Approaches for Revealing Virus and Phage Communities in Healthy and Diseased Individuals

Joseph Petrosino

Director, Center for Metagenomics and Microbiome Research
Department of Molecular Virology and Microbiology
Human Genome Sequencing Center
Baylor College of Medicine

International Human Microbiome Congress
March 10, 2011

Alkek Center for Metagenomics and Microbiome Research

HUMAN GENOME SEQUENCING CENTER

Who we are…some context

- HMP clinical sampling and nucleic acid extraction center
  - Responsible for half the HMP samples (other half WashU)

- HMP sequencing center (NHGRI Large Scale Sequencing Center, BCM-HGSC)
  - Metagenomic samples and reference genomes (poster 122)
  - Analysis

- Numerous metagenomic studies in mice, primates and man ongoing with collaborations in Texas Med Center, in U.S., and abroad (129, 156, 171)

- BCM Alkek Center for Metagenomics and Microbiome Research (CMMR) recently formed
  - provide resources to drive microbiome research and collaboration
  - JOBS AVAILABLE (Tenure track faculty, informatics, project managers)

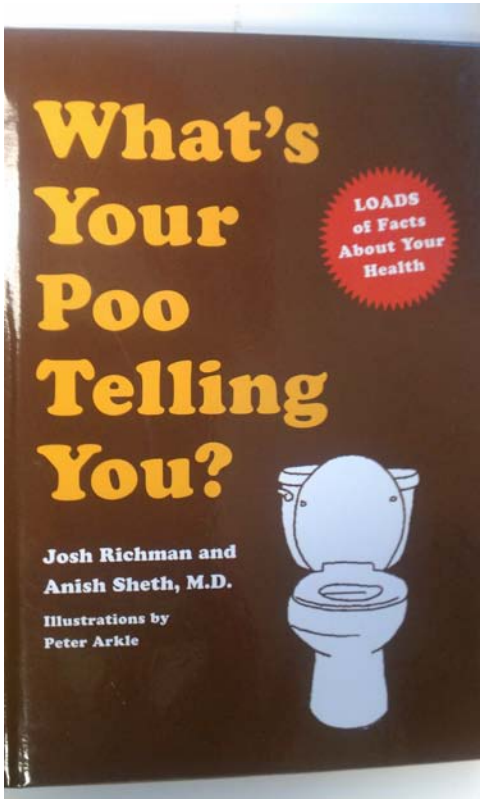
Alkek Center for Metagenomics and Microbiome Research

Viral Metagenomics

- Goal: Detect intact viruses in clinical samples to discover relationships to health and disease (and etiologic agent candidates)
  - Develop/validate methods for low yield samples
    - Nasal Washes (non-HMP)
    - Stool samples
    - Vaginal samples
  - Relate viral/phage data to bacterial data and subject metadata

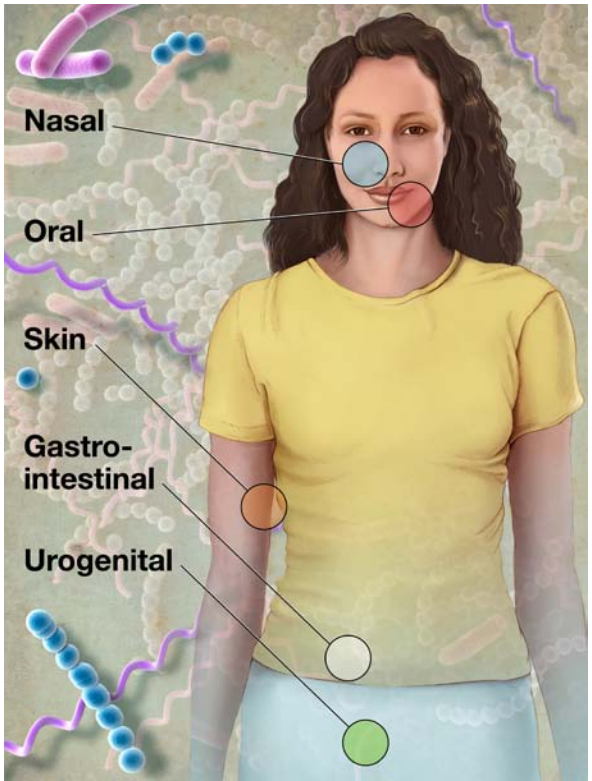
Overview

- Viral Metagenomics
  - Challenges
  - Technical development/strategies employed
  - Initials results/advances made
- Viral metagenomic applications
  - “Virus Hunting”
  - Genome sequencing of uncultivable viruses



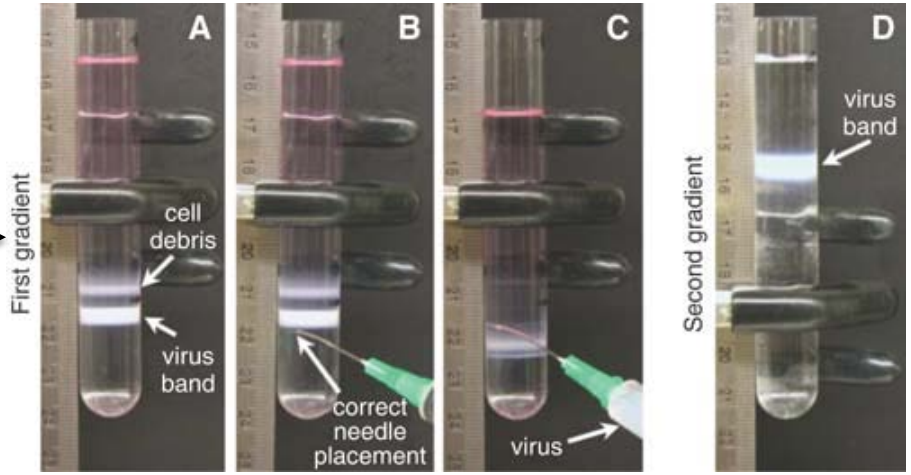
HMP sample sources

- Oral Cavity
  - Saliva, Tongue, Hard Palate, Buccal Mucosa, Keritinized Gingivae, Tonsils, Throat, Supragingival Plaque, Subgingival Plaque
- Skin/Nasal
  - Retroauricular Crease (L,R)
  - Antecubital Fossa (L,R)
  - Anterior Nares
- Vagina
  - Vaginal Introitus
  - Mid-Vagina
  - Posterior Fornix
- GI Tract
  - Stool



The Challenge

Classically…..



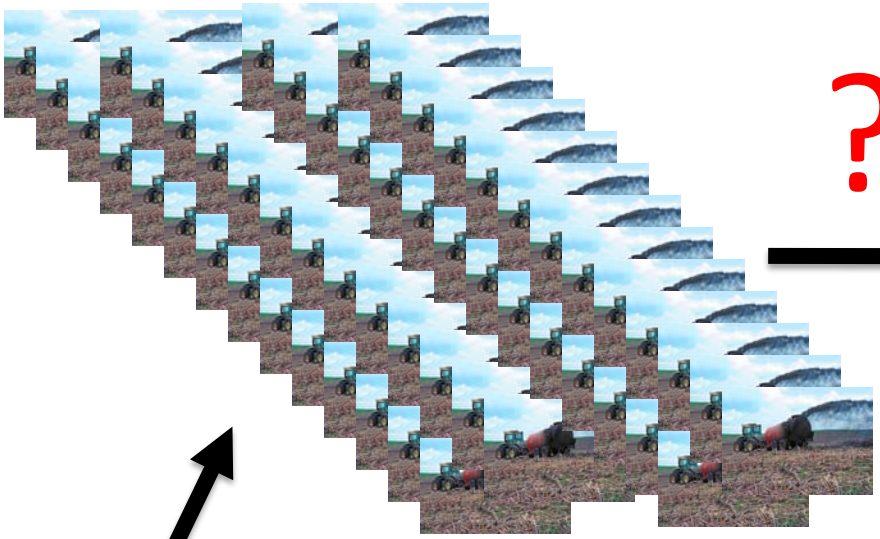
http://www.currentprotocols.com/protocol/ns0423

http://extension.usu.edu/waterquality/htm/agriculturewq/manuresolutions/

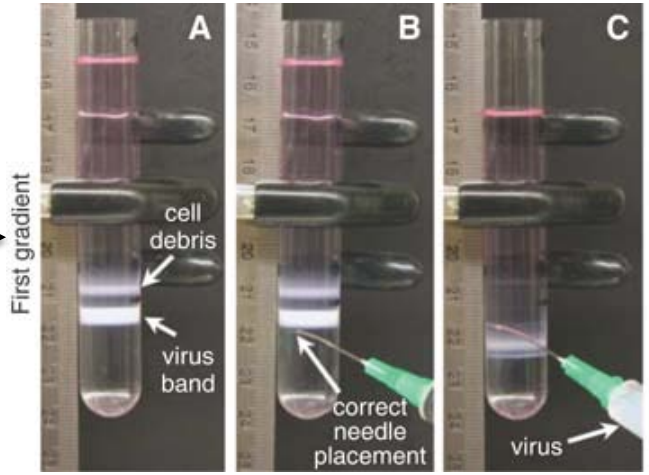
….Few high volume/titer samples are processed per study

The Challenge

BUT NOW…..

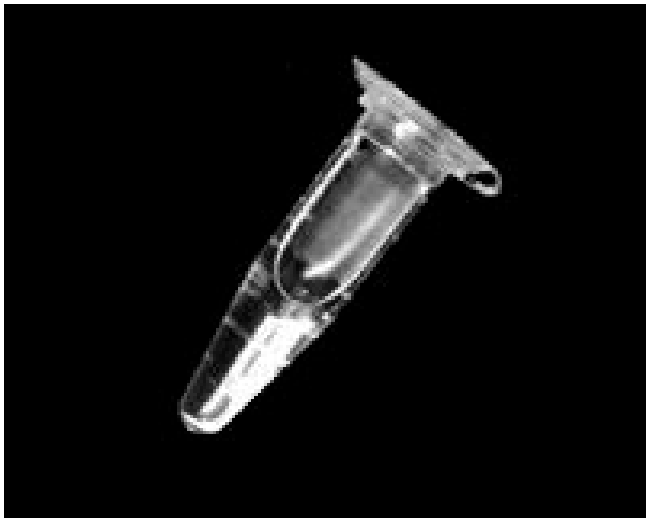


?



The Challenge

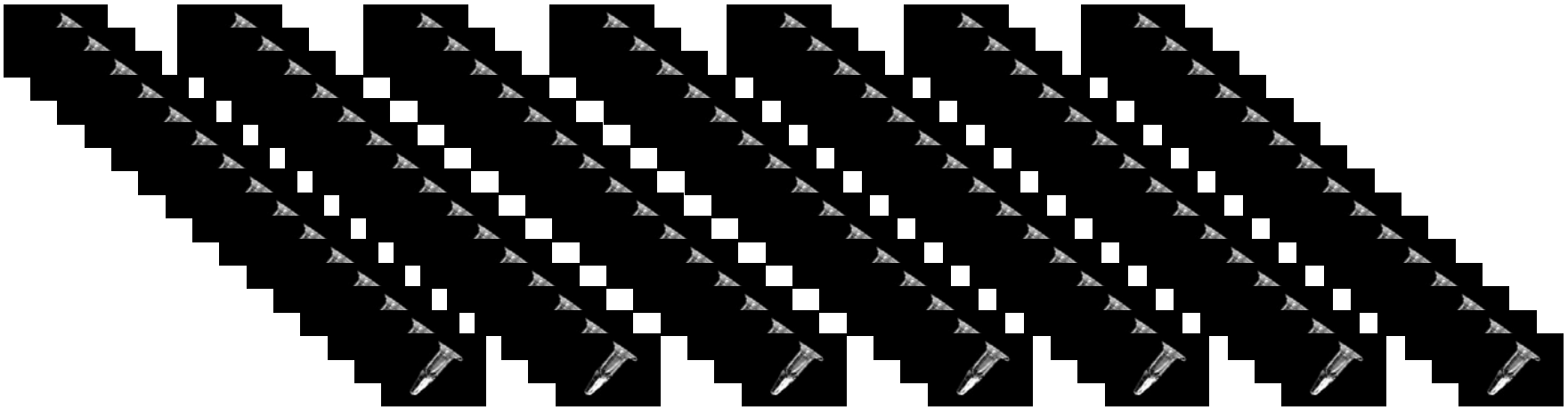
http://www.phillybroadcaster.com/craigslist-philly-homeless-horse-manure/

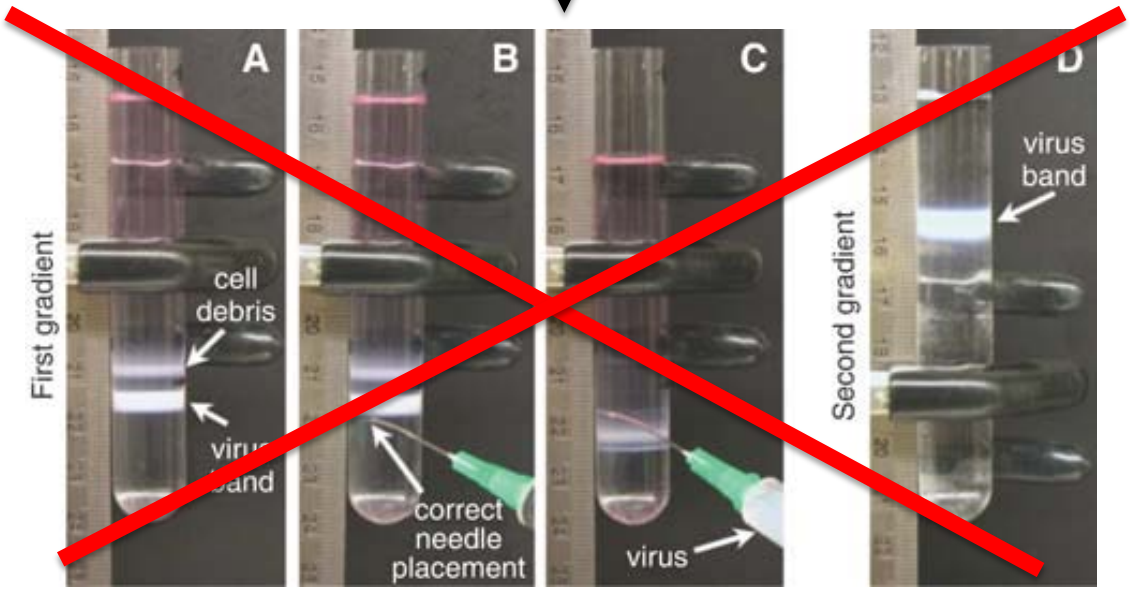


http://www.vmri.hu/fishparasitology/links_en.html

Collecting and processing high volume clinical samples (esp. non-stool) results in lower biomass to work with

The Challenge





Low yield sample considerations

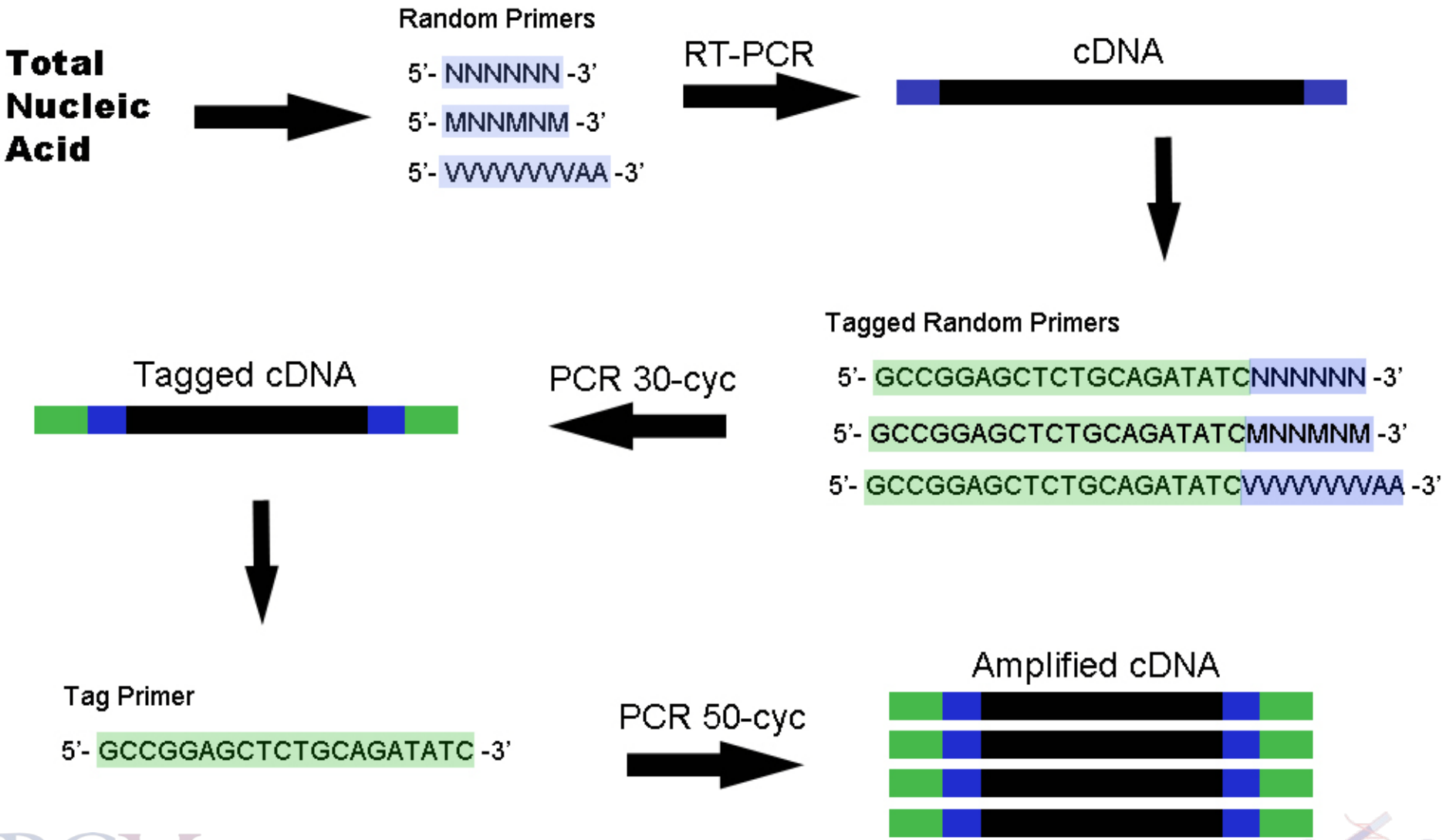
- Less is more
  - More handling = more sample loss
  - Trade off higher background for greater virus retention

- Amplification is often necessary
  - Need ~10ng for HiSeq library construction
  - Viral quantification impaired with random amplification
  - Need an awareness of how random amplification impacts virus detection
  - Data analysis not refined when looking for “dark matter”
    - Trade off looser stringency for more viral hits

Viral nucleic acid prep…(at most)

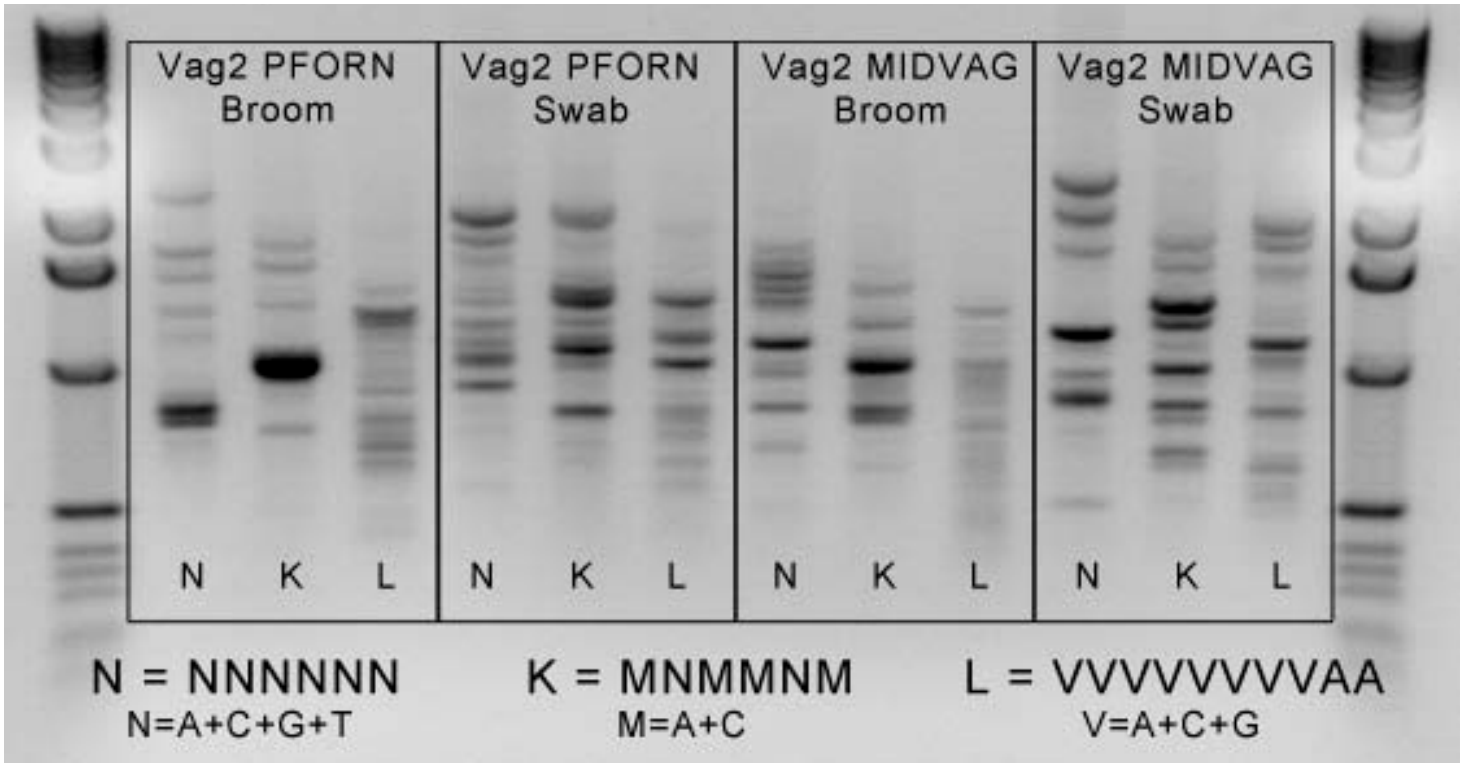
- Enrich clinical samples for virions/VLPs before nucleic acid is extracted
  - Centrifuge and pre-filter (100 micron) to remove large debris and cells.
  - Filter at 0.45 micron to remove cells and aggregates
  - Concentrate via filter centrifugation (100 kD cutoff)
- After concentration...
  - Treat with DNAse/RNAse to remove unprotected NA
  - Extract total nucleic acid
  - Split sample; generate cDNA libraries for sequencing

DNA/cDNA Library Construction

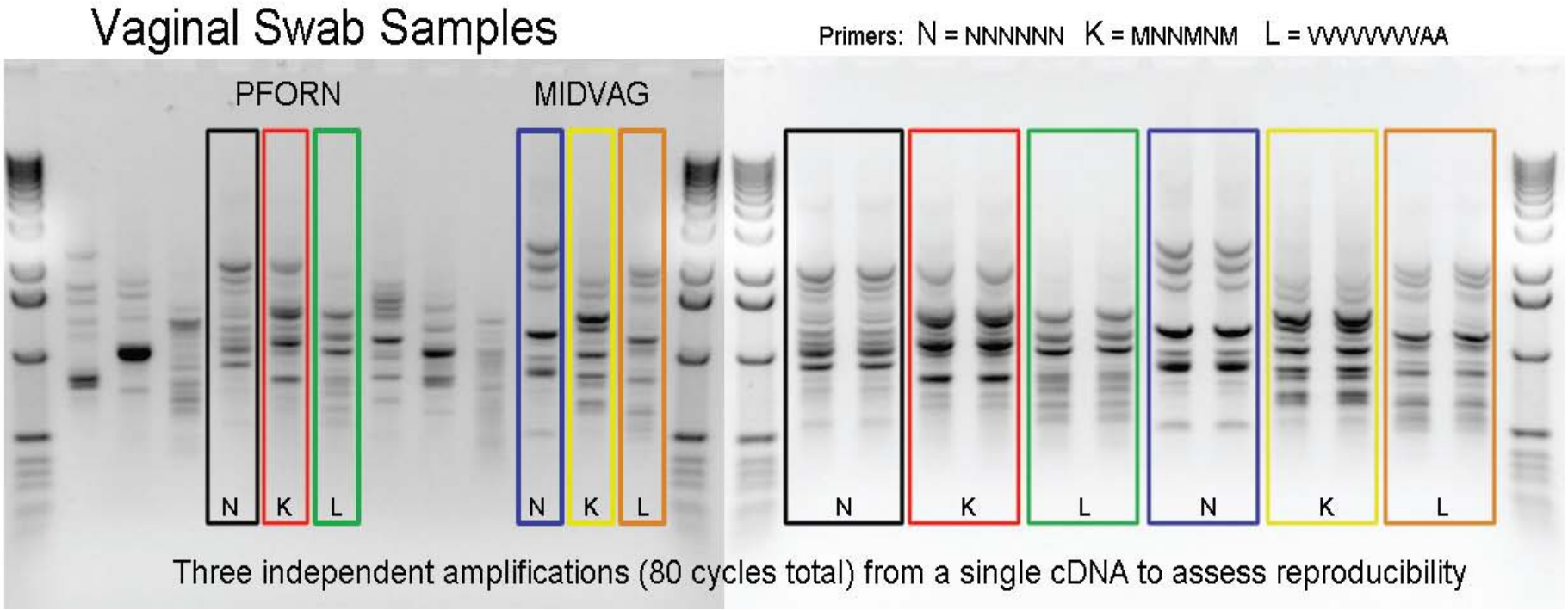


Sampling method and primer impact…

Vaginal samples: 2 sites, 2 collection methods,
3 random primer designs



How reproducible is random PCR?



LEFT: Original cDNA amplification (80 total cycles of PCR)

RIGHT: Two additional PCR attempts, two months later.

Conclusions: Amplification is reproducible. Each random primer produces a distinct banding pattern for the same cDNA template.

Technical questions

- How much depth is needed to viral diversity
  - 454 and Illumina
- Do random primer designs sample viruses equally well
- How do we remove contaminating DNA
- How do we analyze the data in a cost effective manner

Viral detection on two platforms

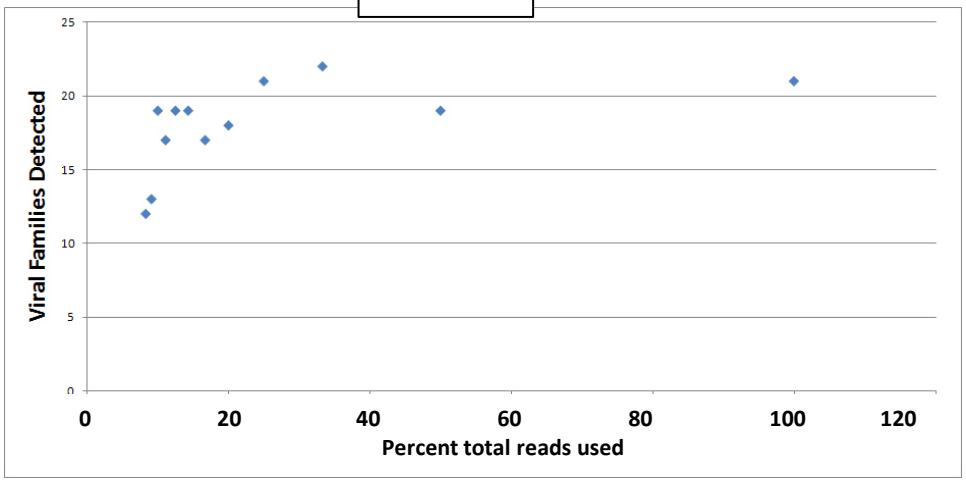
Two stool samples (S6 & S7)

	½ PTP			One lane	
Platform	**454-Ti**			**Illumina GAII**	
Sample	S6	S7		S6	S7
Avg read length (bp)	240	250		95	95
Read # (million)	0.515	0.660		131	151
Total number of bases (Mb)	140.15	182.65		12,445	14,345
Viral families*	17	14		22	31
Unique viruses*	62	71		92	138

* Following assembly with Newbler (454) or Soap or Velvet (Illumina)

How much data captures total detectable diversity…

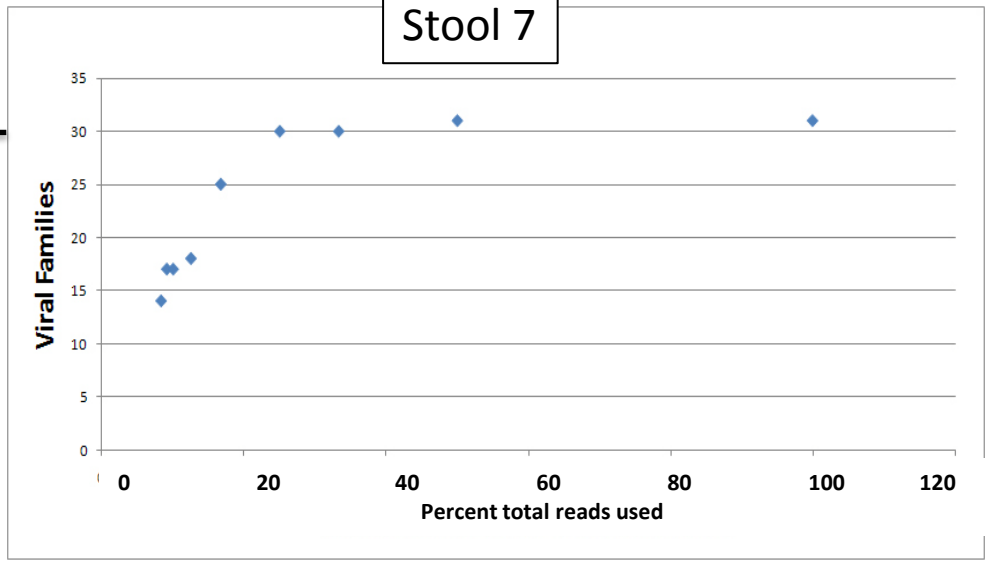
Stool 6



→ 23 viral families

Illumina: ~ 131 million 95 bp reads

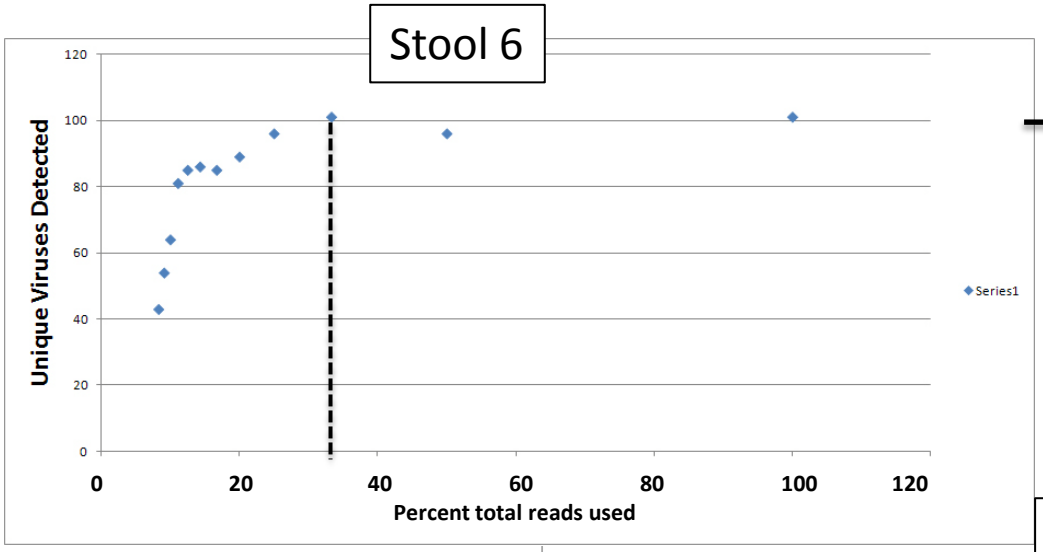
Stool 7



← 31 viral families

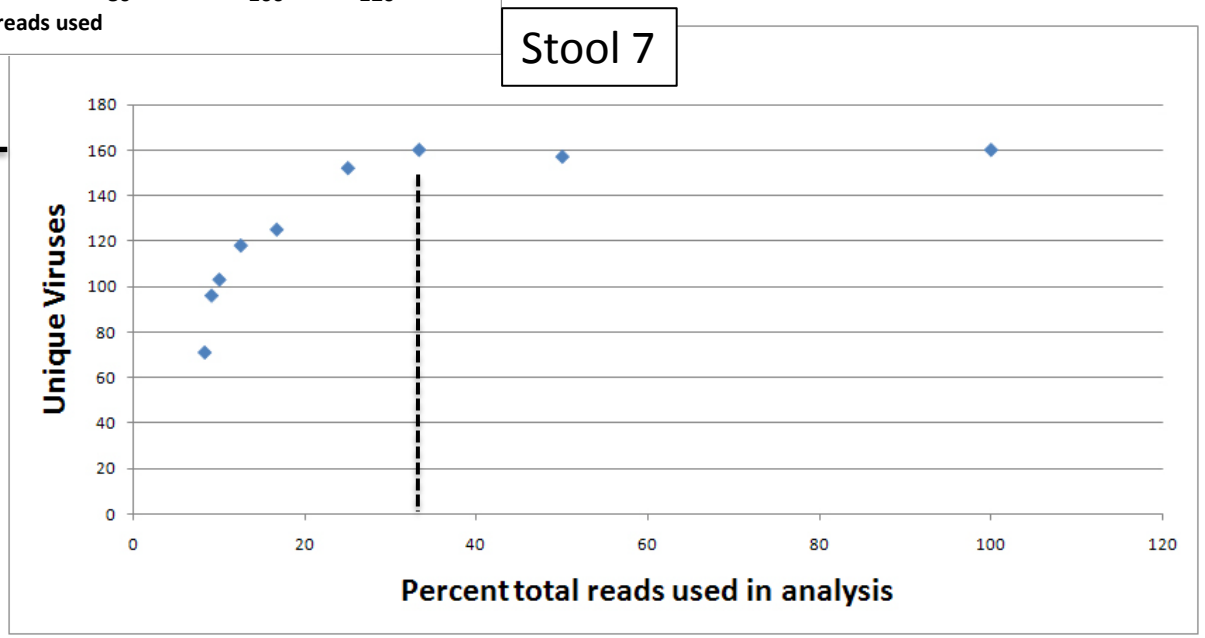
Illumina: ~ 151 million 95 bp reads

…and detectable unique viruses…



101 unique viruses

Illumina: ~ 131 million 95 bp reads

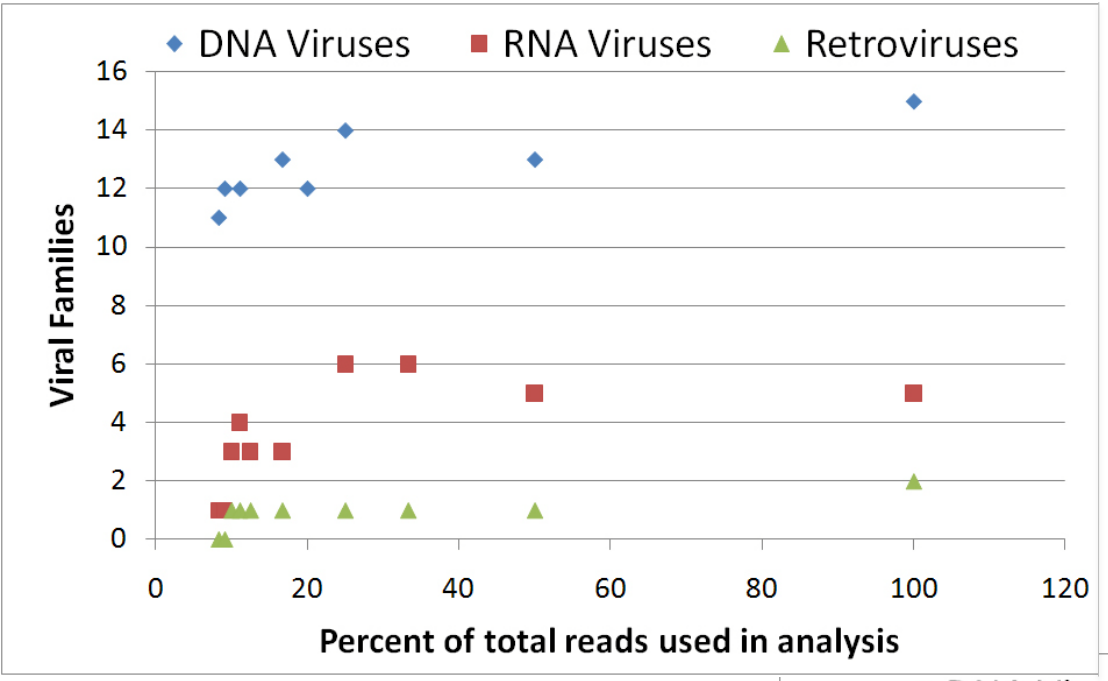


160 unique viruses

Illumina: ~ 151 million 95 bp reads

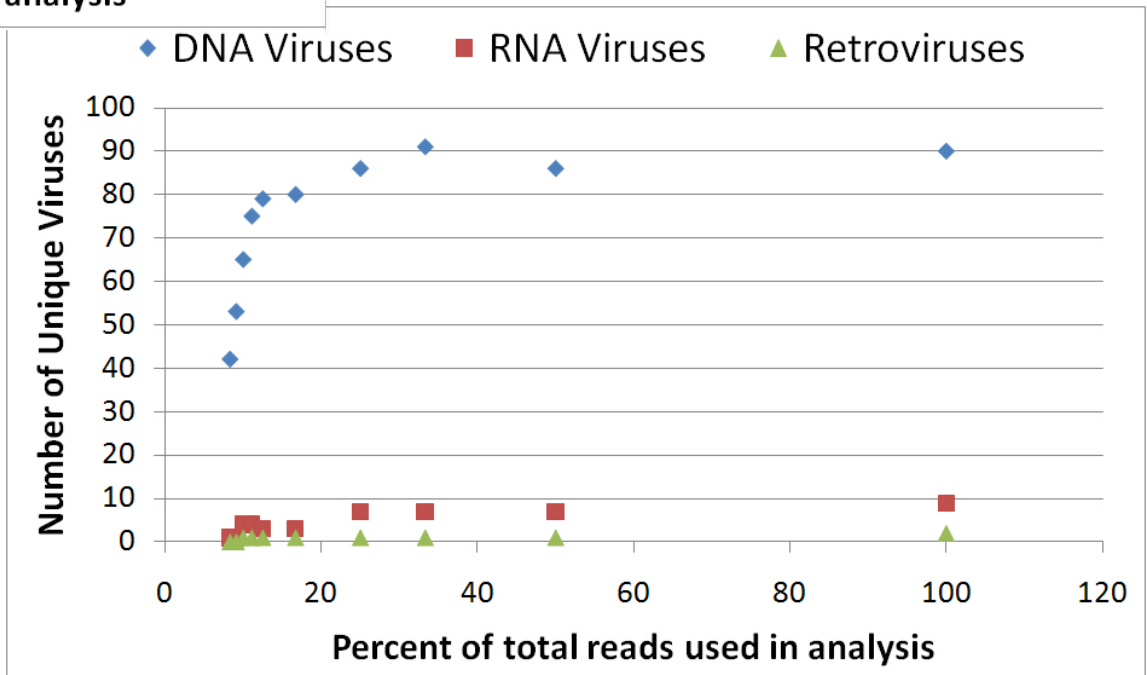
*Results suggest one can multiplex on Illumina and still capture detectable viral community.

Relative representation of DNA/RNA viruses in stool



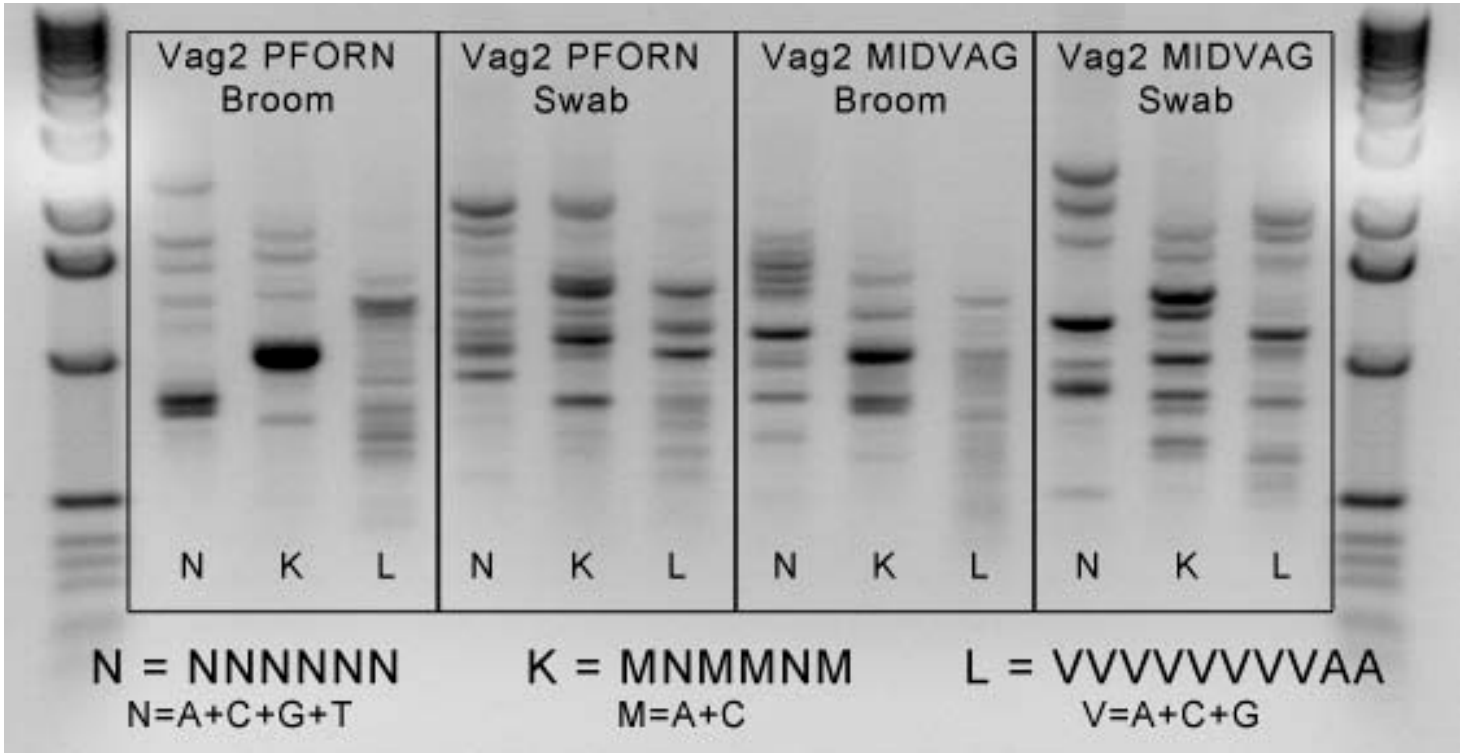
Recovering RNA viruses…

Illumina: ~ 131 million 95 bp reads



Sample quantity and random primer impact…

Vaginal samples: 2 sites, 2 collection methods, 3 random primer designs



How does sample quantity help/hurt viral detection? How well do the random primers capture viral families?

Summary of sequence stats (454-Ti)

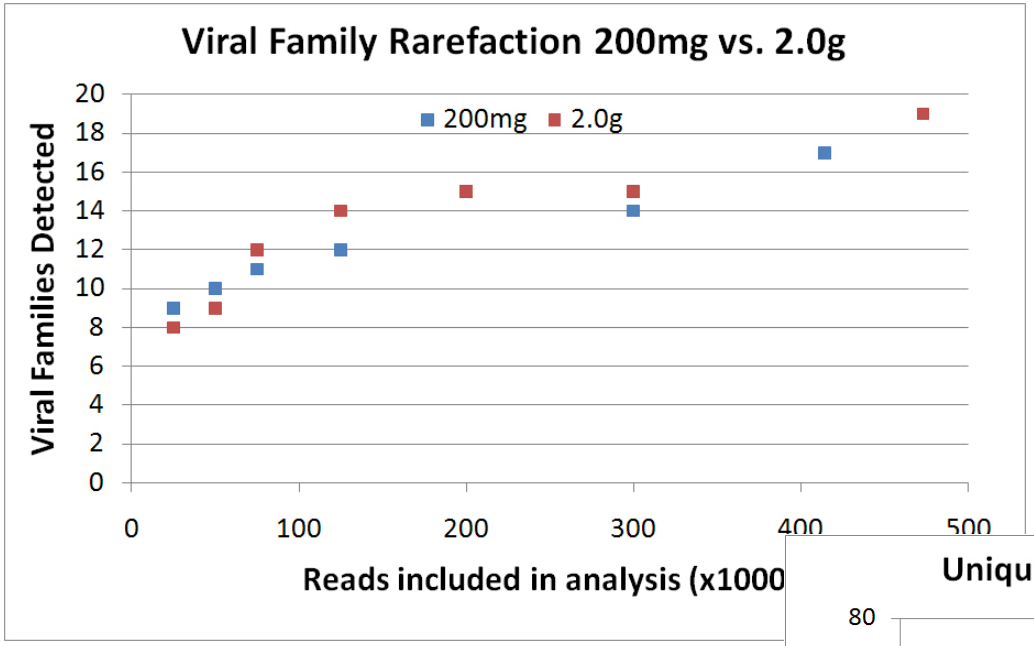
1,280,088 total reads (200mg = 594,335 ; 2.0g = 646,365)
Average read length after trimming = 307 bp

Summary of Assembly stats

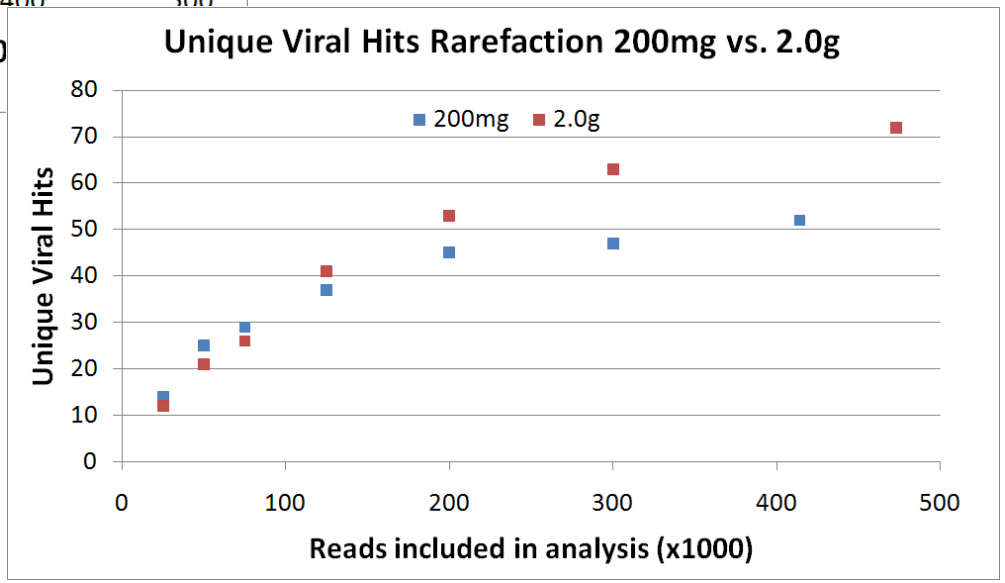
Sample	# of Contigs	Contig N50	Families Found	Unique Viral Hits
200mg Random Hexamer	731	635	13	32
200mg K-Random (MNMNNM)	691	615	14	30
200mg 3'-Locked (VVVVVVVVAA)	2145	599	15	46
2.0g Random Hexamer	1429	571	14	43
2.0g K-Random (MNMNNM)	1728	582	13	42
2.0g 3'-Locked (VVVVVVVAA)	2495	568	19	54

Does more sample = more viruses?

(How low can we go?)



… going from 2g to 200mg doesn't appear to greatly impair recovery



Viral families captured by random primers

	200mg Starting Material			2g Starting Material		
	Random Primers					
Viral Family	NNNNNN	MNMNNM	VVVVVVVVAA	NNNNNN	MNMNNM	VVVVVVVVAA
Adenoviridae			X			
Alloherpesviridae			X	X		X
Anellovirus						X
Ascoviridae		X	X	X	X	X
Baculoviridae	X	X	X	X	X	X
Bunyaviridae						X
Caliciviridae	X		X	X	X	X
Flaviviridae						X
Herpesviridae	X	X	X	X	X	X
Iridoviridae	X	X	X	X	X	X
Mimiviridae	X	X	X	X	X	X
Nimaviridae		X			X	
Papillomaviridae	X	X		X		
Phycodnaviridae	X	X	X	X	X	X
Picobirnaviridae	X	X	X			X
Polydnaviridae						X
Potyviridae	X	X		X	X	X
Poxviridae	X	X	X	X	X	X
Retroviridae			X			X
Tobamovirus	X	X	X	X	X	X
unclassified_dsDNA	X	X	X	X	X	X
unclassified_viruses	X	X	X	X	X	X

X Detected

Not Detected

SUMMARY

9 of 22 families detected by all 3 primers at both starting amounts
10 of 22 families detected by all 3 primers at 200mg starting amount
12 of 22 families detected by all 3 primers at 2g starting amount

5 of 22 families were detected only by the "VVVVVVVVAA" primer .... 4 of the 5 only found in the 2.0g sample.
1 of 22 families was detected only by the "MNMNNM" primer (Nimaviridae)

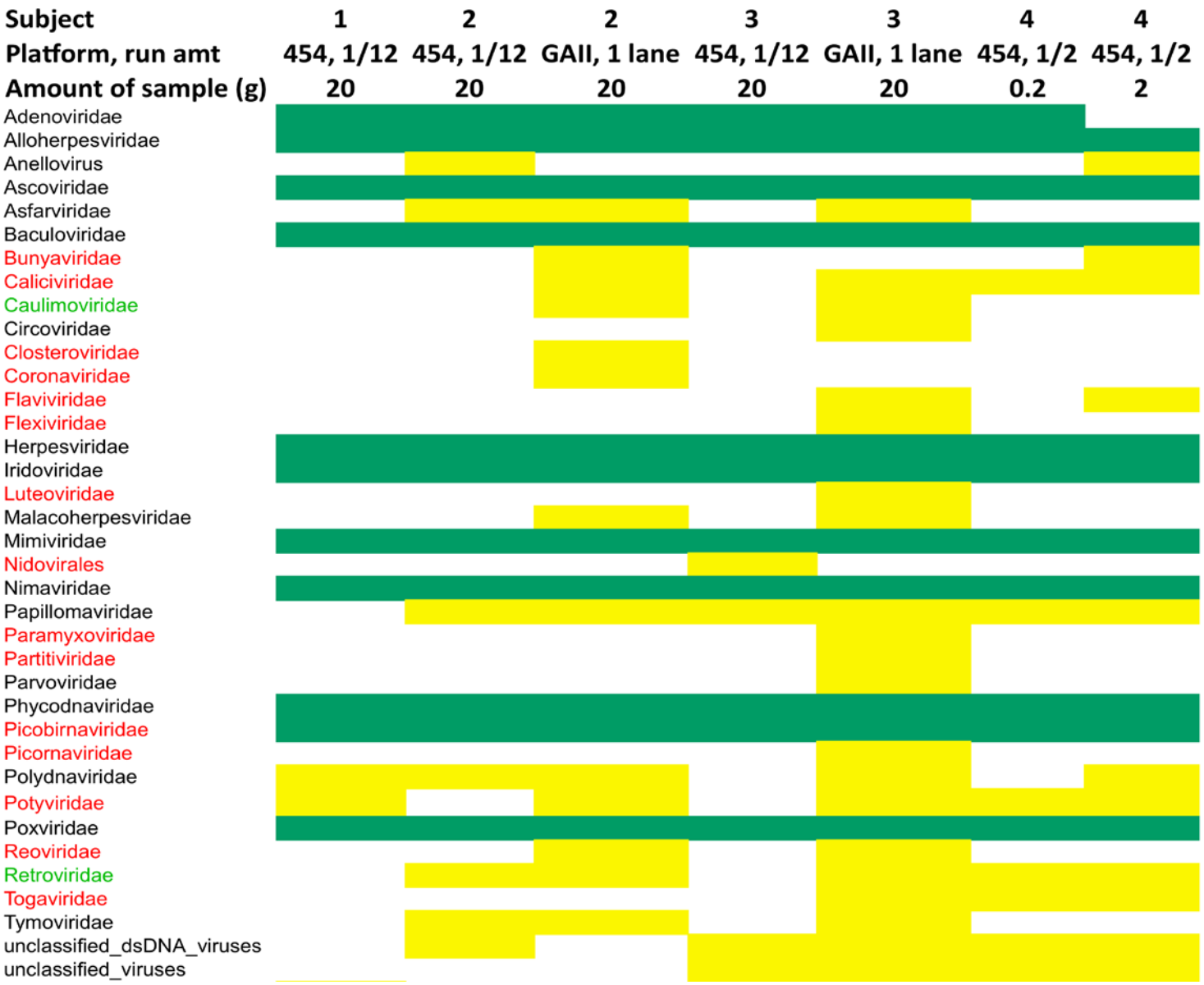
Biological questions

- What viruses are present at different body sites
  - Phages
  - RNA vs DNA
  - Colonize vs passing through
- Do different people have the same viral membership
  - Cannot measure abundance quantitatively with random primers

Viral families detected in 4 subjects

- Patterns emerging
- Assembly helps

- need to verify hits
- colonizing?
- intact?



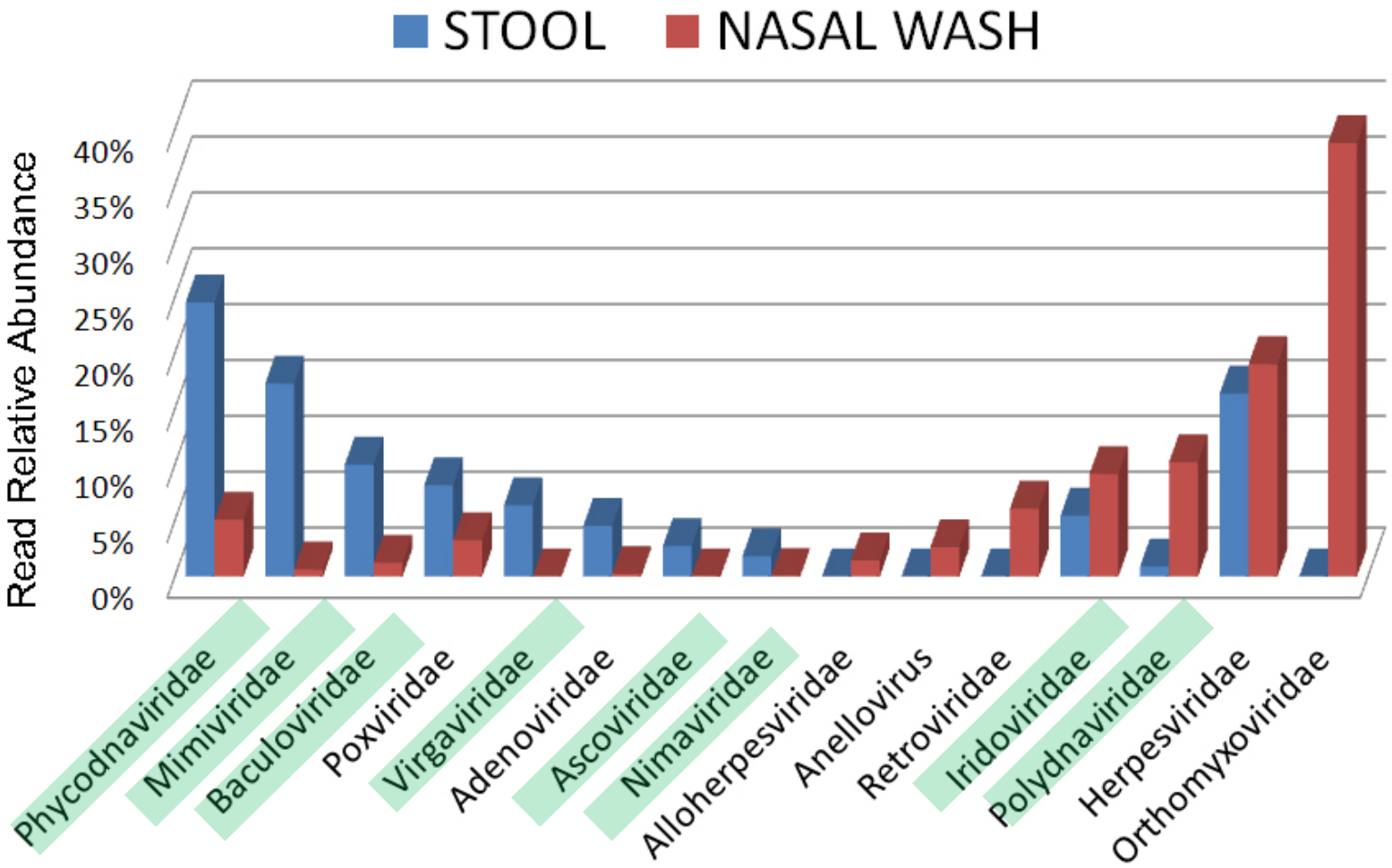
Phage…

48 discovered in 1st pass

Greater than 1% genome coverage
Between 0.25% and 0.99% coverage
Between 0.01% and 0.249% coverage

phage	% of genome covered	Ref Genome Length
Enterobacteria phage phiV10	11.6868	39104
Lactococcus phage CB13	11.1584	32182
Lactococcus phage CB20	11.1406	28625
Lactococcus phage bIBB29	10.4283	29305
Lactococcus phage P008	10.295	28538
Lactococcus phage SL4	7.3728	28144
Lactococcus phage CB14	6.6092	29459
Bacteriophage bIL170	6.3866	31754
Lactococcus lactis phage jj50	6.1888	27453
Lactococcus phage 712	6.0964	30510
Bacteriophage sk1	3.8768	28451
Lactococcus phage bIL67	2.424	22195
Streptococcus phage 858	2.3493	35543
Lactococcus phage CB19	2.0179	28643
Streptococcus phage ALQ13.2	1.2273	35525
Streptococcus thermophilus bacteriophage Sfi11	0.9672	39807
Propionibacterium phage PA6	0.7801	29739
Salmonella typhimurium phage ST64B	0.4981	40149
Streptococcus phage Abc2	0.4902	34882
Streptococcus suis phage SMP	0.4777	36216
Streptococcus thermophilus temperate bacterioph	0.3645	43075
Lactococcus phage phismq86	0.2943	33641
Enterococcus phage phiFL4A	0.2827	37856
Enterobacteria phage P7	0.24	101660
Phage cdtI DNA	0.2148	47021
Geobacillus phage GBSV1	0.1874	34683
Streptococcus phage 5093	0.1748	37184
Staphylococcus prophage phiPV83 proviral DNA	0.1622	45636
Streptococcus pneumoniae bacteriophage MM1 1	0.1208	38893
Enterobacteria phage CUS-3	0.1144	40207
Yersinia phage Yepe2	0.1138	38677
Enterobacteria phage DE3	0.1072	42925
Mycobacterium phage Myrna	0.0668	164602
Mycobacteriophage PLot	0.0633	64787
Rhodococcus phage ReqiPoco6	0.0589	78064
Clostridium phage c-st genomic DNA	0.0474	185683
Bacteriophage SPBc2	0.0372	134416
Pseudomonas phage phiKZ	0.0342	280334
Synechococcus cyanophage syn9	0.0327	177300
Enterobacteria phage AR1 DNA	0.0311	167435
Mycobacterium phage ScottMcG	0.0299	154017
Mycobacterium phage Rizal	0.0286	153894
Acinetobacter phage 133	0.0263	159801
Ralstonia phage RSL1 DNA	0.0259	231255
Aeromonas phage phiAS5	0.0253	225268
Synechococcus phage S-RSM4	0.0247	194454
Pseudomonas phage phiEL	0.0246	211215

Virus protocol differentiates stool and nasal wash viruses



Families boxed in green are not known to infect mammals

Candidate etiologic agent discovery and direct pathogen sequencing

Kawasaki Disease*

- Affects mainly children (6mo-5yrs) of Japanese or Korean descent
- Causes multi-system vasculitis and can cause coronary artery aneurism and other abnormalities
- The cause is currently unknown:
  - Infectious agent ?
    - Seasonal peaks, Acute onset, Self-limited, increased susceptibility of a particular age group (toddlers), defined epidemics
  - Genetic predisposition ?
    - High recurrence within families (10-15 fold greater probability)
    - Incidence rates determined by race and not geographical location
    - Mutation of CCR5 (HIV co-receptor) is associated with 80% reduced risk of KD

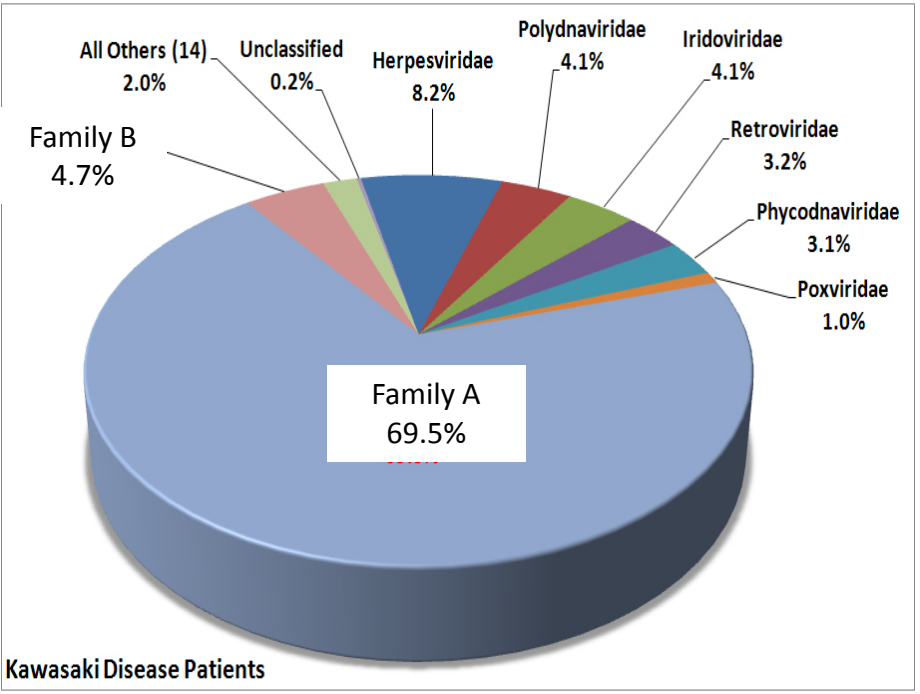
* w/ Dr. Sheldon Kaplan, TCH

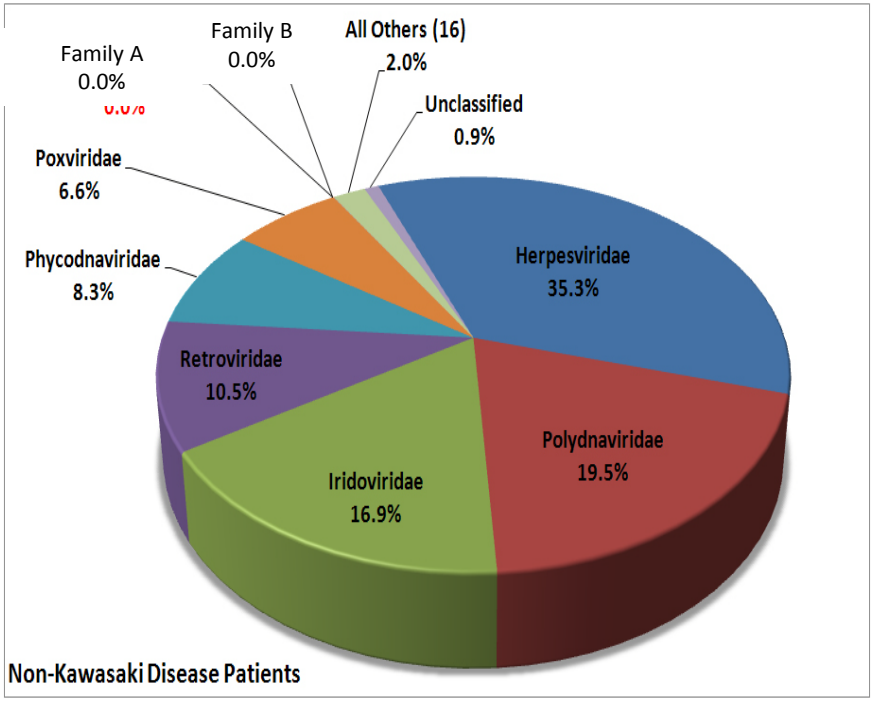
454 analysis of KD samples

- 23 KD patient nasal washes were pooled (groups of 5 and 3)
- 10 non-KD patient nasal washes were pooled
- cDNA Libraries were constructed and 454 adapters were added by PCR
- Data filtered and assembled, contigs and reads examined…

Results from pooled samples

- Currently working on analyzing samples individually
- Evaluating the legitimacy of hits and determining genome coverage of each virus detected





Elephant Herpes

- Herpesviruses are ubiquitous in nature

- A novel Elephant endotheliotropic herpesvirus (EEHV) is causing significant morbidity and mortality in both captive and wild juvenile Asian elephants (endangered species)
    - unable to cultivated outside of host

- Until 2010, all 6 calves born at the Houston zoo in the last two decades have died from EEHV infection

- BCM, in coordination with the Houston Zoo assembled to improve EEHV diagnostics and develop vaccines

Upping the ante: Baylor

In for a penny…

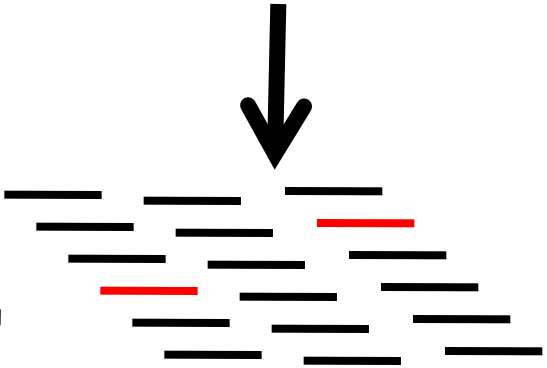
…in for several tons….

Approach: Sequence the genome of EEHV

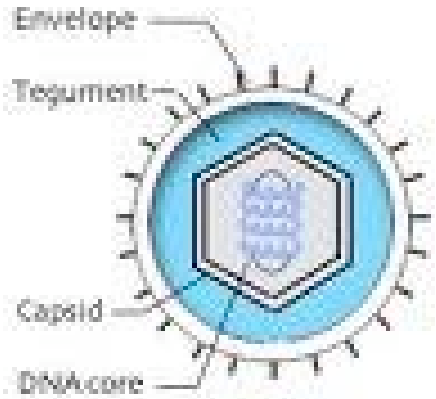
Sample

Mixed DNA

Shotgun Sequence



22 x 10^9 base pairs
(22 Gigabases, mixed)



Bioinformatics-Filter for Viral Genome

Assembly Metrics for EEHV1

Total number of contigs: 882

Total sequence length: 320526 → Refined to ~260 kb

Total number of >=1k sequences: 19

Total >=1k sequence length: 173502

Total number of >=5k sequences: 7

Total >=5k sequence length: 150184

Average sequence length: 363

Largest sequence size: 83680 → **Now: 162 Kb**

Smallest sequence size: 100

N50 size: 2726

N50 node: 10

Next Step: PCR with herpes specific primers, sequence on 454

Other discovery projects…

1. Pediatric encephalitis and meningoencephalitis (CSF)

2. Other neurological syndromes such as acute disseminated encephalomyelitis ADEM (CSF)

3. Culture negative acute osteomyelitis (blood or bone biopsy) or septic arthritis (synovial fluid or blood).

4. Fever in the neutropenic patient generally with leukemia (blood)

5. Community acquired pneumonia-- a large number of cases do not have a proven etiology.

6. Pediatric Acute Liver Failure

Summary

- Viral metagenomic strategies are improving with less sample
  - samples may be multiplexed in GAII (more in HiSeq)
  - RNA and DNA viruses, as well as phage are captured
  - Further enhancement possible

- Areas for immediate attention
  - Improve curation of viral db
  - Improve removal of background contaminating DNA
  - Establish measures to test for colonization

- These strategies are already yielding results in several models

Acknowledgements

Petrosino Lab/CMMR
Brian McWilliams
Stephanie Herb
Matthew Ross
Tulin Ayvaz
David Berken
Lisa Atkins
Diane Smith

BCM/TCH
Sarah Highlander
James Versalovic
Kjersti Aagaard-Tillery
Wendy Keitel
Bonnie Youmans

HGSC
Richard Gibbs
Donna Muzny
Jeffrey Reid
Xiang Qin
Yi Han
Christie Kovar
Christian Buhay
Michael Holder

WashU
George Weinstock
Erica Sodergren

DACC
Owen White
Todd DeSantis

JCVI
Karen Nelson
Barb Methe

Broad Institute
Bruce Birren
Dirk Gevers

Funding
NHGRI, NIH-Roadmap HMP
NIAID
Vivian Smith Foundation