- *Hi all! Jenn Gardy here. This is the collaborative note-taking document for the ASMNGS17 conference, more formally known as the 2nd ASM Conference on Rapid Applied Microbial Next-Generation Sequencing and Bioinformatics Pipelines. The conference hashtag is #ASMNGS, and rather than trying to live-tweet every talk, it's far easier to have people use this document to take more elaborate notes, then just regularly tweet the link to this (https://public.etherpad-mozilla.org/p/asmngs17 or the short URL: http://bit.ly/asmngs17) out on Twitter instead. We have done this at ABPHM and at Virus Genomes & Evolution and it works brilliantly.*

Please jump in and join the note-taking fun here. A couple of tips to get you started:

1. Enter your name over at the side so people can see who you are

2. Keep the text field to notes on the talks and use the Chat box at the side for general chat

3. The whole process unfolds pretty organically, with one person (usually somebody whose ework is in line with the talk's subject) emerging as the uber-note-taker for that talk and others jumping in by expanding on points, adding links to papers, etc..

Also, The Journal of Clinical Microbiology is having the keynote speakers write up some mini-reviews, and has asked me to do an overall conference recap as a standalone mini-review. These notes will help form the basis of that review, so if you include your name on the side and contribute a lot of notes, you will be acknowledged in the article. Heck, maybe they will even let me appoint some co-authors.

Have fun note-taking!

Sunday, October 8: Opening Session

Opening Comments
Dag Harmsen
Conference Chair, University of Muenster, Muenster, Germany

- Everything will be amazing during the next couple of days!
- We had a super-awesome program committee this year and really tried to make this super-fun and inclusive and excellent!

Keynote 1: WGS for Public Health Microbiology - Trials, Tribulations and Triumphs
Catherine Arnold
Head of Genomic Services and Development Unit, National Infection Service, Public Health England, London, UK

- We've come so far so fast - DNA in 1953, 1973 first 24bp sequence, 1977 M&G and Sanger sequencing, 1982 Genbank starts, 1983 PCR, ABI Prism 373 automated sequencer in 1987(Cath loaded the radioactive gels and mouth pipetted), 1998 C. elegans, 17 years and $3B for human genome 2000/3 - now $1000 in a few days
- Nextgen came along i 2005/6, solexa and 454, very interesting to see which machines will be along in the future.
- 2011 single-molecule PacBioRS big giant monster sequencer that eats high-quality DNA in large amounts
- Now, nanopore is along, easy, cheaply and quickly
- Public health NGS uses - microbiological source attribution, presumptive transmission cluster identification, monitor MDR epidemic pathogens, find MDR element outbreaks/epidemics, monitor vaccine strain coverage/predict vaccine preventability for disease, find new hyper-pathogenic strains, find diagnostic escape variants
- Phenotypic methods are what most labs use these days: culture (DSTs, growth characteristics, colony morphology), serotyping, phage typing), molecular epi methods (PHFGE, RFLP< spoligo, sequence-based typing, VNTR/MLVA, ribotyping)
- Many phenotypic methods are low resolution wrt typing
  - PFGE
  - RFLP
  - CRISPRS
  - Sanger
  - VNTR, MLVA
  - Ribotyping¨
- 'Future is NGS
- We need timely analysis, early outbreak detetction, and national/international surveillance (patient: dx, local population: outbreaks, national/intl population: surveillance). Within PHE: Nationa Health Service NHS oversees patient - level, National Infection Service does local/national/international population stuff
- The more analysed, the earlier outbreaks can be detected
- Genomics: one sequencing run can answer many questions, from dx to epidemiology
- Must be easy and robust, discriminatory, reproducible, portable, rapid, appropriately interpreted and assess to guide PH response
- PHE/NIS goal: one common workflow to define lineage, predict resistance, find outbreaks,speciate, assess virulence, spot biomarkers of interest - WGS can do this
- Central or local?
  - Central: easy to manage, expertise concentrated, cost advntages, better national coordination.
  - Local: faster TATs, responsive to local needs, effective interaction with frontline users, multiple sites facilitate business continuity
- PHE: looking at salmonella spp, s.aureus, s.pneumoniae, influenza, hcv, emergency response
- Using illumina for their service, the more they do, the cheaper it gets. Reagents cost per sample, double lane hiseq per sample: 30 pounds
- Practically exponential drop in costs as you batch samples - ideal is to fill a HiSeq. Lowest cost seems to be around £30/genome, achieved on NextSeq and larger instruments at high batch sizes
- 4 day turnaround time
- Customer plate -> Biomek liquid handler pre-PCR step -> Glomax for DNA concentration -> G3 liquid handler -> Biomek post-PCR step -> LabChip GX for fragment size -> Viia7 for library concentration ->cBot ->flowcell ->HiSeq rapid throughput mode

- LIMS updated at every step - key for UKK accreditation process (need to know everything about where sample is, what has happened to it)
- Big goal: minimize human interaction with sample/dna
- NOTE: are using rapid run
- fully automated analysis, if it passes qc it goes through. Note, lims is custom
- Customer receives report via web interface once the process has finished
- Sequence Analysis Viewer allows you to observe the process in real-time - clusters, imaging, indexing, etc...
- Validation: measure accuracy (depth of coverage: average # of reads of Q30+ across bugs with different GC content and genome size), precision (degree of agreement between replicate measurements of the same material, includes repeatability - within-run precision and reproducibility - between-run precision)
- MIx of genuses for validation, e.g. Acinetobacter, E.coli, Chlamydia, Mtb, GBS, Campy, all your public health favourites - ran these as a checkerboard for validation, with a plate full of different bugs and to show that you can put various bugs on one plate and get good data
- KmerID: custom similarity measure between reads and NCBI reference genomes - is there DNA from more than one sample present? Can identify if something went into a stray well.
- Customer workflow for each species selected by user at submission, kmer id checks to ensure what was sequenced matched the requested workflow
- Data automatically released if E. coli positive control and negative control meet particular standards (K12 >99% kmer ID, >150Mb at Q30+ after trim; <100k reads in negative control) and organism:selected workflow kmer mismatch <7%
- Can manually release if positive control <150Mb but K12 still >99%, negative control >100k but all environment/reagent contaminant like P. acnes
    - P acnes also seen in other studies http://jcm.asm.org/content/early/2016/01/21/JCM.02723-15
- No external QA done yet but have used GMI proficient test and results for Salmonella, E. coli, and Staph were good. EUCAST was happy with this choice of PT.
- http://www.globalmicrobialidentifier.org/workgroups/about-the-gmi-proficiency-test-2017
- Had to double the amount of input DNA required when users noticed a funky issue with alignments
- Few labs in the UK accredited
- Are accredited for NGS microbiology, only one
- Examples following
- Salmonella
    - typing can be replaced with sequening- compared MLST and phenotypic serotype for 1593 isolates
    - 84% complete match between MLST and phenotypically dreived serotype, up to 94% when new serotypes considered too
- S. aureus - can you predict resistance?
    - some small errors,
- Gastro bugs have been going for >1year, others coming online
- TB (yeahhhh) all done with WGS - inferred DST, dx, etc...
- Learnings: know your throughput and expected TATs - manage expectations. Kmer ID good to confirm species. WGS data more reliable/comparable to existing methods - finds new molecular serotypes. Routine evaluation of circulating and emerging clones. Added valueL MLST, SNP typing. Detect capsular switching. Find, investyigate, designate AMR mechanisms. Use WGS data to improve PCR primers and probes for rapid dx.
- Summary: centralized service with medium throughput has produced over 100k genomes, with TART of 4-5 days (fastest service Illumina knows of)
- Benefits: migrated certain assays to WGS, rapid outbreak response. More accurate info to inform PH response. Near real-time genomics makes for a huge R&D resource. Positive experiences with initial priority projects, including good DST prediction
- The future: single molecule? Tried E. coli K12 through MAP but Nick Loman got there first ***Ha!***
- Metric of nanopore success: when you BLAST a read and get the nice red line in the results table
- Genomics Services & Development Unit, Bioinformatics Unit, and Reference Laboratories all working together in happy harmony

Keynote 2: Three Decades Riding the Exponential Curve - NCBI Putting Sequences to Use from 1988 to 2017

Jim Ostell

Acting Director, National Center for Biotechnology Information, NLM, NIH, NIH Distinguished Investigator, Bethesda, MD

- Exponential growth, both number of users and number of sequenes out. Loads of data both in and out
- 3000 different groups and individuals provide subissions to CBI daily
- 15PT data every day - work with it every day, comparison and analyzing
- download: 40 TB per day.
- many publications, many journals etc
- Goal fo talk: give a history of the NCBI
- Started in 1988
- They actually had a user interface online! 32 bit processers became available
- Internet became available for commercial users, just opened up
- CD rom 3 years old, www not invented yet
- Was not awareded PHD because people couldn't agree if what he did was biology or comptuer science, he got his degree in 1988,
- first report out re human genome, and one of the things they specified wwas that bioinf was needed
- NCBI creadet by congress 1988 - Reagan
- Lipman recruited him, recruited other people too, low pay, field without future
- Nothing is done, no resources, sat down to think what they wanted to do
- Idea: comparisons a fundamental thing to do in biology, make inferences
- But, complicated to compare functions/phenotype, much easier to compare DNA, is simply letters, methods exist for doing comparisons
- found human sequnce that was similar to human and yeast - mutation in human seq. Found ecoli/yeast seq was DNA repair gene, thus lightbulb
- developed the term frequenciy statistics - distances between sequencies
- AT the time, medline, protein seqs and dna seqs were the domains of separate groups, no connections
- Problem with bibliography info found in protein/dna databases, problems figuring out journal names, paper titles

- created iso language data exchange, made the data publicly available
- created standard: ASN.1
- but: people hated it! commercial companies felt threatened
- Stalinistic to force people to use the same thing
- consequence: Created entrez ssystem for manouvering around in things. Go between medline abstracts to protein sequence to nucleotide sequence.
- still hated it, several companies complained to congress, accused of secrecy, threatened by cuts
- Richard Roberts mobilized people to petition congress
- Came out that the signatures of some of the commercial companies on those letters were faked
- "any commercial company that can't compete with govenment shouldn't be in business anyway"
- Deal with companies etc, only supply data, no software. But, could do whatever they wanted on the internet
- Made Entrez online, before www it was a client/server thing
- Started adding more info and more links to this syste, taxa, 3d structure, and more, 25-26 of these online now
- Data deluge - wrote about this in Science in 1993
- If you're still standing, you're winning
- Question: what do you do with this data?
- First genome: 1995 *Haemophilus influenza* - a genome could be an entity in the db
- HIV was a good target to start with, HIV a lot in the press
- Collaborated with experts, authors of retroviruses book, to get seqs for all of these into genban
- next target: yeast genome, each chromosome separately owned by different institutions
- european not willing to share with anybody else
- incomplete, not annotated the same way
- Decided to create refseq - the ncbi review article of genbank
- Turned out to be a good idea
- 8471 distinct bacterial species, 95k complete genomes consistently reannotation, shrunk the no of genes in many genomes
    - ftp.ncbi.nlm.nih.gov/hmm
    - ftp.ncbi.nlm.nih.gov/pub/something+rules
- working on organizing genomes, connecting from well categorized to their neighbors
- rules for how to annotate available for download for those who are interested, paper coming
- 2010s- saw a leveling off in # of new genomes, and an increase in resequencing, switch from exploration to sequencing being an assay, cheaper to sequence whole to get to one gene
- around the same time: approached by FDA to do typing etc, also the CDC
- now have a pathogen surveillance pipeline, as a collaboration with many different agencies
- running comparisons with existing info, are starting to hit more historical samples, can see that for instance a factory didn't get cleaned enough
- also started to create dbs for AMR genes, started by feldgarden at the NCBI
- over 4000 resistance proteins now, can also accept antibiograms
- have 4000 phenotyped genomes linking between pheno and genotype
- MCR-1 article: https://www.cdc.gov/mmwr/volumes/65/wr/pdfs/mm6536e3.pdf
- have included this into the food pathogen pipeline
- "yes, we're on an exponetial curve, but you should not live in fear, you should try to surf it instead"

Monday, October 9

Morning: Session 1 – Epidemiological Cues: NGS in Clinical and Public Health Microbiology

Introductory Remarks

Jennifer Gardy

No photography is the ASM policy, but it is OK if the speaker says they are OK with it. Speakers are encouraged to announce at the start of their talk if they are OK with it.

Session Keynote: Incorporating Epidemiology and Microbial NGS to Maximize Clinical and Public Health Impact

Yonatan Grad

Assistant Professor of Immunology and Infectious Diseases, Harvard TH Chan School of Public Health, and Division of Infectious Diseases, Brigham and Women's Hospital, Harvard Medical School, Boston, MA)

- Roosevelt had chronic high blood pressure
- Yonatan's favourite president is FDR, but this isn't a polio example
- Hypertension was not at that time considered a disease
- Invention of blood pressure measurement 1881/96, lead to discovery of hypertension, lead to a theory that some had "beneficial" high blood pressure, and that some of this was bad.
- Measurement of blood pressure in 1905, butsphygmomanometer 0 invented in 1881/1896..... what was it invented for and used for before that?
- 1940s - actuarial data led to discovery that mid hypertension was connected with 2x mortality
- Interesting way of looking at a disease - from not a disease to understanding the epidemiology to an oh shit, we should be dealing with this moment
    - also, for my part: not a condition untill we can actually quantify it, measurement makes it real
- 1970's Framingham Heart Study
- NGS as sphygmomanometer- gives us the ability to quantify things that were maybe just circulating undetected/unnoticedorwhose important wasn't recognized - who is sick,who is at risk,with what? How do we reduce disease incidence? How do we reduce AMR rates? Microbiome, networks of spread, host factors, pathogen factors
- What will be NGS's Framingham equivalent?

- RE microbes:
  - no symptoms, healthy
  - colonized and no symptoms - unhealthy
  - colonised and symptoms - unhealthy
- is that how it should be?
- Example, gonorrhea - high burden of disease, imminent threat of AMR - resistance to every antibiotic used in tx. Current recommendation is dual-agent: cef and azithromycin,but this is now failing too
- 468,514 cases of gonorrhea in 2016
- lots of ressistance to amr, have become resistant to everything that is being used
- started with dual therapies to beat resistance development, just led to resistance to those too
- even dual therapy is starting to fail
- Old-timey therapies included urethral irrigation, cocaine and alcohol, and elevating patients' temperatures to 40C to kill the bug
- 1935 paper on fever therapy: https://jamanetwork.com/journals/jama/article-abstract/257884
- How can we slow the spread of resistant gonnorhea?
  - surveillance show that still, a bit over half is still susceptible
  - If we knew resistance profile right away, could prescribe more strategically - use NGS to coupe diagnosis and resistance prediction
- Sequenced 1102 isolates from 2000-13, varying resistance patterns, from different populations e.g. MSM, MSW, MSWM
- Is resistance due to clonal expansion or repeated re-emergence? WGS phylogeny with BAPS population structure - ceph is clonal, azithromycin is re-emergence
- figures etc is from https://www.ncbi.nlm.nih.gov/pu\\bmed/27638945
- redictability genotype phenotypic resistance
- Fricking gyrA. Always screwing antibiotics up for everyone. 98/99% PPV for cipro resistance, but other mutations play a role.
- Multiple azithromycin resistance mechanisms uncovered
- 65 % of resistance can be explained by known resistance mutations
- penA (especially mosaic versions of the gene) for ESC resistance, but other uncovered mutations at work as well
- Here comes the NGS! Multiple groups working on NGS forsusceptibility,BUT our knowledgebase is incomplete, we don't know the implication of mixed infection (how sensitive do tests need to be?), how do we develop new assays when a new antibiotic is introduced?
- ARIBA! - realy like that program
- Problem: we don't know enough about the mechanisms that give resistance, are likely to be many undiscovered still
- how to we then identify novel mechanisms?
- WGS to predict MICs in Neisseria gonorrhoeae (ridiculously long link...): https://watermark.silverchair.com/dkx067.pdf?token=AQECAHi208BE49Ooan9kkhW_Ercy7Dm3ZL_9Cf3qfKAc485ysgAAAc4wggHKBgkqhkiG9w0BBwaggqG7MIIBtwIBADCCAbAGCSqGSIb3DQEHATAeBglghkgBZQMEAS4wEQQMTvJgBB2IZ1BX5pNuAgEQgIIBgUu_iJ5mW5E-j1qN60Fkszi1vHqWzYrfI4NG4jJ4ThOOvMYsIZCISwsjYozRev0ivst4rQas0RsZjPVW9BGMmdNKeIzczSpuPIdYcU3xVa5s5s5yYryNw70EZYZdfWVXytqeKqIZd-s8fFClzYuLeps-30mE54OUeOZsPPo8i6dPRE71XTNuBX8eCzWtE6mFLtAjvuHd63ikKck2jH9qCJWTWWZus28Y0FnkTvao-ag_JY-8k5D970RfBphqiOg7n6_avNpMcyqFCUjMSp2ai2kWTPCS0YKafgXHVCr_mu-EYn9HNjDZvT4GJJlf-dbldixhE2J0-u6eby1fgPYSxef6CMECq407LyshPAqRkRO2nP4ESYlYCV-ZO8ujkZUNDMB3kGmQCmyzozqzOKJGmG78LiKzOKYShco2hqzFqPcrE_jitHCpW9Wt2YiRGAtvKDADySw15fIa_ZeNcRZq74qZe0q791TWiIDYdOCz353Zf8zSm7uTvqqoa9DIshOcd6A
- Resfinder 3.0 https://academic.oup.com/jac/article/doi/10.1093/jac/dkx217/3979530/PointFinder-a-novel-web-tool-for-WGS-based also have an experinemtal (not published yet) feature to predict Neisseria gonorrhoeae from reads https://cge.cbs.dtu.dk/services/ResFinder/ based on known resistance mutations from https://www.ncbi.nlm.nih.gov/pubmed/24982323
- Practical question: if X percent of your populatin is resistant to something, do do still treat witht that something?
- How do we translate genomic data into clinical dx?
- Modelling! Compartmental model describing gonorrhea transmission in MSM under 3-antibiotic regimen. SI, where I can be any combination of the drugs
- POC testing device that reports resistance for A, B, and C gives best results and delays the emergence or AMR to these 3 drugs versus device only reporting results for drug A - diagnose over 25% of cases to achieve best result
- Can we infer fitness cost of AMR given epi/evo data? My pal Xavier! Pull in other data streams https://www.biorxiv.org/content/early/2017/03/28/121418
- https://www.ncbi.nlm.nih.gov/pubmed/28968710
- Higher fitness cost for azithromycin resistance than for quinolone resistance - preferred to use azi instead of quinolones when sensitive to both
- Quinolone resistant lineages of C diff in the UK fell after quinolone prescribing fell. Dingle Lancet Infection Dis. https://www.ncbi.nlm.nih.gov/pubmed/28130063
- NGS informs our biological understanding of resistance, helps us improve prescribing, and informs how we roll out a new antibiotic
- Genomic epidemiology and public health
- If a resistant strain appears, where will it go next? 2014 paper on cefixime resistance spread from west to east and from MSM to MSW networks
- Reconstruct transmission paths and compare to known sexual contacts. Brighton study: 72% of isolates sampled over3mo period linked by transmission
- Sequenced 800+ isolates from NYC - what are the demographic overlaps with the genomic clusters? Can it be used to develop models?
- How do we make this actionable? What % of cases need to be sequenced? From which subpopulations? What additional data do we need? How do we build the infrastructure to integrate different labs, different datasets, facilitate sharing?
- Above the belt and below the belt gonococci overlap somewhat - e.g. recent outbreak of N. mening urethritis in multiple states
- large outbreak of n. menigitidis urethris in multiple states going on these days
- Outbreak mening picked up two anaerobic environment adaptation genes from gono - probably facilitate the urethral growth adaptation. Does this happen often? (Neisseria is super plastic. Lots of flux/).
- Global collection of mening have the genomic factors linked to urogenital niche adaptation
- NGS helps us in our diagnosis, define the epi, and generate hypotheses about the underlying biology
- NGS can improve dx: speed,accuracy, expand range of dx-able conditions, guide clinical decision-making, understand interaction between patient and population; improve prognostics - what will our Framingham be? Develop new clinical and PH interventions; democratization - technology, training for interpretation

Questions:

- Jenn Gardy: What would the NGS "Framingham" study look like?
- Question about Antibiotic resistance prediction
- how sensitive do we need to be? what level of error rate can you tolerate - use phenotypic if always want to be right

Validating a System for Grading Drug-resistance Associated Mutations in *Mycobacterium tuberculosis*: Comparison of WGS Data in the Relational Sequencing TB Data Platform (ReSeqTB) with a Systematic Review

Rebecca Colman
University of California San Diego, San Diego, CA

- WHO 2016 TB Report http://www.who.int/tb/publications/global_report/en/
- 1/3rd of all have tb, but only 10% of these will progress to active tb
- Top 10 cause of death world wide in 2015
- YEAHHHH TB TIME- it's active, it's latent! 10.4M got it last year, 4.3M don't get appropriate care - wrong drugs are being used
- 1.8 million people died from TB in 2015, 400,000 with HIV and TB
- big problem: are not able to provide all with good enough care
- have to be able to identify resistance, so that patients get the right drugs
- Only 1 in 5 MDR patients are dx'd,and <10% of those got second line DST
- also: not just discover that they're resistant to 1st line drugs but all 2nd line drugs
- Conventional TB DST is culture-based: slow, expensive, and needs BSL3 - not available to most settings where MDR is
- normal testing is slow and eexpensive, require BSL3 facilities
- Need a standardized genotype-phenotype database to facilitate clinical interpretation of WGS data in TB
- a lack of user- friendly data analysis and interpretation tools has been frequently cited as a major barrier to routine use of WGS techniques
- What do SNPS mean, what are their impact on drug resistance?
- Urgent need for standardized database
- ReSeqTB Knowledgebase: NGS variant calling pipeline for SNPs in aindels; curated db of genotypic, phenotypic, and clinical data; cloud-based analysis platform to facilitate global networking
  - Poster #105: A standardized and validated NGS variant detection pipeline
- ReSeq: driven by experts and the community, grading criteria for confidence-binningofmutations (frequency ofmutation in R or S strain), every SNP as a score accodring to the available information, five bins, defined using statistical approach
- If anyone wants the notes from the ReSeqTB meeting in London last week, mine are at https://drive.google.com/file/d/0B8yhcYgIvEHeLXZtY3BDem8xSlU/view?usp=sharing
- 6 different bins for SNP classification- High confidence, moderate confidence, minimal confidence, no association with resistance, and indeterminate
- Bins have various confidence levelbasedon likelihood ratios and p-values, and each bin has interpretation - high, medium, low confidence, no association with resistance, and indeterminate
- P values of <0.05 for all bins other than indeterminate, Likelyhood values of >10 for high, 5-10 for moderate, 1-5 for low confidence, <1 for no association with resistance
- Systematic review of the literature to identify genes and drugs of interest - pulled as much information as possible, e.g. medium, critical concentration, mutation frequency within R or S populations, used this to develop the list of mutations and their bins
- Paolo Miotto has a paper coming out soon that describes this whole thing
- 11 drugs, 16 loci, 1748 papers screened, 43 countries and 13,424 isolates
- Graded mutation system improves specificity dramatically (13%) with only modest descrease in sensitivity (5%) - reduces very major errors (when R is reported as S)
- ReSeq had 6,124 isolates at end ofJuly, looks at 96 loci of interest and 178 mutations. ReSeq yielded 19 mutations also found in systematic review, though with varying confidence. 116 SNPs were graded in one dataset but indeterminate in the other, 39 graded in one dataset but not observed in the other, only 4 discrepancies - ReSeq said R-associated and systematic review said not associated - need to look at association between mutations to sort this out
- Some discrepant mutations mentioned: rrs 517 c to t (capreomycin), katg 1388 G to T (isoniazid), gyrA 284 G to C (oflox), pncA 139 A to G (PZA)
- for other bugs where resistance is due to other things than just mutaitons: combine phenotypic and genotypic data
- analytical approach for validating and interpreting drug resistance associated mutations is critical for the advancement of sequencing to be successfully integrated into patient care
- ReSeq TB website: https://platform.reseqtb.org/

Questions:

- Jenn Gardy: TB is a complex disease, but has a simple genome. Do you have recommendations for similar pathogens with resistance caused by changes in gene expression/ efflux pumps and not simple point mutations?
  - A: Combine genotypic and phenotypic information when you have it. Do more phenotypic work to confirm things. Not sure I have recommendations for more complicated organisms, but there's a lot you can do with NGS and gene expression work. Start to look for associations in a knowledge database of phenotypic and genotypic data and bring it all together.
- Dag Harmsen: How do you deal with mixed infections? Are you kicking out minority variants? How do you deal with it?
  - A: Poster 105 will deal with QC on data coming in, and how you get high confidence in SNPs. There are no mixed samples in the database, we're specifically looking to remove those. We're working off of isolate DNA and not including mixtures in analysis, we're not sure how that would affect phenotypic DST. This is a critical factor in patient samples though.

Metagenomic Sequencing for Diagnosis of Brain and Eye Infections

Steven Salzberg
Johns Hopkins University, Baltimore, MD

- TIGR shout-out! All known human pathogens have now been sequenced - 40k bacterial, 80k viral, 250 eukaryotic
- Today: essentially all known human pathogens have been sequenced. Novel pathogens arise periodically, but these of course get sequenced right away! SARS, MERS, Zika (reemergence): the time from emergence to sequencing completion is getting shorter and shorter, from months to weeks!
- Clinical metagenomics: sample infected tissue, send for rapid sequencing, 20M+ reads, find the infectious agenct
- take sample, sequence the heck out of it, search reads to see what you have
- most data will be human, throw those away, see what you have left

- Think about it as an assay that costs a few hundred bucks, don't worry about throwing out 99% of the data
- Speed matters - one samplealigned by BLAST would take 3-5 days, which is not exactly compatible with clinical dx
- created kraken to do this identification
- alignment not really necessary, can settle for looking for exact matches, if they're many enough and long enough, we're happy
- depends on having a specially prepared database
- Alignment shmalignment
  - :-)
- Kraken has the ability for the user to build their own database
- The human genome is being added to the Kraken database, which will make the results more accurate for human pathogen detection
- Kraken (https://ccb.jhu.edu/software/kraken/) can be found here on GitHub: https://github.com/DerrickWood/kraken
- Kraken publication by @DerrickEWood and @StevenSalzberg1 = https://genomebiology.biomedcentral.com/articles/10.1186/gb-2014-15-3-r46
- Simple database building, what sequences of length k are associated with each species, algorithm is kmer lowest common ancestor mapping to taxonomy tree, sequence classified as belonging toleaf of highest root toleaf path (path with most kmers)
  - Typically k = 31 bp
- 1.5M reads per minute classified (naive bayes = 7, megablast = 7k)
- Independent assessment says Kraken is the coolest!
  - https://www.nature.com/articles/srep19233
- Lowest error rate and one of the fastest.
- Despite contamination, sample prep challenges, etc... can we use NGS to dx infections
- Braaaaiiiiins - now done 30+ cases from difficult neurologic cases - suspected infection, brain biopsy - take SMALL amount, get onto the sequencer right away
  - Half of brain infections go undiagnosed, often fatal, different categories of pathogen yield similar symptoms
  - PT8 67yo woman, intracranial, spinalcord and lung lesions, positive for Nocardia, two lesions biopsied, two NGS runs for each (14M and 15M reads)
  - One run - nothing. Second run: 15 reads mapped to Mycobacterium
  - I KNEW IT! I KNEW IT WAS MYCOBACTERIAL. NOCARDIA STAINS ON AFB SO THEY HAD A FALSE POSITIVE MWA HA HA HA
  - INH,PZA, and RIF tx initiated and led to recovery
  - Solved 3/10 cases, another solved a year later when Elizabethkingia outbreak hit the news and a genome was added to NCBI
- An evaluation of the accuracy and speed of metagenome analysis tools (Lindgreen et al. 2016): https://www.nature.com/articles/srep19233.pdf
- Pavian https://github.com/fbreitwieser/pavian
- Great to see Elizabethkingia used as an example of the value of having sequences available for unusual human pathogens.
- https://www.nature.com/articles/ncomms15483
- https://trace.ncbi.nlm.nih.gov/Traces/study/?acc=SRP072035
- It worked for brains, let's do it on eyes - 20 banked corneal samples, usually paraffin-preserved
  - 67yo woman w/ glaucoma, had surgery, 12d after surgery came in with keratitis, corneal edema (not an unusual complication)
  - Cornea is non-sterile, 46M human reads, 28k acanethamoeba, 13k B. cepacia (interesting fact from Yonaton who is sitting next to me - this pathogen causes onion rot!)
  - 17/20 corneal infections diagnosed
- Eukaryotic pathogens are a big challenge: only 250 genomes, all draft, many contaminants within the genomes themselves - toxoplasma (CATS!), sarcocystis
- Cleaning eukaryotic draft genomes - removing the crap to leave you with a better substrate for your clinical metagenomics - massive improvement in classification
- Jenn: "The takeaway is 'clean your data, and clean your contacts!' "
- Question on host removal: no technical reason not to, ethical concerns of leaving human data in the dataset. Steven's lab includes human as an organism in his kraken databases, usually.
-

Prognostic Drug Sensitivity Testing: Targeted Sequencing to Detect "Pre-Resistance" in TB

David Engelthaler
Translational Genomics Research Institute, Flagstaff, AZ (www.tgennorth.org)

- Amplicon sequencing - the next generation of DST for TB
- Concept of pre-resistance, not just diagnostic, but prognostic
- Building out amplicon panels for the classic drug/mutation combos and new drugs, lesser-known mutations, playing with MinION too
- Amplicon sequencing panel consists of eis, gyrA, inhA, katG, rpoB, rrs, and pncA - they found 90-95% association of mutations in these regions with drug resistance
- Heteroresistancee - S and R phenotypes in same population - may come from superinfection of multiple strains at once, but more commonly representing within-host diversity
- Phenotypic DST can pick up 1-5% minor population, but what if R is 1 in 10k? Can we find it? Is it important?
- Minor populations will quickly become dominant in the presence of drug. Theoretically as low as 1% but more likely 5-10% in practice.
- SMOR Single Molecule Overlapping Reads - primer gives you double-coverage on mutation of interest and overcomes sequencing error (average 1% error in Illumina data in high-GC bug like TB) - SMOR reduces error detection by 2 OOMs, achieve 1 in 10k. Detect microheterresistance
- Looked at panel of phenotypic R but genotypic S (at the classic sites) isolates - Sanger didn't spot mutations but SMOR found the crytpic mutations
  Heteroresistance lost in subculture
  Hollow Fiber System (the "glass mouse") – very fine control over how much bug, how much drug goes into a system. Can set up microheteroresistance then apply sub-tx RIF - 1% heteroresistance moves to 60-80-fixed resistance - detect heteroresistance as early as possible to initiative correct tx
- Moldovan sputum samples - found low-frequency katG mutations in phenotypically S strains - pre-resistance?
- Had to look at serial samples from S. Africa (MDR->pre-XDR->XDR)
  Pateint X, samples over 3 years, INH-R, SNP went from 66-100% fixation; AG - began as S, went to R, no SNP in first two, 80% at 3rd sample, fixes by end - no hint of exciting heteroresistance here
  Patient Y, four years and lots of samples, AG resistance only inlast sample, first four = no SNP, fifth sample see rrs1401G at 0.1%, then 4%, then 30% and phenotypically R
  Found 8 patients with pre-resistance detected - rrs mutation dynamics somewhat variable but mutation detected 8 months before phenotype observed
- Genotype is a better predictor of outcome than phenotype

- The TB quasi-species: multiple foci within the lung, each lesion is doing its own thing
- Single cell resistance lineage analysis: are mutations occuring in the same bug?
- GyrA again, such a jerk (Hey, I love that gene!):) It's like a troublesome teenager - changes all the time"How was your day today, gyrA?" "I dunno" *mutates* <-teenagerimpression, thus mutations is the gyra equivalent of getting a tattoo? Or maybe drugs "are you doing drugs again? you seem to be changing a lot" "No, you taught me to be resistant to drugs"
- Sample with 100% gyrA resistance, but whoa, two different alleles inequal proportions; other sample with 10% gyrA - four different R lineages (four different lung foci)
- $30/sample genotypic DST.
- Dynamic changes for personalized medicine- track resistance over time. Important for TB, CF, etc...
- Use SMOR as a prognostic assay - find pre-resistance and manage appropriately
- Is SMOR a sphygmomanometer?

Questions:

- Q: Why amplicon over whole genome? A: it's cheaper, and you can see resistance factors that you can't see in WGS except with high coverage
- Q: Do you ever see coinfections with gyrA mutatants at codon 90 and 94? A: Yes they do, and they see both pathways to get to that, but unless you do serial analysis you wont see the change happen. One mutant will replace the other- typically mutations at 94 replace 90.

Genomic Proof of Probiotic Transmission from Capsule to Blood in Patients with *Lactobacillus rhamnosus* Bacteremia

Christina Merakou
Boston Children's Hospital, Boston, MA

- BCH ICU: 6 cases Lactobacillus bacteremia over five years, all were receiving probiotics with Lb (often prescribed for anitbiotic-associated diarrhea)
- L. rhamnosus - enhance gut immunity, make antimicrobial substances, other nice things
- Patient-level risk factors predisposing to LB bacteremia? Blood strains from the probiotic? Within-host evolution?
- 645 patients received 15k doses of probiotic, 81% with L. rhamnosus GG
- Matched bacteremai patients to kids with similar ICU history, probiotic exposure
- Risk factors: GI disease, meds, illness severity, device utilization - none were significant
- 48 sequences from probiotic and 6 from blood isolates and deep sequenced 500-700x commercial pbx
- Pbx were the source of the bloodstream infections
- 4/6 blood isolates have unique SNPs suggesting within-host evolution - one showed a RIF-resistance mutation, turned out pt had been treated with rifamixin, somewhat enhanced biofilmformation in the bacteremia-derived Lr isolates
- I-TASSER - visualizes structural changes as a consequence of mutations (https://zhanglab.ccmb.med.umich.edu/I-TASSER/)
- Pbx-associated bacteremia happens, but very rare
- How does it get into the blood? Catheter contamination via HCW dirty paws? Inestinal translocation?

Afternoon: Session 2 – From Pipelines to Pixels: NGS Data Integration (Reporting, QC/QA, Accreditation, Training) and Visualization

Session Keynote: Real-time Tracking and Prediction of RNA Virus Evolution

Richard Neher
Associate Professor of Computational Modeling of Biological Processes, Biozentrum of the University of Basel, Basel, Switzerland

- Slides at https://neherlab.org/201710_ASMNGS.html
- Love a person with their slide URL on slide #1!
- Cool timeline of flu from 1910 onwards - H1N1, H2N2, H3N2, and swine origin H1N1: avian, human, and swine histories
- Lovely tree - flu phylogenies have a glorious backbone with little lineages dangling off, one of which moves to dominate the next seasons - gotta keep updating that vaccine!
- CONSTANT VIGILANCE
- nextflu.org with Trevor Bedford (and nextstrain.org!)
- Can we use nextflu for prediction? Who's the viral lineage we should be worried about next? Look at present-day branching order to predict future state
- Model distributions of competing variants - variants from the middle of the distribution were at the rightmost skew of the fitness distribution of the prior generation. Fit variant clonally expands while other variants try to catchup
- Posterior fitness distribution for each node and tip in the tree, find the peak posterior that lies to the right the rest of the distribution
- Yay, WHO is updating our vaccine for the next season (but not this one so wash your damn hands)
- nextstrain extends this framework to other pathogens: ebola, zika, stay tuned for our mumps one coming out this week or next yay new data! OH IT'S OUT YAY
- Go check out my mumps genomes along with their friends from Broad: http://www.nextstrain.org/mumps
- Nextstrain architecture available to all - suck in sequences, they go into fauna db, augur python pipeline processes them, and auspice js package makes the pretty vis
- RethinkDB - https://www.rethinkdb.com/ - open-source real-time JSON database
- Treetime - super-rapid time-scaled phylogenies - never wait a month for your BEAST chain to converge again!
    - https://github.com/neherlab/treetime
- Treetime: you have a tree, the leaves have time labels, you have the underlying data. TT reconstructs the ancestral sequences at each node and thereby estimates branch length, then propogates timing information from leaves back to the root - when did that ancestral node exist? Nice little distribution of time around each node with CIs. You can do the same thing with any sort of tip label, e.g. geography. Fit a population size distribution, iterate, and you get a

super-slick molecular clock-based tree before you can say BEAST! Ha! Actually I love both methods.
- Bacteria! Hooray! Vertical and horizontal transmission, rearrangements, large genomes, variation at different rates throughout genome, varying quality of annotation
- pangenome.de - panX: pan-genome identification pipeline: analyse each COG, do some other stuff I missed
  - https://github.com/neherlab/pan-genome-analysis
- We need tools for interpretation and exploration of increasingly large genomic datasets. Provide breadth and depth - overview->integrate->deep dive
- Actionable outputs require real-time analysis - we need fast pipelines for analysis

Evidence-based Design and Evaluation of a Whole Genome Sequencing Clinical Report for the Reference Microbiology Laboratory
Anamaria Crisan
University of British Columbia, Vancouver, BC, CANADA

- All this good stuff is written up - check the paper (freshly out this weekend) at https://www.biorxiv.org/content/early/2017/10/06/199570, and the webpage with a project abstract and links to all the different pieces: http://www.cs.ubc.ca/labs/imager/tr/2017/MicroReportDesign/
- The report template is free on Overleaf: https://www.overleaf.com/latex/templates/tb-wgs-report-for-reference-lab/psmnzmcnwrwm and we put a load of data and analysis code on GitHub: https://github.com/amcrisan/TBReportRedesign
  - thanks!
- focus is on if the report actually reports out what people need.
- Used design study methodology
  - discover
    - figure out what different people that received data were actually doing with it - interviews with 17 different people of different types
    - speciation and drug ssceptibility the more useful data from wgs in their opinion
  - design
    - created designs and tested them against the current reports
    - new designs were preferred
    - focus on precise interpretations - avoid abbreviations, be careful with language
    - prioritize clinically actionable data
    - for instance, want trees, but preferrably with metadata/context
    - pictures might not be as amazing as we bioinf people want them to be
  - implement
    - put together a report whose template is on overleaf, see above
- People are familiar with one representation of data (trees) but don't always know what they want. A common problem with this type of survey unfortunately. "I don't like it", but no specific suggestions to improve.
- YOU GUYS, MY STUDENTS ARE SO COOL :)
- also includes a template for how to go through such a design process for other things - intuitive for you might not be intuitive for others, and graphical UI does not automatically equal user friendly
- slides: https://goo.gl/9jt625

Whole Genome Sequence Analysis: wgsa.net - Rapid Online Interpretation of Microbial Genomics for Surveillance and Epidemiology
David Aanensen
Centre for Genomic Pathogen Surveillance, Wellcome Genome Campus, Cambridgeshire, UNITED KINGDOM

- Follow @theCGPS - they need some twitter love (https://twitter.com/thecgps)
- Exploring how thing spread between institutes, hospitals, countries, etc
- Sample, sequence, and show similarity = pretty classic genomic surveillance approach. Paint on extra info, like AMR markers
- Focus is on distilling data into information
- Many geographic scales at which info is being collected and visualized
- Here comes the microreact.org.... - this dataset is the MRSA one from David and friends' mBio paper (it was actually from Croucher et al - PMEN 14 work)
- Cheap plug: microreact.org has partnered with the Microbial Genomics journal, on whose editorial board I sit, to include MR visualizations of datasets published in the journal
- CPGs is about providing tools and carrying our structured pathogen population surveys
- epicollect for portable epi data collection and WGSA for doing the pipeline good stuff
- microreact takes a tree (newick) and a linelist (CSV) and gives you lovely viz, whoa .dot format contact network - this is new
  - dot format: https://gephi.org/users/supported-graph-formats/graphviz-dot-format/
- epicollect: make a project,design a questionnaire, load it onto mobile devices, collect data, and visualize online - pathogen agnostic. Snazzy form-builder interface.
  - again, keep wishing this was something we could set up with a different backend, we are frequently precluded from storing stuff outside of our institute network
- David here: you can set this up with a different back end - the clients(mobile) and back end are decoupled - there is a setting on the apps(mobile) to define the location of the 'server' it's then a matter of setting up your server to accept the request.. we have a bunch of groups (mostly African agencies) doing things this way - email for more info (dmaa@imperial.ac.uk)
  - http://www.epicollect.net/instructions/developers.html links to google code, which is offline. Seems like newer version is https://github.com/ImperialCollegeLondon/EpiCollectplus
  -
- wgsa.net for pipeline fun times
- Presentation ended with a picture of a sphygmomanometer to keep with the theme

Enabling Phyletic-based Visualization and Comparison of Genomic Islands for Tens to Hundreds of Microbial Genomes
Claire Bertelli
Simon Fraser University, Vancouver, BC, CANADA

- Genomic islands: they came from horizontal gene transfer! Lods of types - prophage, integron, ICEs, transposons, etc...
- They tend to have different nucleotide composition than the surrounding sequence, have mobility genes, and encode metabolic genes, virulence factors, and AMR genes - they encode 5x as many virulence factors as non-GI regions, and GIs are not generally enriched for AMR genes (work in progress), though some classes of AMR element are preferentially in GIs
- amr ontology: https://ardb.cbcb.umd.edu/go.shtml (not really referred to in the talk, but mentioned briefly)
- Lots of tools for GI prediction - compared 19 tools
- what's the nameof the best one? can't see the screen well enough
  - IslandViewer4, which they developed of course :) P)
- Many tools identify GIs, but few visualize them
- IslandViewer4 - precomputed predictions for RefSeq genomes, and enables custom analysis of your own genome whether draft or complete
  - http://www.pathogenomics.sfu.ca/islandviewer/
- IslandCompare - web-based tool to compare 2-100 genomes' GI content (upload via Django, parsnp for core genome phylogeny, mauve for pairwise alignment, two GI predictors, Mash-MCL to cluster GIs, RGI/CASRD for AMR gene ID, dump into db and visualize with D3 CompareVis)
- IslandPath-DIMOB: https://github.com/brinkmanlab/islandpath
- IslandViewer: www.pathogenomics.sfu.ca/islandviewer
- IslandCompare: https://github.com/brinkmanlab/islandcompare
- https://github.com/brinkmanlab/islandpath

Tuesday, October 10

Morning: Session 3 – Farm-to-Table: NGS in Veterinary, Food, and Environmental Microbiology

Opening Comments
Marc Allard

- The International Standards Organization (ISO) has accepted WGS as something that should be standardized globally.
- More information about FDA Genome Trakr - https://www.fda.gov/food/foodscienceresearch/wholegenomesequencingprogramwgs/ucm363134.htm
- It sounds like the FDA would like to begin making functional annotations in the genomes that are currently a part of Genome Trakr, in addition to the current phylogenetic analyses.
- The next generation of analyses will be metagenomics, but the FDA isn't there yet. There is not a cost effective application and supporting database(s) yet at the FDA. All current draft genomes are WGS from single isolates.

Session Keynote: Single-cell Sequencing: From Diversity to Function
Tanja Woyke
Head, Microbial Genomics Program at DOE Joint Genome Institute, Lawrence Berkeley National Laboratory, Walnut Creek, CA

- NGS in the context of microbial ecology
- DOE funded research
- Continually develop new technologies that are made available to users, e.g., single cell genomics pipelines
- First, a bit on how single cell genomics work
- Tree of life, many clades that are sparsley known
  - https://www.nature.com/articles/nmicrobiol201648
- Two main methods overcome requirements to culture before sequencing: shotgun metagenomics and single-cell genomics
- Single cell dna: old day dilution, but very inaccurate, now flow cytometry (was that it?) I think that's right
- Then lyse the cell
- MDA: Multiple Displacement Amplification, very cool technique for amplifying a whole genome including from a single cell
  - Result is a hyperbranched DNA molecule
  - Least biased method, but still vulnerable to some artifacts, including chimeras and palindromes, some of which can be ameliorated in the bioinformatics steps. Be mindful of these artifacts.
- Will shut down amplicon seqencing pipeline come spring, move solely to shotgun
- FANCY VIDEO! Acousticsc are used to transfer a specific volume from one container to another very accurately and effectively
- using NextSeq platform, recommending to use this one for single cell, apparently some bleed over in hiseq
- assembly - reads are biased, use a single cell assembler, possibly with kmer reduction first, very happy with SPADES
- Bower et al (ETLS) - QC of SC data
- Naturally, contamination with DNA upstream of the whole genome amplification will result in coamplification of the contaminating DNA. QC is important!
  - Tennessen, et al. ProDeGe: a computational protocol for fully automated decontamination of genomes. https://www.nature.com/ismej/journal/v10/n1/pdf/ismej2015100a.pdf

Lux, et al. acdc - Automated Contamination Detection and Confidence estimation for single-cell genome data. https://bmcbioinformatics.biomedcentral.com/track/pdf/10.1186/s12859-016-1397-7?site=bmcbioinformatics.biomedcentral.com

- MDA bias seem to be mostly random
- Genome completeness varies based on number of Single amplified genomes (SAGs) and species analysed in the pipeline
  - Incomplete SAG assemblies complement each other
    - Stepanauskas, et al. Improved genome recovery and integrated cell-size analyses of individual uncultured microbial cells and viral particles. https://www.nature.com/articles/s41467-017-00128-z.pdf
- FISH: Fluorescent in situ hybridization
- More about single cell sequencing here: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4321370/pdf/mbt20008-0038.pdf
- 200 single cell genomes from candidate phyla:
  - Rinke, et al. Insights into the phylogeny and coding potential of microbial dark matter. https://www.nature.com/nature/journal/v499/n7459/pdf/nature12352.pdf
  - These genomes can act as phylogenetic "anchors".
  - Proposed new bacterial and archael superphyla
- Proposed minimal standards for SAGs and metagenome-assembled genomes:
  - Bowers, et al. Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea. https://www.nature.com/nbt/journal/v35/n8/pdf/nbt.3893.pdf

Unbiased Strain-typing of Arbovirus Directly from Mosquitoes Using Nanopore Sequencing: a Field-forward Biosurveillance Protocol

Joseph Russell
MRIGlobal, Gaithersburg, MD

- Detected circulating Venezuelan Equine Encephalitis (VEE) in the Everglades
  - "We got lucky, or unlucky, depending if you think about it from a scientific or public health standpoint."
- Pooled NAAT with Biomeme iPhone Rt-PCR device, followed by nanopore recovery of EVEV from the RNA itself - 33 reads, many hybrid EVEV-Culex, but they really did the hardest things they possibly could, trying to get sequence this way. And they did it! Yay!
- Mini Intel computer in the field could run BWA, LAST, or Kraken against a targeted database
- Many of the SNVs called by Illumina also showed up in the MinION data (10/16) - permitted strain-typing of EVEV against the complex background of the whole mush of what was sequenced
- Biomeme sample prep -> extracted RNA -> Biomeme two2 -> deplete RNA -> amplify whole transcriptome (Sigma kit) and some other stuff but now the slide is gone
- Squish a bug and you can do meta-transcriptomics (with enrichment) and get usable public health info out
- "Mercury Lab": 72h field hand-held molecular device platform - Biomeme, MinION, Intel NUC
- SEQUENCING VAN is the best minivan
- Pooled 25 female mosquitoes per tube; collected in a trap next to an irrigation ditch (biasing towards collection of unfed mosquitoes, so potentially underestimating the prevalence of the virus)
  - Trap based on CO2 and light attraction of the mosquitoes
- Most reads were ~2,000 - 3,000bp from the MinION
- FDA is collaborating with MRIGlobal in future studies

Usage of Organelle Genome Polymorphisms for Differentiation of *Cyclospora cayetanensis* Isolates from Outbreak Samples

Hediye Cinar
U.S. Food and Drug Administration, Laurel, MD

- Let's just abbreviate *Cyclospora cayetanensis* to CC for note purposes
- Ingest an oocyte via poo-laced food or water, no person-to-person transmission
- Many produce-linked outbreaks with thousands of cases. Berries, basil, and cilantro - all the tasty things.
- The pattern of US outbreaks in 2013(?) suggests that there was more than one outbreak
- Endemic areas in places you'd expect, epidemics in travellers, cruise ship peeps, others in NA - little overlap between global endemic and epidemic regions
- Typical complex parasitic life cycle - sporulates in the environment, which is the infectious unit - spore opens in the gut and becomes an oocyst
  - https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2806662/figure/f1/
- The oocyst is the only thing you can rock out on in the lab
- No culture method, hard to get the oocysts from food and poop, oocyst cell wall is a jerk - hard to lyse and get DNA
- They wanted to sequence and de novo assemble a 45Mbp genome from a few tiny nanograms of metagenomic-derived DNA
- Murphy et al. BAM 19b: Molecular Detection of *Cyclospora cayetanensis* in Fresh Produce Using Real-Time PCR. https://www.fda.gov/food/foodscienceresearch/laboratorymethods/ucm553445.htm
- Used to use rRNA/hsp70 for typing but can now use mitochondria (6.3kb) and apicoplast (34.1kb) sequence
- Texas cilantro outbreak: only 14% of outbreak cases could be linked into clusters - rest were a mystery
  - Abanyie, et al. 2013 multistate outbreaks of Cyclospora cayetanensis infections associated with fresh produce: focus on the Texas investigations. https://www.cambridge.org/core/services/aop-cambridge-core/content/view/0EF2268B9CFFAD8CA9556D18975F96DE/S0950268815000370a.pdf/2013_multistate_outbreaks_of_cyclospora_cayetanensis_infections_associated_with_fresh_produce_focus_on_the_texas_investigations.pdf
- Organelles are good molecular epi targets: SNPs not disrupted by recombination, maternal inheritance means stability, multicopy makes them easy to grab
- Purify oocyst (need 10^5), extract DNA, library prep, NGS, bioinformatics to get genome, mt genome, and ap genome
- Booooo a contaminated food item might only have 200 oocysts, so they wash to get some oocysts, then PCR up the mt genome
- Apicoplast genome is larger and AT rich and hard to recover

WGS of Environmental *Staphylococcus aureus* Strains Show Numerous Virulence Determinants

Sandra Tallent
US FDA/CFSAN/ORS, College Park, MD

- Oooh, a closed genome! didn't know they still existed! heheh
- gives various infections, wide temp and ph range, can survive in nacl up to 15% (gah!)
- Have many virulence factors and secreted proteins
- Also causes food poisoning - points out that there's a difference between food poisoning (toxin) and infection
- Virulence factors: lots of surface proteins, trigger reactions from various cells in us
- Also resistance to antimicrobials and other things
- Start: collection of strains isolated from bakery linked to a recurrent outbreak event
- Became involved because bakery shipped goods across state borders
- Found staph from many different surfaces in the bakery
- biochemical profiles, amr testing, immunoassays, pcr enterotoxin screening, pfge strain variability
- Then used wgs: 70 strains sequenced on Illumina MiSeq, 7 of those sequenced and closed on PacBio
- Compare genotype, phenotype, not 100% match
    - ? invitro not sensitive enough
    - ? new genes
    - ? lack of expression
- After looking at the percentage of enterotoxin genes detected by PCR and WGS, it looked like there was incongruence between the two assays in the expected phenotype.
- do we have HGT? Pathogenicity islands, plasmids, phages
- Were able to locate phages in the genome
- Plasmids are persistent and globallywidespread
    - B-lactamase persistance from the 1940s
    - 23 serologically distinct proteins for staphylococcal enterotoxins (SE)
- Prevously: outbreaks > disease, but that is changing with wgs, can now see microbes without them having caused disease yet
    - Yay!
- Q: Are any of the phages lytic? A: We don't know yet, but we will test that

Meta-analysis Describing Genomic Relatedness Among Epidemiologically Well-defined Isolates of Priority Foodborne Pathogens in Canada

Aleisha Reimer
Public Health Agency of Canada, Winnipeg, MB, CANADA

- suffering from sequencing fever, got a lot of data, more than 10k isolates, mostly salmonella and listeria
- How do we make public health decisions when genetic relatedness is grey? This decision has an impact on many companies, as well as the public.
    - We have to be able to stand behind our decisions and interpretation guidelines.
- Public health Decisions are black and white, but genetic relatedness is grey
- Listeria: absolute genetic differences, reimer et al 2017, related 0-1 snps
- Looking at core snps, looked at a graph showing # of differences, found epi related isolates to be wtihin 18 snps apart, with a valley until the next set
- 3700 genomes from 3 pathogens on a MiSeq?
- SNVPhyl: A Single Nucleotide Variant Phylogenomics pipeline for microbial genomic epidemiology
    - Preprint: https://www.biorxiv.org/content/early/2016/12/09/092940
    - GitHub: https://github.com/phac-nml/snvphyl-galaxy
    - IRIDA Workflow Description: http://www.irida.ca/workflows/
- 40% of Listeria PFGE clustes do not appear to be truce clusters by WGS
    - False PFGE clusters = dead end investigations
    - This is why we need money to do real time WGS!
- Salmonella enterica Serovar Enteritidis
    - Y'all shouldn't order your kids baby chicks for Easter. This is a bad decision. At least they poked holes in the box? :)
    - 12 cases of infection from adorable baby chicks
    - Of 485 flock isolates sequenced, 482 matched the outbreak clade!
- Chicken on a leash!!
- Killing kids' pets. That's my Canada. Yeesh :(
- 105 Salmonella WGS clusters were posted from surveillance activities, exceeding more than all other organisms posted in previous years combined
    - ~50 of these are active in any moment in time
    - Trees are being generated every hour
- Retrospectively sequenced ~360 Salmonella outbreak isolates that previously had PFGE
    - Created a nice phylogenetic tree showing how the outbreak clustered in different geographic regions
    - Results were comparable between SNV and wgMLST
    - wgMLST is done first, and SNP trees are done less frequently
- Interpretation guidelines were developed based on the data collected to date
    - It is unlikely that a single range will be able to consistently predict whether isolates will be related or not
    - Interpretation guidelines will be provided on a case by case basis

- Always need supporting epidemiological evidence!

Afternoon: Session 4 – Drugs & Thugs: NGS to Combat AMR

Session Keynote: Tracing Ancient Origins of Hospital Adaptation using NGS

Ashlee Earl
Group Leader, Bacterial Genomics, Infectious Disease & Microbiome Program, The Broad Institute of MIT & Harvard University, Cambridge, MA

- Enterococci are some of the leaders in AMR, especially recently
- They love hospitals and cause all sorts of bad things, but hey! They're in our guts too. Just chillin.
- Apparently they like freighting AMR around
- Lebreton, et al. Emergence of Epidemic Multidrug-Resistant *Enterococcus faecium* from Animal and Commensal Strains. http://mbio.asm.org/content/4/4/e00534-13
- Apparently the bad ones (hospital srains) weren't coming from humans, but from animals.
- Schloissnig, et al. Genomic variation landscape of the human gut microbiome. http://www.nature.com/nature/journal/v493/n7430/full/nature11711.html
- During evolution from animal strains they picked up a prophage that gave it the ability to live on mannose, increased its fitness-adaptation to living in the gut
- Two main problem children in hospitals: *E. faecium* and *E. faecalis*
- Q: where do enterococci come from?
- started with a bunc of carnobacterium pleistocenium, can grow anaerobly in cold
  - then vagococcus
    - then enterococci
- compared enterococci genomes, vary a lot in size, but are still monophyletic
  - Lebreton, et al. Tracing the Enterococci from Paleozoic Origins to the Hospital. http://www.cell.com/cell/pdf/S0092-8674(17)30478-6.pdf
- This part is very cool, IMO
- What makes an enterococci - they all share 1k genes
- took 100 of these and compared tthem to representative genes
- found a set that were genus unique
- examined 126 for function, reinforce cell wall, respond to cell wall stress and other stress, and undknown (aka tougher bastards)
- interim conclusion: think these come from an ancestor from mondern day carnobacterium, then to vagococcus (fish gut), then moved to enterococcus - land animal gut
- Next question: why do they then form new species?
- Looking at gene gain/loss - lebreton et al 2017 cell paper again
- enrichment of carbohydrate transport and metabolism genes
- Inference: As new host evolves, gut microbes adapt to new available nutrients, forming new species
- When did species diverge? How similar are their ecologies?
  - All vs. all pairwise comparison using ANI
  - found a gap between... what? probably carnobacteria and the vagococci
- Tried to date when this gap appeared
- Molecular clock approach
- think this one is part of the basis of what tthey did http://www.pnas.org/content/96/22/12638.full
- big extinction event, diversification started moving again
- everything really started 2 billion years ago with the first big freeze. carnobacteria were selected for.
- Then the cambrian explosion, oceans with amazing life in it. the vagococci found its niche at this time
- Tough for a microbe, basically swimming in shit, kind of easy, didn't have to deal with harsh conditions
- then: teresstrialization. Bacteria crawls onto land, lots of oxygen. birth of enterococci. Harsher conditions, had to develop more defense systems
- Dinosaurs: http://masahatto2.p2.bindsite.jp/
- "I would love an *Enterococcus* from a woolly mammoth."

Machine Learning-based Antimicrobial Resistance Prediction in PATRIC and RAST

James Davis
Argonne National Laboratory and the University of Chicago, Argonne, IL

- https://www.patricbrc.org
  - Currently contains ~116,003 bacterial and archaeal genomes
- A couple years ago they got pushed from NIH to start studying AMR
  - Started with *de novo* annotations for AMR-related proteins
- Antimicrobial Resistance Prediction in PATRIC and RAST - https://www.nature.com/articles/srep27930
- S/R prediction, not only ones who do this, but still
- Uses machine learning to predict S/R, put in S and R, create classifyer
- Consideratons
  - need info on S/R

need genomes (around 200 in most cases)
    - snapshotting the diversity of strains and mechanisms
    - careful of overfitting
    - cannot use on all AMs and species
- Advantages:
    - once trained, almost instantanious
    - good at finding snps in housekepping and itnergenic regions
    - no a priori knowledge needed
    - can reduce clinical response time
- Protocol
    - find nonredundant k-mers (15-31 bp k-mer size works well in their experience)
    - use counts or presence/absence
    - merge to form matrix
    - use an ml method to make prediction
    - current method
        - adaboost
    - other methods also work
- code available on github, tutorial.theseed.org
- have spent a lot of time curating genomes adding amr info to them, now have 15k isolates
- ftp://ftp.patricbrc.org/patric2/current_release/RELEASE_NOTES/PATRIC_genomes_AMR.txt
- Real goal, MIC prediction
    - Developing an *in silico* minimum inhibitory concentration panel test for *Klebsiella pneumoniae* - https://www.biorxiv.org/content/early/2017/09/25/193797
    - GitHub: https://github.com/PATRIC3/mic_prediction
    - around 90% accuracy for K. pneum
    - note, usually don't have any much of the isolates around the breakpoints
    - also, skewed datasets, because that's more interesting for hospitals
- Davis, et al. Antimicrobial Resistance Prediction in PATRIC and RAST. https://www.nature.com/articles/srep27930.pdf
- Antonopoulos, et al. PATRIC as a unique resource for studying antimicrobial resistance. https://academic.oup.com/bib/article-lookup/doi/10.1093/bib/bbx083
- Long, et al. Population Genomic Analysis of 1,777 Extended-Spectrum Beta-Lactamase-Producing *Klebsiella pneumoniae* Isolates, Houston, Texas: Unexpected Abundance of Clonal Group 307. http://mbio.asm.org/lookup/pmid?view=long&pmid=28512093
- Also collaborating with NARMS, salmonella
- 4600 isolates with mic data
- applied same protocol as for klebsiella to build a model for those
- more accurate than for the klebsiella stuff
- " for tool builders it's the metadata that matters"
- moving this into patric and rast shortly
- want to make patric a repository for models

Using 'Insider Information' to Identify Novel Antibacterial Targets

Laura Nolan
Imperial College London, London, UNITED KINGDOM

- Laura is open to new collaborators if anyone is interested in working with her! Her research makes for a very interested biological story.
- *Pseudomonas aeruginosa* = PA
    - On the WHO Priority 1 list of pathogens which require R&D for new Abs
- The PA Type VI secretion system (T6SS) is bacteriophage-like in its activity, delivering toxins into a host cell
    - T6SS is used as a weapon for interbacterial competition - an effective bacterial killing machine!!
        - Basler, et al. Tit-for-tat: type VI secretion system counterattack during bacterial cell-cell interactions. http://www.cell.com/cell/pdf/S0092-8674(13)00134-7.pdf
    - Highly active within biofilms
- Bacteria in biofilms are very resistant to antibiotics and immune system factors
- T6SS toxin targets are validated antibacterial targets - lots of things to exploit here for novel antibiotic targets
- Three distinct clusters of genes are underlying T6SS in PA
  One of these gene clusters is H1, and this is the target of the study described here
  Used transposon directed insertion-site sequencing (TraDIS) to identify novel H1-T6SS effectors
  A TraDIS Tn mutant library was generated and sequenced on the MiSeq
- Identified known and novel immune variants and confirmed them in subsequent experiments
    - Tse8 is a novel H1-T6SS associated toxin
      Tsi8 is the cognate immunity for Tse8
    - Tse8 is delivered by the VgrG1a/1c tip complex
      Structural predictions showed that Tse8 targets the GatABC transamidosome complex
- Bacteria have alternative amino acid synthesis pathways
- Hypothesis: Tse8 replaces GatA within the transamidosome complex and is inactive
    - Tse8 does not have the same substrate as GatA
    - Asparagine tRNA synthase can rescue Tse8 toxicity

- Selective toxicity of Tse8
  - Agrobacterium tumefaciens requires GatABC transamidosome
  - Escherichia coli doesn't require GatABC transamidosome
- Tse8 is not only a new T6SS toxin, it also acts in a completely novel manner to all other T6SS toxins
- How widespread would the effectiveness of this be as a therapeutic?
  - <<insert bioinformatics analyses here>> ... Result: the transamidosome is a selective antibacterial target!
- TraDIS cna be utilized to uncover the full repertoire of all T6SS toxins
- They are speaking with pharmaceutical companies to perform high throughput screens and nail down what the specific interaction is, but it's in the early stages.
- Another research question - which bacteria have Tse8? Most of the *Pseudomonas* have it. It was probably around with the dinosaurs.

Encoding the Efflux Pump Phenomena
Kara Tsang
McMaster University, Hamilton, ON, CANADA

- http://card.mcmaster.ca
- 3 mechanisms for amr development, target modification, inactivation and efflux
- 5 major superfamilies of pumps, both mobile and chromosomally encoded,
- drug specific or multi resistant, ne or multi component
- many studies on them, but no database, no detection tools, thus no phenotype prediction
- Goal: Detect AMR phenotype from genotype
- solution to db problem: card db, ontology driven
- YAY ontologies!
- have encoded efflux pumps in the ontology
- created detection models for each specific pump, decided to aggregate them into a higher level model
- test case: decided on ecoli, added new genes with annotation
- AcrAB-TolC a major efflux pump for many AB (when overexpressed)
- created Resistance Gene Identifier, three groups, perfect match, strict, loose
- have seen that without mutations pumps are normally expressed, with mutations more expressed
- lots of ecoli and Pseudomonas aeruginosa, compared predicted to antibiograms
- pumps are ubuquitous
- over, under and perfect pediction
- pseudomonas: Complete tetracycline resistance prediction, but over/underprediction of several other ab
- ecoli: minimal over/underprediction for genwtamicin and trim-sulpha, heavy in others
- A common platform for antibiotic dereplication and adjuvant discovery: http://www.cell.com/cms/attachment/2074523756/2069028571/mmc3.pdf

Wednesday, October 11

Morning: Session 5 – Pipe Dreams: Analytical Methods, Bioinformatics Tools, and Pipelines

Session Keynote: Rapid End-to-End Workflows for Hypothesis-Free, NGS-Based Pathogen Detection in a Diagnostic Laboratory
Robert Schlaberg
Medical Director, Microbial Amplified Detection, Virology, and Fecal Chemistry Laboratories, and Assistant Medical Director of the Molecular Infectious Disease Laboratories at ARUP; Assistant Professor of Pathology, University of Utah School of Medicine, Salt Lake City, UT
S

- ARUP labs recently launched an NGS dx text - Robert is going to tell the story of what went down
- Current tests don't solve all the dx problems out there - we usually run panels of 10+ tests but often don't get a result (more than half the time!) and the docs end up treating empirically
- NGS: hypothesis-free testing and you can get some host information at too, but it's in the POC stage and implementation is challenging
- Taxonomer plots showing some patient results, e.g. WNV infection causing encephalitis in an immunosuppressed patient, ZIKV in fatal case - http://www.nejm.org/doi/full/10.1056/NEJMc1610613#t=article
- Plot antiviral resistance and viral load with NGS - 3yo immunocompromised child, saw increasing viral load and resistance mutation emerging over four samples/7 days
- For every nice published case, there are loads of unsuccessful ones behind the scene - we have to move from case reports to taking a larger look at the method
- Respiratory infections: NGS >90% agreement with multiplex PCR panel targeting 8 viruses
- Kids with pneumonia of unknown etiology, found a range of pathogens in 30% of kids - these had previously been missed - you will find pathogens and commensals (interpretative challenges), but you can improve the yield over conventional testing
- Take the expert out of the process - standardize and automate workflows
- Considerations before making the NGS leap: interpretation is challenging (commensals, novel pathogens, what are you comfortable giving the doc), samples are highly variable, often NGS is a "test of last resort", balance the issues of generating confident results with a promising new technology
  - You have to decide what "not working" means for a given sample
- Accurate, validated, 24-48 TAT workflow on NextSeq:
  - Patient sample, positive control, and negative control loaded (positive - would like it to be same analyte but this isn't possible, negative - should be same matrix)
  - Spike 2 bacteria/2 viruses/2 fungi into specimens - process controls, helps flag samples that may have too much host DNA

Extraction, both RNA and DNA libraries, sequence, review, report
    - Many steps, many vendors - need to QC all the reagents every time you get a new shipment, new lot, many steps involve multiple kits, some kits have reagents from different lots. What a mess!
- Search tools, classification engines, and pathogen detection bioinformatics approaches
- Search tools: comprehensive result summaries, but match tables require interpretation, performance depends on reads/taxonomic group - example from Strep pneumo - six species come up in the hit list
- Classification engines (Kraken, SURPI, CLARK, Taxonomer, etc...): fast, scalable, open source, can be customized, classify at read level, but performance depends on databases, still needs interpretation, challenging to implement in routine clinical practice
- Sequential subtraction used by many tools - map to human and discard. Speed depends on sample decomposition, misclassification can occur (using partial databases at many steps, heterogenous query sequences)
    - Taxonomer - ultra-fast hypothesis-free testing and discovery
        - Binning step - reads compared to customizable database - every read compared to all databases
        - Classification step - read assigned to a single taxon using databases tailored to each bin
- If you don't know what you're looking for, it's difficult to define the thresholds for what a good match is.
- Pathogen detection: you have to go from the list of "hey, here are all the things" to "this is what we're confident in and is meaningful to the clinician" - diagnostic engine sits on top of Taxonomer and generates a clinical report (this requires a lot of validation and tuning)
- Validation: need positive and negative specimens (known results), add external and internal controls, QC sequencing and samples, assess different databases, algorithms, cutoff values
- Helpful validation strategy: "virtual specimens" - real sequencing data from pathogen-negative samples (equivalent of pooling residual specimens, a common lab QC approach), add sequences from pathogens, pathogen near neighbours, commensals, and contaminants
- A first validated test! Challenge/solution: rapid TAT/optimized workflow (<48h), range of bugs/sequence RNA and DNA, variable sample quality/internal controls to ensure sample adequacy, complex workflows/workflow management and process controls, variable data quality/validated QC criteria, contaminated reagent/matrix-matched negative control, accurate results/curated databases and validated cutoffs, actionable results/list of respiratory pathogens, semi-quantitative result, expert result review
- Workflow manager for wet banch work: scan barcoded reagents and sample, outputs protocols and barcode labels for downstream steps
- Sequencing metrics: run yield, cluster density, library size, read count, %PF, bases >q30, phiX error
- Timing: 8-12h sample and library prep, 15h sequencing, <1h for downstream steps
- BAL from immunocompromised kids with pneumonia in ICU, prior negative results on all tests, found pathogens in 40% of cases - around 70% of patients had one pathogen, around 1/3 had 2 pathogens
- Remaining challenges: curated, comprehensive, balanced databases - need to update these often, re-validate, include version control
- Wish list: automated sample/library prep (reagents from same source and lot, closed system, DNA/RNA-free), rapid sequencing with flexible batch size (consistently performing and stable reagents with few lot changes), rapid automated data analysis (validated databases, algorithms, and cutoffs), and go beyond yes/no to quantification, resistance, epi typing, and host-based diagnostics
- Q: Large variation of samples, and therefore need for customized workflow, so how did you address this? First validated for BAL samples. Then customize worflows for other samples, with adaptation in the preparation steps.

Pathogen Detection and Resistome Characterization by the Application of Next Generation Sequencing and Bioinformatics

Rita Colwell
University of Maryland, College Park, MD

- Start point: human microbiome project, 16S data gave them a good idea of the different flora on humans
- Starting to discover the roles that all of the microbes play - involved in many diseases, like parkinsons, cycstic fibrosis etc
- Can undertand human disease by understanding the microbiome
- Wrote her first program in the 60s to identify phenotypes, on the IMB650 (mind boggles!)
- Nowadays analysis is much more elegant
- Will be using the V. cholera as an example
- Have a progression from the environmental strains to the infectious disease strains, can see it in the core genome
- facinating organism, environmental but also kills plenty of us
- Have genome re-assortment, onot serotype, that define an epidemic clone
- interesting, because the O1 that we make vaccines against, might be laterally transferred
- Now moving from classical tests to NGS based tests
- Have an AMR problem here too
- 2007: wanted to create automated detection mechanisms based on NGS
- want an universal method, without culturing (I think?)
- spent approx 10 years building a highly curated database for various bacteria, viruses etc, 65k genomes
- How it works:
    - biological specimen
    - commuity dna
    - dna sequencing
    - reads
    - genbook database
    - genobook biomarker matching
    - identified micromes
    - genboookk AR/VF library
    - leads to microbial identification adn pathogen info
- important to get to subspecies/strain level, strain really matters
  another example: ecoli, both good and bad
  have to pay attention to AMR
- concept of community resistome suggeste by wright, 2007
  http://www.nature.com/nrmicro/journal/v5/n3/full/nrmicro1614.html
  collaborating with CONSERVE project, to understand waste water reuse

- analysis of various water samples, characterize the presence of AMR
- with AM treatment you loose some species (less diversity) and gain someothers
- study on cholera patients coming in in calcutta
- they used their standard methods, sent samples to them for further analysis
- never just a single pathogen
- amr common, also in non-patients
- woman in somewhere that got an Aeromonas infection in her leg, spread to liver and pancrreas,
- mixed infection, but only one bacteria spread to other areas of her body
- strains were working together to create disease
- recycled drinking water, took samples from the entire process to show safety
- Microbial quality across the water manufacturing process: three steps
- 1. Microfiltration
- 2. Reverse sosmosis + H202
- 3. Ultraviolet irradiation then decarbonation, lime
- Stepwise reduction of bacterial genera along the process, to end up with normal microbial composition of “clean water”, occasionally opportunists such as Klebsiella are found.
- metagenomics showed a lot of microbes before processing, but after there's nothing in it, i.e. clean
- amr genes are present before processing, very few remain after

ProxiMeta: a Hi-C-based Metagenome Assembly Platform and Microbial Discovery Platform

Ivan Liachko
Phase Genomics, Seattle, WA

- Shows THAT CURVE but points out we're nearing the point whereIllumina will pay US to sequence genomes :)
- Shotgun metagenomics is cool and all, but there are still issues in sorting out what DNA belongs to what donor
- Hi-C: genomes are actually blobby, spherical 3D structures. Hi-C captures this: take your cell, cross-link the chromosomes - links 3D-proximal sequences, fragment, proximity ligation to seal the ends, then sequence the junctions
- You can scaffold genomes with Hi-C and get sweeeeeet assemblies
- Goat genome! Papadum the Goat! Had 3k contigs, reduced to 1600 with optical mapping, got it down to 31 goat chromosomes with HiC
- Not just good for goats, but cancer tumours too!
- Doesn't need hmw DNA and is low-input friendly (except for lobsters and weed, in which case they say send lots)
- Sequences interacting from the same cells interact on Hi-C
- Metagenomics: can tell which contigs came from which species, connectes plasmids to hosts, connects eukaryotic microbes with multiple chromosomes
- https://www.ncbi.nlm.nih.gov/pubmed/24855317
- You can get reference-quality, scaffolded, awesome genome assemblies from mixed microbial populations
- Can quickly determine how many centromeres, and therefore how many chromosomes, an organism has, even if you've never seen it before
- Assembled a hybrid yeast from a metagenome - got it from one of Seattle's hipster breweries that uses open-top fermentors. Saw a yeast with a genome 2x the size of usual, half looked like a known yeast, half was mystery. Hi-C enabled the detection of this cool hybridization event.
- Other neat application: assemble infected/contaminated genomes - e.g plant that was riddled with fungus - went from 13k scaffolds to 11 chromosomes after proximity-guided assembly, enabled the separation of the fungal DNA
- Unculturable organisms from bacterial vaginosis - caused by dysbiosis rather than a single pathogen - method even enabled the resolution of unique strains of a single pathogen
- Nice reports that give you completeness, contamination, and novelty of each genome cluster detected
- They just Hi-C'd a poop, got 50 genomes out including some novel stuff
- Cattle rumen - LOADS of novel genomes - 63 high-quality genomes from one run (

Machine-Learning Based Identification of Pathogenic Species in Next Generation Sequencing Experiments

Carlus Deneke
Federal Institute for Risk Assessment, Berlin, GERMANY

- Metagenomics has some issues, like mapping failing due to poor quality, ambiguity, wrong reference DB and oh yeah, sometimes the organism is new and isn't in the reference database! What's a bioinformatician to do?
- Can we predict pathogens in metagenomic data? Even if they're novel? Notes that "pathogen" is a bit of a messy definition - they are more interested in genomes with "pathogenic potential"
- Supervised machine-learning approaches, per-read predictions -> species predictions (majority rule: average prediction of all reads)
- 2836 bacterial strains from 422 species, labels (human pathogen, human-associated non-pathogen) from JGI's IMG
- Used one random strain per species and did 5-fold cross-validation
- Used 948 features for classification: DNA-based (kmers), protein-based (mono/dimer frequencies, amino acid property frequencies)
- Random forest: 61% pathogens, 39% non-pathogens
- Method accuracy is robust to closely-related genomes in the reference database
- Pathogenicity prediction for bacterial genomes (paprbag) software on GitHub: https://github.com/crarlus/paprbag
- PaPrBaG paper: https://www.nature.com/articles/srep39194
- Extensions: explore a specific taxonomic niche, like just the E.coli species, use non-genomic features like cell morphology, try deep learning to learn features automatically, predict other phenotypes like resistance or host range

NCBI Pathogen Detection: Facilitating Traceback and Outbreak Investigation of Foodborne Pathogen Genome Sequences in Real-time using Automated wgMLST and SNP Analysis Pipelines

William Klimke
NCBI/NLM/NIH/HHS, Bethesda, MD

- Contact Bill Klimke <klimke@ncbi.nlm.nih.gov> or Lee Katz <gzu2@cdc.gov> if you have data that you'd like to donate for their pipelines
- Bill can also help you setup a BioProject for your data on NCBI
- Have come along quite a bit since Lipman spoke here 2 years ago
- FDA approached them to do salmonella, was also asked to do listeria
- now have a collaboration with several agencies for pathogen reception, then analyses it in their pipeline
- want to have a 24hr turnaround data
- turn it into actionable data
- Questions:
- are they related? - snp pipeline, clustered things less than 50 SNPs apart, salmonella broke apart all of their systems
- pipeline: kmer analysis, assembly, annotation, genome placement, clustering, snp analysis, tree construction
- want to replace parts with skesa assembly and wgmlst, nearest neihbors, cluster on wgmlst
- skesa: strategic kmer extension for scrupulous assemblies
- breaks at repeats, great contigs, but may not be very long
- wgmlst complementary to snp analysis
- linear relationship between snps and wgmlst, good agreement
- skesa and wgmlst allele call is very fast
- about wgmlst clustering
- bteween genomes: only interested loci with identical alleles, and those with different alleles
- something about clustering that I didn't catch....
- can then cluster genomes on wgmlst
- NCBI Pathogen Detection Website: https://www.ncbi.nlm.nih.gov/pathogens/
-
-
-
-

Software

- Mashtree https://github.com/lskatz/mashtree (Poster 27, generating WGS trees)
- Omniscope https://github.com/mjmiossec/omniscope (poster 16, ID pipeline w/ reference+denovo assembly)
- Benchmark datasets https://peerj.com/articles/3893/ (both for datasets and for a script that downloads datasets)

Oh