Abstracts

Metagenomic assembly with viral genomes
Authors: Martin Ayling, Livia Patrono, Kirsten McLay, Pablo Murcia, Richard M. Leggett

In recent years, metagenomic studies - that is sequencing the genomes of a heterogeneous population of species from an environment - have been performed with increasing frequency. The introduction of shotgun sequencing in this field offers the potential of assembling viral genomes direct from the environment, something impossible with earlier 16S amplicon sequencing approaches. As a result, assembly tools have appeared following algorithms designed to overcome the particular obstacles of metagenomics which are not addressed by existing genomic assemblers. Despite this, the new generation of assemblers have continued to focus on reconstructing the genomes of bacterial rather than viral species, and a standardised comparison between the performance of these assemblers has been lacking. Here we present an assessment of several available metagenomic assemblers (IDBA-UD, MetaVelvet, MEGAHIT, Omega, Ray Meta, VICUNA) in reconstructing viral genomes from heterogeneous environmental samples. The samples consisted of cattle tissue samples artificially infected with known viruses, as well as human samples that had tested clinically positive for viruses. In each case, total nucleic acids were extracted and sequencing was performed on the Illumina Hi-Seq platform.

Alignment by numbers: Sequence assembly using compressed numerical representations
Avraam Tapinos1*, Bede Constantinides1, Douglas B Kell2,3, David L Robertson1*
1 Computational and Evolutionary Biology, Faculty of Life Sciences, The University of Manchester, Manchester, M13 9PT, UK.
2 School of Chemistry and 3Manchester Institute of Biotechnology, The University of Manchester, Manchester, M1 7DN, UK
*Correspondence to: david.robertson@manchester.ac.uk,avraam.tapinos@manchester.ac.uk

DNA sequencing instruments are enabling genomic analyses of unprecedented scope and scale, widening the gap between our abilities to generate and interpret sequence data. Established methods for computational DNA sequence analysis generally consider the nucleotide-level resolution of sequences. While these approaches are sufficiently accurate, increasingly ambitious and data-intensive analyses are rendering them impractical for demanding applications such as genome and metagenome assembly.
Increases in the length of the DNA short reads generated by the current sequencing technologies, combined with the overall size of the datasets introduce a number of analytical challenges. Due to size of the datasets (usually several gigabytes), not all reads can be loaded in RAM memory simultaneously, thus data have to be accessed from hard disk memory during the analysis, slowing down the overall execution time of the analysis process. In addition, due to the high dimensionality of DNA reads, data similarities can be overlooked, hindering the statistical significance of the analysis; a phenomenon related to the curse of dimensionality[1]. Future sequencing technologies will provide considerably longer reads (several thousand bases long), intensifying these problems.

Comparable challenges involving high dimensional sequential data have been investigated thoroughly in fields such as signal processing and time series analysis,and a number of effective dimensionality reduction methods have been proposed.Methods include the discrete Fourier transform (DFT)[2], the discrete wavelet transform (DWT)[3], and piecewise aggregate approximation (PAA)[4]. These transformation methods are commonly used for compressing sequence data to a lower dimensional space and reduce the dataset's size prior to the analysis process. Due to the smaller size of the transformed dataset more data can be loaded in RAM memory simultaneously, hence speeding the analysis process. Furthermore, data approximations at a lower dimensional space lessen the effects introduced by the high dimensionality of the data.

Representing DNA sequences as numerical sequences, allows these application of the transformation and approximation methods (fig. 1). Proposed DNA sequences numerical

representation techniques include the Voss method, the integer number representation, the real number representation (fig. 1), and DNA walk [5].

To explore the applicability of data transformation and approximation techniques for sequence assembly, we implemented a short read aligner in C++ and evaluated its performance against simulated high diversity viral sequences alongside four existing aligners[6].

Despite using heavily approximated sequence representations, our implementation yielded alignments of similar overall accuracy to existing aligners, outperforming all other tools tested at high levels of sequence variation. Furthermore, the performance of our prototype reference based implementation was comparable to the rest of the tools. Our prototype approach was also applied to the de novo assembly of a simulated diverse viral population. Our approach demonstrates that full sequence resolution is not a prerequisite of accurate sequence alignment and that analytical performance may be retained or even enhanced through appropriate dimensionality reduction of sequences.

Figure 1. A numerically represented DNA sequence (x-axis) transformed at various levels of resolution using the discrete Fourier transform, DFT(A), the Haar discrete wavelet transform, DWT(B), and piecewise aggregate approximation, PAA(C). A 30 nucleotide sequence (x-axis) is represented as a numerical sequence (black lines) using the real number representation method (y-axis: T = 1.5, C = 0.5, G = - 0.5 and A = -1.5) and three approximation types: DFT with 5 (red), 3 (blue) and 1 (green) Fourier frequencies(A). DWT with 8 wavelets (red), 4 wavelets (blue) and 2 wavelets (green) (B).And, PAA with 8 (red), 5 (blue) and 3 (green) coefficients (C).

References

1. Verleysen M, François D: The curse of dimensionality in data mining and time series prediction. In: Computational Intelligence and Bioinspired Systems. Springer; 2005: 758-770.
2. Agrawal R, Faloutsos C, Swami A: Efficient similarity search in sequence databases: Springer; 1993.
3. Woodward AM, Rowland JJ, Kell DB: Fast automatic registration of images using the phase of a complex wavelet transform: application to proteome gels. Analyst 2004, 129(6):542-552.
4. Keogh E, Chakrabarti K, Pazzani M, Mehrotra S: Locally adaptive dimensionality reduction for indexing large time series databases. ACM SIGMOD Record 2001, 30(2):151-162.
5. Kwan HK, Arniker SB: Numerical representation of DNA sequences. In: Electro/Information Technology, 2009 eit'09 IEEE International Conference on: 2009. IEEE: 307-310.
6. Tapinos A, Constantinides B, Kell DB, Robertson DL: Alignment by numbers: sequence assembly using compressed numerical representations. bioRxiv 2015.

Microbial genome assembly using synthetic error-free reads

Mohammed-Amin Madoui1, Stefan Engelen1, Corinne Cruaud1, Caroline Belser1, Guillaume Gautreau1, Benjamin Istace1, Laurie Bertrand1, Adriana Alberti1, Arnaud Lemainque1, Patrick Wincker123 and Jean-Marc Aury1
1Commissariat à l'Energie Atomique (CEA), Institut de Génomique (IG), Genoscope, Evry, BP5706, 91057, France
2Université d'Evry Val d'Essonne, UMR 8030, Evry, CP5706, 91057, France
3Centre National de Recherche Scientifique (CNRS), UMR 8030, Evry, CP5706, 91057, France

The technology of long-read sequencing offers different alternatives to solve genome assembly problems which cannot be resolved adequately by short-read sequencing. Here, we present a hybrid approach developed to take advantage of long (MinION or PacBio) and short (Illumina) reads. Our method is able to generate synthetic long reads up to 90kb with no error and that span large repetitive regions. The method was applied to several bacterial and small eukaryotes read sets to generate the error-free synthetic reads that were used to produce highly contiguous and accurate genome sequences. For bacterial genomes, our method outperformed the existing methods of reads correction and the whole strategy (including the sequencing) enabled the release of a near perfect genomes in less than three days.

Running and Reading in Real Time: Looking at Squiggles on the Oxford Nanopore MinIon

Martin Blythe, Sunir Mall, Mike Stour, Fei Sang and Matt Loose

The utility of the Oxford Nanopore minION sequencer is becoming clear. However, to fully exploit its potential requires a shift in thought in both experimental design and analysis. The conventional model of sequence, map and analysis leading to final result has been replaced by instant access to sequence data before the completion of a sequencing run. In the extreme case of the ONT minION it is possible to analyse squiggle data before a read has even completed. We have developed a platform of tools, minoTour, to analyse minION data in real time and extended it to exploit both 'Run Until' and 'Read Until'. Run until allows the sequencer to switch off after achieving a specific goal, such as depth of coverage. Read until allows individual reads to be rejected from the pore and free that pore to sequence an alternative preferred read. We have applied run and read until in a number of different scenarios including selective small genome sequencing and barcode normalisation. Limitations and challenges to implementing read until will be discussed along with the challenge of Fast Mode whereby speed of processing will be of vital importance. Finally we demonstrate how methodologies for analysing squiggles may help to reduce the reliance on base calling in the field.

The Cloud Infrastructure for Microbial Bioinformatics (CLIMB)

Tom Connor[1], Nick Loman[2], Simon Thompson[3], Matt Ismail[4], Sam Sheppard[5], Mark Pallen[6]
[1] Cardiff School of Biosciences, Cardiff University, UK
[2] School of Biosciences, University of Birmingham, UK
[3] Research Computing, University of Birmingham, UK
[4] Centre for Scientific Computing, University of Warwick, UK
[5] Medical Microbiology and Infectious Diseases, College of Medicine, Swansea University
[6] Warwick Medical School, University of Warwick, UK

Genome sequencing has made it possible to examine fundamental biological questions over a huge range of scales; from bacteria to man. Since the first bacterial genome was published 20 years ago, there has been an explosion in the production of sequence data, fuelled by next-generation sequencing, placing biology at the forefront of data-driven science. As a consequence, there is now huge demand for the physical infrastructure to produce, analyse and share software and datasets. This need is compounded by the continuing lack of trained bioinformaticians to analyse the data. In 2014 the Medical Research Council made a ~£50m investment in "big data" to support the development of new research infrastructures. The £8.5m CLoud Infrastructure for Microbial Bioinformatics (CLIMB) was the only award to a microbial consortium and is one the largest investments in microbial genomics bioinformatics ever made. CLIMB will provide bioinformatics infrastructure as a service to the academic UK medical microbial community. CLIMB is spread across four sites (Birmingham, Cardiff, Swansea and Warwick) and we provide a single sign-on, distributed computing and storage infrastructure. The total investment in hardware is £3.6m, which provides 7680 virtual CPUs, 500 terabytes of local high performance storage per site and 7000 terabytes (7PB) of replicated object storage. This facility is sufficient to provide over a thousand simultaneously running virtual machines to users. The service is implemented using the open-source OpenStack framework. Users will be able to login using their eduroam account and start up virtual machines instantly with access to large microbial genome datasets. We will present case studies of using CLIMB to integrate large microbial datasets from diverse sources and provide live demos of the service. CLIMB is currently in a beta testing state and we are looking for early access participants to help us develop the service. In parallel with the development of the system itself we are running an active programme of bioinformatics training workshops hosted in newly commissioned space at Warwick, Swansea and Birmingham

NanoOK: Flexible, multi-reference software for pre and post-alignment analysis of Nanopore sequencing data, quality and error profiles

Richard M. Leggett, Darren Heavens, Mario Caccamo, Matthew D. Clark, Robert P. Davey

The recent launch of the Oxford Nanopore Technologies MinION® Access Program (MAP) resulted in the rapid development of a number of open source tools [1, 2, 3] aimed at extracting reads and yield information from the HDF5 format files produced by the platform. poretools [1] and poRe [2] are designed to extract FASTA/Q files and to provide plots of yield and utilisation stats. minoTour [3] aligns reads and provides coverage and variation information too. However, none of the tools provides error profiling based on detailed alignment analysis of Nanopore reads, a methodology that is critical in order to understand the applicability of the platform to a new problem area and is often performed ad hoc. With the platform evolving as rapidly as it is, this kind of analysis is also crucial to understanding the performance of new flowcells, chemistries, base callers and aligners. NanoOK has been written to address this gap.NanoOK processes the raw HDF5 files output by the MinION® basecaller, extracts FASTA/Q format files, aligns to references, calculates a wide range of QC and error metrics, and finally consolidates information into a PDF report. Crucially, it is designed to support multiple concurrent references, enabling analysis of metagenomic samples and pooled libraries. NanoOK produces in-depth data on a variety of key metrics including: number of reads aligning; quality of alignment; coverage and perfect kmer plots for template, complement and 2D reads; analysis of longest perfect sequence; statistics on types of error (substitutions, indels); analysis of over- and under-represented kmers; location of error; error motifs, such as preceding n-mers before observed errors. Here, we present a description of the tool and associated sequence data as well as a some analysis of real world samples. Full source code is available from the TGAC GitHub site.

1. Loman and Quinlan (2014). doi:10.1093/bioinformatics/btu555
2. Watson et al. (2014). doi:10.1093/bioinformatics/btu590
3. Loose (2014). doi:10.6084/m9.figshare.1159099

New Online Resource Provides Millions of Induced Mutations for Gene Function Analysis in Bread Wheat

Paul Bailey[1], Leah Clissold[1], James Simmonds[2], Hans Vasquez-Gross[4], Martin Trick[2], Chris Wilson[2], Andy Phillips[3], Ksenia Krasileva[1], Jorge Dubcovsky[4], Cristobal Uauy[2], Sarah Ayling[1]

[1]The Genome Analysis Centre (TGAC), Norwich Research Park, Norwich, UK, NR4 7UH
[2]John Innes Centre, Norwich Research Park, Norwich, UK, NR4 7UH
[3]Rothamsted Research, Harpenden, Hertfordshire, AL5 2JQ
[4]Dept. Plant Sciences, University of California, Davis, CA 95616, USA

Recently chromosome survey sequences (CSS) of the 17 Gb hexaploid genome of bread wheat have been assembled by the IWGSC (International Wheat Genome Sequence Consortium). In this work we have used this assembly to develop an exome capture resource for identifying gene mutations in a TILLING (Targeted Induced Local Lesions In Genomes) population of bread wheat for use in functional genomics. Previously we assembled a set of 82,511 wheat protein coding sequences (Krasileva et al, 2013). These sequences were aligned to the CSS contigs to identify exon-intron gene models and the exon regions plus flanking intronic regions were used as the target for the preparation of exome capture bait sequences (NimbleGen).

We will present the results from the exome capture procedure for 1000 TILLING lines, showing the effectiveness of read mapping and mutation calling to the CSS reference. Extra steps were taken to improve the quality of mutations called and to maximise the return of mutations so that each gene was represented in the mutation set. These steps included supplementing the genome reference used for read mapping with missing homoeologues (on average only 2 out of 3 homoeologue copies exist in IWGSC version 1 reference) and performing reference masking so that mutations in reads that map to multiple locations could still be detected.

To date more than four million EMS-induced mutations for bread wheat have been identified and are present primarily in gene exons. These mutations will be available soon for browsing via a website where the corresponding lines can also be ordered. The identification of mutations that abolish or modify protein activity will provide an extremely useful functional genomics resource for both basic and applied wheat research.

Characterisation of snake venom gland transcriptomes using the Oxford Nanopore MinION
portable single-molecule nanopore sequencer
Adam D Hargreaves and John F Mulley

Snake venom toxin proteins are encoded by a number of different gene families which have expanded via repeated gene duplication, giving rise to a large number of highly similar paralogs. It is these expanded gene families in particular that prove difficult to characterise using short-read “next-generation” sequencing and de novo assembly. We have therefore tested the feasibility of single-molecule cDNA sequencing using the Oxford Nanopore MinION portable DNA sensing device in order to overcome these issues and to accurately characterise the venom gland transcriptome of the painted saw-scaled viper, Echis coloratus. We find the raw error rate of R7.3 flowcells using the Nanopore Sequencing Kit SQK-MAP005 to be around 11% when compared to an Illumina HiSeq dataset from the same tissue samples assembled using Trinity. However, we are able to improve this to 0-2% using proovread and around 2-5% using nanocorrect. Our proovread-corrected Nanopore data provides full coding sequences and 5' and 3' UTRs for 29 of the 33 candidate venom toxins detected, far superior to Illumina alone (13/40 complete) or even Sanger-based ESTs (15/29 complete). We suggest that, should the current rapid pace of improvement to the MinION device, flowcells, sample preparation and chemistry continue, it will become the default method of cDNA sequencing in a variety of species, but will be particularly useful for the characterisation of snake venom gland transcriptomes.

Using genomics to define, explore and expand our understanding of the species *Shigella flexneri*
Thomas R Connor[1,2,], Clare R Barker[1], Kate S Baker[2], François-Xavier Weill[3], Claire Jenkins[4], Nicholas R Thomson[2,5]
Affiliations:
[1] Cardiff School of Biosciences, Sir Martin Evans Building, Museum Avenue, Cardiff CF10 3AX, United Kingdom
[2] Pathogen Genomics, Wellcome Trust Sanger Centre, Cambridge CB10 1SA, United Kingdom
[3] Institut Pasteur, Unité des Bactéries Pathogènes Entériques, Paris, France
[4] Gastrointestinal Bacteria Reference Unit, Public Health England, 61 Colindale Ave, NW9 5HT
[5] The London School of Hygiene and Tropical Medicine, London, United Kingdom

Shigella flexneri is thought to be the most frequent cause of the estimated 165 million cases of bacterial dysentery that occur every year in low-income countries around the world. Despite this, *S. flexneri* remains largely unexplored from a genomic standpoint and is still described using a vocabulary based on serotyping reactions developed during the middle of the 20th century. Using a species-wide dataset, we use whole genome sequencing to subdivide and characterise this species on a genomic basis, redefining what is understood about the population structure of the species, while also revealing a strong phylogeographic signal. In contrast to its relative, *S. sonnei*, *S. flexneri* appears to have colonised regions over the long term, suggesting that it has a distinct lifestyle that involves a measure of environmental colonisation. This work demonstrates the importance of genomics in re-evaluating what is understood about the population structure and the diversity of key pathogens, as well as the potential new questions that the use of genomics can pose, questions that could have a major impact on public health priorities around key pathogens.

The distribution and composition of *Campylobacter jejuni* plasmid pan-genome

Andrea Gori, James Harrison, Katie Luckes, Olivia Champion, Rick Titball, David J. Studholme
School of Biosciences, College of Life and Environmental Sciences, University of Exeter, UK

With about 500 million cases per year *Campylobacter jejuni* is the worldwide leading cause of food poisoning. This bacterium is commonly found in chicken and other farm animals, and it is able to survive in contaminated stocks of water and milk. Over the last few decades 15 genomes of *C. jejuni* have been fully sequenced, with many others being sequenced to draft status. This has yielded complete sequences of a set of 18 *C. jejuni* plasmids. We sequenced a further 19. Two of these plasmids have been described in a clinically isolated strain 81-176: the pVir plasmid, believed to be involved in pathogenicity and the pTet plasmid, whose function is unclear except for the presence of a tetracycline resistance and a conjugative machinery.

Collectively, all the available plasmids encode 127 unique proteins. We surveyed all available genome sequence datasets for *C. jejuni* to determine the presence or absence of each of the corresponding plasmid-associated gene across 3398 *C. jejuni* isolates.

Approximately 1% of the sequenced isolates show the presence pVir-associated genes, while pTet-associated genes are present and in more than 20% of all the sequenced *C. jejuni* isolates, suggesting this plasmid might have a previously unappreciated importance in the lifestyle of the organism. Both the pTet-like and pVir-like sequences sequences were present in clinical and environmental isolates of *C. jejuni*, including the ones isolated from animals.

The results of this analysis refute the hypothesis that retention of pTet is maintained by selection for tetracycline resistance gene, as the presence of the *tetR* gene is not strictly coupled with the presence of other plasmidic genes. Several possible plasmid structures are also highlighted for the pTet-like plasmids, with “core” and “accessory” genes across a mosaic structure of the plasmid.

Finally, it is interesting to notice that three proteins present in the plasmid pan-genome are found in about a half of the *C. jejuni* strains analysed, independently from the presence of any plasmid: the analysis of the sequence of two of these proteins shows a possible involvement in the production of biofilm and processing bacteriocins, both important traits in the lifestyle of the organism.

Should everyone have their genomes sequenced? Insights from in-depth interviews with personal genomics research participants

Saskia C. Sanderson1,2, Michael D. Linderman1, Sabrina A. Suckiel1, Randi Zinberg1, Andrew Kasarskis1, Noura S. Abul-Husn1, Melissa Wasserstein1, George A. Diaz1, Eric E Schadt1
1 Department of Genetics and Genomic Sciences, Icahn School of Medicine, New York, NY
2 Health Behaviour Research Centre, University College London, London

Background: Sequencing everyone’s genomes has the potential to dramatically change the medical landscape by making information about medication response available to physicians at the point of prescribing, identifying individuals with previously unsuspected disease-causing rare variants, and providing risk information about relatively common diseases that might empower individuals to take more control over their health. However, challenges include interpreting the clinical significance of rare variants, how to educate patients and ensure they make informed decisions, how to communicate the potentially vast amounts of personal information arising from genome sequencing to patients in ways that are not
overwhelming for them, and avoiding psychological harm such as anxiety, confusion and false reassurance.

Methods: The HealthSeq project is a study in which personal genome sequencing (PGS) was offered to ostensibly healthy individuals from a range of backgrounds. Participants (n=35) received personal health-related results from PGS including pharmacogenomics, rare disease

and carrier variants, and common disease risks. They also received PGS results relating to ancestry, physical traits, number of unique variants identified, sequencing coverage, and their raw data. In-person and telephone interviews were conducted at baseline (T1), immediately after (T2), one week after (T3) and six months after (T4) results were returned. Qualitative in-depth interviews were audio-recorded, transcribed and analysed using thematic analysis in NVivo 10.

Results: In the follow-up interviews, the majority of participants reported positive emotional reactions to their results, including feeling happy and relieved. Some participants reported having experienced more mixed feelings, including one whose results included a rare 'pathogenic' variant associated with Brugada syndrome, and increased risk of Alzheimer's disease (e4/e4). Many had discussed their results with family members, and several had discussed their results with their physicians. Few had made lifestyle changes based on their results. While many did not feel the results were of immediate clinical value to them, several
said they planned to keep the raw data and/or specific results, in particular the pharmacogenetic and in some cases carrier results, so that these could be used in their future healthcare if and when necessary. Participants were divided about wanting their results in their medical records, with some viewing this as beneficial, and others havingconcerns about privacy and insurance.

Discussion: Patients are key stakeholders in debates about the future directions and value of offering personal genome sequencing (PGS) more widely. The findings from this in-depthstudy with ostensibly healthy individuals suggest the possibility that currently neither the benefits nor the harms of PGS are significant for most individuals, but that there may be important exceptions to this that warrant further investigation. The impact of PGS when implemented on a larger scale remains to be seen.

Challenges of fine and broad scale diversity in the genome of the malaria vector *Anopheles gambiae*

Nicholas J Harding, Alistair Miles, Giordano Botta, Dominic Kwiatkowski on behalf of the Ag1000G consortium.

The closely related mosquito species *A nopheles gambiae* a nd *A nopheles coluzzi* a re the predominant vectors of malaria in Africa, where the burden of disease remains high. Vector control measures are being undermined by the emergence and spread of insecticide resistance in mosquito populations.

The *A nopheles gambiae* 1 000 genomes project (Ag1000G) has been established to provide a foundation for the next generation of research into malaria vector control by surveying genetic variation in mosquito populations in sub-Saharan Africa. The Ag1000G consortium comprises members from 13 institutions, and has to date completed deep sequencing of over 1700 whole genomes of mosquitoes sampled from 13 different countries. Phase 1 of the project has recently been publicly released, an initial cohort of 765 samples from 8 countries, and has provided the first whole-genome view of the astonishing natural diversity within this species.

This enormous diversity presents many challenges in the identification of good quality genotypes. In raw aligned sequence reads we observe a putative SNP every 2 bases. This makes alignment and genotype calling extremely challenging. We are probably approaching the limit of the reference genome paradigm for short reads. There are likely to be common structural variants divergent from the reference genome that lead to misalignments and subsequent genotyping errors. We developed and applied rigorous filters and an accessibility mask, helping to attain genotype concordance rates of 99.8% upon validation, while identifying 39m SNPs in a ~260Mb genome.

A genome that only has 60% accessibility is also challenging when it comes to estimating important parameters for vector control monitoring such as autozygosity. We developed an HMM based method to identify runs of homozygosity in our samples. This allows us to look at the extent of inbreeding in mosquito populations, and monitor the genetic signals of

population crashes potentially caused by vector control. The degree of inbreeding in Ag1000G populations varies hugely, with strong evidence for a recent severe population bottleneck in Kenya.

Haplotype estimation is an important endpoint to increase the power of our searches for loci under selection and identification of resistance origins. Population diversity makes this much harder to resolve. We use recent haplotype estimation tools to phase our data, the read aware phasing functionality in SHAPEIT2 is especially informative. Pedigrees of lab crossed mosquitoes are used to evaluate the performance of haplotype estimation on autosomes. For the X chromosome we evaluate using synthetic females generated from male samples. Using the resulting high quality haplotypes, we report selective sweeps at various loci and

present evidence for adaptive introgression between species and populations, most notably at the *k dr* l ocus of the *p ara* v oltage gated sodium channel (VGSC) gene.