Bacterial genomes: haploid, 0.1% as big as the human genome, 10% as many genes as we do.

Great paper: <http://www.omicsonline.org/biodefense-oriented-genomic-based-pathogen-classification-systems-challenges-and-opportunities-2157-2526.1000113.pdf>

Cost to generate a draft genome using short read technologies is many times lesser than the costs associated with long read 3rd gen technologies like PacBio.

PacBio can sequence long sequence reads of up to 30 kbp. However, there is a dearth of bioinformatics tools designed for long reads with a high error rate.

**Hybrid technologies for generating contiguous assemblies (Koren et al, 2012)**4. Combines long read data with short, high-accuracy sequences to correct the error inherent in long, single-molecule sequences. Improves read accuracy from as low as 80% to over 99.9%.

* De novo assembly will also overcome the potential risk of missing any synthetic sequences if we map to a reference genome. Note that denovo assembly is definitely much more computationally intensive (i.e. bigger computational resources).
* Increase in N50 (see citation 4, Figure 3)
* Long reads are capable of producing better assemblies, even at greatly reduced coverage.
* Paper (4) finds PacBio CCS long read sequences to be the best combination with Illumina short-range paired data in their Celera adaptation. (did comparison with Illumina, and 454). PBcR algorithm. *Illumina short-range libraries double the sequencing time, and long range libraries are difficult to construct. So long, single molecule sequencing si a practical alternative to both.*
* CCS yields 1-2 kbp reads. Applications in shotgun metagenomics, 16S rRNA

De novo assembly of bacterial genome using long read data (PacBio or Nanopore) and Celera assembler 3

**Dna repeat sequences:** Terminal inverted repeats. CRISPRs. Role in regulatory systems.

Recent paper on hybrid approaches: <http://www.nature.com/articles/srep08747>

Compares PBcR pipeline with other assemblers.

An ideal system would screen DNA fragments down to 50 nt lengths from which genes are made (but this means accidental match chances increase, as gene length declines).

*Current screening practices fail to identify roughly 15% of all orders (which already include a Genbank search).*

**Industrial code of conduct**: Companies should screen sequences against the Australia Group, Select Agent, and national export lists at both DNA and protein levels (include all known genomic and plasmid transcripts).

the technical hurdles for *de novo* synthesis of Select Agents and Toxins from double-stranded DNA are much lower than for *de novo* synthesis of these agents from single-stranded oligonucleotides. ***2***

InterproScan: annotate entire bacterial genome. Computationally very expensive.

Prokka:

MAKER: part of GMOD project.

**Genes that code for surface proteins rather than enzymes are not considered virulent because they cannot harm the host.** They might be considered useful in stabilizing pathogens against environmental stresses and evading the human immune system though. **Genes may also confer virulence only in the presence of other genes and/or certain biological contexts, so it may be impossible to know when isolated genes pose a threat.** ***1***

Research has shown that many virulence factors are located within ‘pathogenicity islands’. Adding sequence location data to current homology scores could hence dramatically improve virulence prediction. ***1***

*Examine predicted proteins against an HMM****1***

2008: SARS virus recreated using synthetic biology

2005: 1918 flu virus reconstructed

1. https://gspp.berkeley.edu/assets/uploads/page/Maurer\_IASB\_Screening.pdf

2. <https://www.gpo.gov/fdsys/pkg/FR-2009-11-27/pdf/E9-28328.pdf>

3. <http://www.nature.com.ezproxy.library.ubc.ca/nmeth/journal/v12/n8/full/nmeth.3444.html>

4. <http://www.nature.com.ezproxy.library.ubc.ca/nbt/journal/v30/n7/full/nbt.2280.html>

**ALLPATHS-LG (2010)**

* Requires a minimum of two libraries (short, and jumping)
* Not tested with long reads
* Review shows accuracy but gives gapped assemblies in the absence of long reads
* Requires customized implementation to get complete accurate bacterial genomes
* *Sees multiple paths through the data, making possible more accurate assemblies. Applied to mammalian genomes (large, repeat-rich)*

**PacBio**

* Read depth, specification of genome size, Celera pramaters
* Expected genome size should be specified in long read correction, 25X longest PBcR should be used for assembly, and contigs fewer than 100 mapped PacBio corrected reads should be discarded

**SPAdes**

* single cell assemblies
* adjusts errors iteratively; 1 hour for assembly (for E.coli, needs max 8 gb displace, 5 gb ram)

**SSPACE-LongRead**

- required pre-assembled contigs (will have to combine with one of the previous)