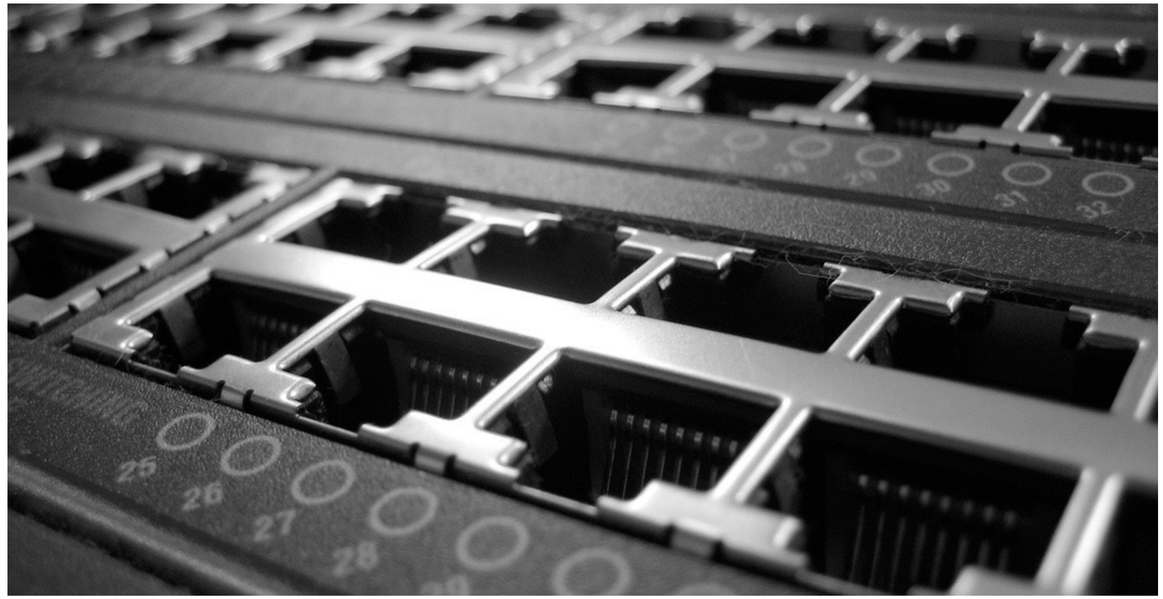
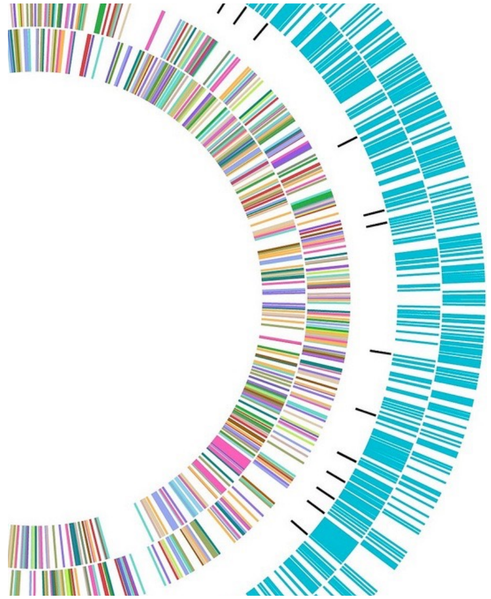




Truthiness in genomic variation

Oliver Hofmann

Harvard T.H. Chen School of Public Health

University of Glasgow

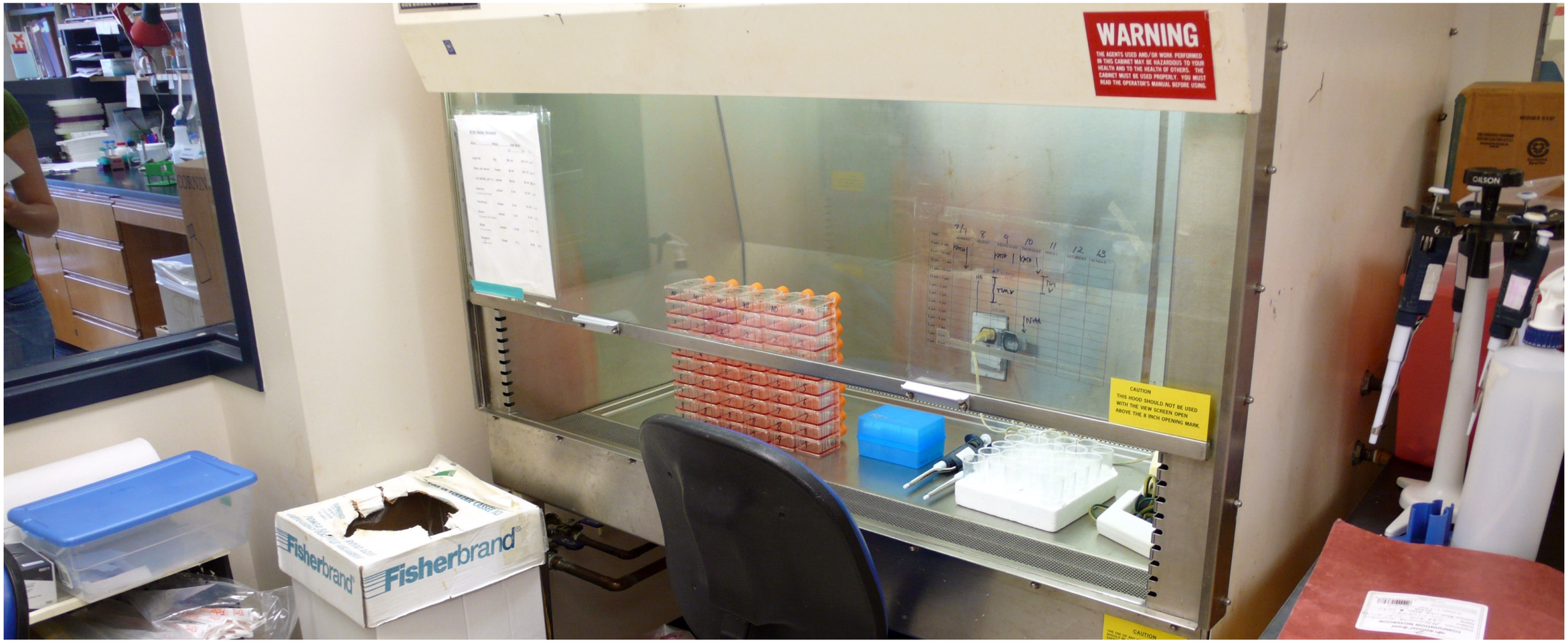
HARVARD | SCHOOL OF PUBLIC HEALTH

Powerful ideas for a healthier world

<intro>

Training as bench scientist

Why am I here… ?


WARNING
THE AGENTS USED AND/OR WORK PERFORMED IN THIS CABINET MAY BE HAZARDOUS TO YOUR HEALTH AND TO THE HEALTH OF OTHERS. THE CABINET MUST BE USED PROPERLY. YOU MUST READ THE OPERATOR'S MANUAL BEFORE USING.
CAUTION
THIS HOOD SHOULD NOT BE USED WITH THE VIEW SCREEN OPEN ABOVE THE 8 INCH OPENING MARK.
Fisherbrand
Fisherbrand
CORNING
GILSON

From bench…


commodore VIC 20
POWER
CTRL
RESTORE
RUN STOP
SHIFT LOCK
RETURN
SHIFT
CLR HOME
INST DEL
CRSR
f 1
f 3
f 5
f 7

… to desk

Methods

For validation purposes we used three biologists to annotate a corpus of domain-specific literature.

Collaborate with computer scientists?

PhD project: text mining

ochronosis… caused by an inherited lack of homogentisic acid oxidase

Natural Language Processing

ochronosis… caused by an inherited lack of homogentisic acid oxidase

Renal fibrosis biopsies were digested with trypsin

Natural Language Processing

ochronosis…	caused by	an inherited lack of	homogentisic acid oxidase
Disease or Syndrome	*Finding*		Protein
Renal fibrosis	biopsies	were digested with	trypsin

Natural Language Processing

ochronosis…	caused by	an inherited lack of	homogentisic acid oxidase
Disease or Syndrome	*Finding*		Protein
Renal fibrosis	biopsies	were digested with	trypsin
Disease or Syndrome		*Experimental Procedure*	Protein

Natural Language Processing

Methods

For validation purposes we used three biologists to annotate a corpus of domain-specific literature.

Methods, redux

</intro>

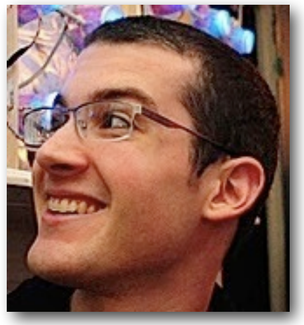
Oliver Hofmann

Shannan Ho Sui

John Hutchinson

Lorena Pantano

Meeta Mistry



John Morrissey

Rory Kirchner

Brad Chapman

Radhika Khetani

Mary Piper

Andreas Sjödin

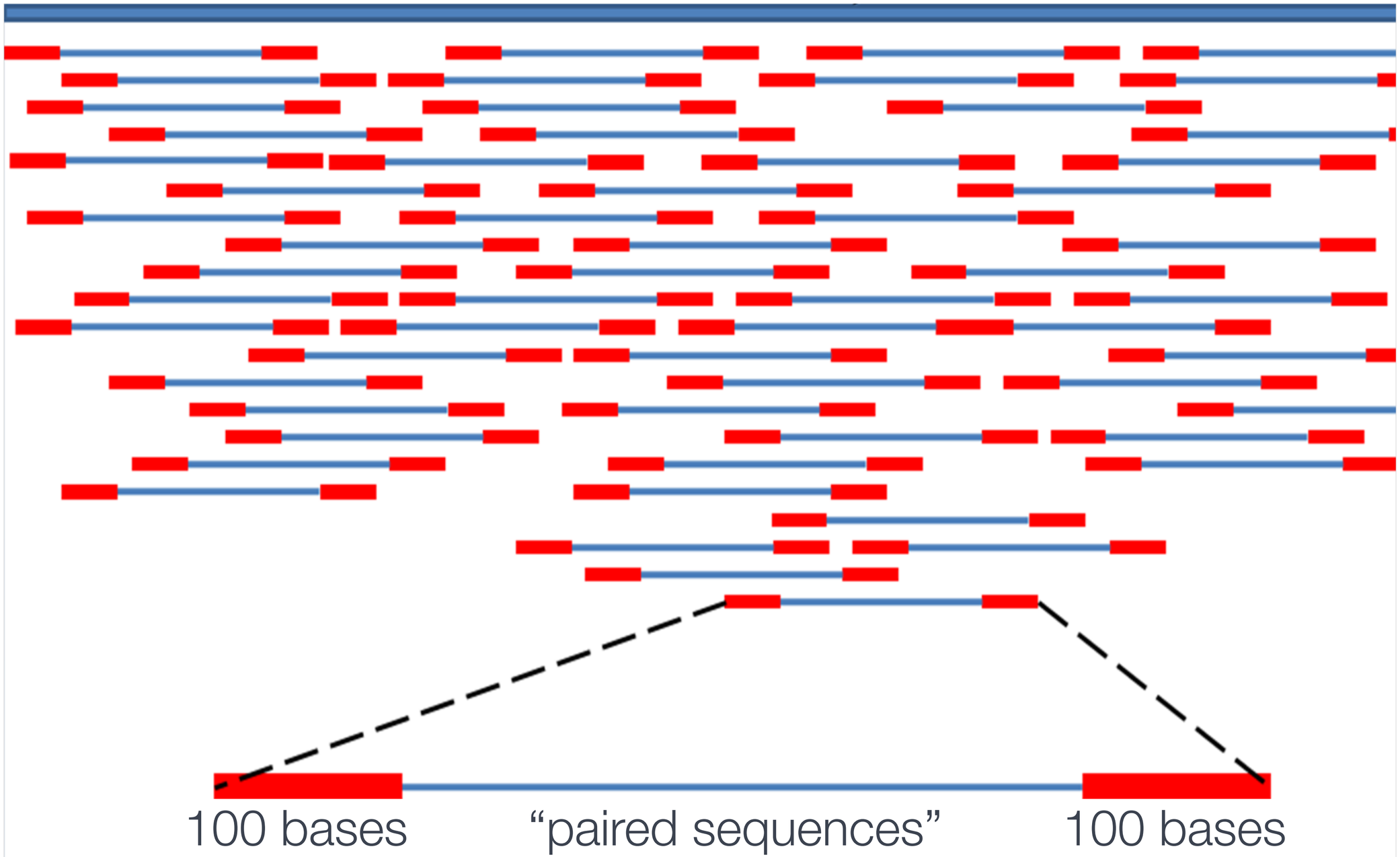
Winston Hide

Finding systematic differences

Choi, Genetic diagnosis by whole exome capture and massively parallel DNA sequencing, PNAS

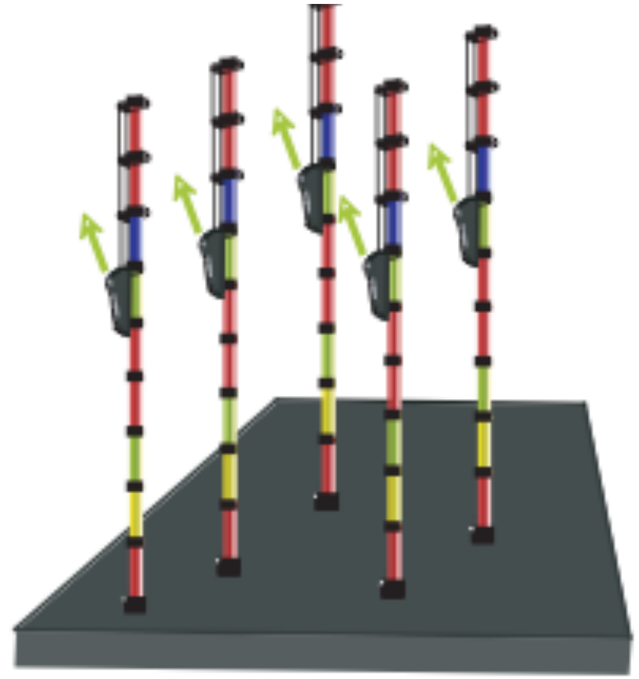
Precision Medicine… with Errors

Reference genome

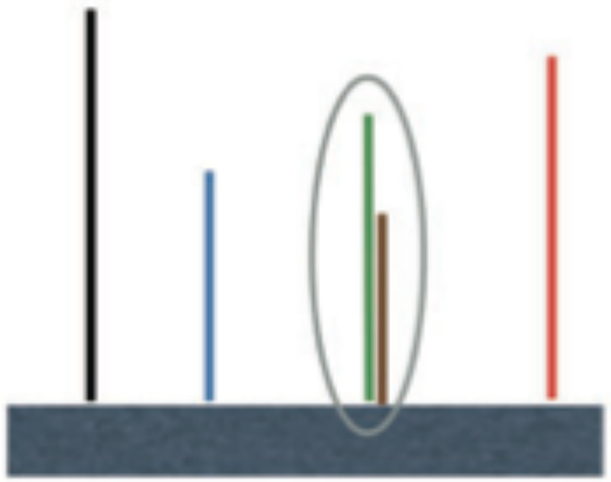


Error profiles

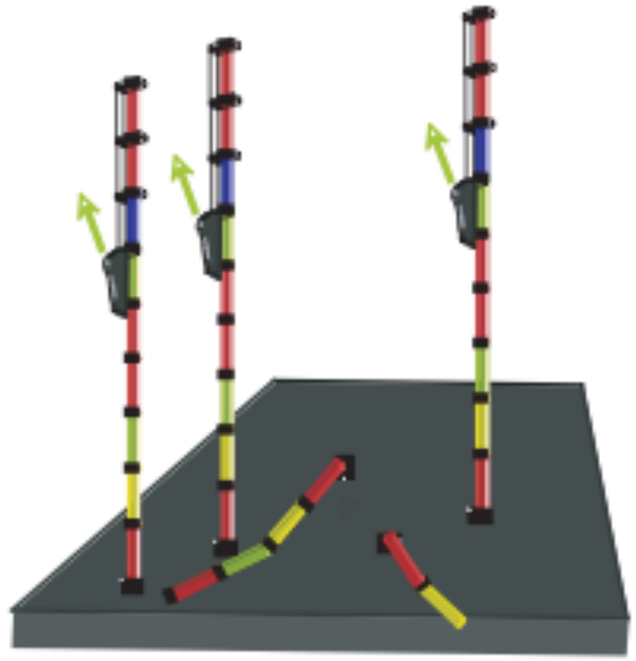
- PCR artifacts
- Error dependency on technology

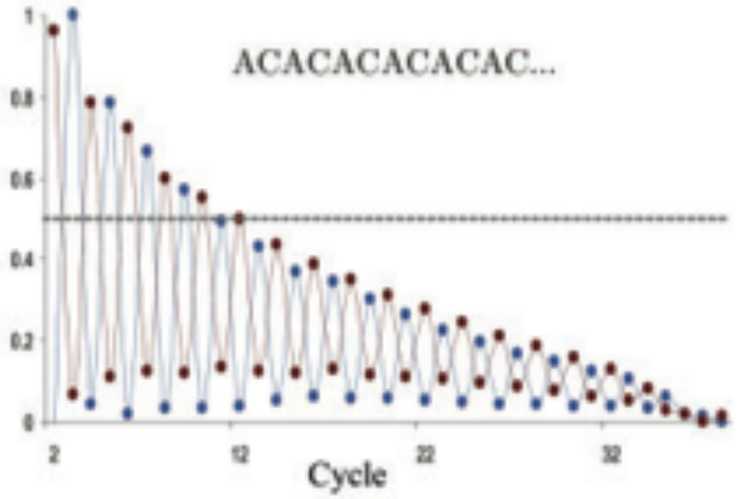


Illumina

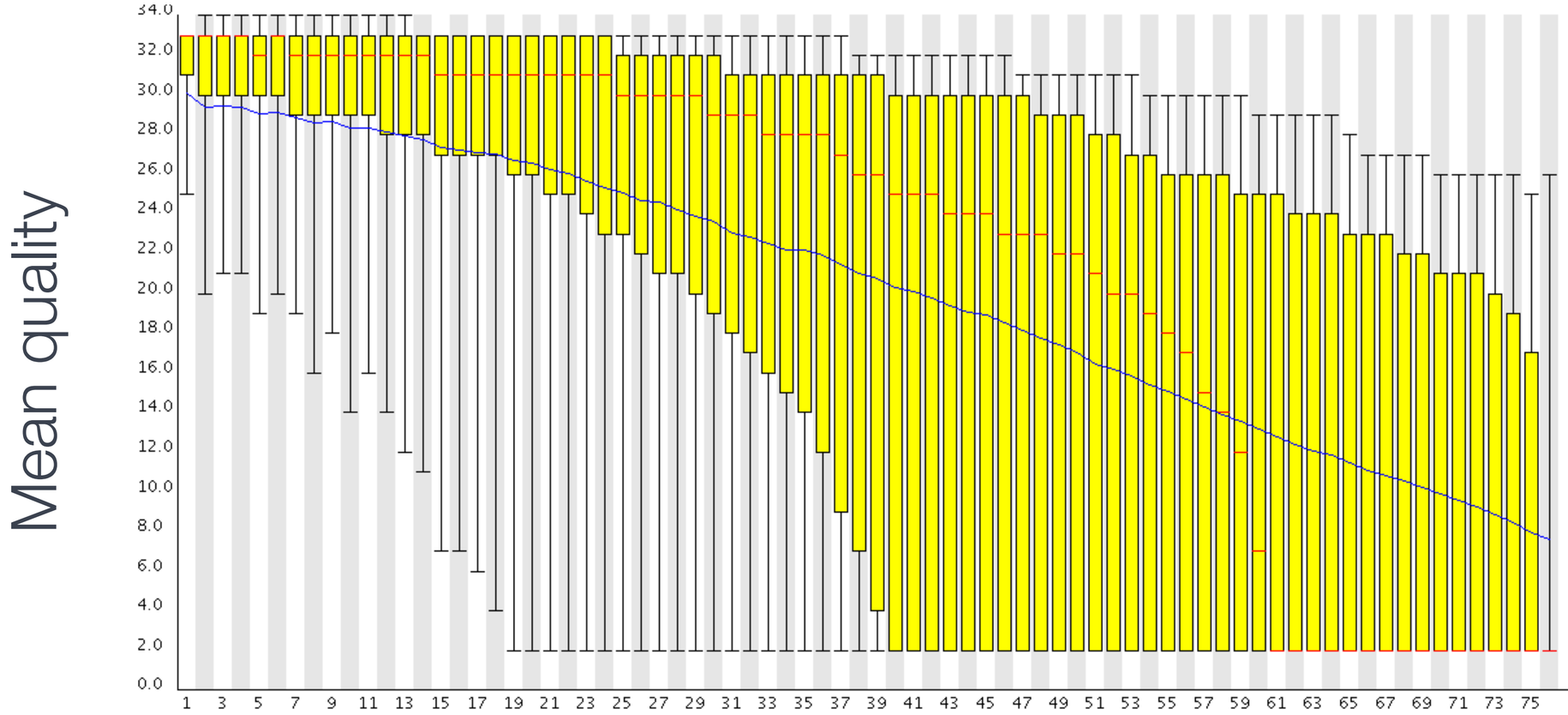


Illumina: mixed clusters

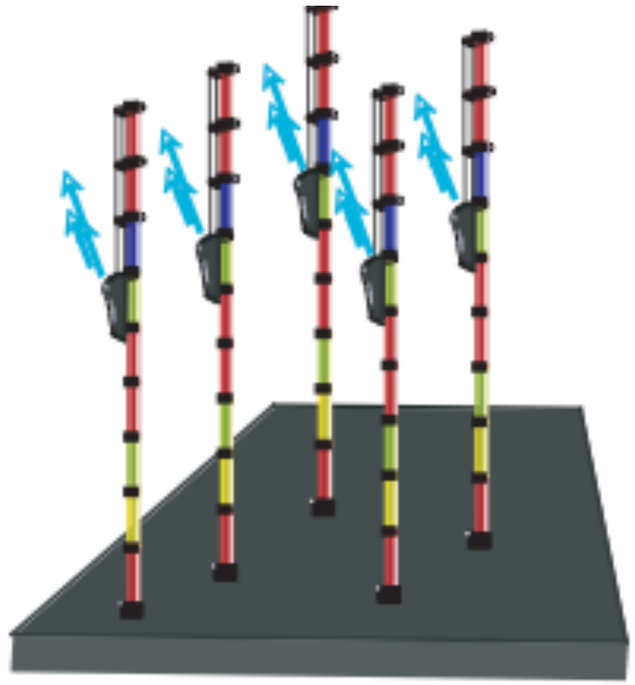


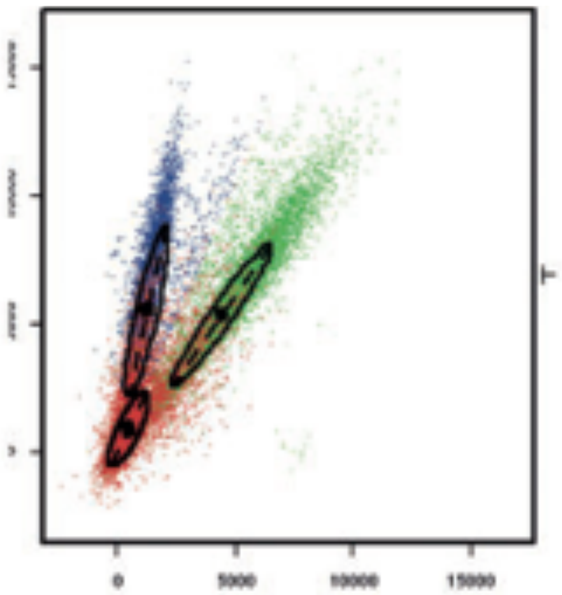


Illumina: signal decay

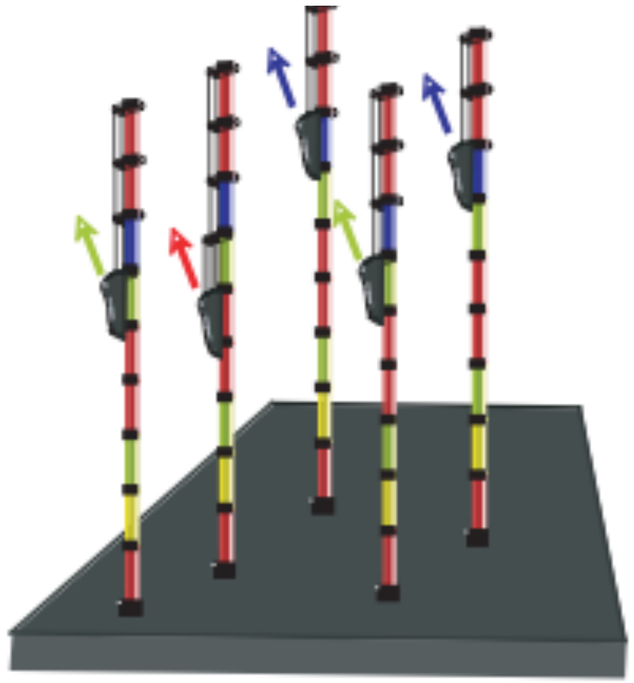


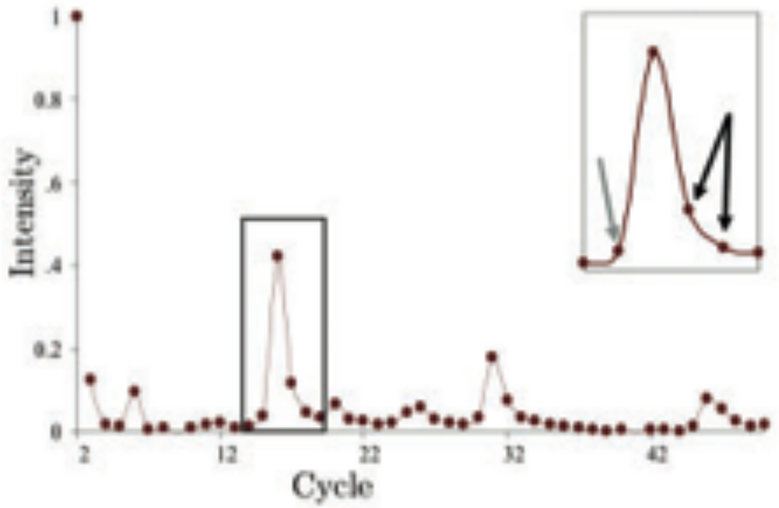
Illumina: signal decay



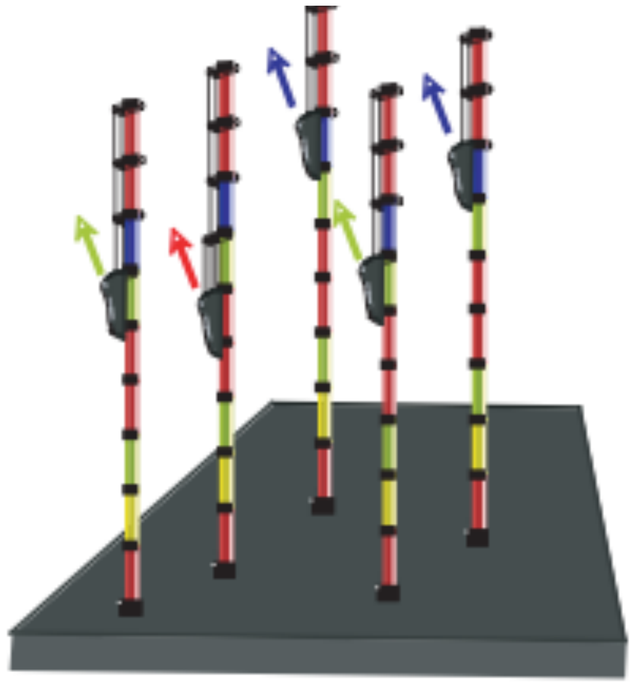


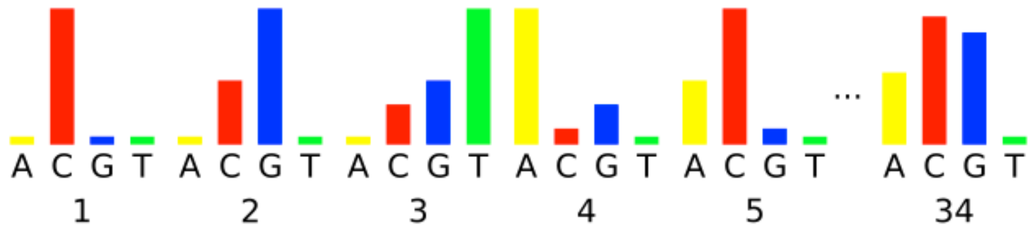
Illumina: cross-talk





Illumina: phasing





Illumina: phasing correction

Software

Highly accessed Open Access

Improved base calling for the Illumina Genome Analyzer using machine learning strategies

Martin Kircher, **Udo Stenzel** and **Janet Kelso***

* Corresponding author: Janet Kelso kelso@eva.mpg.de

Author Affiliations

Department of Evolutionary Genetics, Max Planck Institute for Evolutionary Anthropology, Deutscher Platz, 04103 Leipzig, Germany

For all author emails, please log on.

Genome Biology 2009, **10**:R83 doi:10.1186/gb-2009-10-8-r83

The electronic version of this article is the complete one and can be found online at: http://genomebiology.com/2009/10/8/R83

Illumina: phasing correction

The need for test data

Genome in a Bottle

Highly confident variant calls for **NA12878**

Reference bioinformatic workflows

genomeinabottle.org

Genome in a Bottle Consortium

Marc Salit

Justin Zook

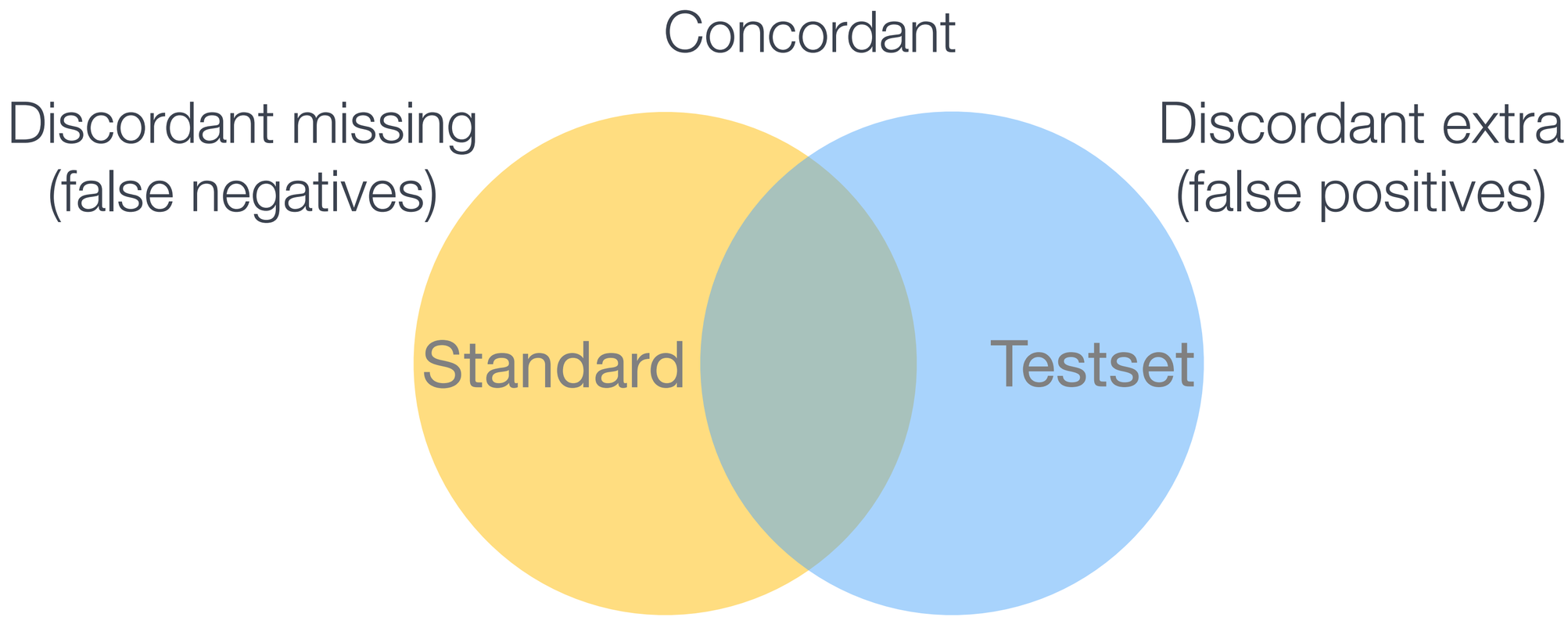
Genome in a Bottle

11 WGS

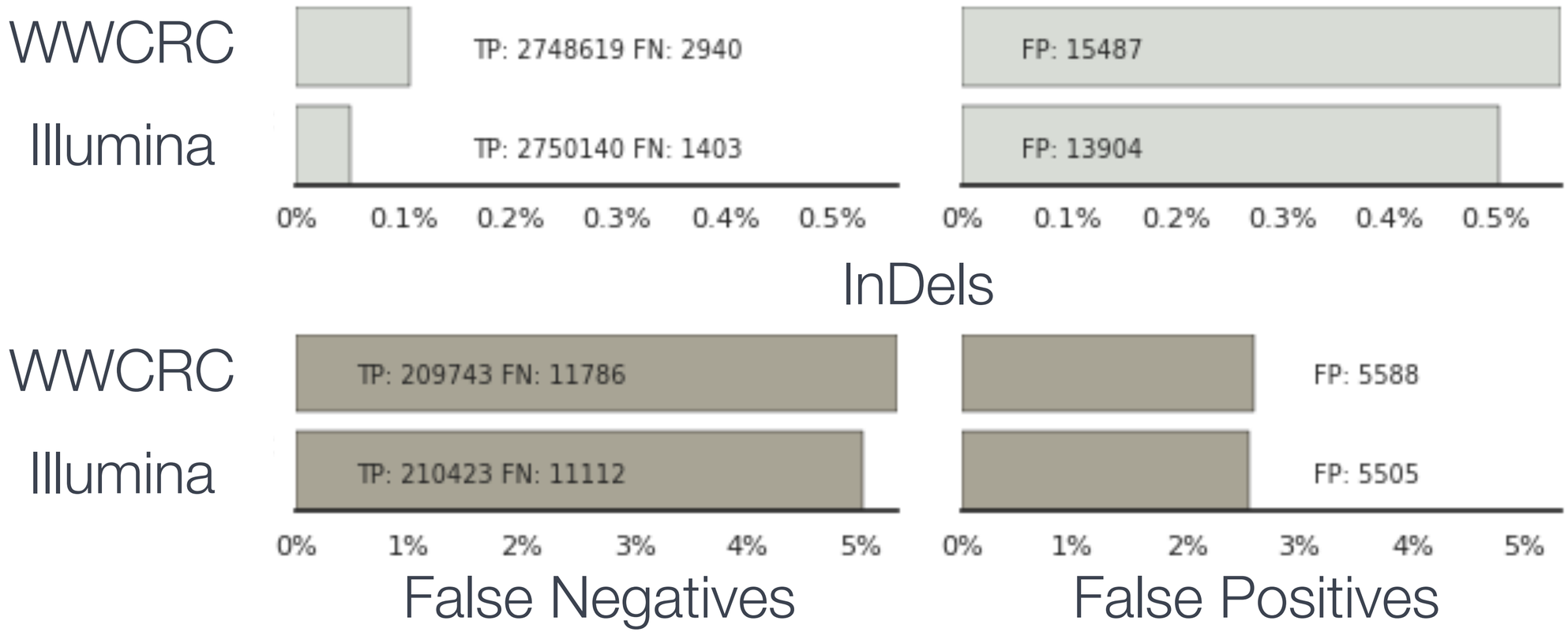
3 Exomes

Illumina, SOLiD, 454, IonTorrent, CGI

~3 million highly confident SNPs



Gold standards



Evaluate different workflows

Found with n/n callers

Without training data

Train SVM on total depth, called allele depth and the posterior likelihoods from the variant callers

Found with n/n callers

Supported by only 1 caller
Indels in low complexity regions
Novel, non-dbSNP variants with low reads
…

Without training data

Train SVM on total depth, called allele depth and the posterior likelihoods from the variant callers

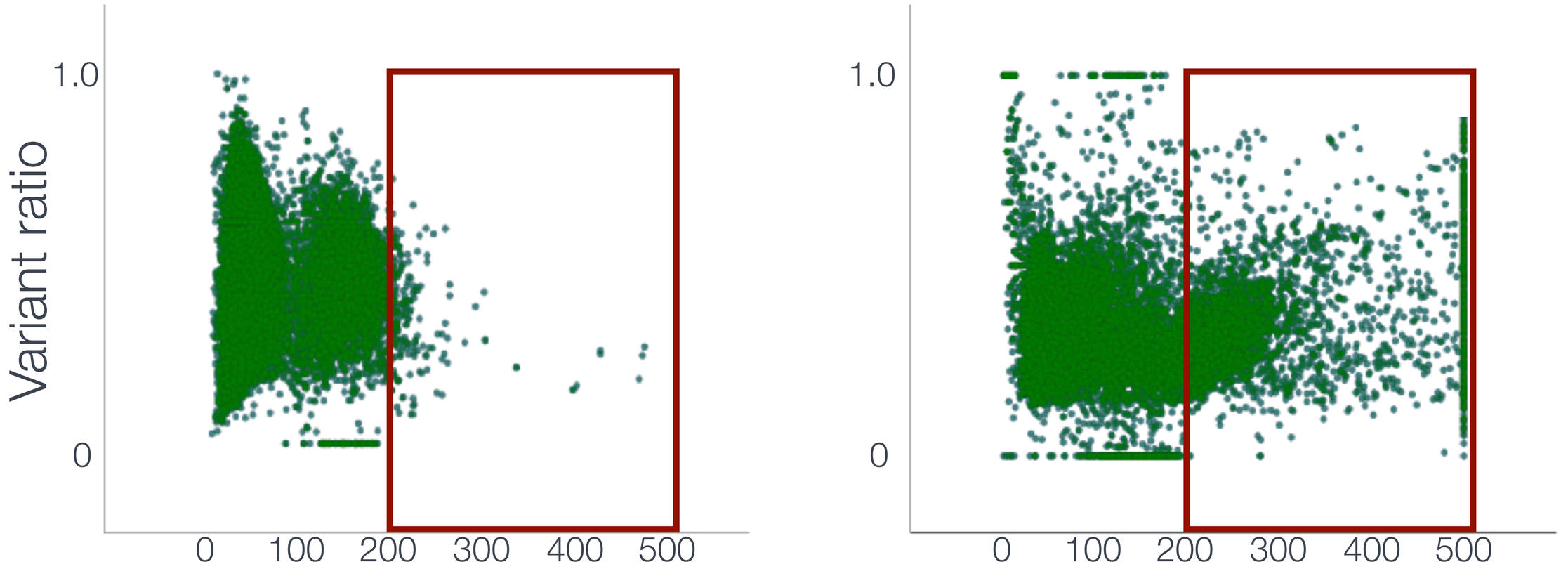
SNP concordance vs GiaB: 86.6% -> 87.4%

Marginal benefits

Train SVM on total depth, called allele depth and the posterior likelihoods from the variant callers

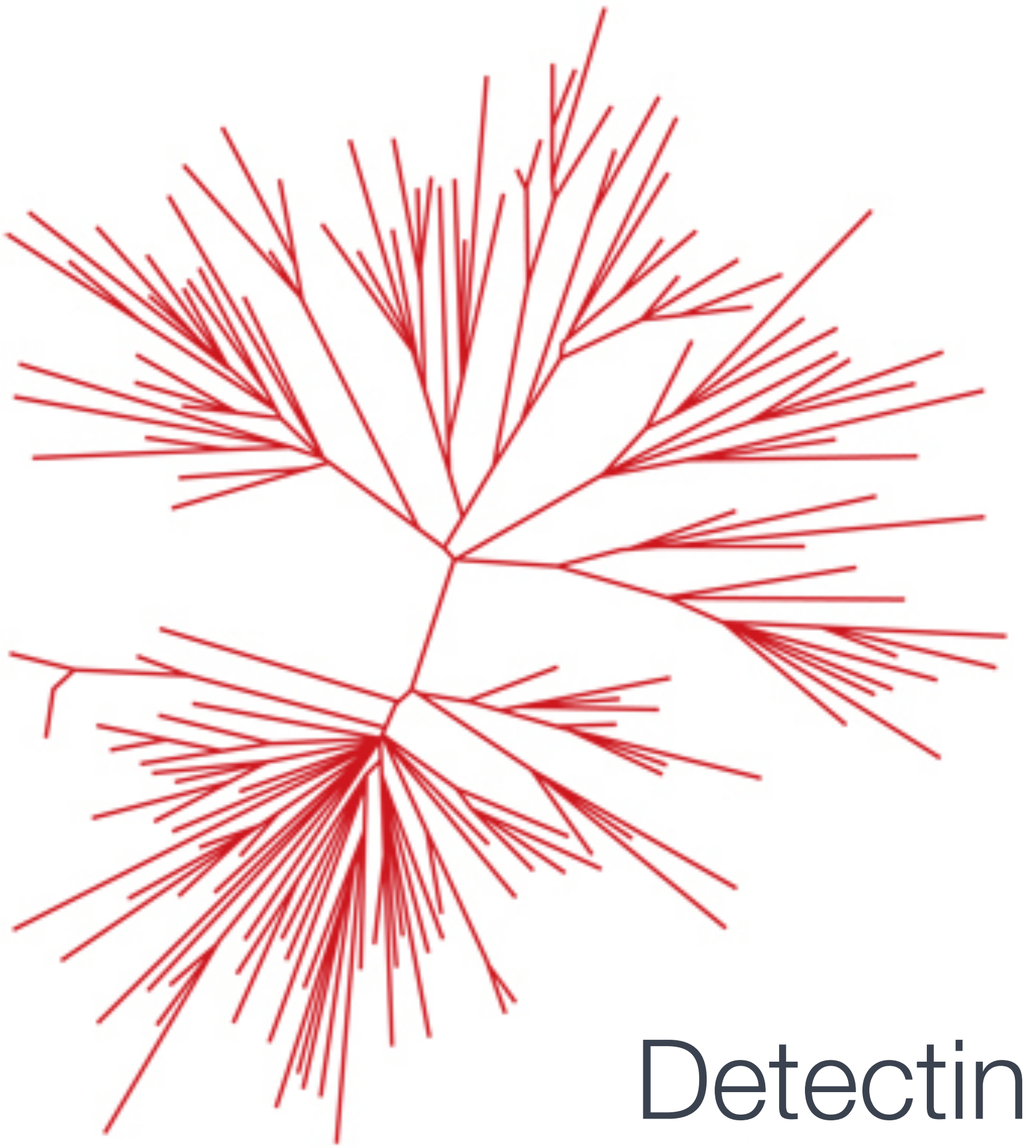
True positive heterozygous calls

False positive heterozygous calls



Understand classification attributes

http://arxiv.org/abs/1404.0929

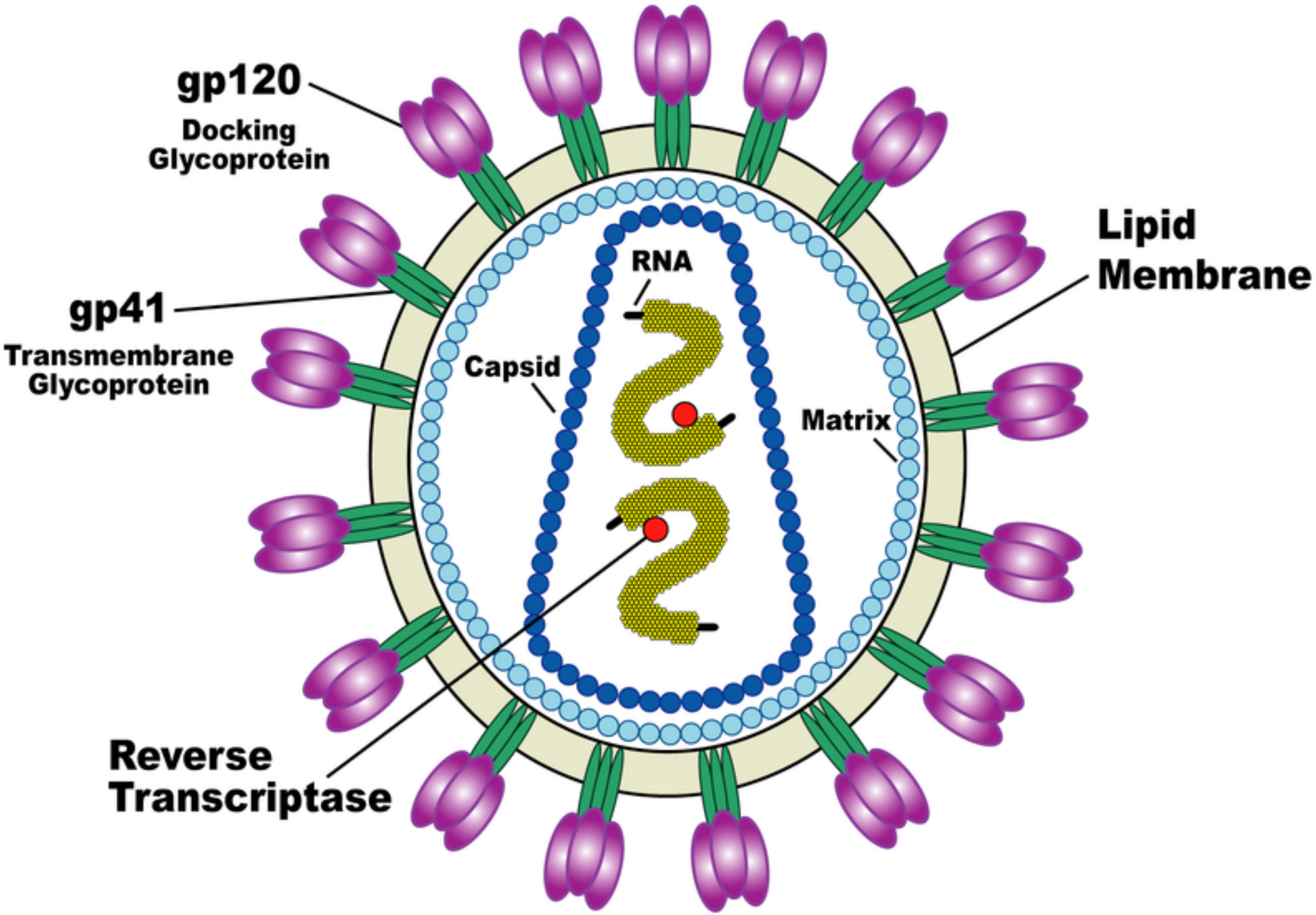


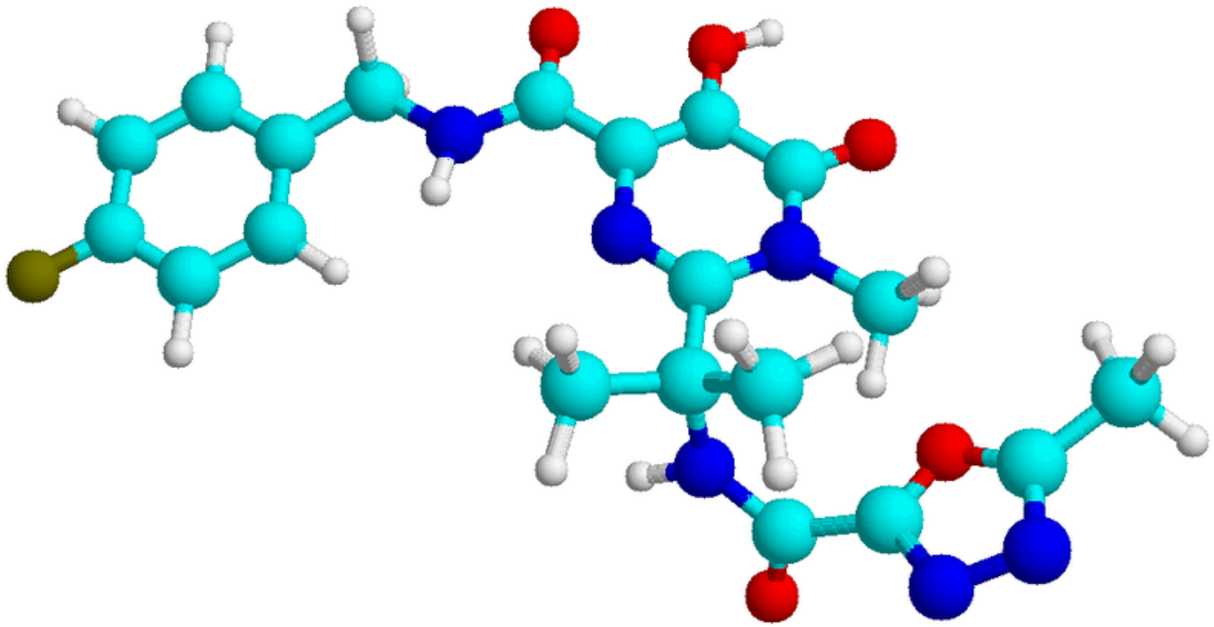
Daniel Kuritzkes

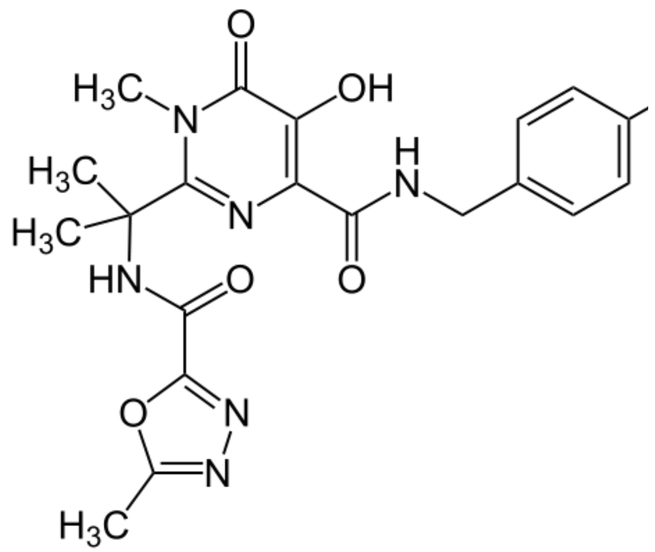
Jonathan Li

Detecting minority variants in HIV

Assess raltegravir-resistant minority variants in ACTG A526 using Illumina


gp120
Docking
Glycoprotein
gp41
Transmembrane
Glycoprotein
Lipid
Membrane
RNA
Capsid
Matrix
Reverse
Transcriptase





raltegravir

Targets HIV *integrase* gene

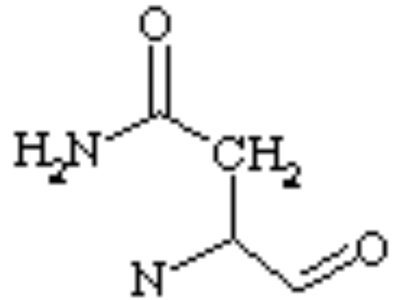
...CGTCCCTCAGAATGGAAACCTCGCTT...

HIV genome snippet

...CGTCCCTCAGAATGGAAACCTCGCTT...
...CGTCCCTCAG**C**ATGGAAACCTCGCTT...

HIV genome snippet

Asparagine

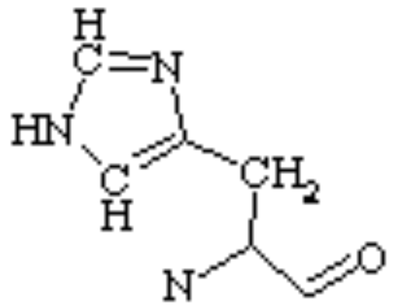


[Asn]

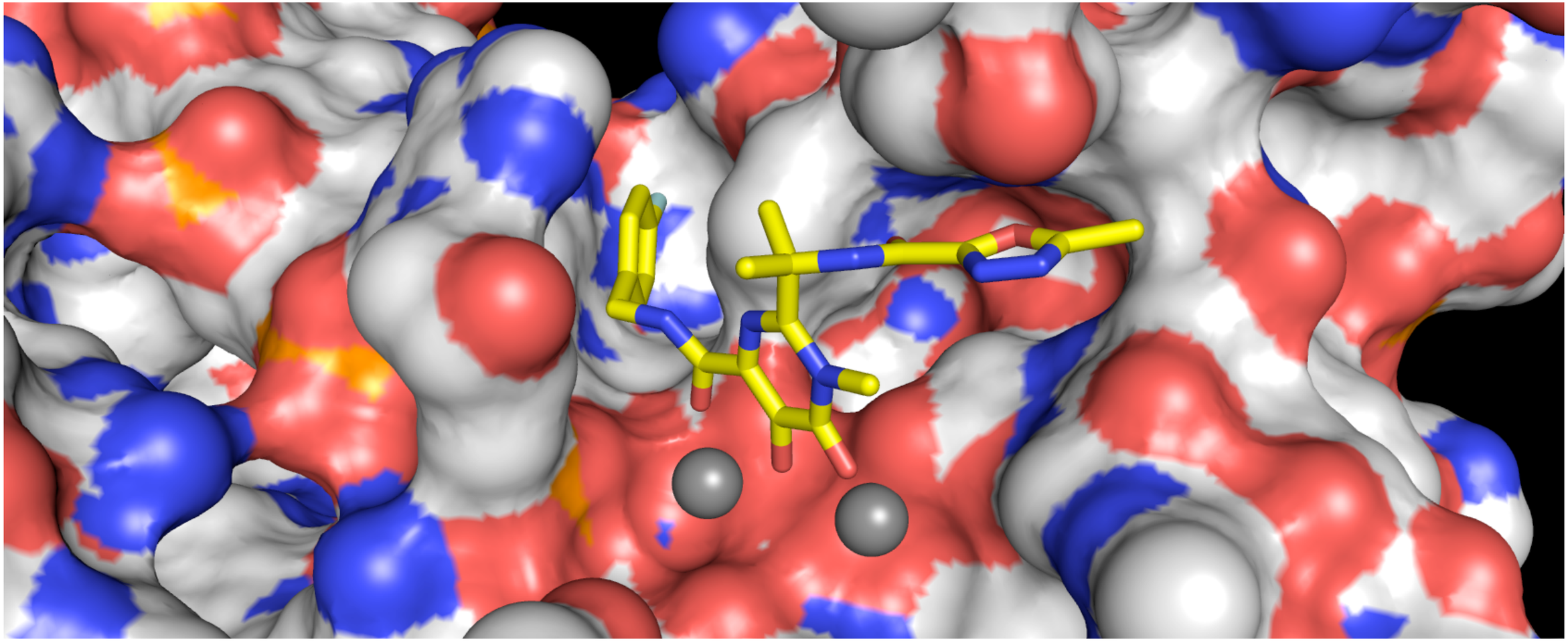
...CGTCCCTCAGAATGGAAACCTCGCTT...

...CGTCCCTCAG**C**ATGGAAACCTCGCTT...

Histidine



[His]



Variation changes binding to integrase

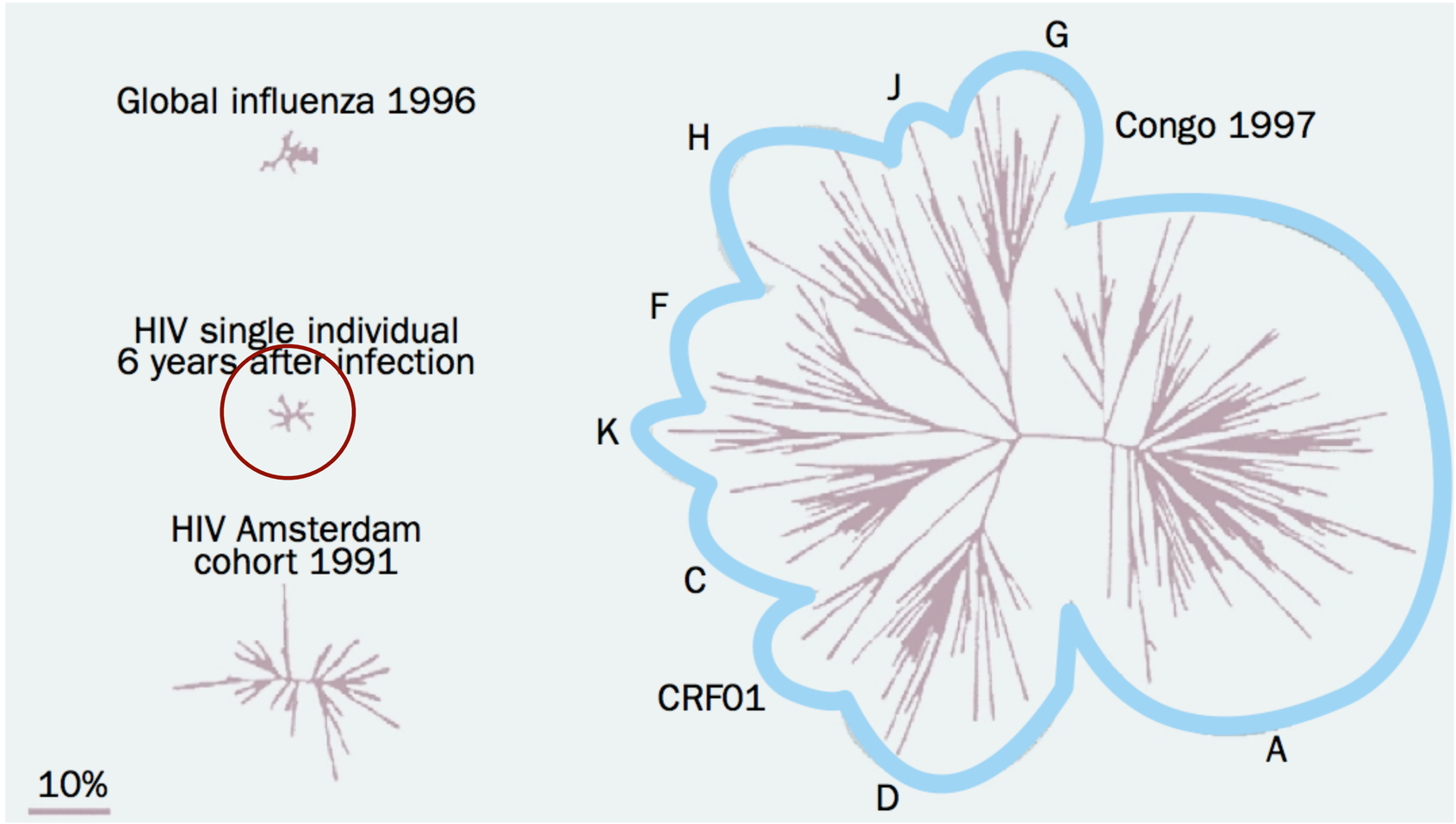
IAS-USA* Drug Resistance Mutations Group

2011 Update of the Drug Resistance Mutations in HIV-1

Victoria A. Johnson, MD, Vincent Calvez, MD, PhD, Huldrych F. Günthard, MD, Roger Paredes, MD, PhD, Deenan Pillay, MD, PhD, Robert Shafer, MD, Annemarie M. Wensing, MD, PhD, and Douglas D. Richman, MD

	E	Y	Q	N
Raltegravir	**92**	**143**	**148**	**155**
	Q	R H	H K	H

International Antiviral Society



HIV diversity within a patient

Garber, David A, Guido Silvestri, and Mark B Feinberg. “Prospects for an AIDS Vaccine: Three Big Questions, No Easy Answers..” The Lancet Infectious Diseases 4, no. 7: 397–413. doi:10.1016/S1473-3099(04)01056-4.

...CGTCCCTCAG**C**ATGGAAACCTCGCTT...

Minority variants at different frequencies

...CGTCCCTCAG**C**ATGGAAACCTCGCTT...
...CGTCCCTCAG**C**ATGGAAACCTCGCTT...
...CGTCCCTCAG**C**ATGGAAACCTCGCTT...
...CGTCCCTCAG**C**ATGGAAACCTCGCTT...

Minority variants at different frequencies

...CGTCCCTCAG**C**ATGGAAACCTCGCTT...
...CGTCCCTCAG**C**ATGGAAACCTCGCTT...
...CGTCCCTCAG**C**ATGGAAACCTCGCTT...
...CGTCCCTCAG**C**ATGGAAACCTCGCTT...
...CGT**C**CCTCAG**C**ATGGAAACCTCGCTT...
...CGT**C**CCTCAG**C**ATGGAAACCTCGCTT...

...CGTCCCTCAG**C**ATGGA**A**ACCTCGCTT...

...CGTCCCTCAG**C**ATGGA**A**ACCTCGCTT...

...CGTCCCTCAG**C**ATGGA**A**ACCTCGCTT...

...CGTCCCTCAG**C**ATGGAAACCTCGCTT...

...CGTCCCTCAG**C**ATGGAAACCTCGCTT...

...CGTCCCTCAG**C**ATGGAAACCTCGCTT...

...CGTCCCTCAG**C**ATGGAAACCTCGCTT...

...CGT**C**CCTCAG**C**ATGGAAACCTCGCTT...

...CGT**C**CCTCAG**C**ATGGAAACCTCGCTT...

...CGTCCCTCAG**C**ATGGA**A**ACCTCGC**T**T...

...CGTCCCTCAG**C**ATGGA**A**ACCTCGCTT...

...CGTCCCTCAG**C**ATGGA**A**ACCTCGCTT...

...CGTCCCTCAG**C**ATGGA**A**ACCTCGCTT...

...CGTCCCTCAG**C**ATGGAAACCTCGCTT...

...CGTCCCTCAG**C**ATGGAAACCTCGCTT...

...CGTCCCTCAG**C**ATGGAAACCTCGCTT...

...CGTCCCTCAG**C**ATGGAAACCTCGCTT...

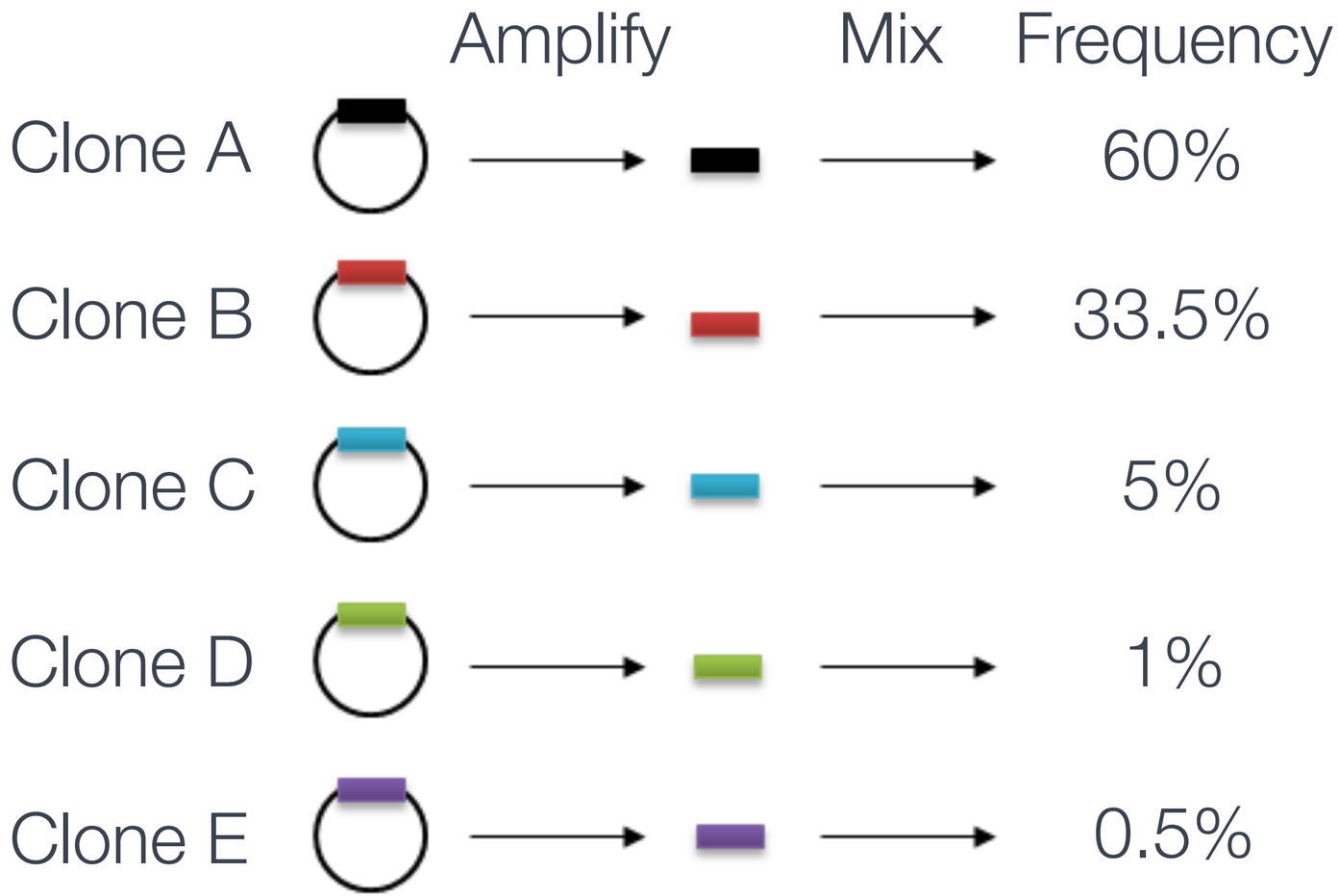
...CGT**C**CCTCAG**C**ATGGAAACCTCGCTT...

...CGT**C**CCTCAG**C**ATGGAAACCTCGCTT...

20% 40% 10%

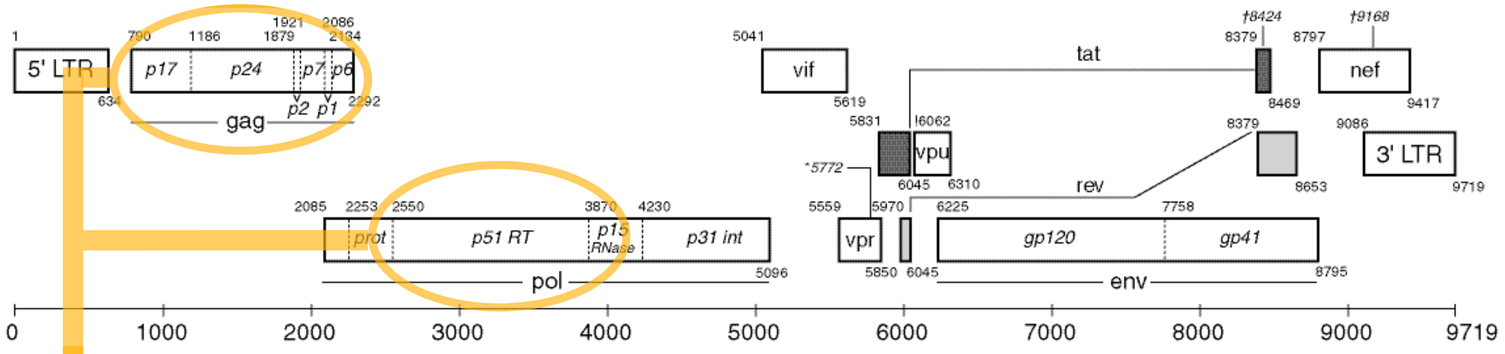
	DRUG			PHENOSENSE™ SUSCEPTIBILITY		Evidence of Susceptibility		Net Assessment	
	Generic Name	Brand Name	Cutoffs (Lower - Upper)	Fold Change	Increasing ← Drug Susceptibility → Decreasing (1 · 10 · 100)	Pheno Sense	Gene Seq		
NRTI	Abacavir	Ziagen	(4.5 - 6.5)	2.31		Y	Y	Sensitive	
	Didanosine	Videx	(1.3 - 2.2)	1.05		Y	Y	Sensitive	
	Emtricitabine	Emtriva	(3.5)	5.41		N	Y	Resistant	19
	Lamivudine	Epivir	(3.5)	3.68		N	Y	Resistant	19
	Stavudine	Zerit	(1.7)	1.58		Y	N	Sensitive	16
	Zidovudine	Retrovir	(1.9)	10		N	N	Resistant	
	Tenofovir	Viread	(1.4 - 4)	1.68		P	Y	Partially Sensitive	19
	NRTI Mutations		D67N, T215Y, K219K/R						
NNRTI	Delavirdine	Rescriptor	(6.2)	0.56		Y	N	Sensitive	13,22
	Efavirenz	Sustiva	(3)	3.21		N	N	Resistant	
	Etravirine	Intelence ®	(2.9)	0.60		Y	Y	Sensitive	
	Nevirapine	Viramune	(4.5)	2.78		Y	N	Sensitive	13,22
	NNRTI Mutations		V106M						

Patient-specific recommendations



The need for test data: internal control

HIV Genome



Defined mix of *gag* and *RT* clones

Source	Sequence
HXB2	... G A A A G C A T T A G G A C C A G C A G C T A C A C T A G A A ...
60%	... G A A G G C A T T A G G A C C A G C A G C T C C A C T A G A T ...
33.5%	... G A A A G C A T T A G G A T C A G C A G C T A C A C T A G A G ...
IUPAC	... G A A R G C A T T A G G A Y C A G C A G C T M C A C T A G A D ...
5%	... G A A A G C A T T A G G A C C A G C A G C T A C A T T A G A A ...
1%	... G A A A G C G T T A G G A C C A A C A G C T A C A C T A G A A ...
0.5%	... G A A T G C A T T A G G A C C A G C A G C T A C A C T A G A A ...

Sequence & align vs majority sequence

Re-alignment

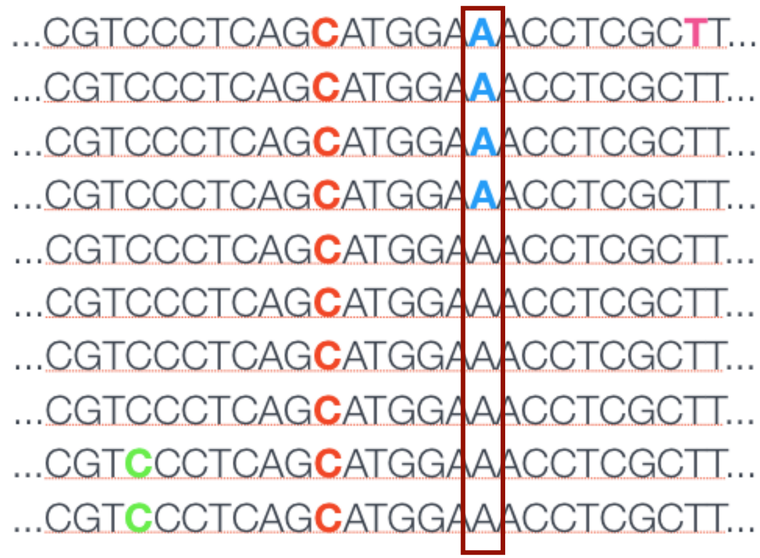
via http://bioinformatics.ca/workshops/2011/informatics-high-throughput-sequencing-data

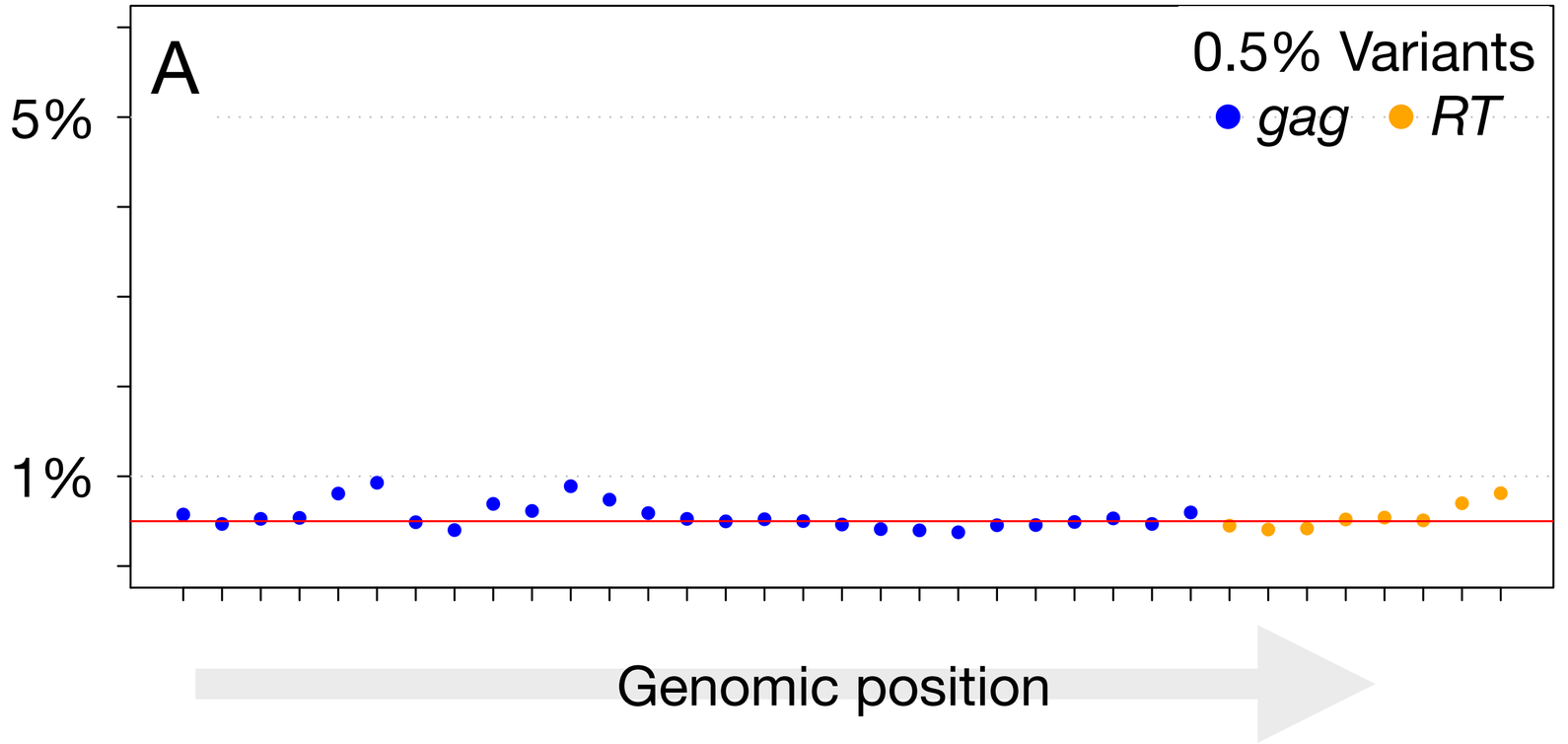
Re-alignment

via http://bioinformatics.ca/workshops/2011/informatics-high-throughput-sequencing-data

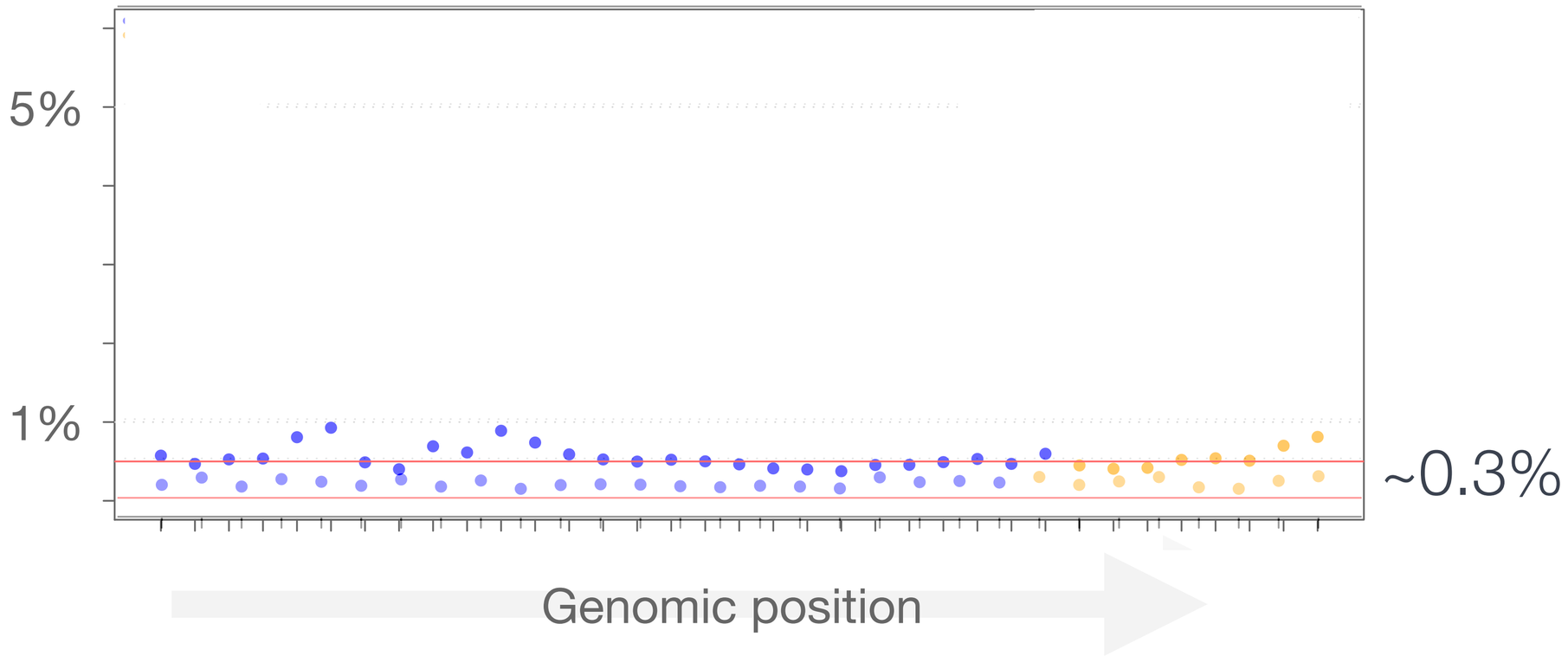
Filtering reads before calling frequencies

- Base quality score
- Read mapping score
- kmer abundance (13 nucleotides around position)
- Combined into linear classifier





Monitor variation: 0.5% frequency

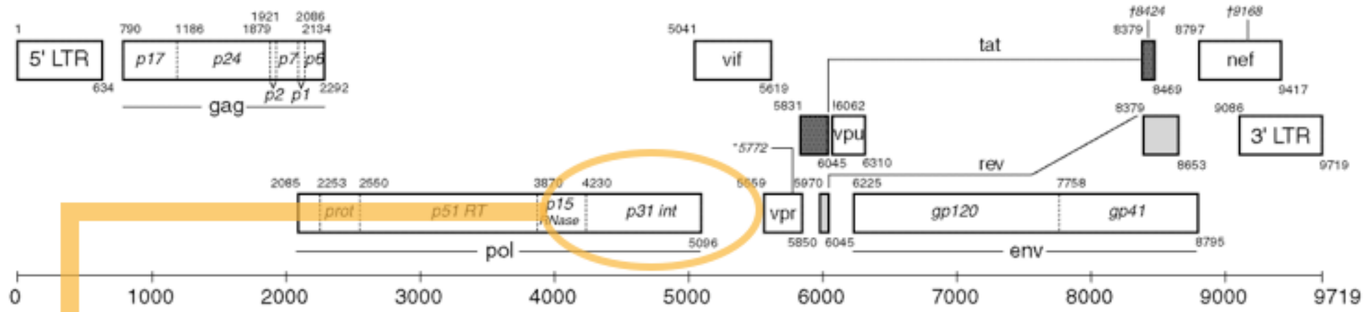


Background noise

Nagging question...

- ... can we do better?

HIV Genome



Defined mix of clones

Source	Sequence
HXB2	... G A A A G C A T T A G G A C C A G C A G C T A C A C T A G A A ...
60%	... G A A G G C A T T A G G A C C A G C A G C T C C A C T A G A T ...
33.4%	... G A A A G C A T T A G G A T C A G C A G C T A C A C T A G A G ...
IUPAC	... G A A R G C A T T A G G A Y C A G C A G C T M C A C T A G A D ...
5%	... G A A A G C A T T A G G A C C A G C A G C T A C A T T A G A A ...
1%	... G A A A G C G T T A G G A C C A A C A G C T A C A C T A G A A ...
0.5%	... G A A T G C A T T A G G A C C A G C A G C T A C A C T A G A A ...
0.1%	... G A A A G C A T T A G G A C C A G C A G C T A C A C T A G A A ...

Add a 0.1% frequency clone

HARVARD CATALYST | THE HARVARD CLINICAL AND TRANSLATIONAL SCIENCE CENTER

- About Harvard Catalyst
- National CTSA Consortium
- Contact Us
- News
- Events
- Spotlights

Search Harvard Catalyst

People & Collaboration | Consulting & Advice | Education & Training | Funding | Research Resources | Programs

Solve health-related algorithmic problems through crowdsourcing.

Algorithm Development Through Crowdsourcing

Overview

Contacts

Solving Algorithmic Problems Through Crowdsourcing

At a glance

Key Features
- Pilot service to develop a solution for your computational algorithmic problems
- Solutions are developed by engaging a world-wide community of elite software developers who compete to solve your problem on a crowdsourcing platform
- Full funding for platform access and prize money for competing software developers
- Full support for problem statement development
- Solutions ready in about a month

Useful for
- Obtaining a computational algorithm for advancing one's research

Submissions Due
- November 18, 2011, with rolling decisions as of October 30, 2011

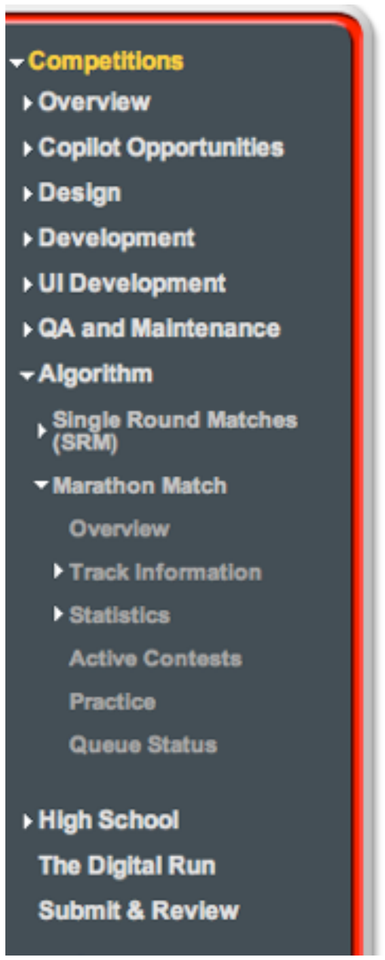
Apply

Note: The submission process has closed. Please check back for future opportunities.

Sponsoring Program

Novel Clinical and Translational Methodologies

Optimization: Crowd sourcing



Marathon Match

Problem Statement

Printable view

Contest: HMS Challenge #1

Problem: MinorityVariants

Problem Statement

We have a mixture of genomes. Each genome can be treated as a sequence of nucleotides denoted using characters 'A', 'C', 'G', 'T'. Consider a certain fixed position within each genome and nucleotides at that position. It is known that one of two possible cases holds:

1. Each genome has the same nucleotide at this position. In this case, we call this position *constant*.
2. There are two different nucleotides at this position. Let's call them X and Y. Suppose that X occurs more often than Y. In this case, nucleotide Y is called *minority variant* at this position. The fraction of nucleotides that contain Y at this position is called the *frequency* of this minority variant.

Given a lot of sequencing calls for this mixture, we would like to identify which positions are constant and which have minority variants. Also, we are interested in frequencies of minority variants. This seems to be trivial to do, however, some sequencing calls have low quality and we can't trust to their results. The goal of this problem is to develop an approach to separate low quality sequencing calls from real calls, and thus improve the detection ability of minority variants.

Training data

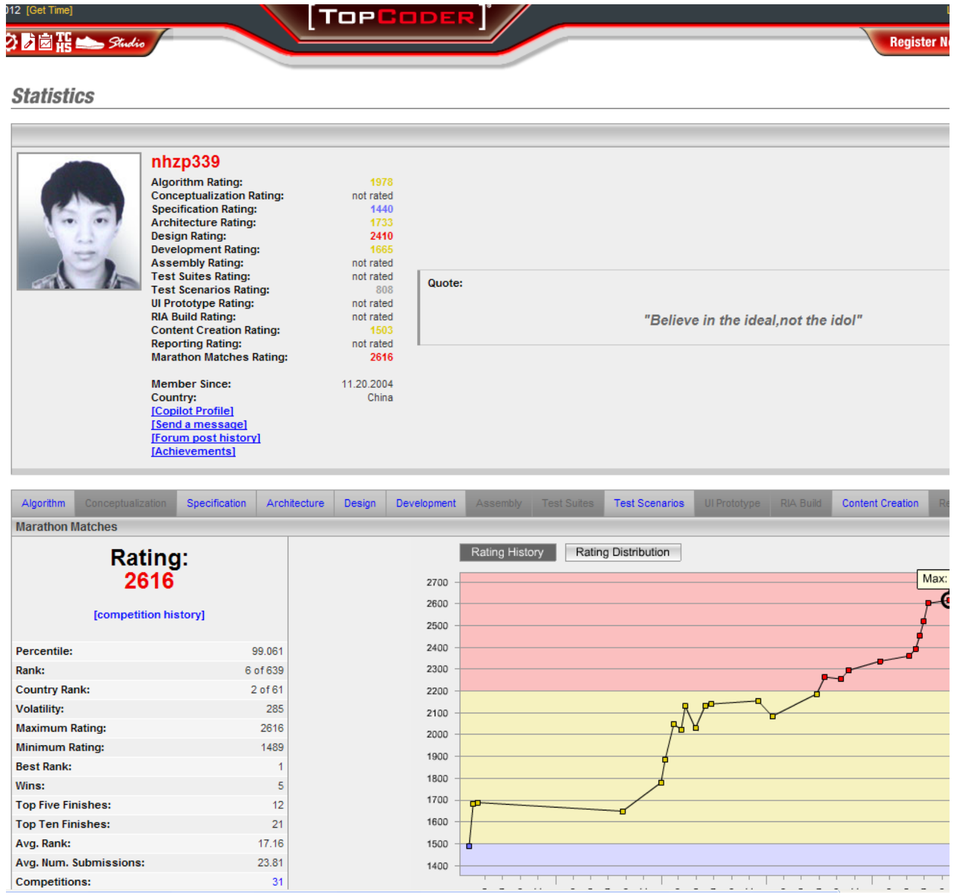
We provide some data in order to help you develop the solution. It consists of 3 files, all in TSV format.

Problem definition & test set

TopCoder Challenge

Crowd Sourcing at work

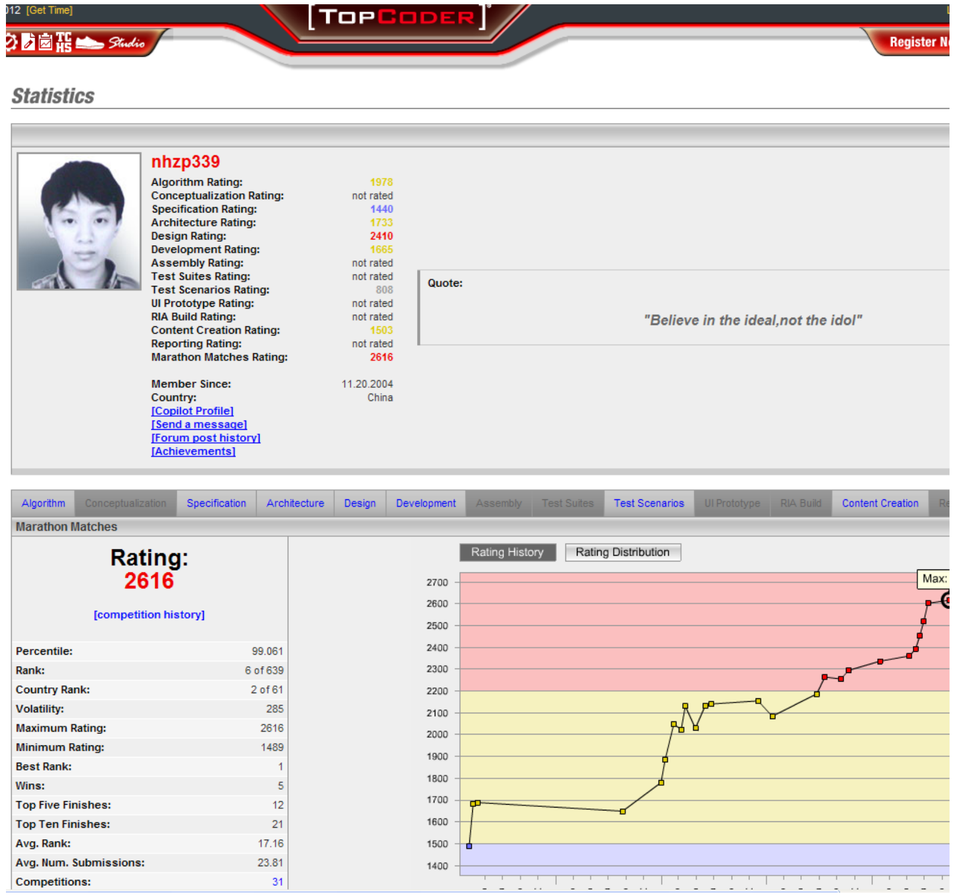
- > 200 competitors
- > 600 submissions



Crowd Sourcing at work

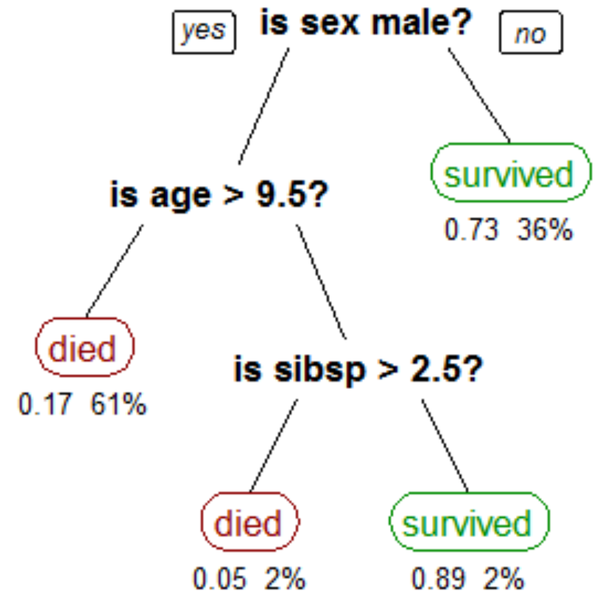
Version 2.0

- > 200 competitors
- > 600 submissions



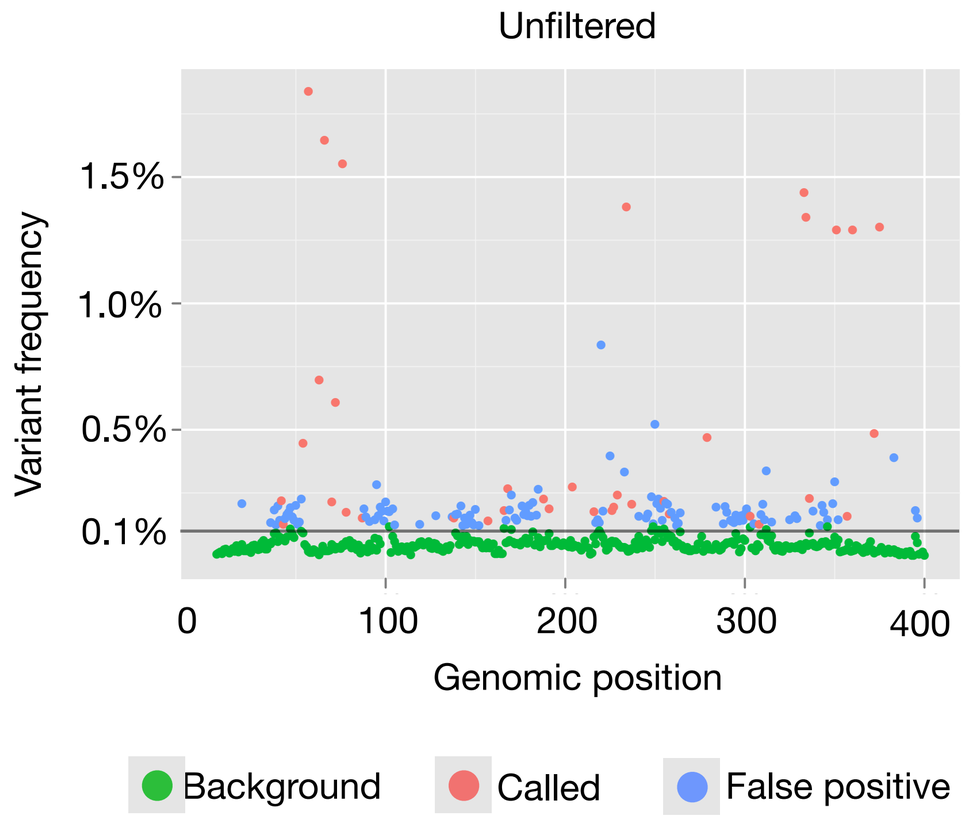
Random forest classifier

- Decision tree-based
- Expanded 3 metrics to 14 features
- Random forest of 120 trees
- Reference implementation provided



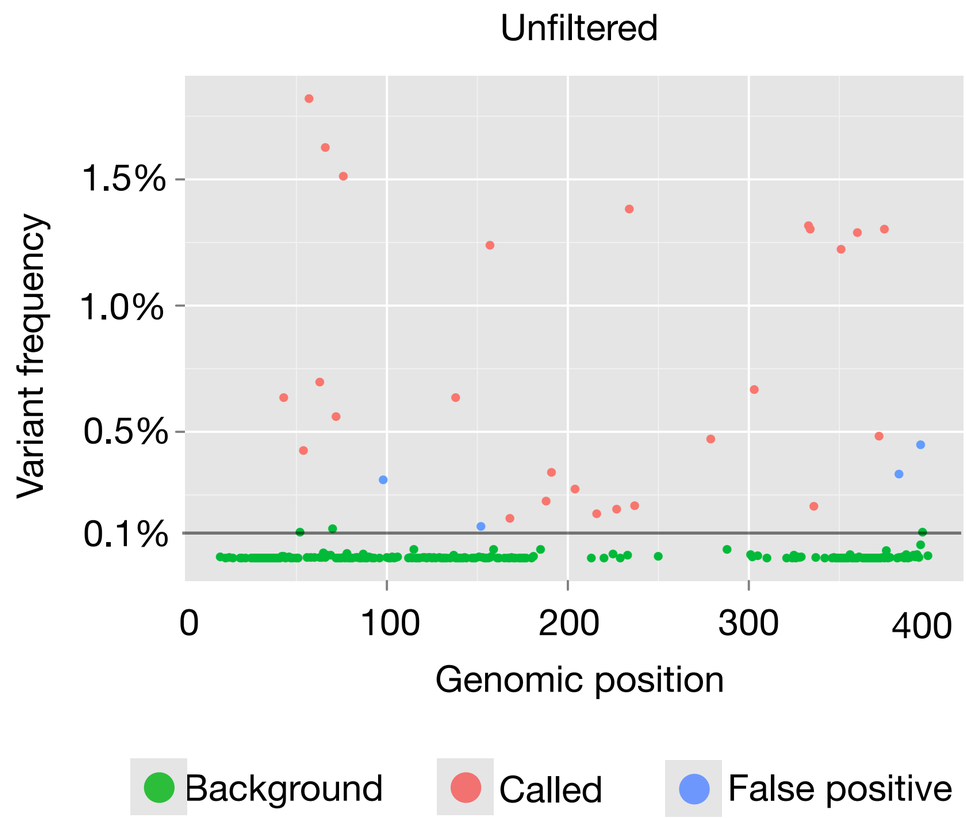
Random forest classifier

- Determine filtering cutoff through downsampling



Random forest classifier

- RF with 15 input parameters
- Determine filtering cutoff through downsampling
- Cutoff at **0.095%** removes 99% false positives
- **Stable** between experiments



BRIGHAM AND WOMEN'S HOSPITAL

A Teaching Affiliate of Harvard Medical School

HOME | ABOUT BWH | DEPARTMENTS AND SERVICES | DISEASES AND CONDITIONS | HEALTH INFORMATION | PATIENTS AND VISITORS

Home > Departments and Services > Medicine > Services > Division of Infectious Diseases

Infectious Diseases

- Overview
- About Us
  - Our Division
  - Our Physicians
  - Faculty
- Our Services
- For Patients
  - Staff Physicians
  - Referral Information
  - Your First Visit
  - While You Are Here

Division of Infectious Diseases

The Division of Infectious Disease at Brigham & Women's Hospital (BWH) is a part of the Brigham Medical Specialties (BMS) and is affiliated with Harvard Medical School. We provide the highest quality patient care and exceptional consultative service to the Brigham and Women's Hospital community, while also advancing the hospital's goals of creative biomedical research and teaching. The division consists of a diversified group of clinicians, clinical and research investigators, epidemiologists, and social scientists.

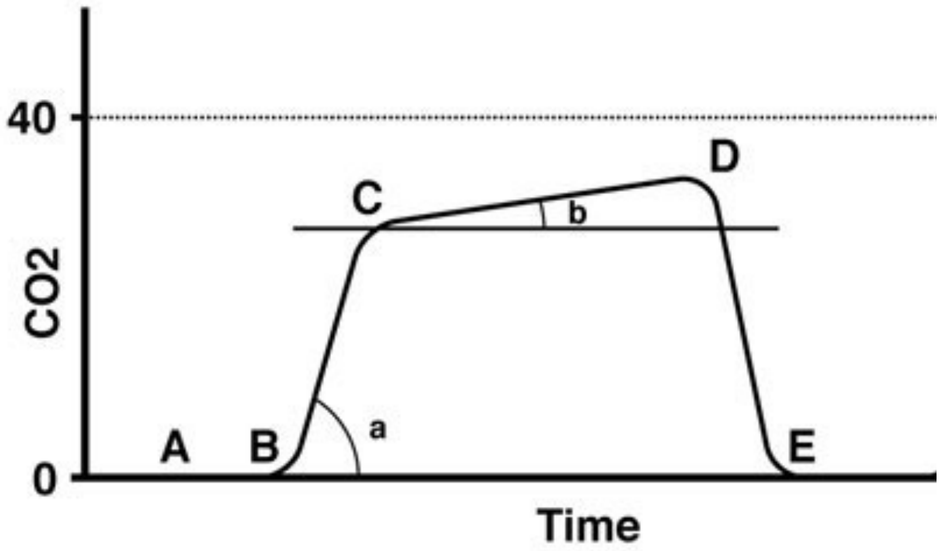
Lynn Bry

Ongoing work: to the clinic

Pediatric Intensive Care Unit Monitoring

Classifier to predict apnea events

- O2
- CO2
- Blood pressure
- Heart rate
- Waveform of measurements



A-B: Inhalation Trough (Baseline)
B-C: Initial Expiratory Phase
C-D: Expiratory Plateau
D-E: Initial Inhalation Phase
D: ETCO2 Value
a: Takeoff Angle
b: Elevation Angle

Plenty of fun problems to solve

Take-home messages

Wide applicability for ML approaches

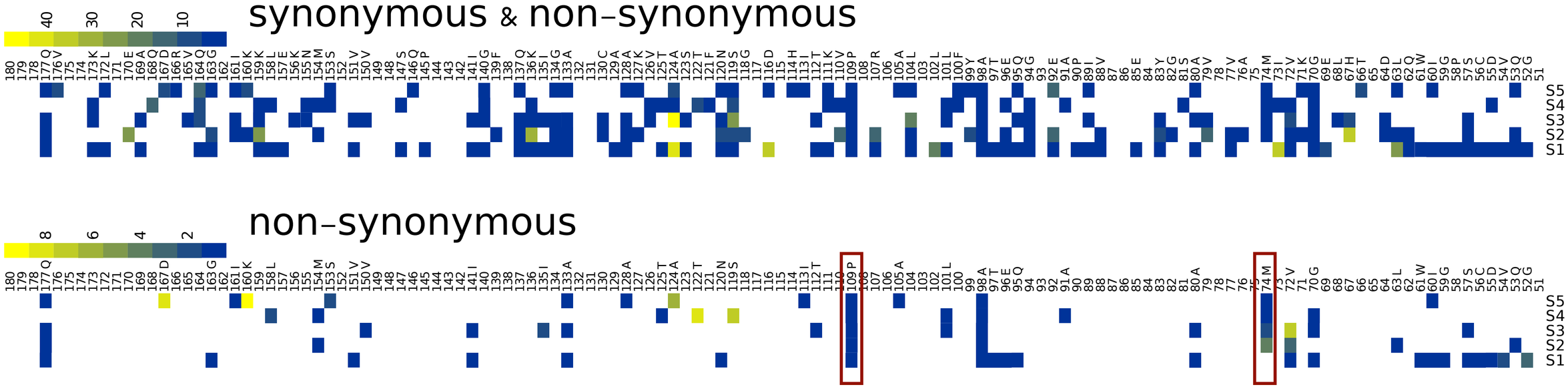
Good training sets still a problem

Talk with the community

Thank you for listening!

oliver.hofmann@glasgow.ac.uk

@fiamh



Variation in patient HIV *int* genes