Estimation of Viral Population Structure from Amplicon-Based Reads

Nicholas Mancuso

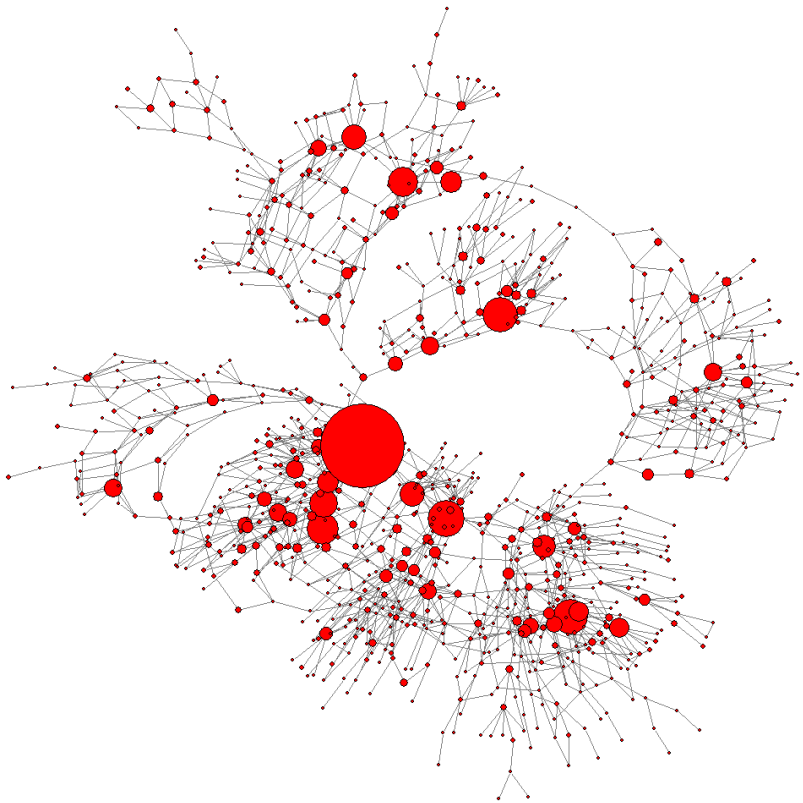
Department of Computer Science
Georgia State University

June 8, 2012

Overview

Viral Quasispecies

- RNA virus replication relies on RNA polymerase
- High mutation rate ($\approx 10^{-4}$)
- Recombination events occur
- HIV, HCV, Influenza

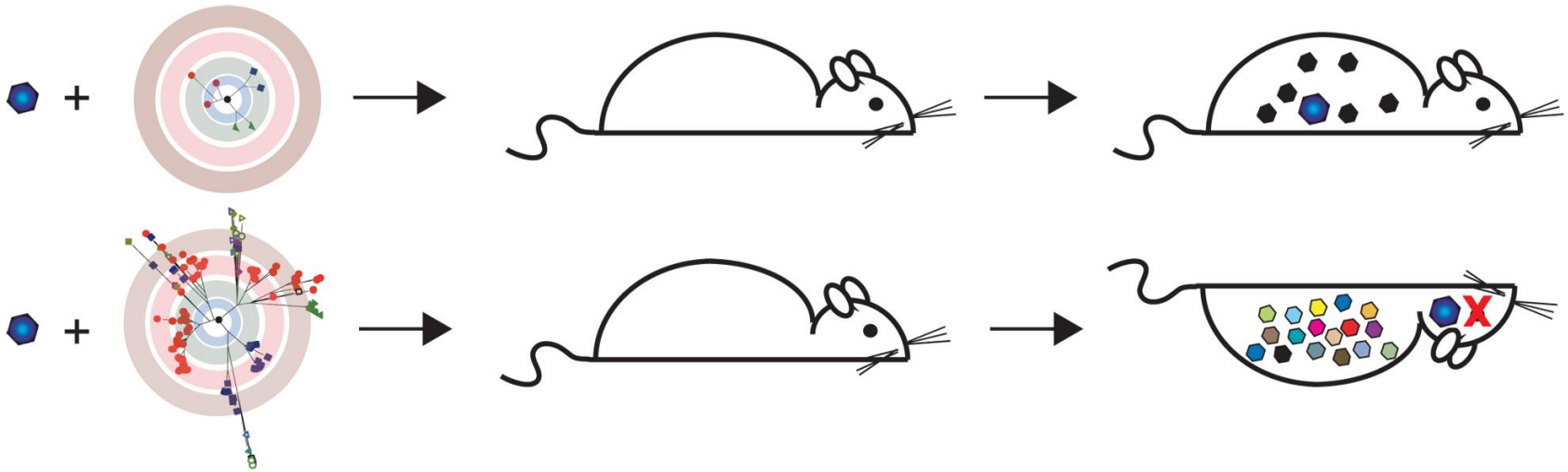


Georgia State University

Viral Quasispecies

Populations may differ in

- Virulence
- Escape immune response
- Resistance to antiviral therapies



Lauring AS, Andino R. PLoS Pathog. 2010

Georgia State University

Hepatitis C

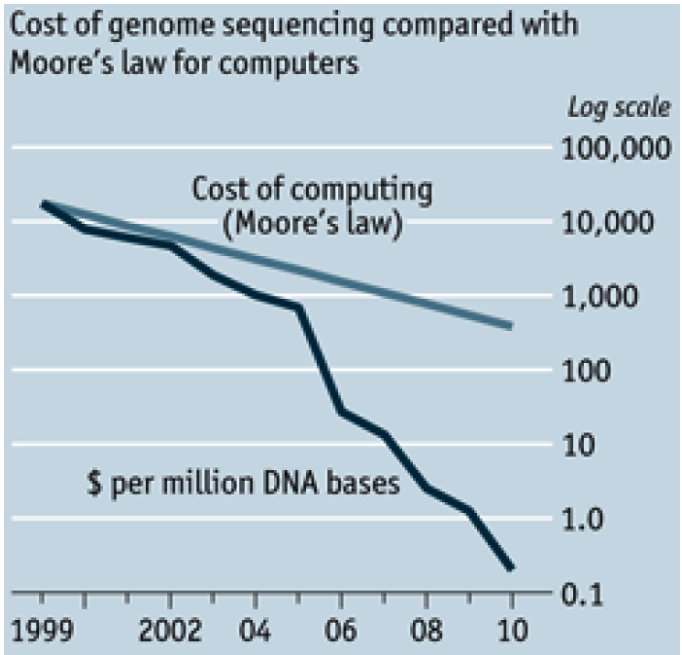
HCV infects 2.2% of the world's population

- No vaccine
- Current interferon and ribavirin therapy effective in 50%-60% of patients
- Therapy is expensive and uncomfortable

Skums et. al., CAME 2011

- Prediction method for interferon outcome
- Highly dependent on accuracy of viral population structure

High Throughput Sequencing



http://www.economist.com/node/16349358

Illumina HiSeq 2000
Up to 6 billion PE reads/run

35 — 100bp read length

Ion Proton Sequencer
Up to 10 billion reads/run

20 — 200bp read length

Roche/454 FLX Titanium
1 million reads/run

400 — 600bp read length

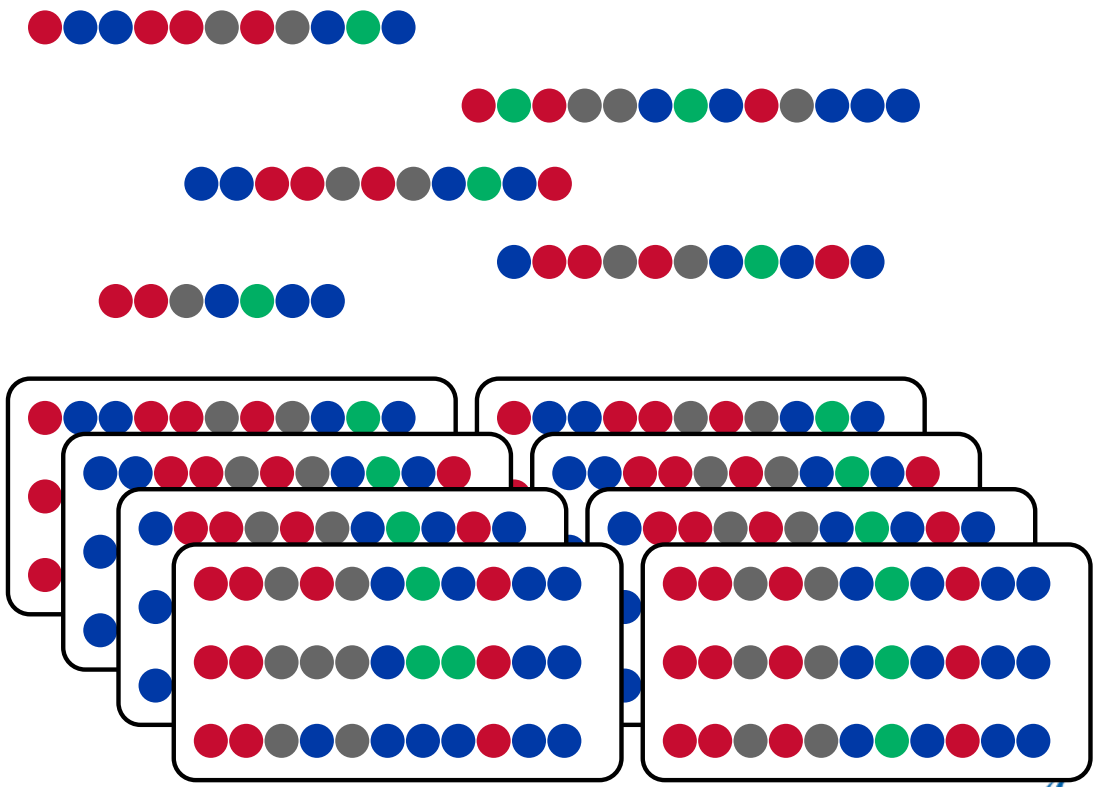
SOLiD 4
1.4-2.4 billion PE reads/run

35 — 50bp read length

Georgia State University

Shotgun and Amplicon Reads

- Shotgun reads have start positions distributed uniformly
- Amplicon reads have start/end positions determined by allele-specific primers

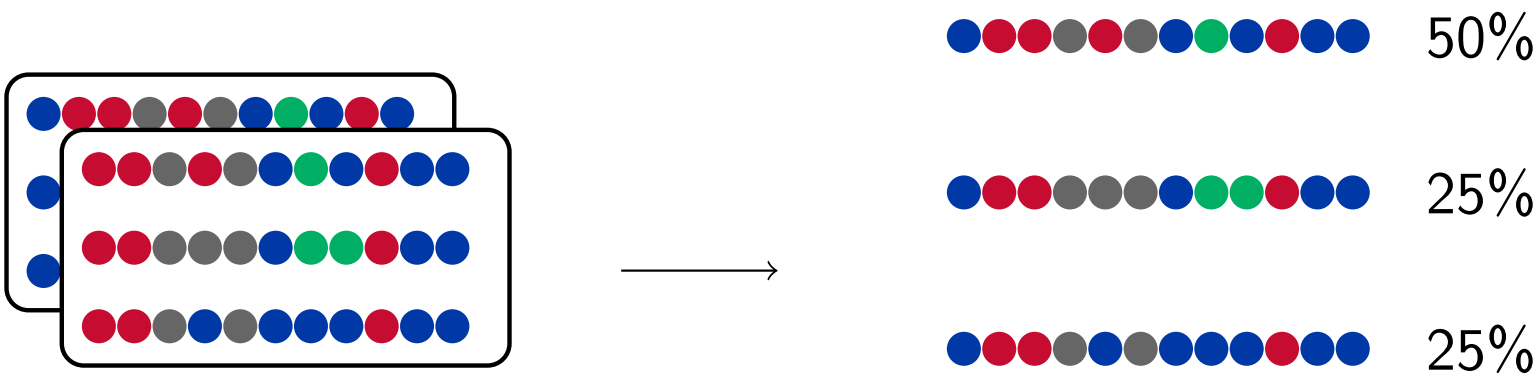


Georgia State University

Viral Quasispecies Reconstruction Problem

Problem

Given a collection of amplicon reads generated from a viral sample, assemble the quasispecies, i.e., the set of sequences and respective frequencies of the sample population.



Viral Quasispecies Reconstruction

Local Reconstruction

- Focus on primer-flanked region
- KEC, QuasiRecomb, *k*GEM

Global Reconstruction

- Focus on larger genomic regions
- Typically use read-graph approach to "stitch" locally reconstructed regions together
- ShoRAH, ViSpA, QuRe

VirA is a tool for global quasispecies reconstruction

Global Viral Reconstruction Challenges

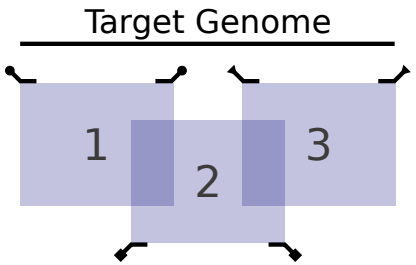
Conserved Regions

Relatively few mutations in long regions obfuscate true population

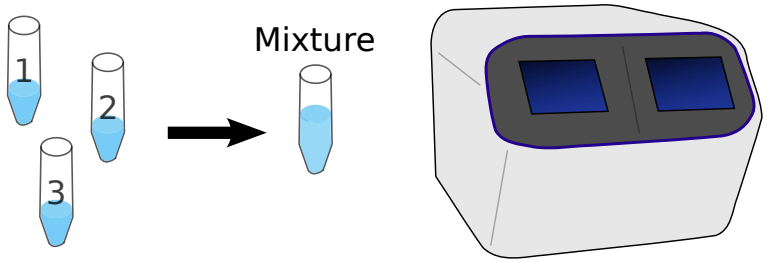
Sequencing Errors

- Homopolymer errors
- Base call errors
- Insertion errors
- Deletion errors

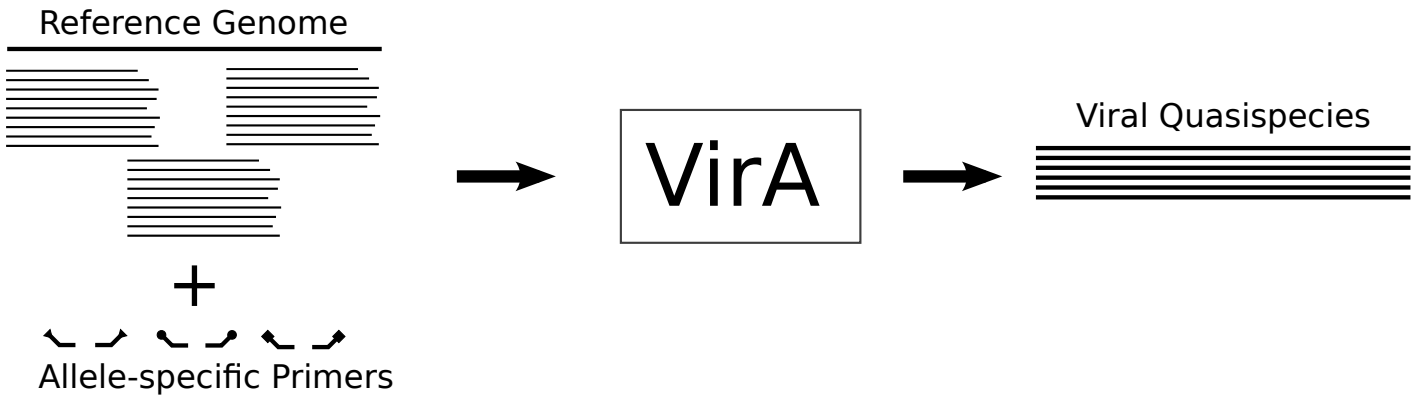
Library Preparations & Workflow



a)



b)



c)

VirA Workflow

1. Align reads to reference
2. Align allele-specific (target) primers to reference
3. Infer amplicon intervals from primers
4. Locally reconstruct in each amplicon
5. Globally reconstruct over read-overlap graph

Georgia State University

Amplicon Inference

Infer amplicons from flanked regions

- Each pair forms interval
- Impose ordering over intervals
- Read belongs in interval if covers significant sub-interval & overlaps with neighboring intervals

Local Reconstruction/Error Correction

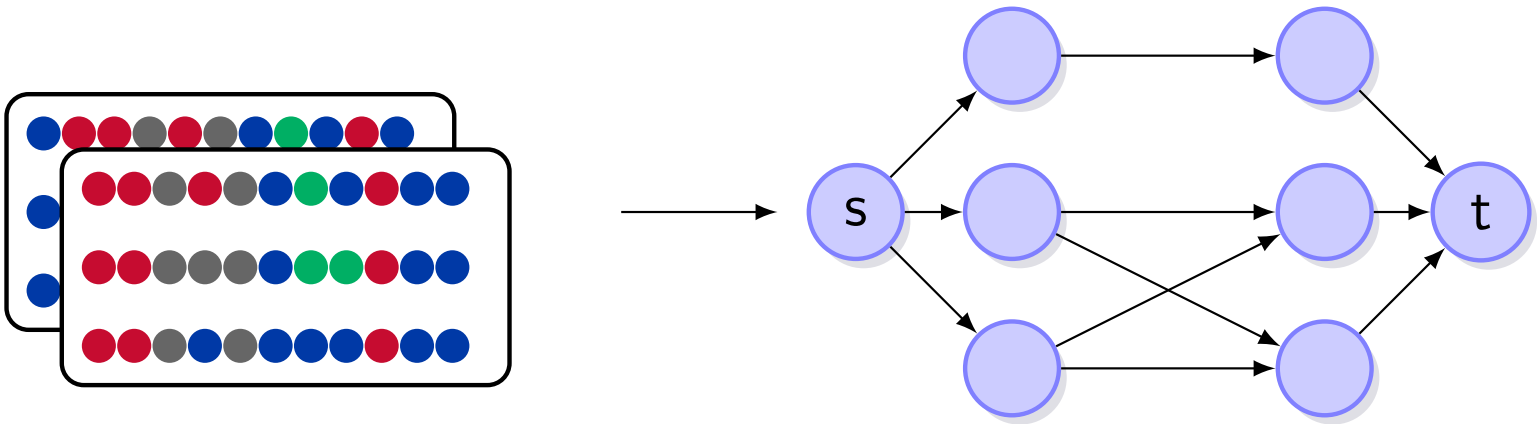
Correct errors with kGEM

- Cluster reads by hamming distance
- Produce local consensuses
- Estimate consensus frequencies
- Estimate allele frequencies
- Repeat until convergence

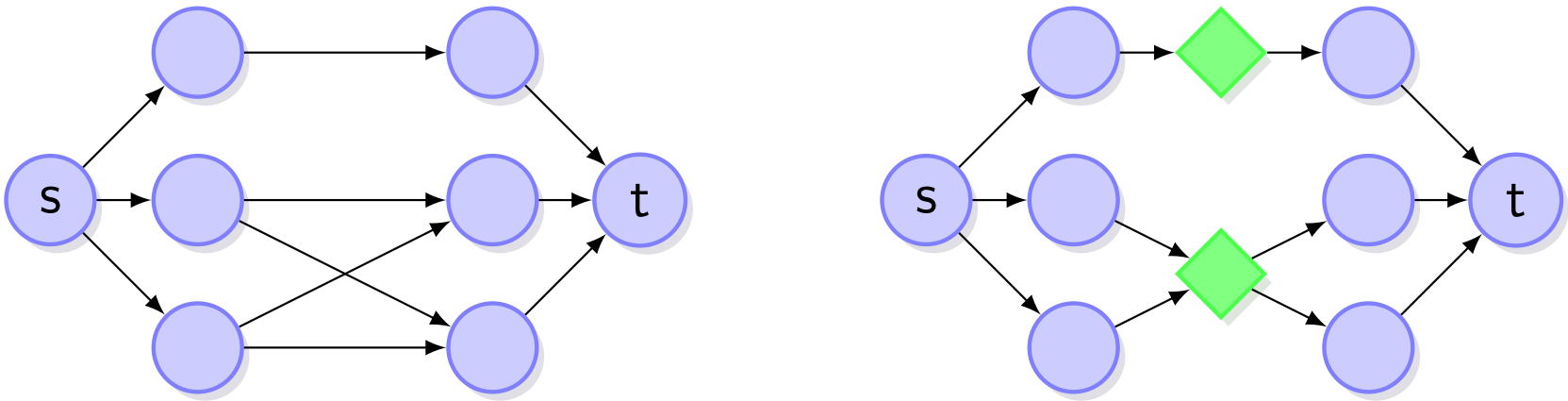
Read Graph

K amplicons represented by K-staged read graph

- Vertices $\Leftrightarrow$ distinct reads
- Edges $\Leftrightarrow$ reads with consistent overlap
- Vertices have count function $c(v)$

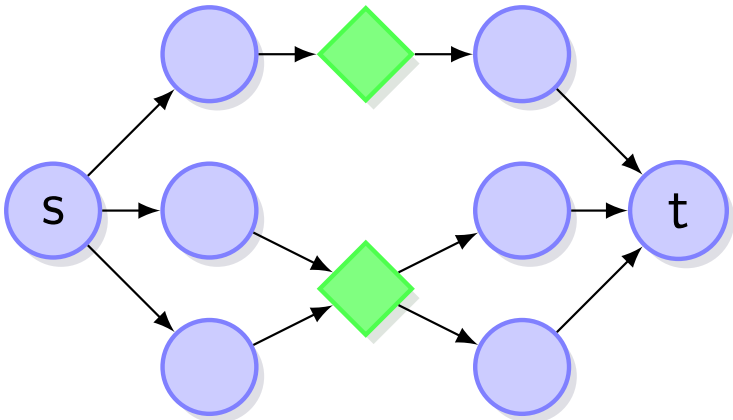


Graph Transformation



Maximum-Bandwidth Paths

- Find simple path containing most possible flow
- Repeat until graph is saturated
- Modified Dijkstra's algorithm
- Mancuso et al, In Silico 2012



Georgia State University

Multi-commodity Flow Formulation

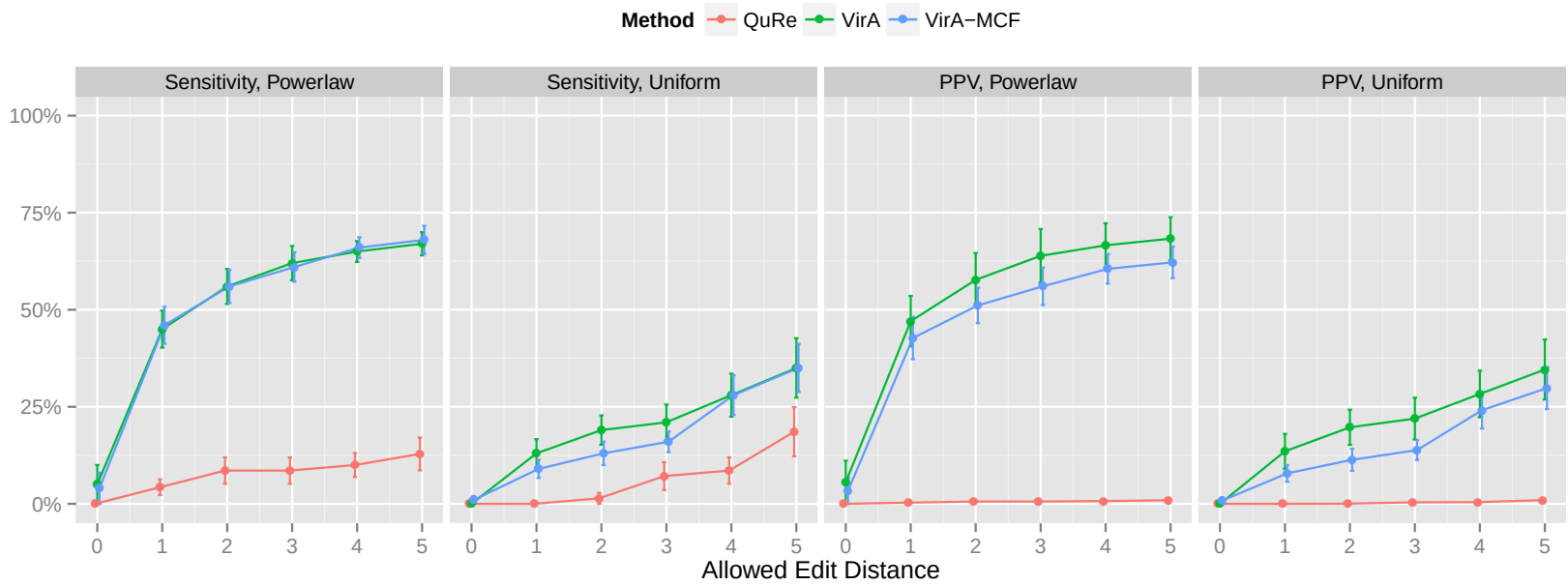
- Finding maximum-bandwidth paths
- Multi-commodity flow k = upper bound on variants
- Minimizes total flow while covering all reads
- ILP on CPLEX
- Skums et al, BMC Bioinformatics 2013

Experimental Setup

- 1734bp HCV E1E2 region
- 43 sequences $\rightarrow$ 10 datasets of 10 variants
- Abundance followed powerlaw ($\alpha = 2$) and uniform distributions
- 7-12 amplicons to cover region
- Reads generated with Grinder version 0.5
- Compared with QuRe [Prosperi et al]

Georgia State University

Results


Method
QuRe
VirA
VirA−MCF
Sensitivity, Powerlaw
Sensitivity, Uniform
PPV, Powerlaw
PPV, Uniform
100%
75%
50%
25%
0%
0
1
2
3
4
5
Allowed Edit Distance

Conclusions and Future Work

- Global quasispecies reconstruction is difficult
- VirA
- `http://alan.cs.gsu.edu/vira`

Georgia State University

Thanks

University of Connecticut

Dr. Ion Măndoiu

Centers for Disease Control and Prevention

Dr. Pavel Skums

Dr. Yuri Khudyakov

Georgia State University

Dr. Alex Zelikovsky

Alex Artyomenko

Bassam Tork

Rest of NGS group

Funding

Q and A

Thank you!

Questions?