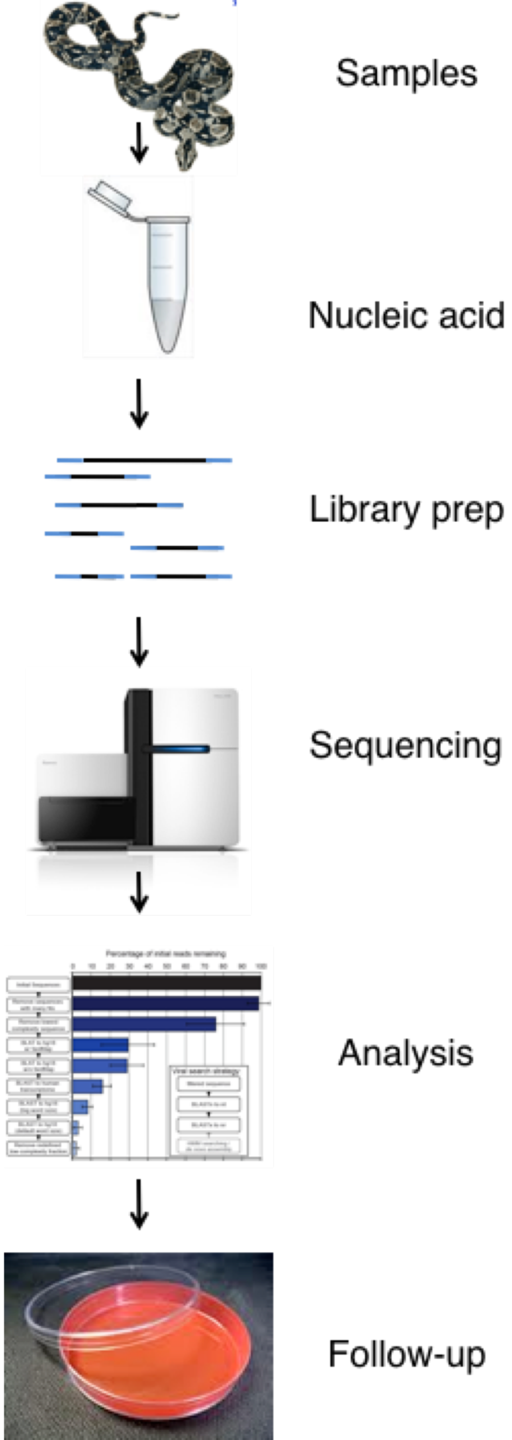


Samples

What groups & controls?
What source material?
N?

Nucleic acid

RNA vs. DNA
Whole genome vs. subset
Pool samples to make library?

Library prep

Details determined by rest

Sequencing

Read length
Paired vs. Single end
Platform
Multiplexing: Depth vs. N*$

Analysis

Analyze:
DNA sequences
Presence vs. Absence of taxa
Quantitative Comparison

Follow-up

Group

Ovary

Group

Treatment 1

Treatment 2

Control 1

Control 2

Group

Pop A

Pop B

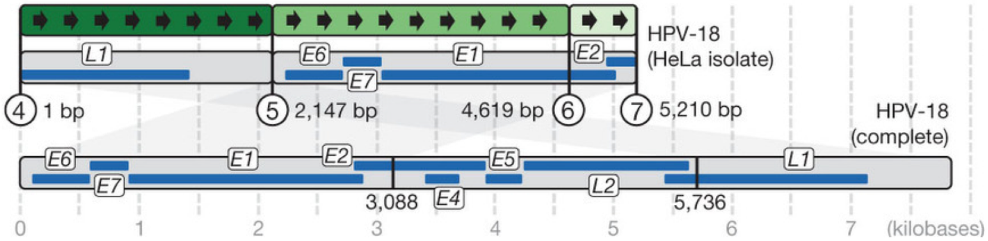
Pop C

Analyze:
DNA sequences
Presence vs. Absence of taxa
Quantitative Comparison

Positive and Negative Controls

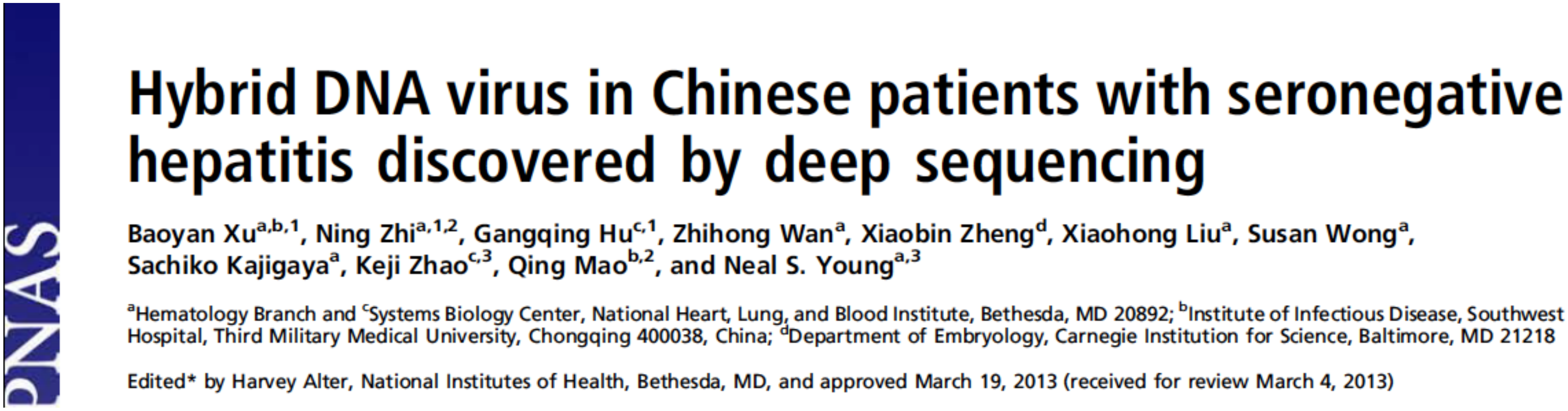
Most relevant for Presence / Absence (Detection)

Positive control:
HeLa total RNA
'Mock community'



Negative control:
Water
Field collection & lab

Why water?



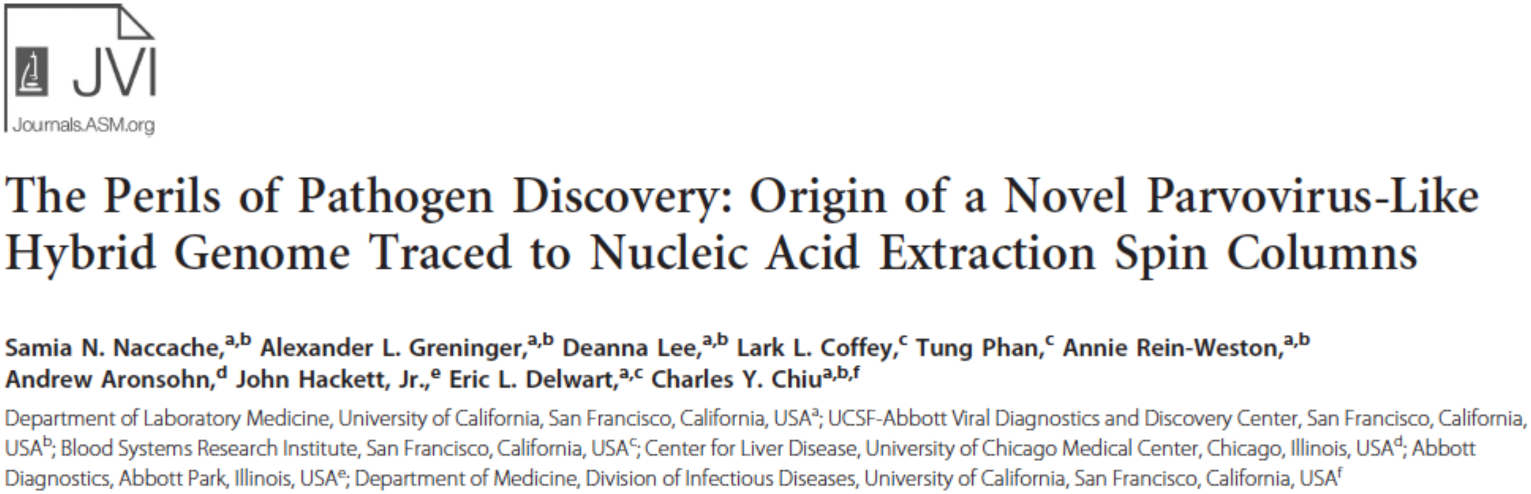
PNAS

Hybrid DNA virus in Chinese patients with seronegative hepatitis discovered by deep sequencing

Baoyan Xu[a,b,1], Ning Zhi[a,1,2], Gangqing Hu[c,1], Zhihong Wan[a], Xiaobin Zheng[d], Xiaohong Liu[a], Susan Wong[a], Sachiko Kajigaya[a], Keji Zhao[c,3], Qing Mao[b,2], and Neal S. Young[a,3]

[a]Hematology Branch and [c]Systems Biology Center, National Heart, Lung, and Blood Institute, Bethesda, MD 20892; [b]Institute of Infectious Disease, Southwest Hospital, Third Military Medical University, Chongqing 400038, China; [d]Department of Embryology, Carnegie Institution for Science, Baltimore, MD 21218

Edited* by Harvey Alter, National Institutes of Health, Bethesda, MD, and approved March 19, 2013 (received for review March 4, 2013)



JVI
Journals.ASM.org

The Perils of Pathogen Discovery: Origin of a Novel Parvovirus-Like Hybrid Genome Traced to Nucleic Acid Extraction Spin Columns

Samia N. Naccache,[a,b] Alexander L. Greninger,[a,b] Deanna Lee,[a,b] Lark L. Coffey,[c] Tung Phan,[c] Annie Rein-Weston,[a,b] Andrew Aronsohn,[d] John Hackett, Jr.,[e] Eric L. Delwart,[a,c] Charles Y. Chiu[a,b,f]

Department of Laboratory Medicine, University of California, San Francisco, California, USA[a]; UCSF-Abbott Viral Diagnostics and Discovery Center, San Francisco, California, USA[b]; Blood Systems Research Institute, San Francisco, California, USA[c]; Center for Liver Disease, University of Chicago Medical Center, Chicago, Illinois, USA[d]; Abbott Diagnostics, Abbott Park, Illinois, USA[e]; Department of Medicine, Division of Infectious Diseases, University of California, San Francisco, California, USA[f]

Reagents can be a source of nucleic acid

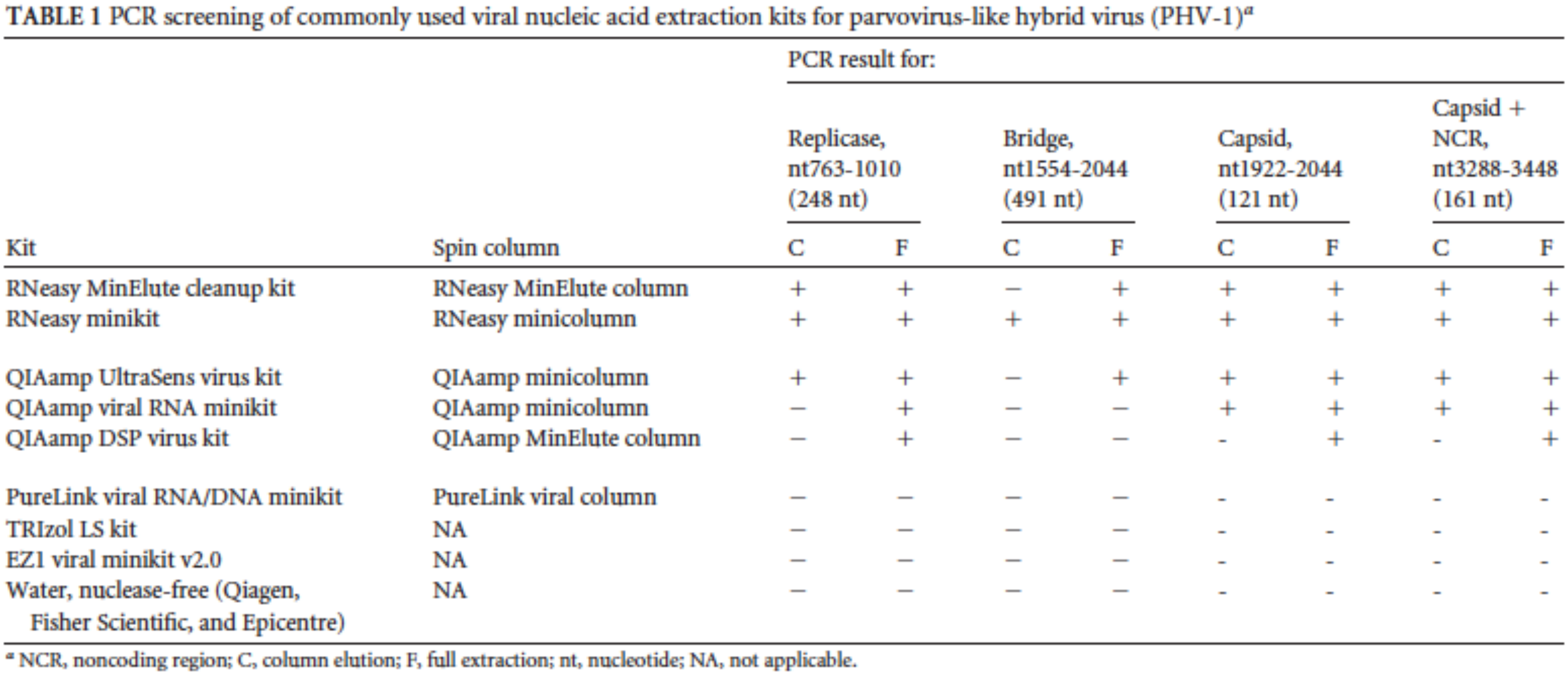


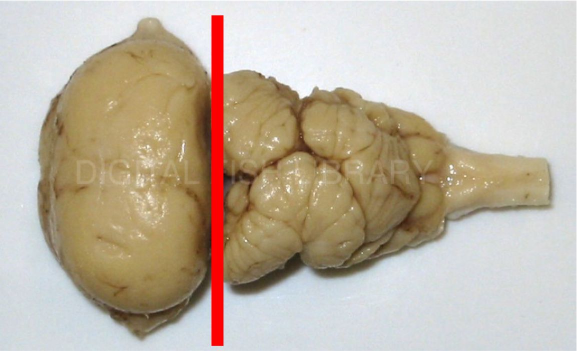
TABLE 1 PCR screening of commonly used viral nucleic acid extraction kits for parvovirus-like hybrid virus (PHV-1)[a]

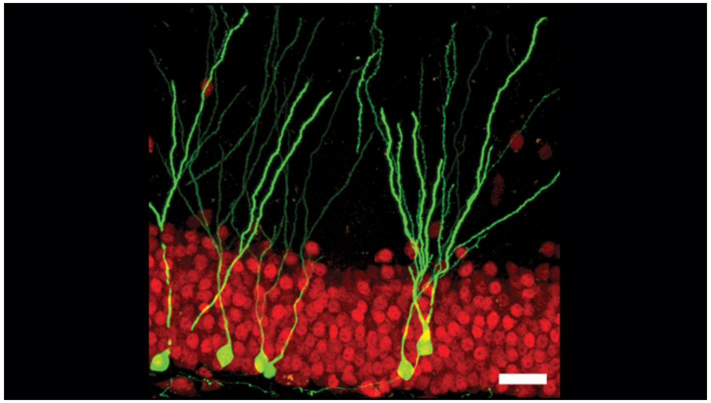
		PCR result for:							
		Replicase, nt763-1010 (248 nt)		Bridge, nt1554-2044 (491 nt)		Capsid, nt1922-2044 (121 nt)		Capsid + NCR, nt3288-3448 (161 nt)	
Kit	Spin column	C	F	C	F	C	F	C	F
RNeasy MinElute cleanup kit	RNeasy MinElute column	+	+	−	+	+	+	+	+
RNeasy minikit	RNeasy minicolumn	+	+	+	+	+	+	+	+
QIAamp UltraSens virus kit	QIAamp minicolumn	+	+	−	+	+	+	+	+
QIAamp viral RNA minikit	QIAamp minicolumn	−	+	−	−	+	+	+	+
QIAamp DSP virus kit	QIAamp MinElute column	−	+	−	−	-	+	-	+
PureLink viral RNA/DNA minikit	PureLink viral column	−	−	−	−	-	-	-	-
TRIzol LS kit	NA	−	−	−	−	-	-	-	-
EZ1 viral minikit v2.0	NA	−	−	−	−	-	-	-	-
Water, nuclease-free (Qiagen, Fisher Scientific, and Epicentre)	NA	−	−	−	−	-	-	-	-

[a] NCR, noncoding region; C, column elution; F, full extraction; nt, nucleotide; NA, not applicable.

Source material considerations:

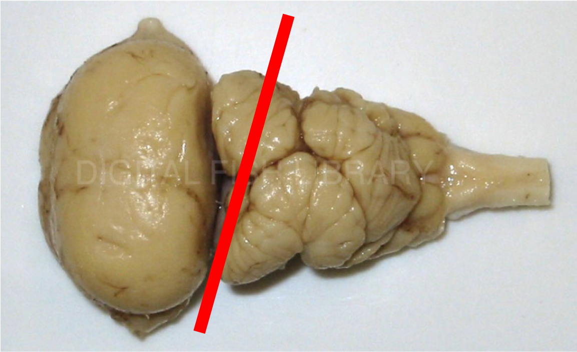
Generality vs. Specificity

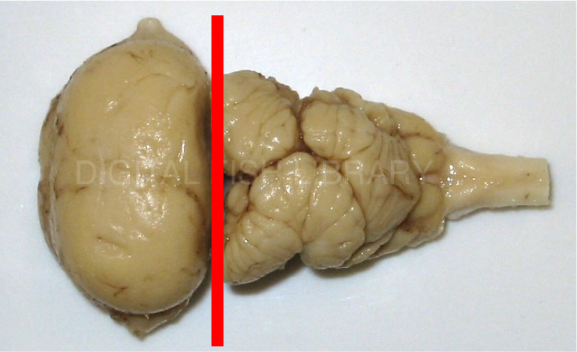




Source material considerations:

Generality vs. Specificity
Heterogeneity vs. Consistency





Source material considerations:

Generality vs. Specificity
Heterogeneity vs. Consistency

	7 am	**7:10**	**7:20**
Pop A	7:30	7:40	7:50
Pop B	1 pm	1:10	1:20
Pop C	1:30	1:40	1:50

Go to class

Source material considerations:

Generality vs. Specificity
Heterogeneity vs. Consistency

OPTION 1

	7 am	**1 PM**	**4 PM**
Pop A	7	1	4
Pop B	7	1	4
Pop C	7	1	4

OPTION 2

7 am	**7 AM**	**7 AM**
7	7	7
7	7	7
7	7	7
M	T	W

Source material considerations:

Generality vs. Specificity
Heterogeneity vs. Consistency

Homogeneous within blocks as much as possible

OPTION 1

	7 am	1 PM	4 PM
Pop A	7	1	4
Pop B	7	1	4
Pop C	7	1	4

OPTION 2

7 am	7 AM	7 AM
7	7	7
7	7	7
7	7	7
M	T	W

Sample size – Biological Replicates

RNAseq –

Depends on power you want, effect sizes you want to detect, risk of false positives you can tolerate

$N \geqq 3$ preferred for ANOVA designs, larger for smaller differences between groups & high confidence.

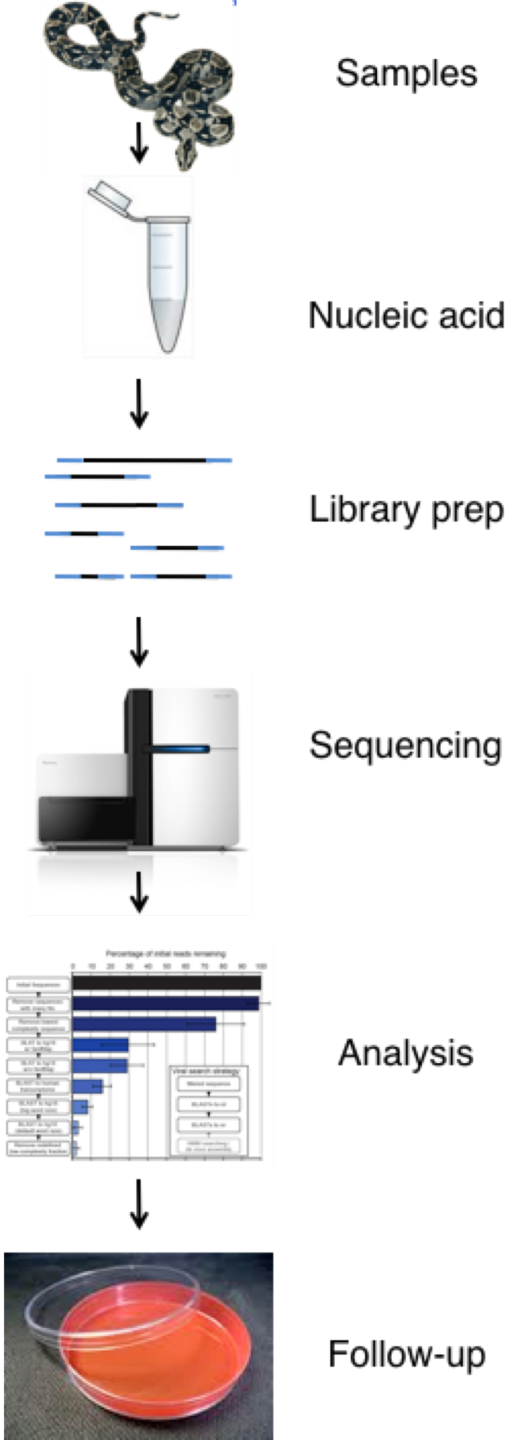
$N \geqq 20$ recommended for most transcriptome network or population genomic analyses, with more better

Sample size – Biological Replicates

Population genomics
- Represent populations?

Genome assembly
- None needed

Microbial quantitative comparisons
- Biological replicates needed



Samples

Nucleic acid

Library prep

Sequencing

Analysis

Follow-up

RNA vs. DNA
Whole genome vs. subset
Pool samples to make library?

Subset of genomes –

Immunoprecipitation
– RNAs bound to activated ribosomes, or DNAs in regions that are methylated

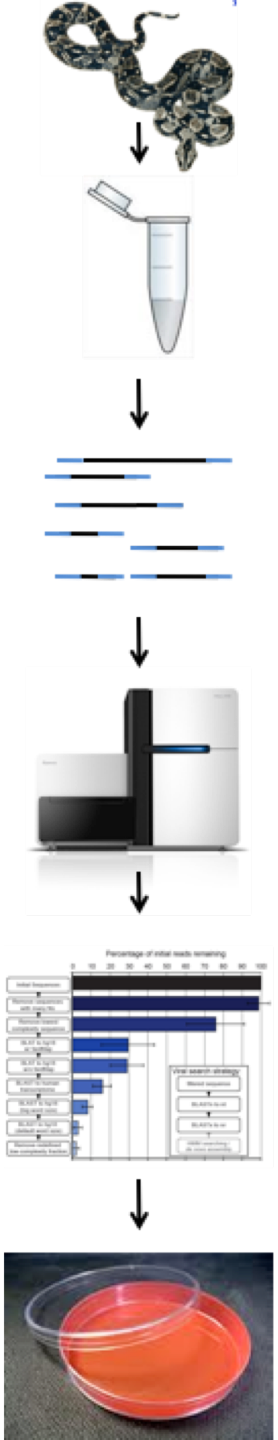
Population or species comparisons
– Amplicon sequencing
– Reduced representation libraries via sequence capture techniques

Pooling before sequence prep:

Enough tissue?

Generality vs. individual differences

{pooling decreases weight of outlier individuals, but still need multiple pools if RNAseq}



Samples

Nucleic acid

Library prep

Sequencing

Read length
Paired vs. Single end
Platform
Multiplexing: Depth vs. N*$

Analysis

Follow-up

Read length & Paired vs. Single end

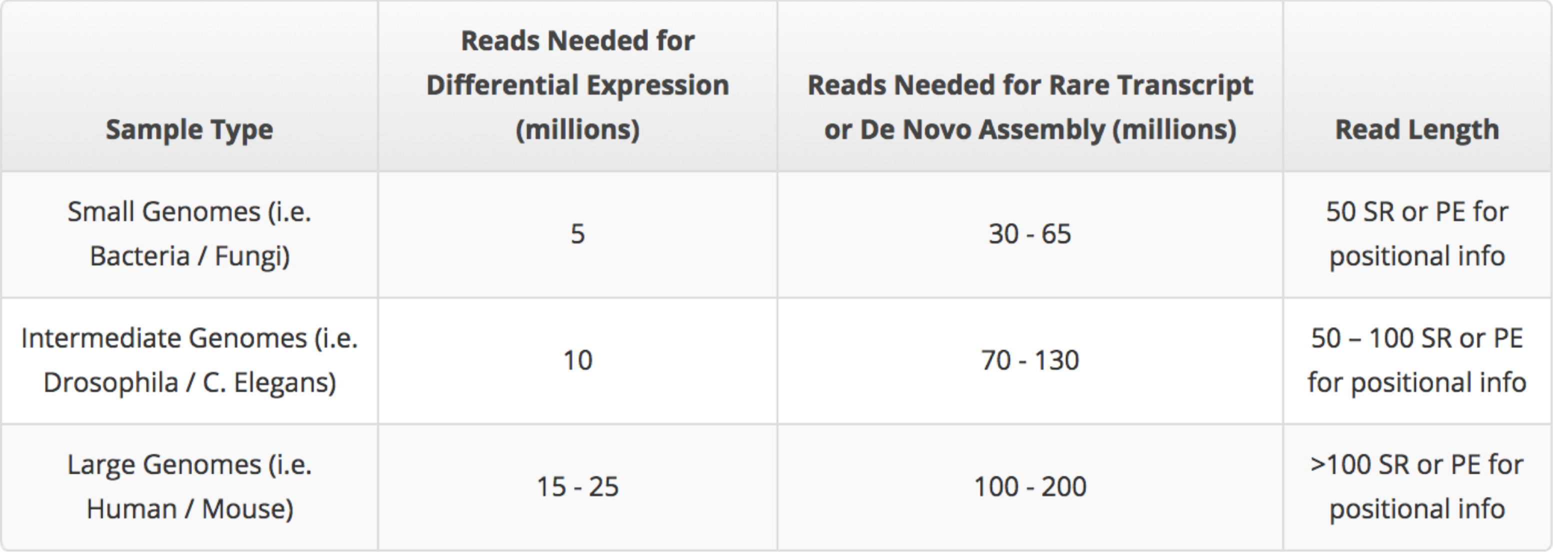
- Gene expression quantification in species with high quality genomes: shorter reads & single end okay {maximize read number/$}

- Otherwise, paired end & 100-150 bp {maximize bp/dollar}

RNA-seq for measuring gene expression levels

More reads per sample -> better quantification of low abundance transcripts {filter out low-count transcripts?}

Greater library complexity -> need more reads



Sample Type	Reads Needed for Differential Expression (millions)	Reads Needed for Rare Transcript or De Novo Assembly (millions)	Read Length
Small Genomes (i.e. Bacteria / Fungi)	5	30 - 65	50 SR or PE for positional info
Intermediate Genomes (i.e. Drosophila / C. Elegans)	10	70 - 130	50 – 100 SR or PE for positional info
Large Genomes (i.e. Human / Mouse)	15 - 25	100 - 200	>100 SR or PE for positional info

https://genohub.com/next-generation-sequencing-guide/

Sequence analysis

De novo genome assembly

50-100x coverage

Variant calling – heterozygosity (diploid genomes)

30x

Variant calling – haploid genome

20x

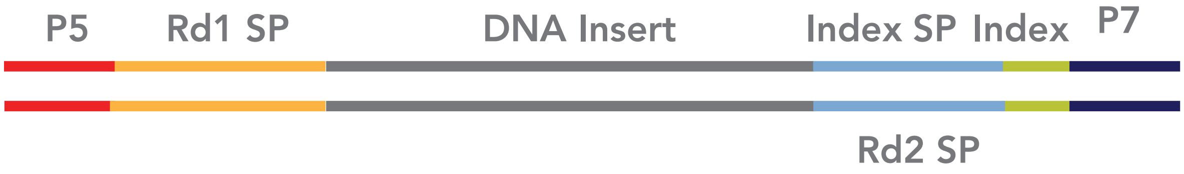
Microbial presence/absence

At least 5000-10,000 reads per sample for 16S

How many rare taxa do you want to detect?

Empirically determined

Multiplexing strategies

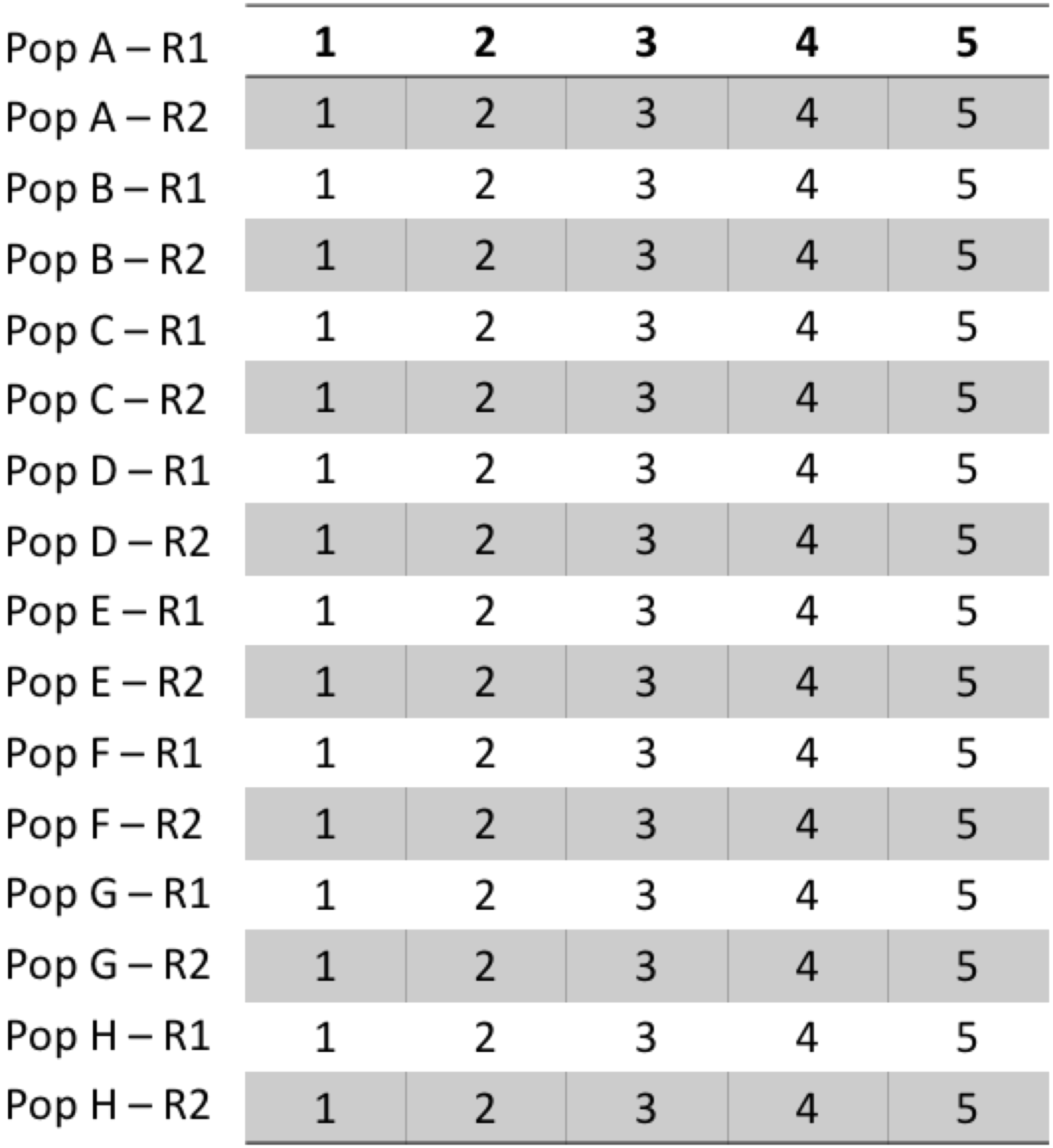


Quantitative comparisons:
Samples pooled in one sequencing lane are most comparable

Sequence comparisons and presence / absence:
Samples pooled in one lane can cross-contaminate

In all cases:
Not all libraries equally represented – be conservative

Multiplexing strategies: Quantification



Pop A – R1	**1**	**2**	**3**	**4**	**5**
Pop A – R2	1	2	3	4	5
Pop B – R1	1	2	3	4	5
Pop B – R2	1	2	3	4	5
Pop C – R1	1	2	3	4	5
Pop C – R2	1	2	3	4	5
Pop D – R1	1	2	3	4	5
Pop D – R2	1	2	3	4	5
Pop E – R1	1	2	3	4	5
Pop E – R2	1	2	3	4	5
Pop F – R1	1	2	3	4	5
Pop F – R2	1	2	3	4	5
Pop G – R1	1	2	3	4	5
Pop G – R2	1	2	3	4	5
Pop H – R1	1	2	3	4	5
Pop H – R2	1	2	3	4	5

Maintain blocks throughout library prep and sequencing

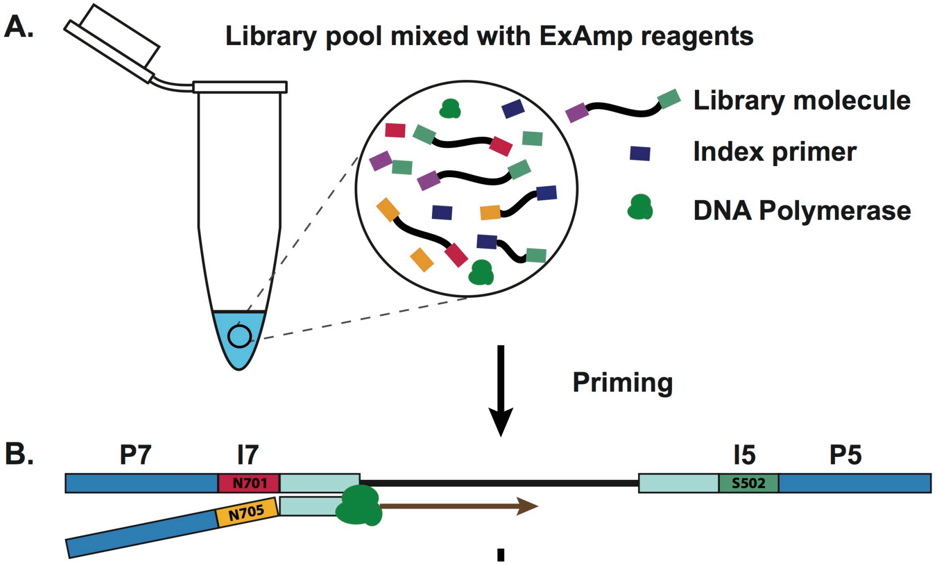
OPTION 1

	7 am	1 PM	4 PM
Pop A	7	1	4
Pop B	7	1	4
Pop C	7	1	4

OPTION 2

7 am	7 AM	7 AM
7	7	7
7	7	7
7	7	7
M	T	W

Cross-contamination: Index hopping



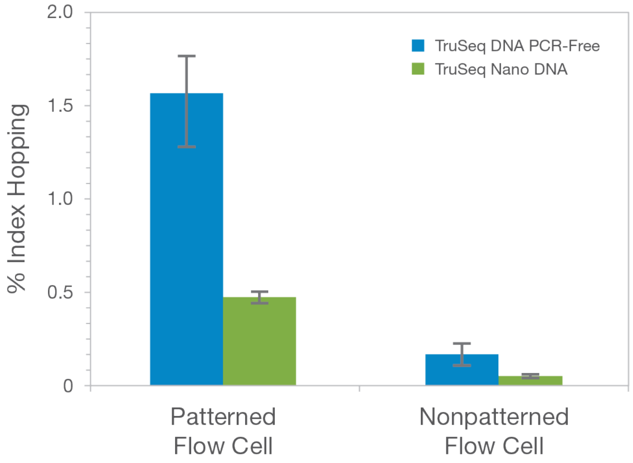


Image Sources: Illumina; Sinha et al. 2017 bioRxiv

Cross-contamination: Index hopping

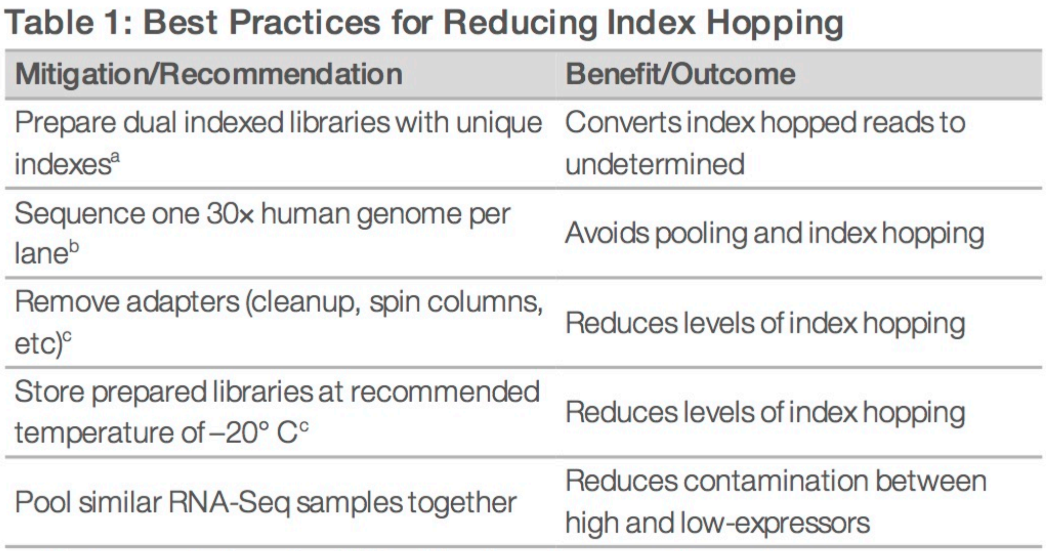


Table 1: Best Practices for Reducing Index Hopping

Mitigation/Recommendation	Benefit/Outcome
Prepare dual indexed libraries with unique indexes[a]	Converts index hopped reads to undetermined
Sequence one 30× human genome per lane[b]	Avoids pooling and index hopping
Remove adapters (cleanup, spin columns, etc)[c]	Reduces levels of index hopping
Store prepared libraries at recommended temperature of –20° C[c]	Reduces levels of index hopping
Pool similar RNA-Seq samples together	Reduces contamination between high and low-expressors

https://www.biorxiv.org/content/early/2017/04/09/125724

https://www.biorxiv.org/content/early/2017/08/16/177048

https://www.biorxiv.org/content/early/2017/09/01/182659

https://www.biorxiv.org/content/early/2017/10/10/200790

https://www.illumina.com/content/dam/illumina-marketing/documents/products/whitepapers/index-hopping-white-paper-770-2017-004.pdf

Image Sources: Illumina; Sinha et al. 2017 bioRxiv

Cross-contamination: Sequencer

Because of close position of clusters on a flow-cell index reads get misassigned at a high rate: ~0.3% (Kircher et al. 2011, Nucleic Acids Res.)

When this matters a lot:

- Single-cell genomics
- RNA-seq (especially comparative transcriptomics)

When it is more tolerable:

- Genome sequencing

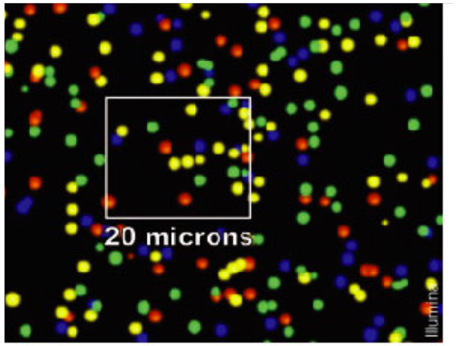


Image Source: TUCF Genomics

Reduce cross-contamination impacts

Figure 1 Single-Indexed Sequencing

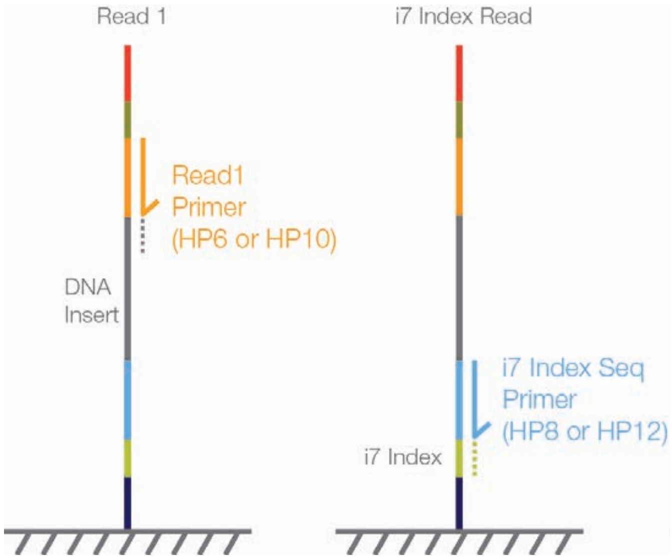
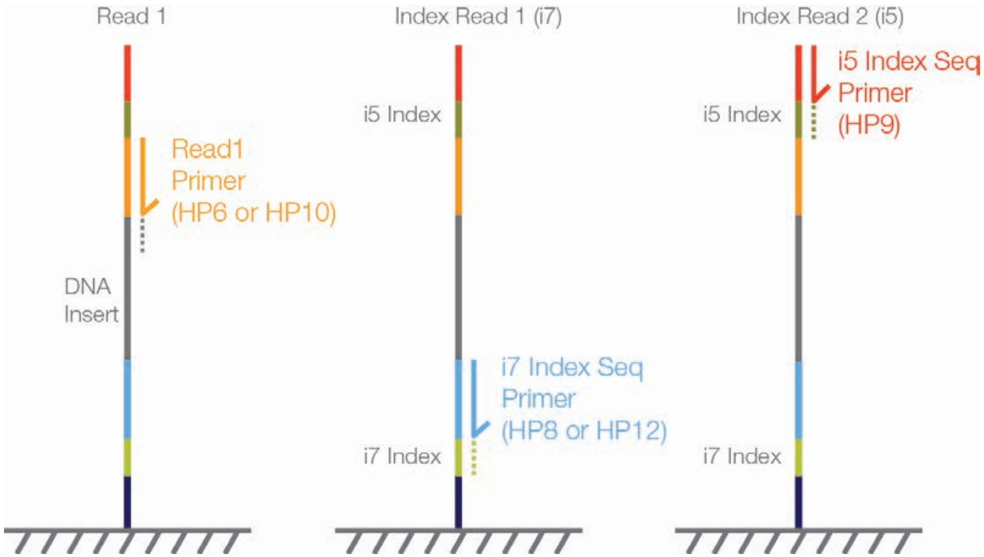


Figure 2 Dual-Indexed Single-Read Sequencing



Reduces cluster misassignment if indexes are used in a redundant fashion

Increases degree of multiplexing if indices are used in a combinatorial fashion