani, V., Mora, M., Scarselli, M., et al., 2005. Identification of a universal group B *Streptococcus* vaccine by multiple genome screen. Science 309, 148–150.

Medini, D., Donati, C., Tettelin, H., Masignani, V., Rappuoli, R., 2005. The microbial pan-genome. Curr. Opin. Genet. Dev. 15, 589–594.

Miller, J.R., Koren, S., Sutton, G., 2010. Assembly algorithms for next-generation sequencing data. Genomics. 95, 315–327.

Montigiani, S., Falugi, F., Scarselli, M., Finco, O., Petracca, R., Galli, G., et al., 2002. Genomic approach for analysis of surface proteins in *Chlamydia pneumoniae*. Infect. Immun. 70, 368–379.

Nakamura, S., Yang, C.-S., Sakon, N., Ueda, M., Tougan, T., Yamashita, A., et al., 2009. Direct metagenomic detection of viral pathogens in nasal and fecal specimens using an unbiased high-throughput sequencing approach. PLoS ONE 4, e4219.

Passalacqua, K.D., Varadarajan, A., Ondov, B.D., Okou, D.T., Zwick, M.E., Bergman, N.H., 2009. Structure and complexity of a bacterial transcriptome. J. Bacteriol. 191, 3203–3211.

Paszkiewicz, K.H., Studholme, D.J., 2010. De novo assembly of short sequence reads. Brief. Bioinformatics. 11, 457–472.

Payne, D.J., Gwynn, M.N., Holmes, D.J., Pompliano, D.L., 2007. Drugs for bad bugs: confronting the challenges of antibacterial discovery. Nat. Rev. Drug Discov. 6, 29–40.

Pizza, M., Scarlato, V., Masignani, V., Giuliani, M.M., Arico, B., Comanducci, M., et al., 2000. Identification of vaccine candidates against serogroup B Meningococcus by whole-genome sequencing. Science 287, 1816–1820.

Pop, M., Phillippy, A., Delcher, A.L., Salzberg, S.L., 2004. Comparative genome assembly. Brief. Bioinform. 5, 237–248.

Rasko, D.A., Rosovitz, M.J., Myers, G.S.A., Mongodin, E.F., Fricke, W.F., Gajer, P., et al., 2008. The pangenome structure of *Escherichia coli*: comparative genomic analysis of *E. coli* commensal and pathogenic isolates. J. Bacteriol. 190, 6881–6893.

Rondon, M.R., August, P.R., Bettermann, A.D., Brady, S.F., Grossman, T.H., Liles, M.R., et al., 2000. Cloning the soil metagenome: a strategy for accessing the genetic and functional diversity of uncultured microorganisms. Appl. Environ. Microbiol. 66, 2541–2547.

Sanger, F., Air, G.M., Barrell, B.G., Brown, N.L., Coulson, A.R., Fiddes, J.C., et al., 1977. Nucleotide sequence of bacteriophage phiX174 DNA. Nature 265, 687–695.

Schoen, C., Blom, J., Claus, H., Schramm-Glück, A., Brandt, P., Müller, T., et al., 2008. Whole-genome comparison of disease and carriage strains provides insights into virulence evolution in *Neisseria meningitidis*. Proc. Natl. Acad. Sci. **105**, 3473–3478.

Seib, K.L., Dougan, G., Rappuoli, R., 2009. The key role of genomics in modern vaccine and drug design for emerging infectious diseases. PLoS Genet. 5, e1000612.

Servant, F., Bru, C., Carrère, S., Courcelle, E., Gouzy, J., Peyruc, D., et al., 2002. ProDom: automated clustering of homologous domains. Brief. Bioinform. 3, 246–251.

Simpson, J.T., Wong, K., Jackman, S.D., Schein, J.E., Jones, S.J.M., Birol, Ä. n., 2009. ABySS: a parallel assembler for short read sequence data. Genome Res. 19, 1117–1123.

Sommer, D., Delcher, A., Salzberg, S., Pop, M., 2007. Minimus: a fast, lightweight genome assembler. BMC Bioinformatics 8, 64.

Tettelin, H., Masignani, V., Cieslewicz, M.J., Donati, C., Medini, D., Ward, N.L., et al., 2005. Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: implications for the microbial "pan-genome". Proc. Natl. Acad. Sci. U.S.A. 102, 13950–13955.

The C. elegans Sequencing Consortium, 1998. Genome sequence of the nematode *C. elegans*: a platform for investigating biology. Science 282, 2012–2018.

Thorpe, C., Edwards, L., Snelgrove, R., Finco, O., Rae, A., Grandi, G., et al., 2007. Discovery of a vaccine antigen that protects mice from *Chlamydia pneumoniae* infection. Vaccine 25, 2252–2260.

Toledo-Arana, A., Dussurget, O., Nikitas, G., Sesto, N., Guet-Revillet, H., Balestrino, D., et al., 2009. The *Listeria* transcriptional landscape from saprophytism to virulence. Nature 459, 950–956.

Tomb, J.-F., White, O., Kerlavage, A.R., Clayton, R.A., Sutton, G.G., Fleischmann, R.D., et al., 1997. The complete genome sequence of the gastric pathogen *Helicobacter pylori*. Nature 388, 539–547.

Trapnell, C., Pachter, L., Salzberg, S.L., 2009. TopHat: discovering splice junctions with RNA-Seq. Bioinformatics 25, 1105–1111.

van Vliet, A.H.M., 2010. Next generation sequencing of microbial transcriptomes: challenges and opportunities. FEMS Microbiol. Lett. 302, 1–7.

Van Voorhis, W.C., Hol, W.G.J., Myler, P.J., Stewart, L.J., 2009. The role of medical structural genomics in discovering new drugs for infectious diseases. PLoS Comput. Biol. 5, e1000530.

Yoder-Himes, D.R., Chain, P.S.G., Zhu, Y., Wurtzel, O., Rubin, E.M., Tiedje, J.M., et al., 2009. Mapping the *Burkholderia cenocepacia* niche response via high-throughput sequencing. Proc. Natl. Acad. Sci. 106, 3976–3981.

Zerbino, D.R., Birney, E., 2008. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res. 18, 821–829.