National Institute of Allergy and Infectious Diseases

Things you can do with the MiSeq

Analysis of Viral Genome Populations Using Next-Generation Sequencing

May 23, 2014

NIAID

NIH National Institute of Allergy and Infectious Diseases

Andrew Oler, PhD
High-throughput Sequencing Bioinformatics Specialist
BCBB/OCICB/NIAID/NIH

Outline

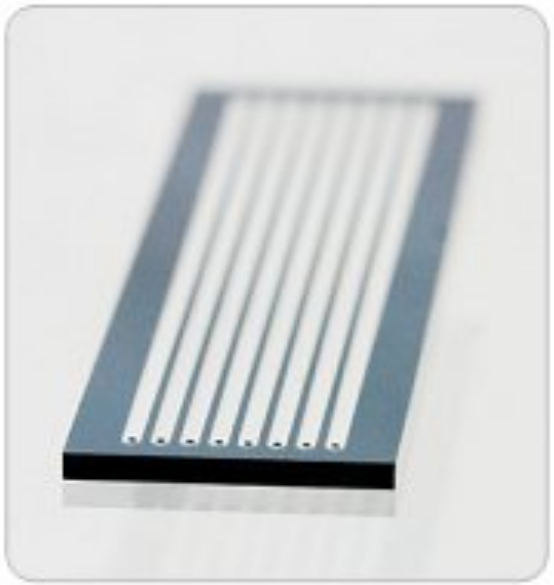
- MiSeq
- Variant analysis
  - PrimerID amplicon sequencing
    - Custom pipeline
  - Whole genome sequencing
    - Vprofiler, Vphaser
- Full genome assembly
  - VICUNA
  - SOAPdenovo2

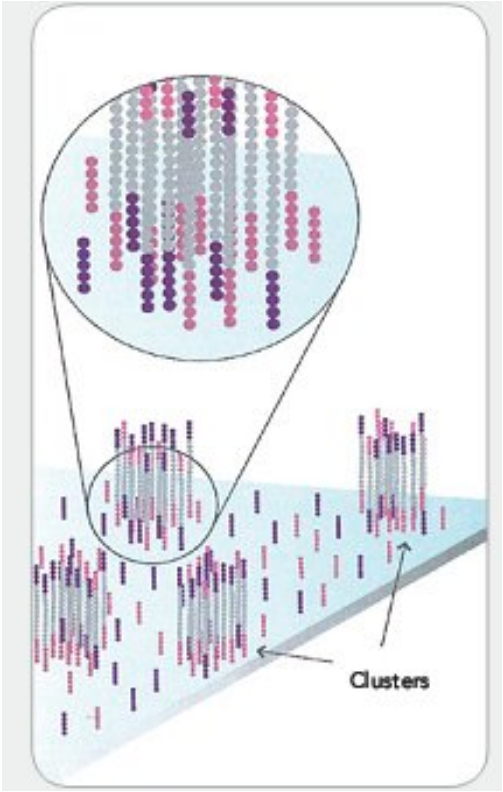
Outline

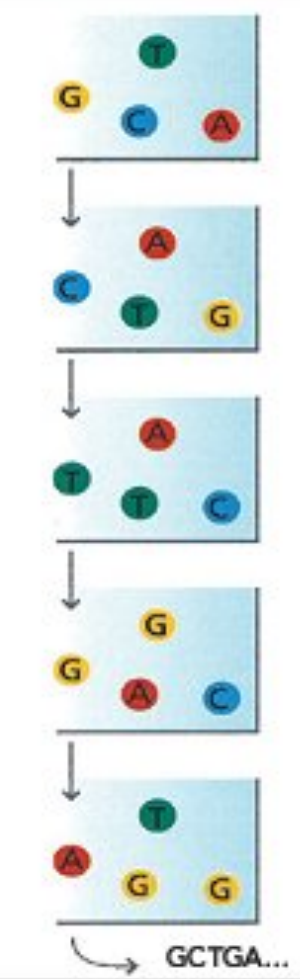
- MiSeq
- Variant analysis
  - PrimerID amplicon sequencing
    - Custom pipeline
  - Whole genome sequencing
    - Vprofiler, Vphaser
- Full genome assembly
  - VICUNA
  - SOAPdenovo2

Illumina sequencing

Sample DNA library







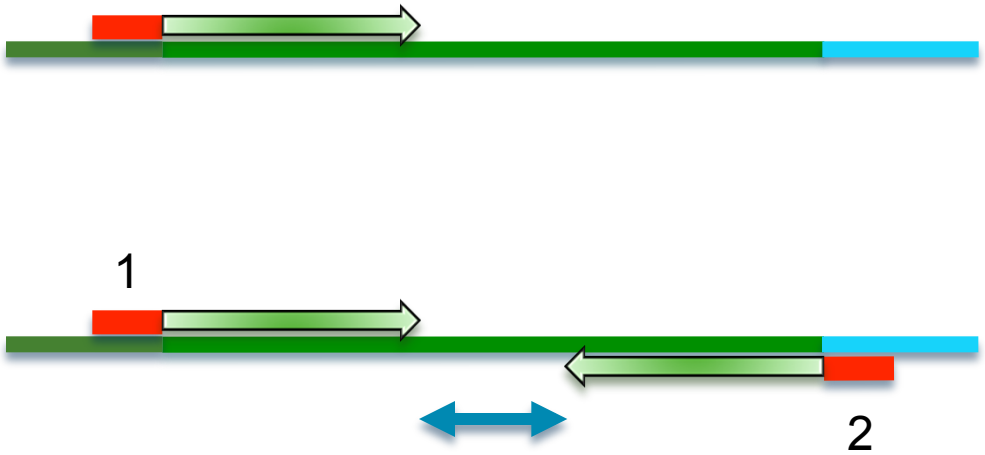
Illumina

National Institute of Allergy and Infectious Diseases

NIAID

MiSeq Sequencing Reads

- Types of reads
  - Single-end
  - Paired-end



(**or make them **overlap**)

- Length of reads (150 – 300 bp)
- 15-25 million reads (read pairs) per run

Outline

- MiSeq
- Variant analysis
  - PrimerID amplicon sequencing
    - Custom pipeline
  - Whole genome sequencing
    - Vprofiler, Vphaser
- Full genome assembly
  - VICUNA
  - SOAPdenovo2

PrimerID Method Paper

- **Accurate sampling and deep sequencing of the HIV-1 protease gene using a Primer ID**
- Proc Natl Acad Sci U S A. Dec 13, 2011; 108(50): 20166–20171.
- Cassandra B. Jabara,a,b,c Corbin D. Jones,a,d Jeffrey Roach,e Jeffrey A. Anderson,b,c,f,1 and Ronald Swanstromb,c,g,2

34 Citations in Pubmed for PrimerID Method

Accurate sampling and deep sequencing of the HIV-1 protease gene using a Primer ID

Cassandra B. Jabara, Corbin D. Jones, Jeffrey Roach, Jeffrey A. Anderson, Ronald Swanstrom

Proc Natl Acad Sci U S A. 2011 December 13; 108(50): 20166–20171. Published online 2011 November 30. doi: 10.1073/pnas.1110064108

PMCID: PMC3250168

Article PubReader PDF–1.1M Supplementary Material

Is Cited by the Following 34 Articles in this Archive: << Previous Page 1 of 2 Next >>

HIV-1 Quasispecies Delineation by Tag Linkage Deep Sequencing

Nicholas C. Wu, Justin De La Cruz, Laith Q. Al-Mawsawi, C. Anders Olson, Hangfei Qi, Harding H. Luan, Nguyen Nguyen, Yushen Du, Shuai Le, Ting-Ting Wu, Xinmin Li, Martha J. Lewis, Otto O. Yang, Ren Sun

PLoS One. 2014; 9(5): e97505. Published online 2014 May 19. doi: 10.1371/journal.pone.0097505

PMCID: PMC4026136

Article PubReader PDF–1.2M Supplementary Material

Risks of double-counting in deep sequencing

Michael W. Schmitt, Edward J. Fox, Jesse J. Salk

Proc Natl Acad Sci U S A. 2014 April 22; 111(16): E1560. Published online 2014 March 20. doi: 10.1073/pnas.1400941111

PMCID: PMC4000836

Currently embargoed: Free in PMC on Oct 22, 2014; PubMed

Comparison of Illumina and 454 Deep Sequencing in Participants Failing Raltegravir-Based Antiretroviral Therapy

Jonathan Z. Li, Brad Chapman, Patrick Charlebois, Oliver Hofmann, Brian Weiner, Alyssa J. Porter, Reshmi Samuel, Saran Vardhanabhuti, Lu Zheng, Joseph Eron, Babafemi Taiwo, Michael C. Zody, Matthew R. Henn, Daniel R. Kuritzkes, Winston Hide, and the ACTG A5262 Study Team, Cara C. Wilson, Baiba I. Berzins, Edward P. Acosta, Barbara Bastow, Peter S. Kim, Sarah W. Read, Jennifer Janik, Debra S. Meres, Michael M. Lederman, Lori Mong-Kryspin, Karl E. Shaw, Louis G. Zimmerman, Randi Leavitt, Guy De La Rosa, Amy Jennings

PLoS One. 2014; 9(3): e90485. Published online 2014 March 6. doi: 10.1371/journal.pone.0090485

PMCID: PMC3946168

Article PubReader PDF–744K Supplementary Material

Primer ID: What and Why

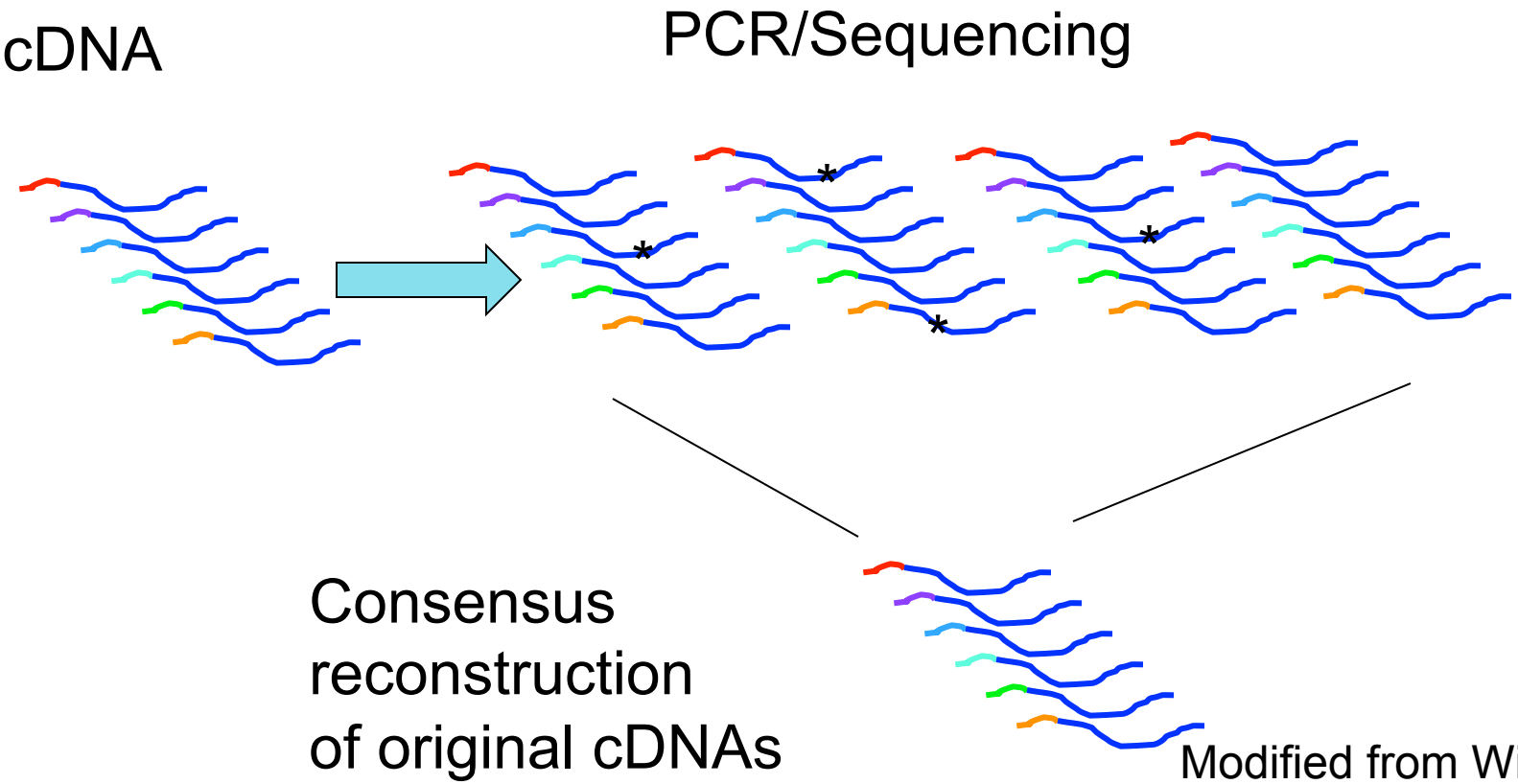
- What is a Primer ID?
- A Primer ID is a degenerate string of nucleotides introduced into a primer during the oligonucleotide synthesis reaction. As primers are synthesized *de novo*, a population of primers will contain unique combinations at that degenerate block. For example, a Primer ID containing a block of 8 degenerate bases will have 65,536 (48) unique combinations.
- Why use a Primer ID?
- Next generation high-throughput sequencing protocols require a large amount of starting genomic material. PCR is typically a necessary first step in sequencing viral populations, as templates are limiting. **During PCR, the polymerase will introduce errors into the viral population**. These errors will be reported by the high resolution of next generation sequencing platforms. **A Primer ID allows for tracking of individual viral genomes through the PCR and sequencing protocol and direct error correction**. Without a Primer ID, artifactual errors have to be removed from biological diversity through statistical means.

A unique barcode for each cDNA

Illumina Adaptor | Sample ID | Primer ID | Flu Sequence

CCATCTCATCCCTGCGTGTCTCCGACTCAG [NNN]GCNNNNNNNNNNAAGCAGTTTTTACAGAAATTTGC

10mer = 1,048,576 unique barcodes per sample



Modified from Will Ince

PrimerID Sequencing Library Preparation

Gene-specific Forward Primer with PrimerID:

ACACTCTTTCCCTACACGACGCTCTTCCGATCT **NNNNNNNNNNN** CA [Region-specific forward]

PrimerID

Gene-specific Reverse Primer:

GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT NNNN [Region-specific reverse]

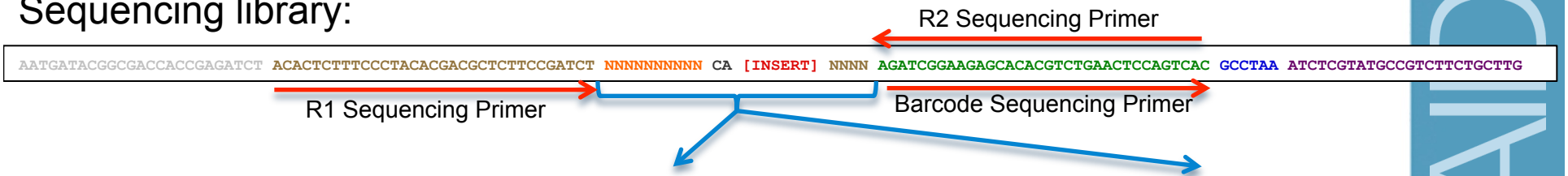
Universal Forward Graft Primer:

AATGATACGGCGACCACCGAGATCT ACACTCTTTCCCTACACGACGCTC

Barcoded Reverse Graft Primer:

CAAGCAGAAGACGGCATACGAGAT TTAGGC GTGACTGGAGTTCAGACGTGTGCTC

Sequencing library:



Paired-end reads:

R1

NNNNNNNNNN CA [INSERT (part)]
NNNNNNNNNN CA [INSERT (part)]
NNNNNNNNNN CA [INSERT (part)]
NNNNNNNNNN CA [INSERT (part)]
NNNNNNNNNN CA [INSERT (part)]
NNNNNNNNNN CA [INSERT (part)]
NNNNNNNNNN CA [INSERT (part)]

R2

[INSERT (part)] NNNN
[INSERT (part)] NNNN
[INSERT (part)] NNNN
[INSERT (part)] NNNN
[INSERT (part)] NNNN
[INSERT (part)] NNNN
[INSERT (part)] NNNN

National Institute of Allergy and Infectious Diseases

NIAID

PrimerID Analysis

1. Merge Paired-end reads

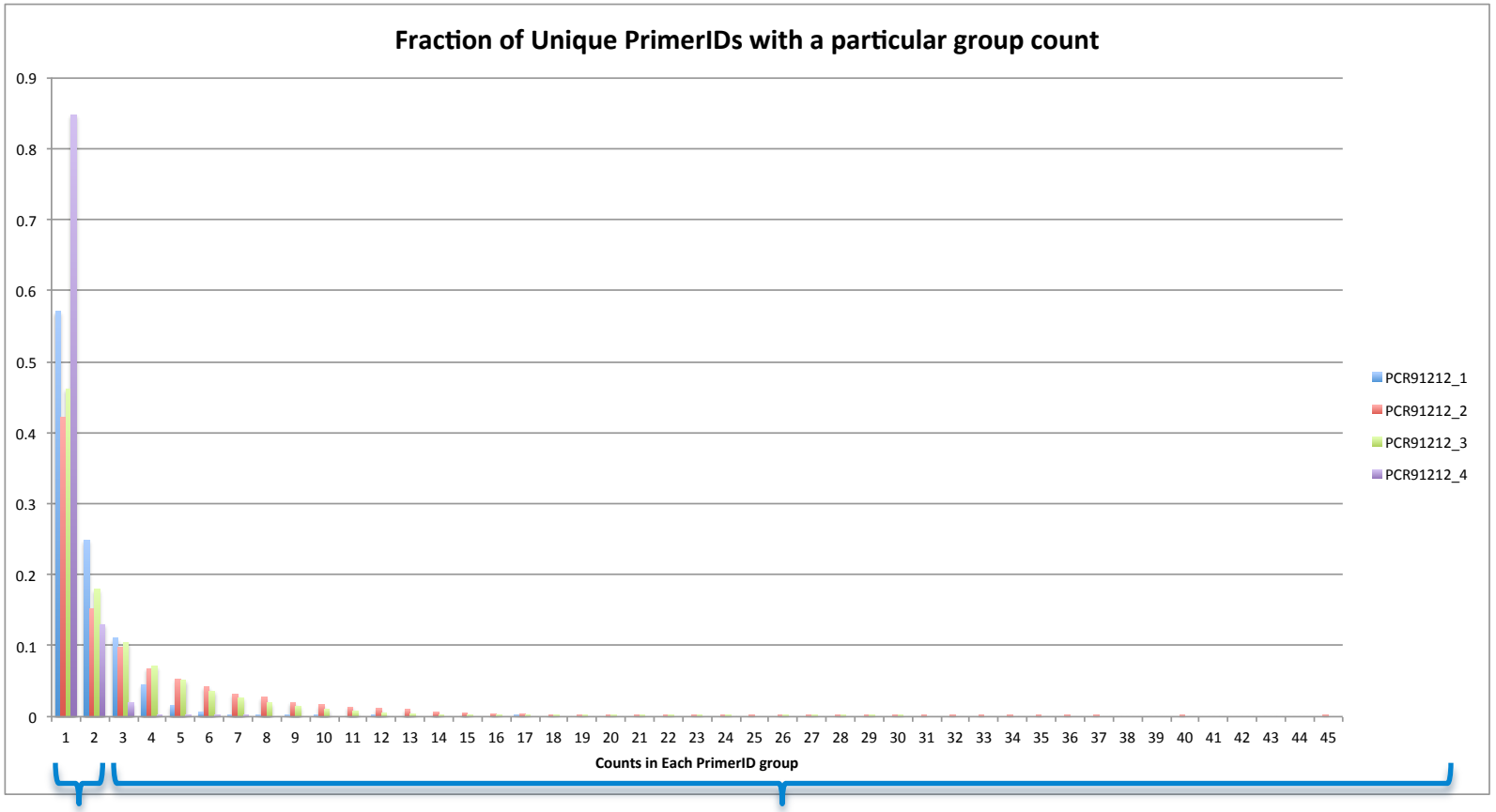
2. Create a consensus for each barcode group (Remove PCR and sequencing errors)

3. Call variants and determine linkage, etc.

National Institute of Allergy and Infectious Diseases

NIAID

A minimum of 3 reads per group is required to call a consensus sequence



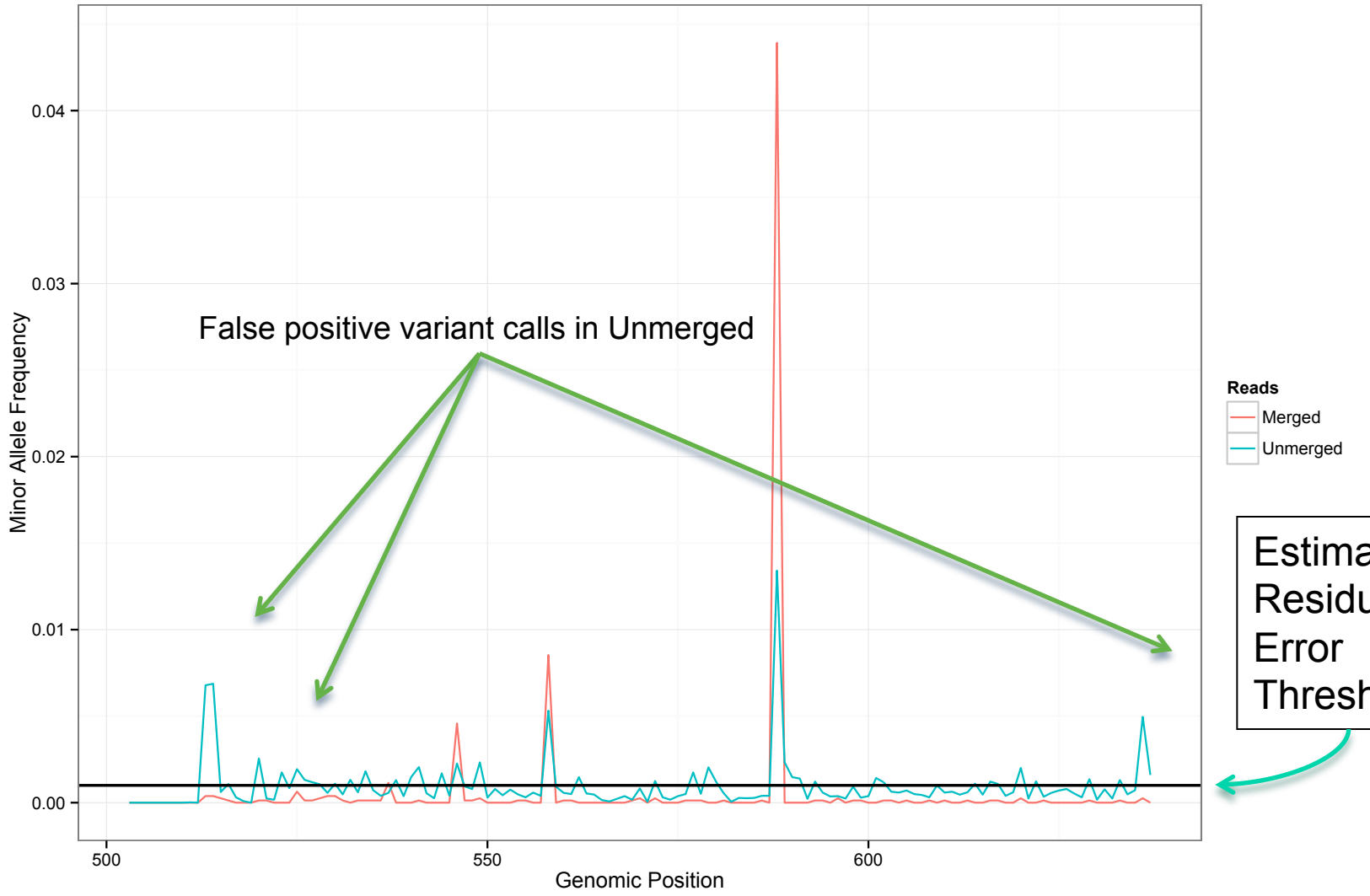
Toss these

Keep PrimerID groups with minimum = 3, maximum = ? (default max = 50)

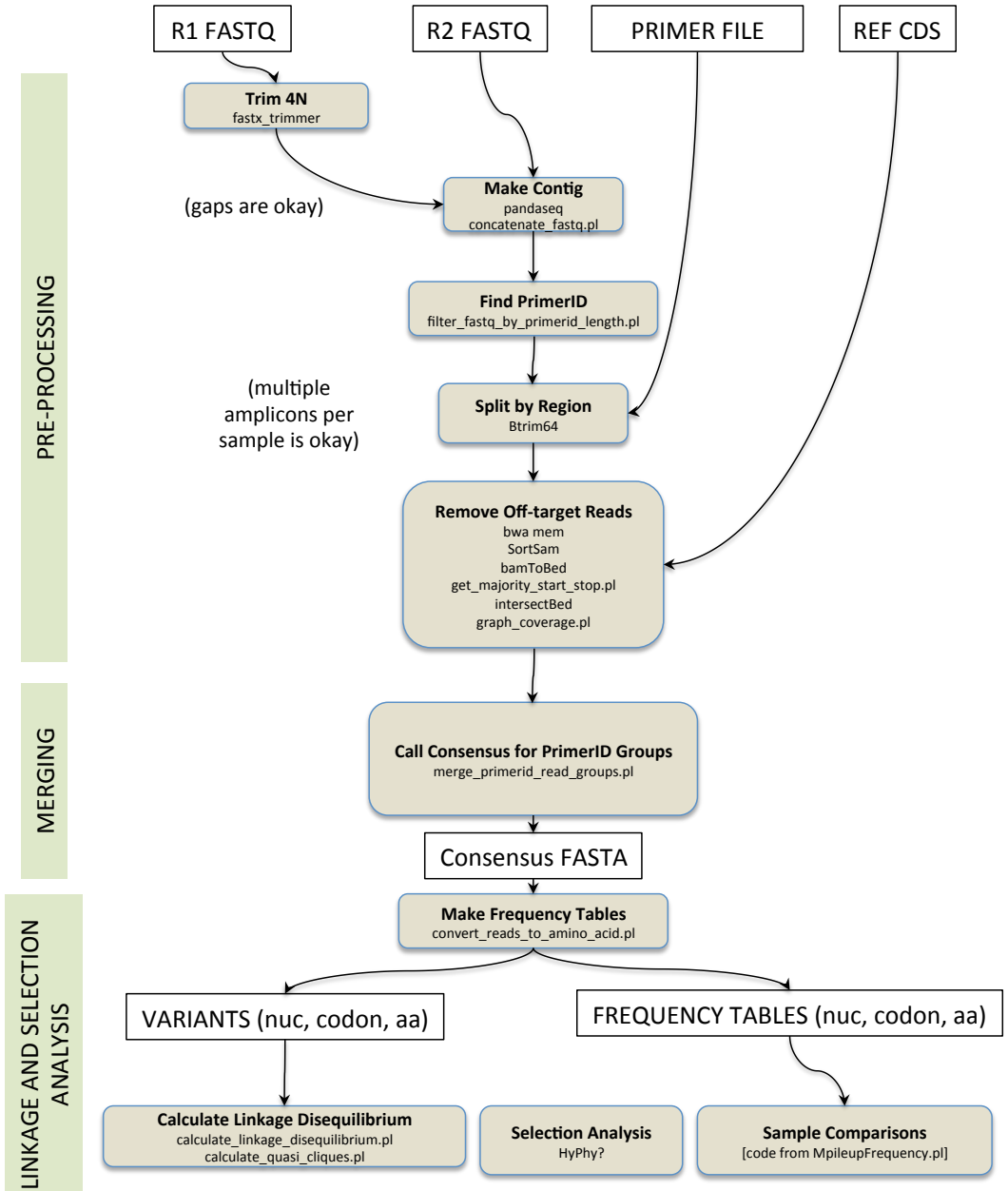
NIH Institute of Allergy and Infectious Diseases

NIAID

Merging Reads with PrimerID Reduces Noise



PrimerID Workflow



National Institu
Allergy and
Infectious Disea

NIAID

Steps in the Pipeline

- Trim R2 reads
- Merge R1 & R2
  - Concatenate overlapping reads (pandaseq), OR
  - Concatenate non-overlapping reads
- Find reads containing primerID-linker
- Split into barcode regions and remove primer sequences
- Align to reference with BWA, convert to BAM
- Graph output coverage files
- Merge the reads by PrimerID and call consensus
- Convert merged reads to codons and amino acids and get frequency tables and cleaned alignment files
- Statistical analysis for linkage of variants, etc.
- Positive selection analysis with HyPhy (still working on this...)

Configuration files for input

Samples file (required):

Sample_prefix	CDS_Reference_file	Primer_file	Group
Sample_1	Seq12_093009_HA_cds.fa	Seq12_primers.txt	E1
Sample_2	Seq12_093009_HA_cds.fa	Seq12_primers.txt	E1
Sample_3	Seq12_093009_HA_cds.fa	Seq12_primers.txt	E1
Sample_4	Seq12_093009_HA_cds.fa	Seq12_primers.txt	Parent
Sample_5	Seq12_093009_HA_cds.fa	Seq12_primers.txt	Parent
Sample_6	Seq12_093009_HA_cds.fa	Seq12_primers.txt	B7_E2
Sample_7	Seq12_093009_HA_cds.fa	Seq12_primers.txt	B7_E2
Sample_8	Seq12_093009_HA_cds.fa	Seq12_primers.txt	B7_E2
Sample_9	Seq12_093009_HA_cds.fa	Seq12_primers.txt	Parent_B7
Sample_10	Seq12_093009_HA_cds.fa	Seq12_primers.txt	Parent_B7
Sample_11	CAL0409_HA_cds.fa	CAL0409_primers.txt	Mock_12_1_12
Sample_12	CAL0409_HA_cds.fa	CAL0409_primers.txt	Mock_12_1_12
Sample_13	CAL0409_HA_cds.fa	CAL0409_primers.txt	Mock_12_1_12
Sample_14	CAL0409_HA_cds.fa	CAL0409_primers.txt	Immun_12_1_12
Sample_15	CAL0409_HA_cds.fa	CAL0409_primers.txt	Immun_12_1_12
Sample_16	CAL0409_HA_cds.fa	CAL0409_primers.txt	Immun_12_1_12

Comparisons file (optional):

Comparison	Treatment	Control
1	E1	Parent
2	B7_E2	Parent_B7
3	Immun_12_1_12	Mock_12_1_12

Needs to match Samples file

Reference coding sequence (required), e.g., Seq12_093009_HA_cds.fa

Primers file(s) (required), e.g., Seq12_primers.txt:

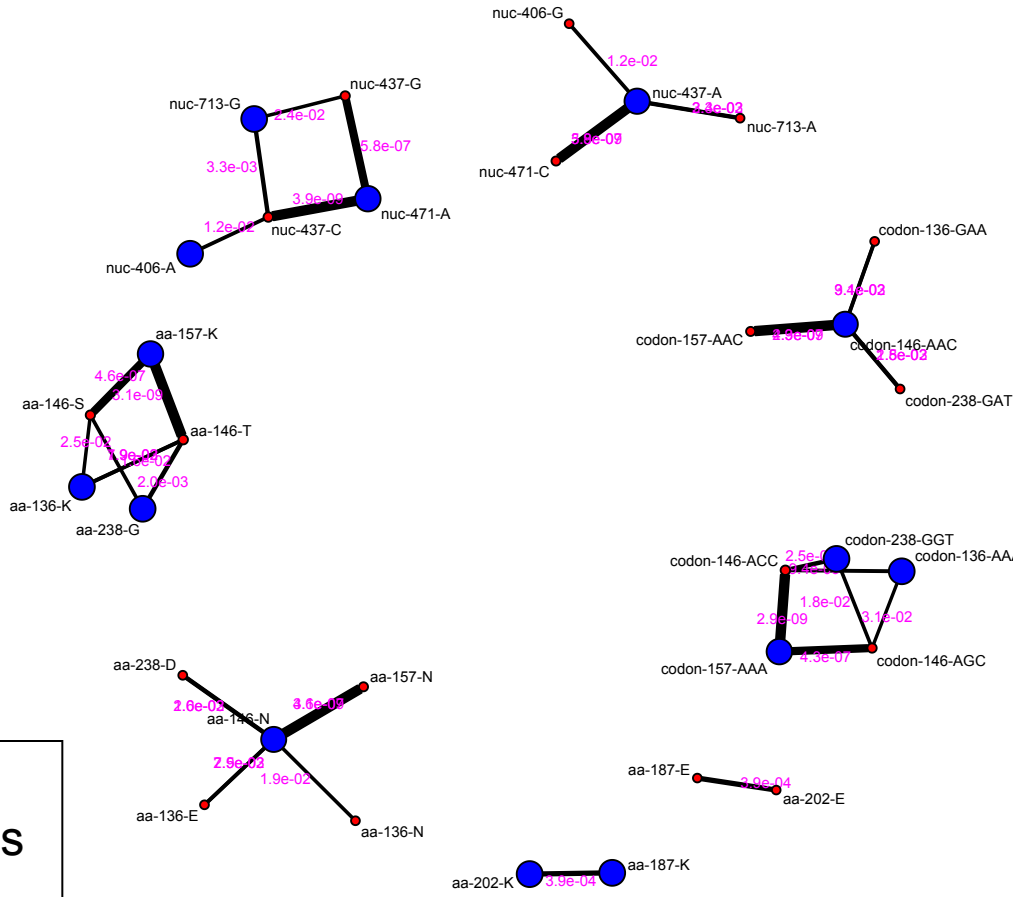
AGCAGGGGAAAATAAAAACAACCAAAATG	GTAACGGCAGCATGCTCCCATGAGGGGAAA	Seq12_093009_miseq_amp4_441
GCTGAGGGAGCAATTGAGCTCAGTGTCATC	GCACTGAGTAGAGGCTTTGGGTCCGGCATC	Seq12_Amplicon2_381_813

- (Still a work-in progress... formats may change)

Output files

- Frequency tables (for each sample)
  - Nucleotide
  - Codon
  - Amino Acid
  - Merged
- Variant (nuc, codon, amino acid; for each sample)
- Linkage between variants above a threshold (with p-values)
  - graphical representation
- Sample comparisons for changes in nuc/codon/amino acid (with p-values) between samples

Graphical Depiction of Linked Sequences



Red = variant
Blue = consensus

Thickness of edge
~ significance

NIAID

PrimerID Pipeline “To do” list

- Deal with replicates in linkage analysis and in sample comparisons
- Add more graphs to be printed out automatically
- Add tree-building with RAXML and positive selection analysis using HyPhy
- Other ideas?

Outline

- MiSeq
- Variant analysis
  - PrimerID amplicon sequencing
    - Custom pipeline
  - Whole genome sequencing
    - Vprofiler, Vphaser
- Full genome assembly
  - VICUNA
  - SOAPdenovo2

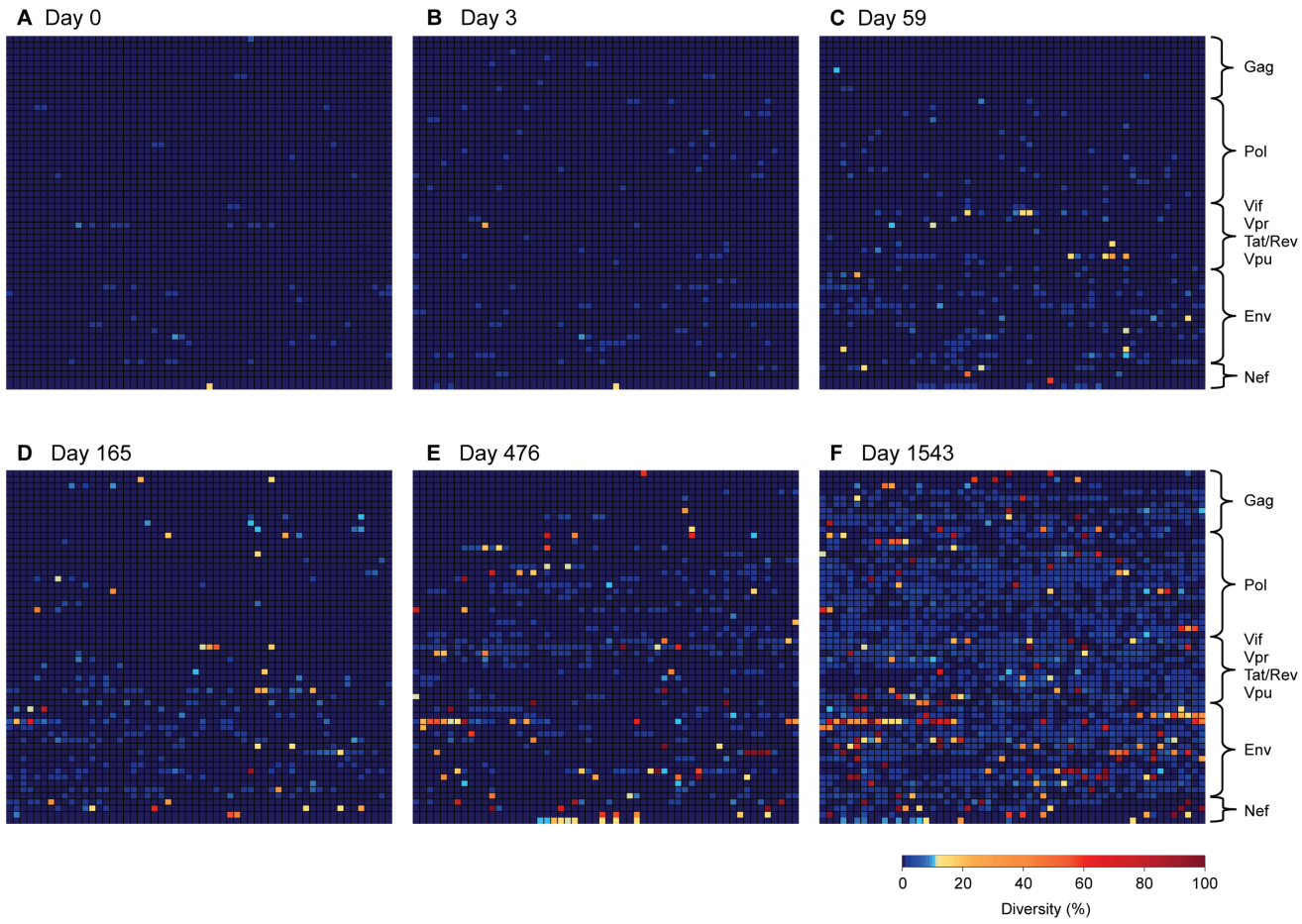
Vprofiler output table

1	Nt Position	AA Position	Coverage(HQOnly)	ConsensusCo	Primary Codon	Secondary Codon	Tertiary Cod	4th Codon	5th Codon	6th Codon	7th Codon	8th Codon	% Primary Codon	% Secondary	% Tertiary Co
2	>PCR1	HA													
161	507	159	70230(55972)	AGT(S)	S (100.00%)(AGT)								100		
162	510	160	70195(55856)	TTT(F)	F (99.99%)(TTT)	S (0.01%)(TCT)							99.99	0.01	
163	513	161	1218(55454)	TAC(Y)	Y (87.93%)(TAC)	I (7.14%)(ATC)	N (2.63%)(AA	T (1.64%)(AC	H (0.66%)(CAC)				87.93	7.14	2.63
164	516	162	1481(34103)	AGA(R)	R (100.00%)(AGA)								100		
165	519	163	1590(34952)	AAT(N)	N (100.00%)(AAT)								100		
166	522	164	69561(243)	TTG(L)	L (97.75%)(TTG)	F (2.25%)(TTT)							97.75	2.25	
167	525	165	67852(31696)	CTA(L)	L (99.83%)(CTA)	L (0.15%)(TTA)	L (0.01%)(CTT)						99.83	0.15	0.01
168	528	166	67842(47947)	TGG(W)	W (99.99%)(TGG)	C (0.01%)(TGC)							99.99	0.01	
169	531	167	67854(50456)	CTG(L)	L (99.99%)(CTG)	V (0.01%)(GTG)							99.99	0.01	
170	534	168	67854(54250)	ACG(T)	T (100.00%)(ACG)								100		
171	537	169	67873(58420)	GAG(E)	K (99.74%)(AAG)	E (0.26%)(GAG)							99.74	0.26	
172	540	170	68286(68457)	AAG(K)	K (99.12%)(AAG)	R (0.88%)(AGG)							99.12	0.88	
173	543	171	68032(67532)	GAG(E)	E (99.77%)(GAG)	G (0.23%)(GGG)							99.77	0.23	
174	546	172	67826(64840)	GGC(G)	S (99.58%)(AGC)	G (0.41%)(GGC)	A (0.00%)(GCC)						99.58	0.41	0
175	549	173	67746(64605)	TCA(S)	S (99.75%)(TCA)	P (0.25%)(CCA)							99.75	0.25	
176	552	174	67687(65840)	TAC(Y)	Y (99.97%)(TAC)	S (0.03%)(TCA)	* (0.01%)(TA	S (0.00%)(TCC)					99.97	0.03	0.01
177	555	175	67753(66410)	CCA(P)	P (99.92%)(CCA)	P (0.08%)(CCG)							99.92	0.08	
178	558	176	67533(68325)	AAG(K)	E (99.66%)(GAG)	K (0.19%)(AAG)	G (0.15%)(GG	R (0.00%)(AGG)					99.66	0.19	0.15
179	561	177	67376(67479)	CTG(L)	L (99.90%)(CTG)	L (0.05%)(TTG)	L (0.05%)(CTA)						99.9	0.05	0.05
180	564	178	69967(65751)	AAA(K)	K (97.23%)(AAA)	E (2.76%)(GAA)	N (0.02%)(AAT)						97.23	2.76	0.02
181	567	179	67900(67150)	AAT(N)	N (98.63%)(AAT)	I (1.37%)(ATT)							98.63	1.37	
182	570	180	66920(65817)	TCT(S)	S (99.82%)(TCT)	S (0.18%)(TCC)							99.82	0.18	
183	573	181	66919(23810)	TAT(Y)	Y (100.00%)(TAT)								100		
184	576	182	66826(610)	GTG(V)	V (100.00%)(GTG)								100		
185	579	183	69724(23159)	AAC(N)	N (95.67%)(AAC)	E (2.66%)(GAA)	T (1.40%)(AC	D (0.21%)(GA	N (0.04%)(AA	K (0.02%)(AA	T (0.00%)(ACC)		95.67	2.66	1.4
186	582	184	68147(16056)	AAA(K)	K (99.64%)(AAA)	K (0.22%)(AAG)	R (0.09%)(AG	E (0.05%)(GA	I (0.01%)(ATA	N (0.00%)(AAT)			99.64	0.22	0.09
187	585	185	70220(62813)	AAA(K)	K (98.63%)(AAA)	K (1.11%)(AAG)	R (0.12%)(AG	R (0.07%)(AG	E (0.07%)(GA	G (0.00%)(GG	* (0.00%)(TAA)		98.63	1.11	0.12
188	588	186	70206(64301)	GGG(G)	K (93.70%)(AAG)	K (3.05%)(AAA)	R (1.74%)(AG	E (0.95%)(GA	G (0.39%)(GG	G (0.15%)(GG	R (0.01%)(AG	E (0.00%)(GA	93.7	3.05	1.74
189	591	187	66357(59538)	AAA(K)	E (97.93%)(GAA)	K (2.07%)(AAA)							97.93	2.07	
190	594	188	1418(858)	GAA(E)	E (100.00%)(GAA)								100		

National Institute of Allergy and Infectious Diseases

NIAID

Vphaser output graphic



National Institute of Allergy and Infectious Diseases

Henn et al., PLoS Pathogens, 2012

NIAID

Outline

- MiSeq
- Variant analysis
  - PrimerID amplicon sequencing
    - Custom pipeline
  - Whole genome sequencing
    - Vprofiler, Vphaser
- Full genome assembly
  - VICUNA
  - SOAPdenovo2

De novo assembly

- VICUNA
  - from BROAD
  - quasi reference-based assembly, designed for Viral population
- SOAPdenovo2
  - Had good success with this using whole-genome Influenza RNA-seq (1-2000x coverage?)

Thanks!

Questions?

andrew.oler@nih.gov