Is it practical to implement NGS technologies in NCPDN labs for virus detection

Dimitre Mollov
USDA ARS National Germplasm Resources Laboratory

What is NGS and how does it work?

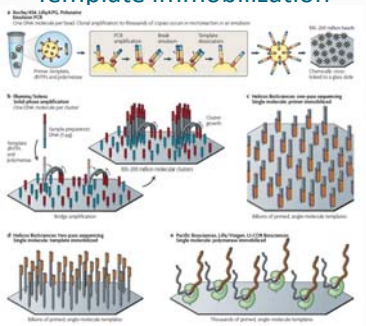
Next Generation Sequencing Technology

- Routine detection of viruses and other pathogens
- Extremely sensitive
- Highly reliable
- Identify unknown or poorly characterized viruses
- Whole genomes sequencing (most feasible approach)
  - Pathogen diversity; mutations; new strains
  - Ability to improve detection techniques

Template immobilization

NGS platforms

- 454
- Illumina
- SOLiD
- Polonator
- Helicos



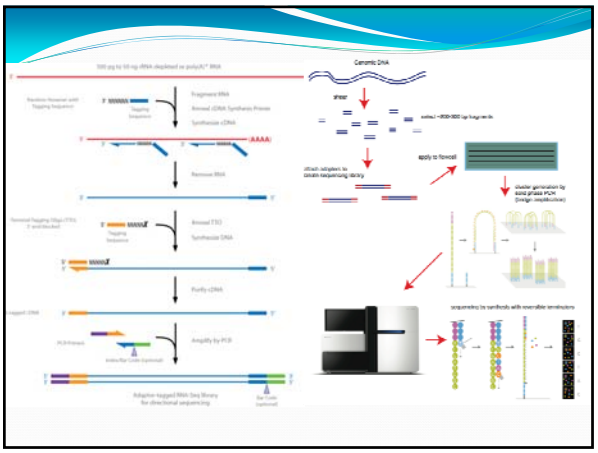
Metzker, 2010

NGS target

- Total RNA (or mRNA) or DNA
- siRNA
- dsRNA
- Virion Nucleic Acid

Materials and methods

RNA Extraction



NGS data

- *De novo* assembly
- NCBI database
  - Nucleotide
  - Protein
- E-probes
  - EDNA
  - Custom database
    - Nucleotide
    - Protein
- Mapping
  - Raw reads
  - Contigs

Materials and methods

Data analysis

Lab conformation

Workflow

Sequence reads	Millions
Contig assembly	CLC Workbench
BLAST	NCBI Local database; BLASTN, BLASTX
Viral contigs	Computer vs. biology
Lab validation	PCR, Sequence
Map to reads	Final assembly

Sequence reads

25-50 million reads per sample,

~5-10% removed due to low quality

Length of reads: ~100 (<50-600)

Contig assembly

<100,000 to >750,000 contigs per sample
Why so many?

Not all are useful: 500-10,000 nt long => 10-20,000 contigs

NCBI Local database; BLASTN, BLASTX

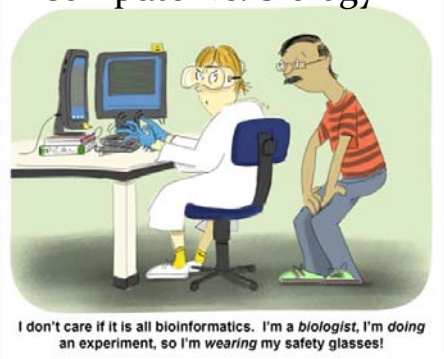
I appreciate your dedication, but genome browsers just aren't really designed for phones.

Sequence reads and contigs

Virus related contigs: 0 to >100
Plant virus related contigs

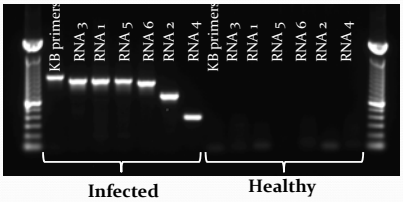
How about uncharacterized viruses

Computer vs. biology

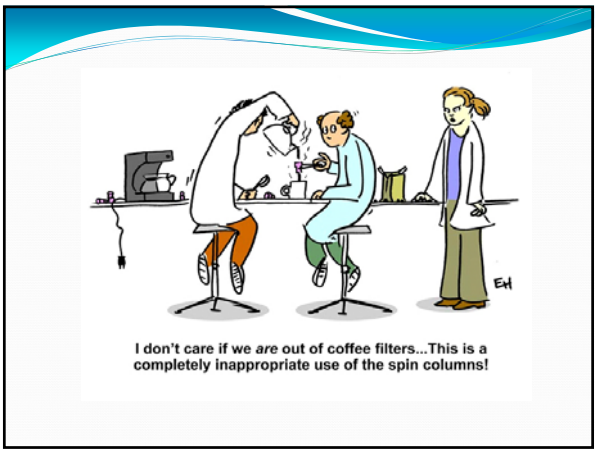


I don't care if it is all bioinformatics. I'm a *biologist*, I'm *doing* an experiment, so I'm *wearing* my safety glasses!

Lab validation

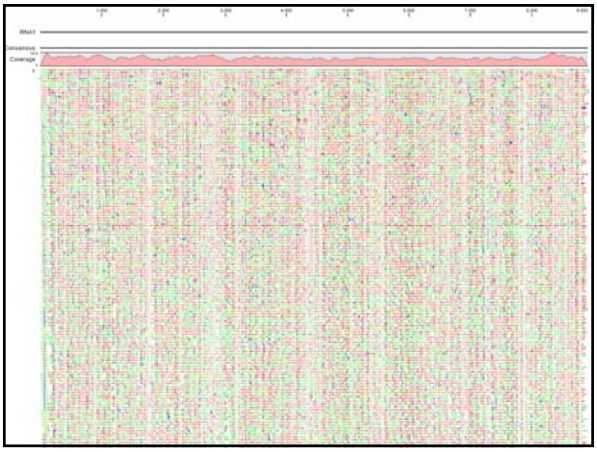


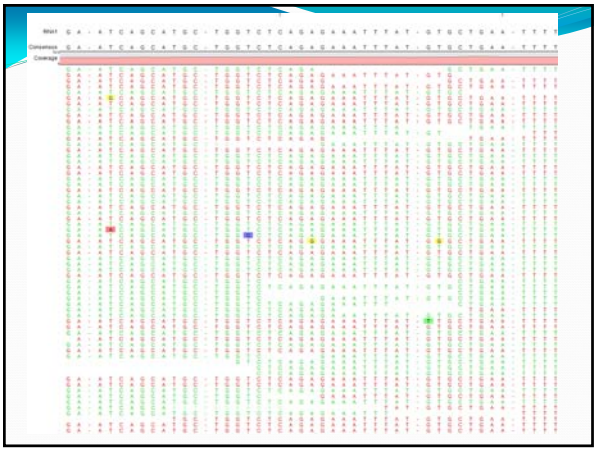
PCR, Sequence



Sequence reads mapping

	Sample 1		Sample 2		Healthy control
	number	percent	number	percent	number
RNA1	373,519	0.96	340,701	0.91	26
RNA2	952,707	2.44	870,223	2.31	66
RNA3	309,844	0.79	285,005	0.76	11
RNA4	767,414	1.97	681,795	1.81	16
RNA5	455,429	1.17	420,168	1.12	26
RNA6	2,306,848	5.92	2,103,686	5.59	109

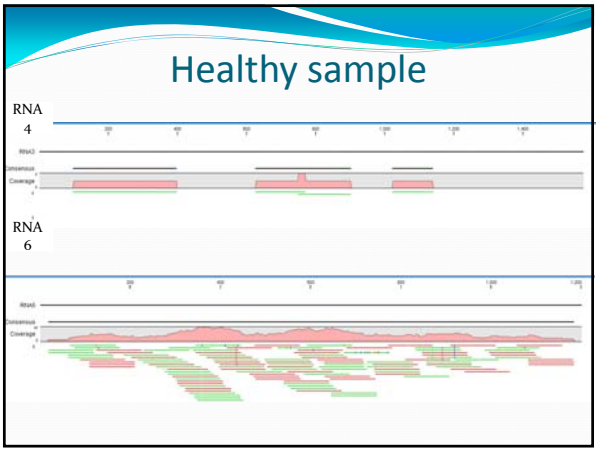




Healthy sample

RNA 4

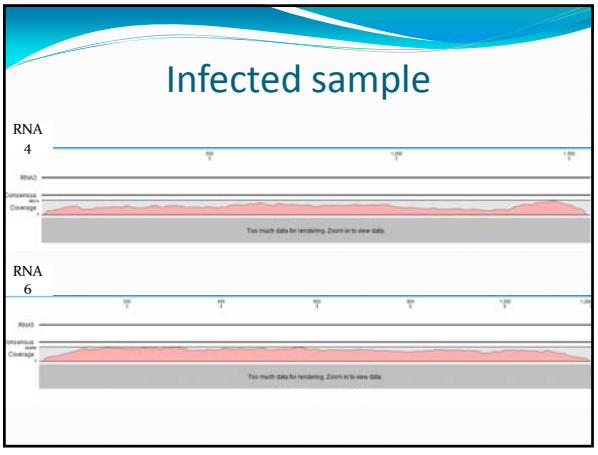
RNA 6



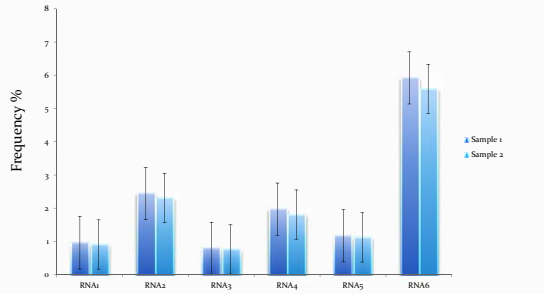
Infected sample

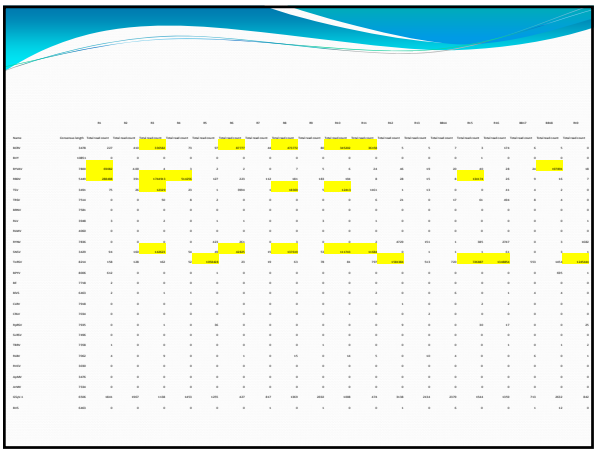
RNA 4

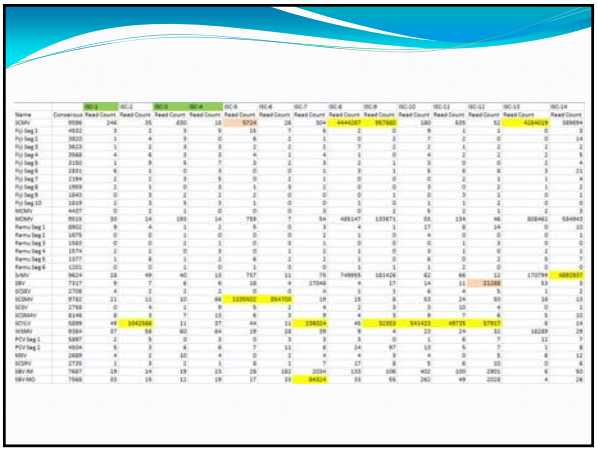
RNA 6



RNAs sequence reads map frequency







Do I see NCPDN using NGS?

- RNA extraction
- DNase treatment ~$50
- rRNA deplition ~$100
- Library preparation ~$100-200
- Sequence run ~200
- Total cost ~$525-650 per sample
- Bioinformatics support
- Lab validation – PCR and Sequence

Next Generation Sequencing Technology: Why do it?

- Routine detection of viruses and other pathogens
- Extremely sensitive
- Highly reliable