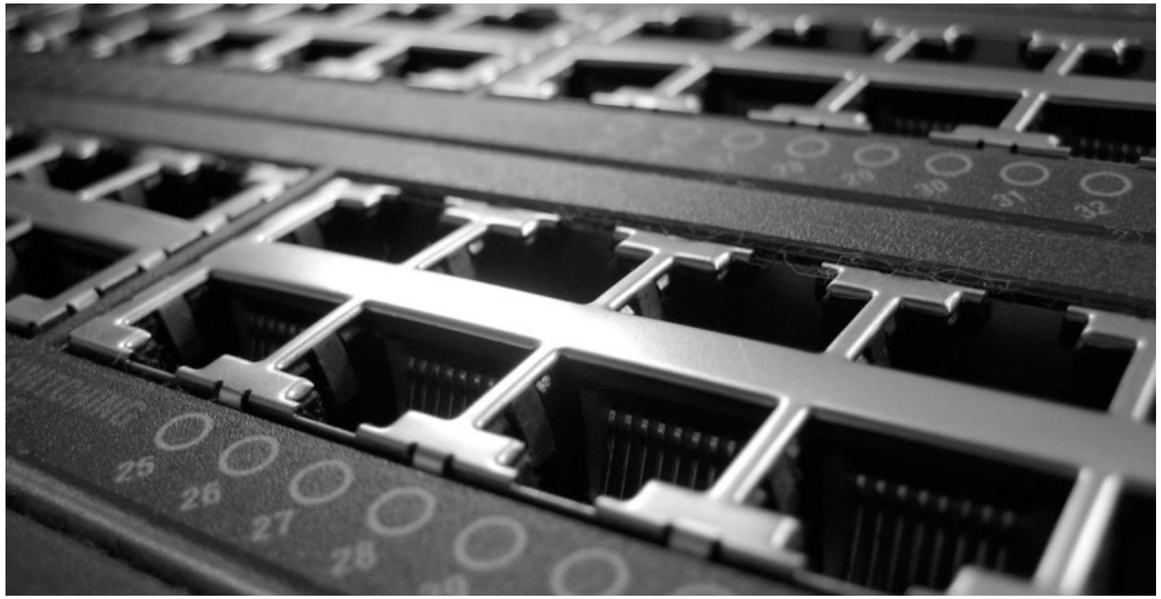
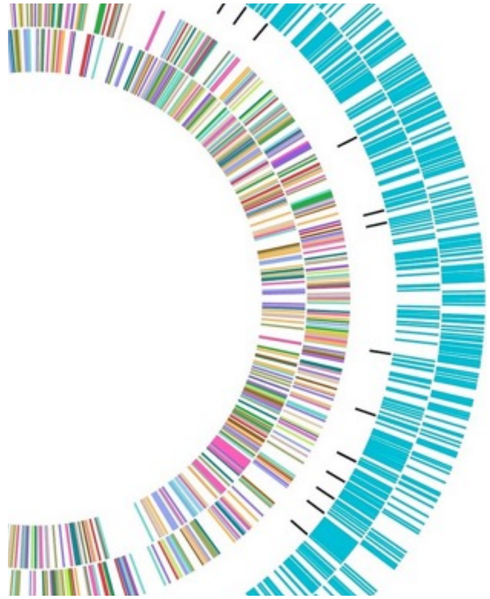




Bioinformatics Update from the Harvard Chan Bioinformatics Core

Shannan Ho Sui, John Hutchinson, Brad Chapman, Rory Kirchner, Meeta Mistry, Lorena Pantano, Radhika Khetani, Mary Piper, Victor Barrera, Michael Steinbaugh, Kayleigh Rutherford

HARVARD T.H. CHAN
SCHOOL OF PUBLIC HEALTH

<intro>

HARVARD
T.H. CHAN
SCHOOL OF PUBLIC HEALTH

Shannan Ho Sui

John Hutchinson

Brad Chapman

Rory Kirchner

Peter Kraft

Lorena Pantano

Meeta Mistry

Mary Piper

Radhika Khetani

Victor Barrera

Michael Steinbaugh

Kayleigh Rutherford

HARVARD T.H. CHAN SCHOOL OF PUBLIC HEALTH

HARVARD STEM CELL INSTITUTE

HARVARD CATALYST
THE HARVARD CLINICAL AND TRANSLATIONAL SCIENCE CENTER

HARVARD MEDICAL SCHOOL
TnT/HNDC

NIEHS / CFAR Bioinformatics Core

Center for Stem Cell Bioinformatics

Harvard Catalyst Bioinformatics Consulting

Harvard Medical School Bioinformatics Core

HARVARD T.H. CHAN SCHOOL OF PUBLIC HEALTH

HARVARD STEM CELL INSTITUTE

HARVARD CATALYST
THE HARVARD CLINICAL AND TRANSLATIONAL SCIENCE CENTER

HARVARD MEDICAL SCHOOL
TnT/HNDC

NIEHS / CFAR Bioinformatics Core

Center for Stem Cell Bioinformatics

Harvard Catalyst Bioinformatics Consulting

Harvard Medical School Bioinformatics Core

More than 500 consults from almost all Harvard-affiliated institutions

Services

- Transcriptomics: RNA-seq, small RNA-seq, single cell RNA-seq
- DNA accessibility and binding: ChIP-seq, ATAC-seq
- Genetic variation: WGS, re-sequencing, exome-seq, structural variation, CNV
- Genome-wide methylation: 450k methylation arrays, RRBS, WGBS
- Data integration
- Functional enrichment analysis
- Experimental design and grant support

HARVARD T.H. CHAN
SCHOOL OF PUBLIC HEALTH

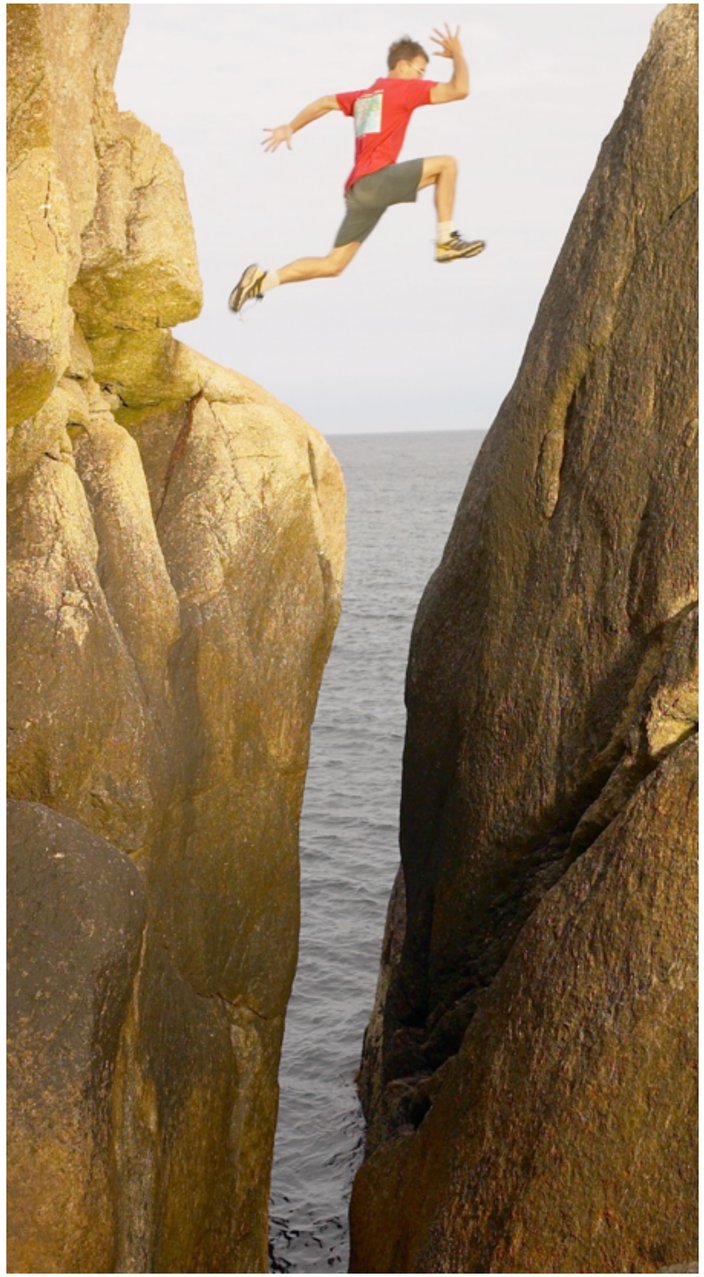
What we don’t do

- Sequencing
- Proteomics / metabolomics
- Metagenomics
- Custom scripts
- One-on-one training
- Survival analysis, imputation, GWAS and statistical genetics
- No longer doing microarrays

HARVARD T.H. CHAN
SCHOOL OF PUBLIC HEALTH

<our approach>

NGS Data



Results

The installation hurdle

github.com/StanfordBioinformatics/HugeSeq

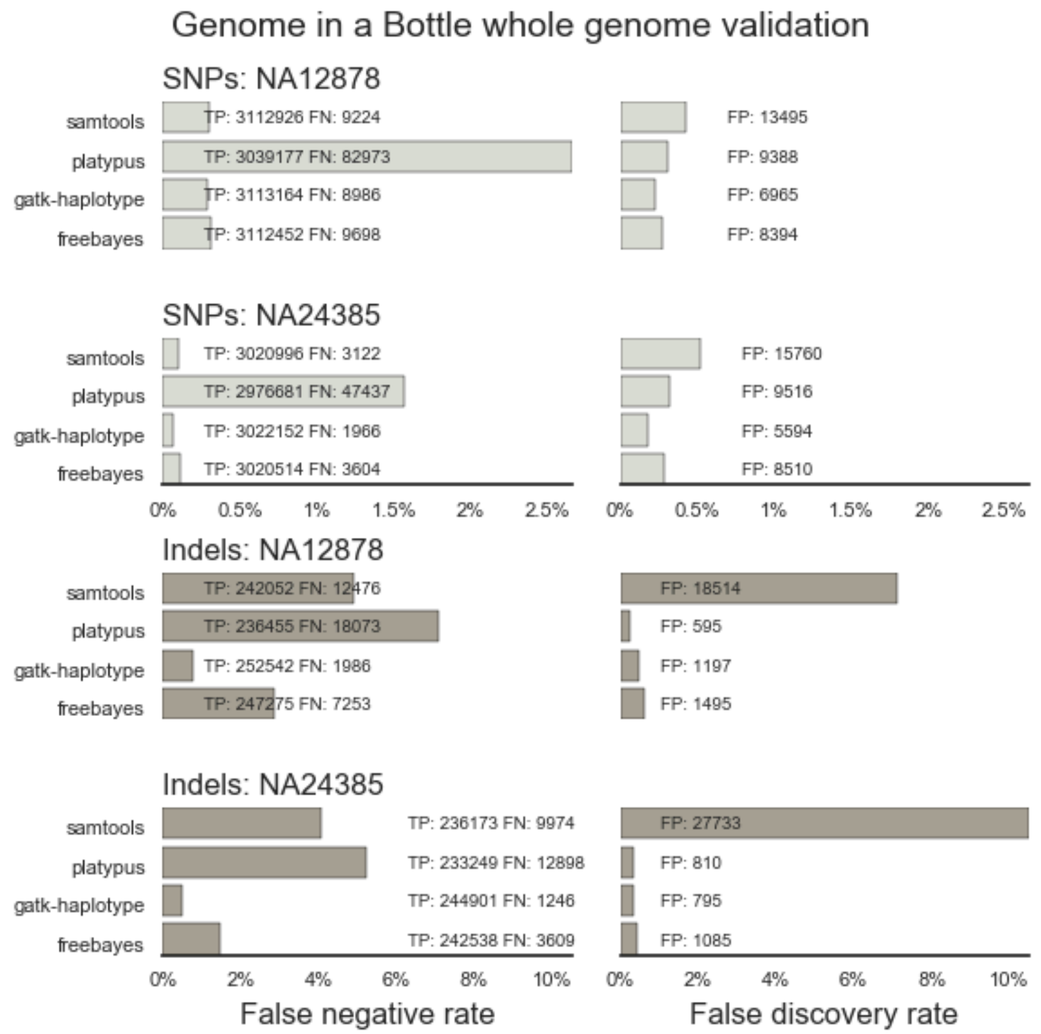
```
########################################
# HugeSeq                              #
# The Variant Detection Pipeline       #
########################################

-- DEPENDENCIES

+ ANNOVAR version 20110506
+ BEDtools version 2.16.2
+ BreakDancer version 1.1
+ BreakSeq Lite version 1.3
+ BWA version 0.6.1
+ CNVnator version 0.2.2
+ GATK version 1.6-9
+ JDK version 1.6.0_21
+ Modules Release 3.2.8
+ Perl
+ Picard Tools version 1.64
+ Pindel version 0.2.2
+ Plantation version 2
+ pysam version 0.6
+ Python version 2.7
+ Simple Job Manager version 1.0
+ Tabix version 0.1.5
+ VCFtools version 0.1.5
```

HARVARD T.H. CHAN
SCHOOL OF PUBLIC HEALTH

IT platform diversity

Method differences - what's best

https://imgur.com/a/xXyXi



Whole genome, deep coverage v1

Best Practice Variant Detection with the GATK v2

RETIRED: Best Practice Variant Detection with the GATK v3

Best Practice Variant Detection with the GATK v4, for release 2.0 [RETIRED]

Mark_DePristo Posts: 153 Administrator, GSA Member admin

July 2012 edited February 4 in Methods and Workflows

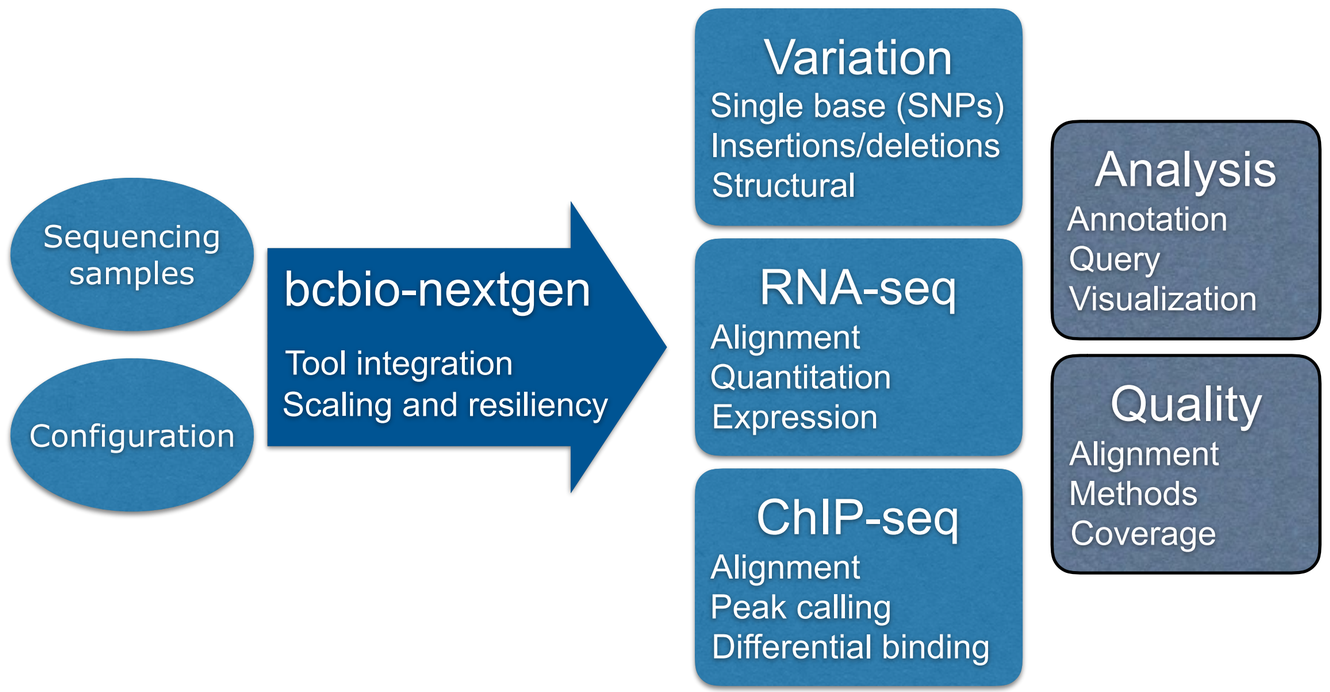
The Best Practices have been updated for GATK version 3. If you are running an older version, you should seriously consider upgrading. For more details

GATK 4.0 will be released Jan 9, 2018

Posted by Geraldine_VdAuwera on 16 Oct 2017

Complex, rapidly changing pipelines

GATK Best Practices



github.com/chapmanb/bcbio-nextgen

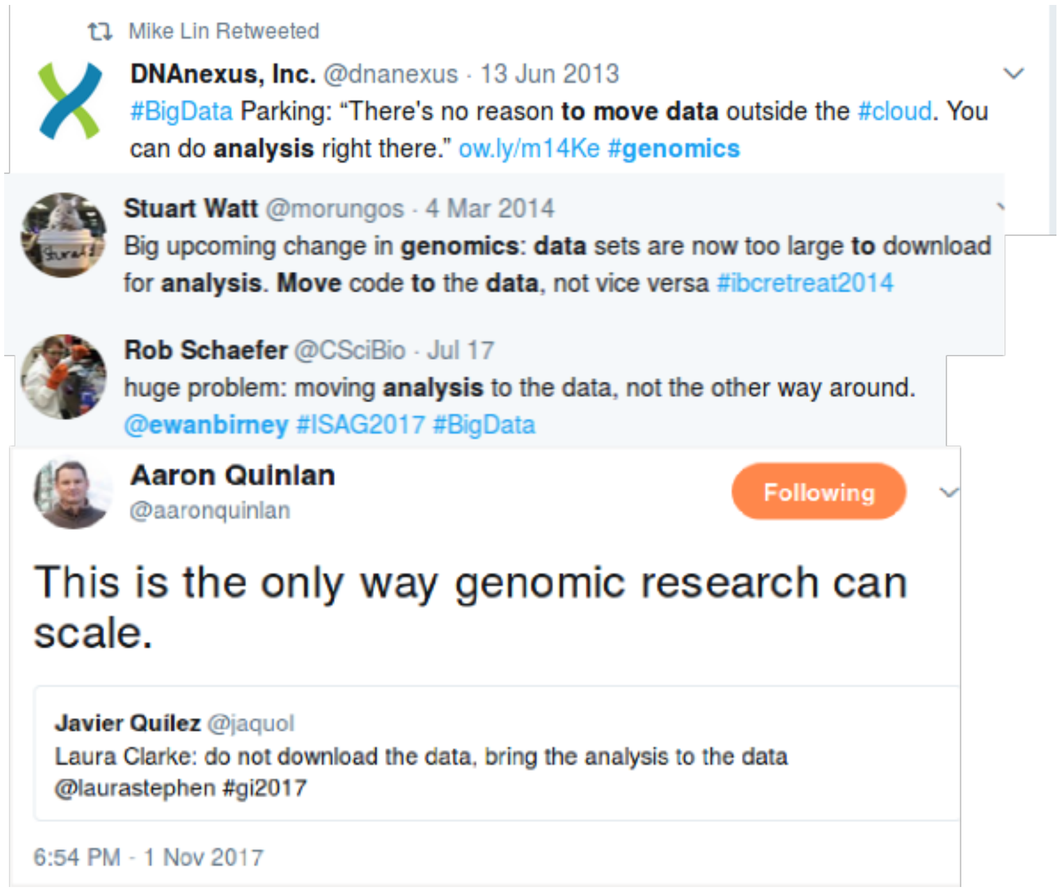
HARVARD T.H. CHAN
SCHOOL OF PUBLIC HEALTH

the bcbio-nextgen project

- **handles all dependencies**
- **scalable** works on a local machine, any cluster setup and on the cloud
- **validated** against known datasets
- **open source**
- well **documented**

Smooth transition to O2

- whole genome and exome **variant calling** including structural calling, CNVs, cancer with paired tumor/normal
- **RNA-seq**, splicing analysis, small RNAs, single cell, 3' DGE, transcriptome assembly, fusion gene calling
- **ChIP-seq** (improved flexibility and QC, removing questionable peaks)



Interoperable infrastructure

https://bcbio-nextgen.readthedocs.io/en/latest/contents/cwl.html

bcbio supports these CWL-compatible tools:

- toil – parallel local and distributed cluster runs on schedulers like SLURM, SGE and PBSPro.
- rabix bunny – multicore local runs.
- Arvados – fully parallel distributed analyses. We include an example below of running on the public Curoverse instance running on Microsoft Azure.
- Seven Bridges – parallel distributed analyses on the Seven Bridges platform and Cancer Genomics Cloud.
- cwltool – a single core analysis engine, primarily used for testing.

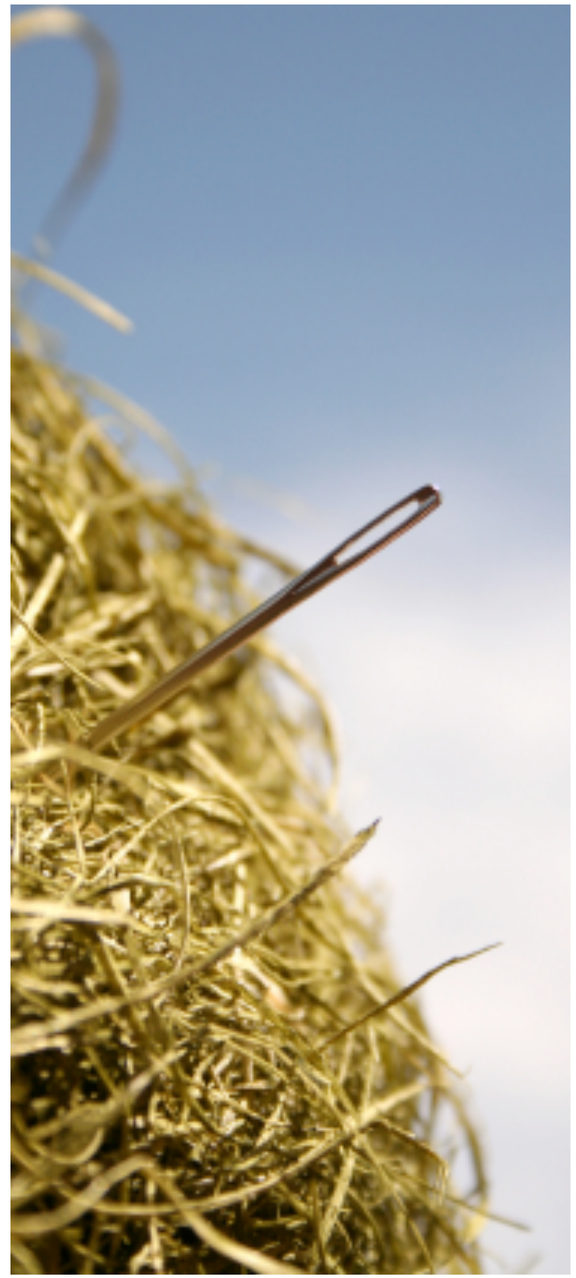
We plan to continue to expand CWL support to include more components of bcbio, and also need to evaluate the workflow on larger, real life analyses. This includes supporting additional CWL runners. We're working on supporting DNAnexus, evaluating Galaxy/Planemo for integration with the Galaxy community, and generating inputs for Broad's Cromwell WDL runner.

Getting started

bcbio-vm installs all dependencies required to generate CWL and run bcbio, along with supported

https://bcbio-nextgen.readthedocs.io/en/latest/contents/cwl.html

<applications>



Cancer variant calling

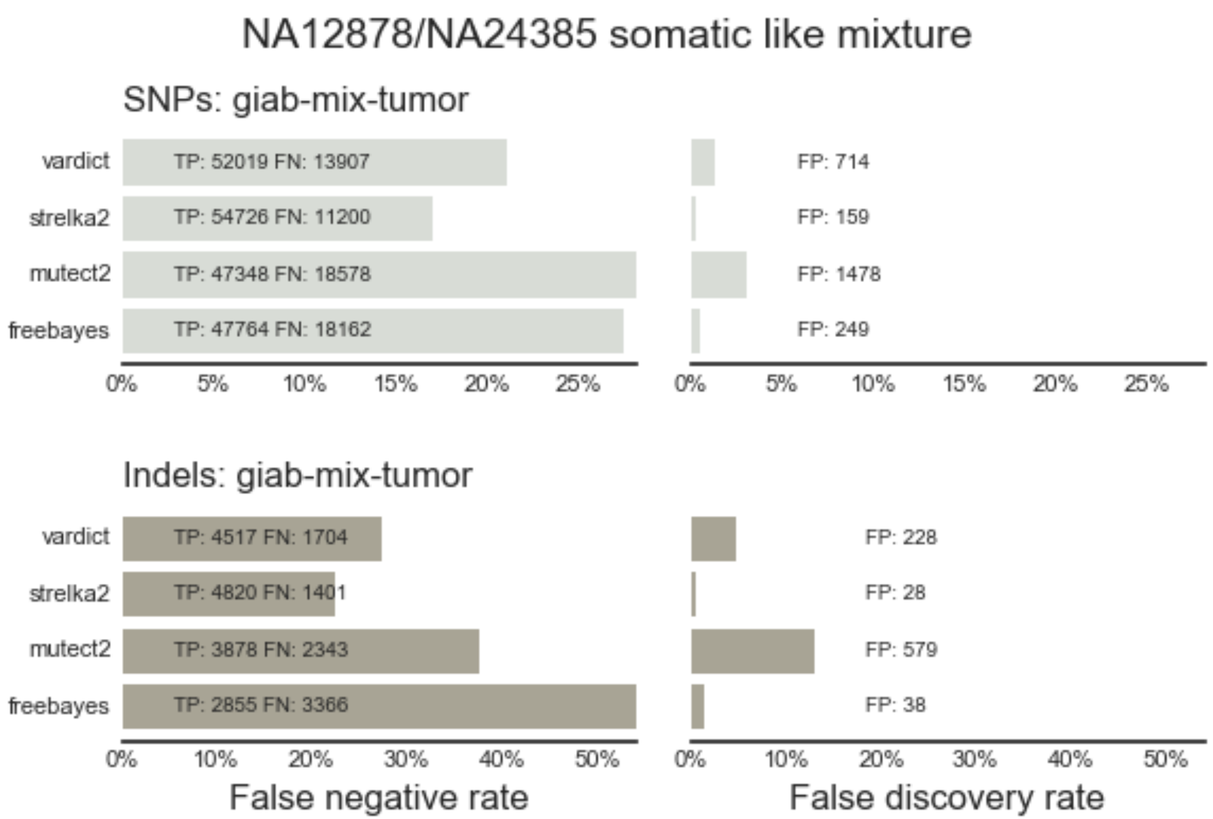
bcbio supports somatic cancer calling with tumor and optionally matched normal pairs using multiple SNP, indel and structural variant callers. A full evaluation of cancer calling validates callers against synthetic dataset 3 from the ICGC-TCGA DREAM challenge. bcbio uses a majority voting ensemble approach to combining calls from multiple SNP and indel callers, and also flattens structural variant calls into a combined representation.

The example configuration for the Cancer tumor normal validation is a good starting point for setting up a tumor/normal run on your own dataset. The configuration works similarly to population based calling. Supply a consistent batch for tumor/normal pairs and mark them with the phenotype:

```
- description: your-tumor
  metadata:
    batch: batch1
    phenotype: tumor
- description: your-normal
  metadata:
    batch: batch1
    phenotype: normal
```

Other Somatic variant calling configuration options allow tweaking of the processing parameters.

Low frequency somatic calling: Multiple truth sets



- Multiple truth sets (synthetic, mixtures and real tumors)
- Multiple variant callers
- Improvements of filters and collaboration with tool authors

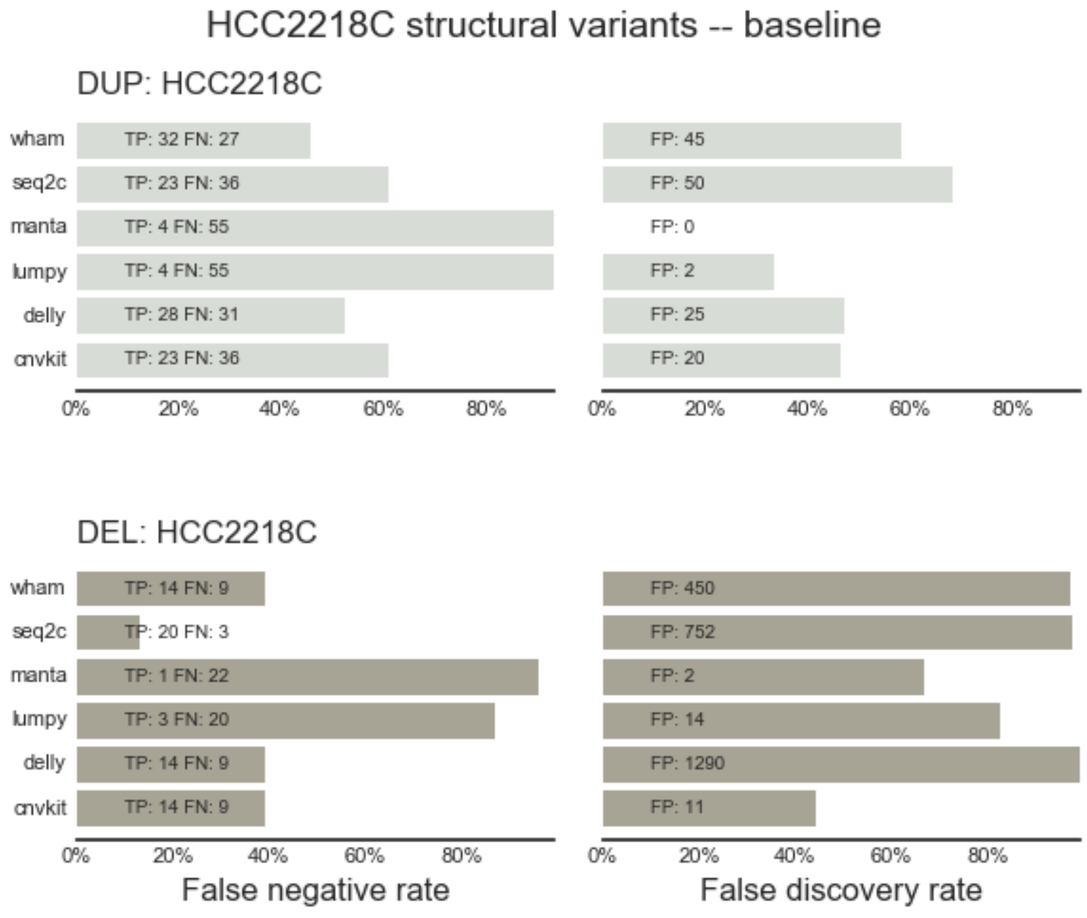
https://github.com/bcbio/bcbio_validations

HARVARD T.H. CHAN
SCHOOL OF PUBLIC HEALTH

Difficult samples: ctDNA and FFPE

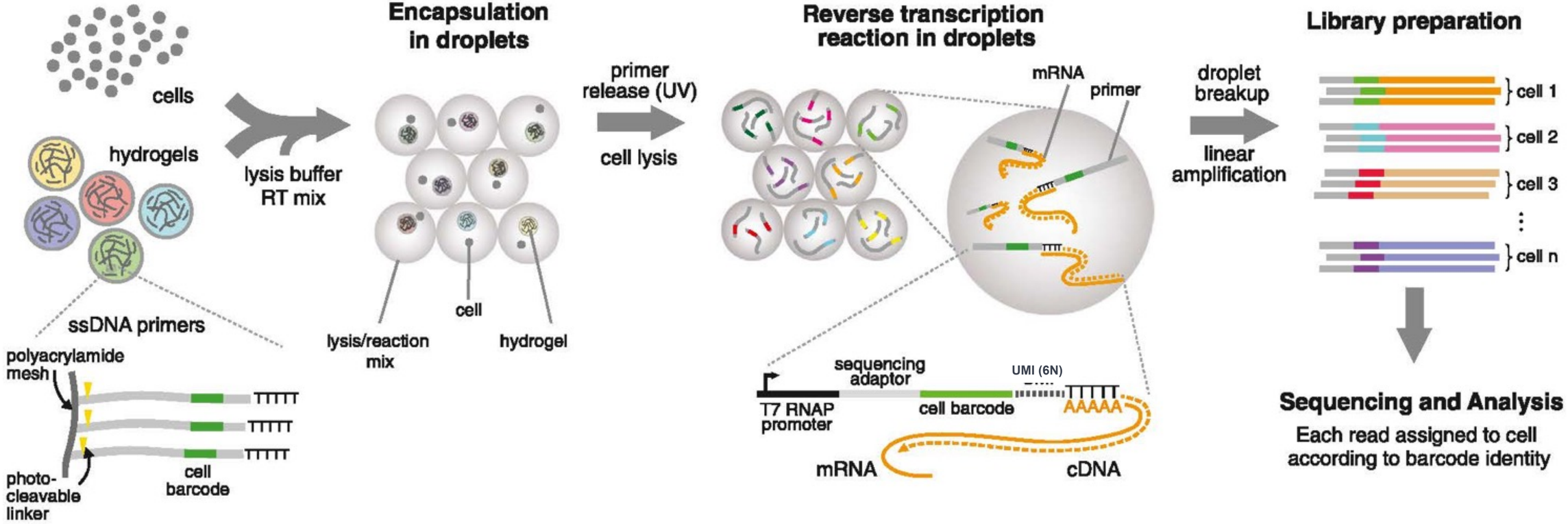
- ctDNA low frequency detection
  - UMI barcodes, distinguish duplications from biologically unique
  - Identify UMIs, consensus call and use in variant calling
  - Low frequency detection to <0.5%
  - http://fulcrumgenomics.github.io/fgbio/
- FFPE damage
  - Persistent source of false positives in low frequency calls
  - Detect strand specific bias using triplet sequence context
  - Flagging likely damage artifacts
  - https://github.com/eilslabs/DKFZBiasFilter

Structural variant validation

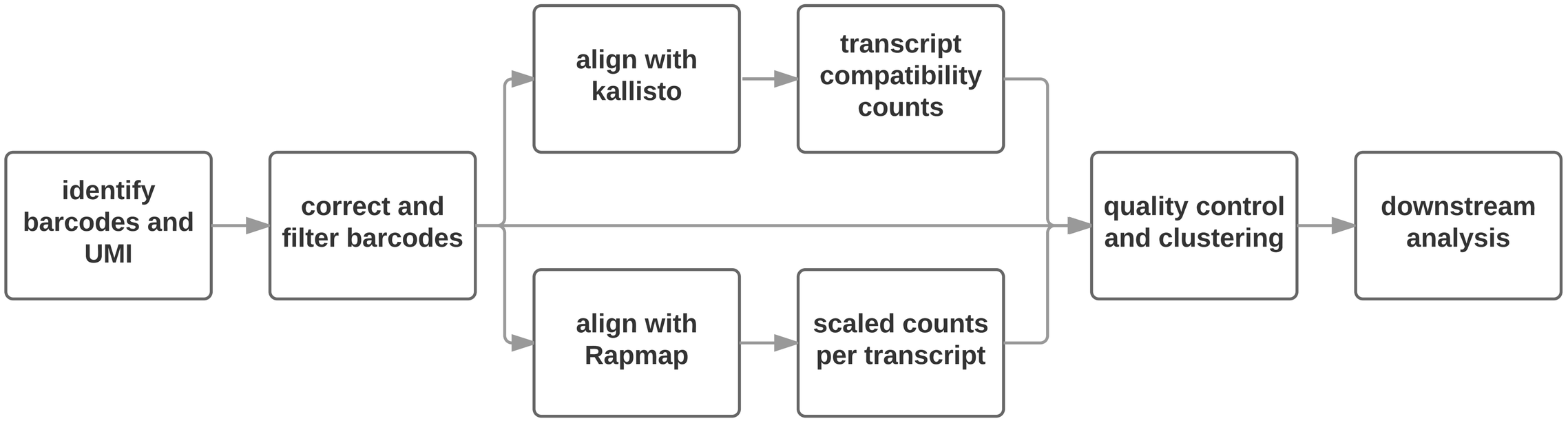


- Multiple callers
- Multiple truth sets
  - Genome in a Bottle germline
  - HCC2218 exome
- Challenges
  - High false positive/negative rates
  - Imperfect truth sets
- Prioritization in cancer genes
  - https://peerj.com/articles/3166/

HARVARD T.H. CHAN
SCHOOL OF PUBLIC HEALTH



Single cell RNA-seq analysis

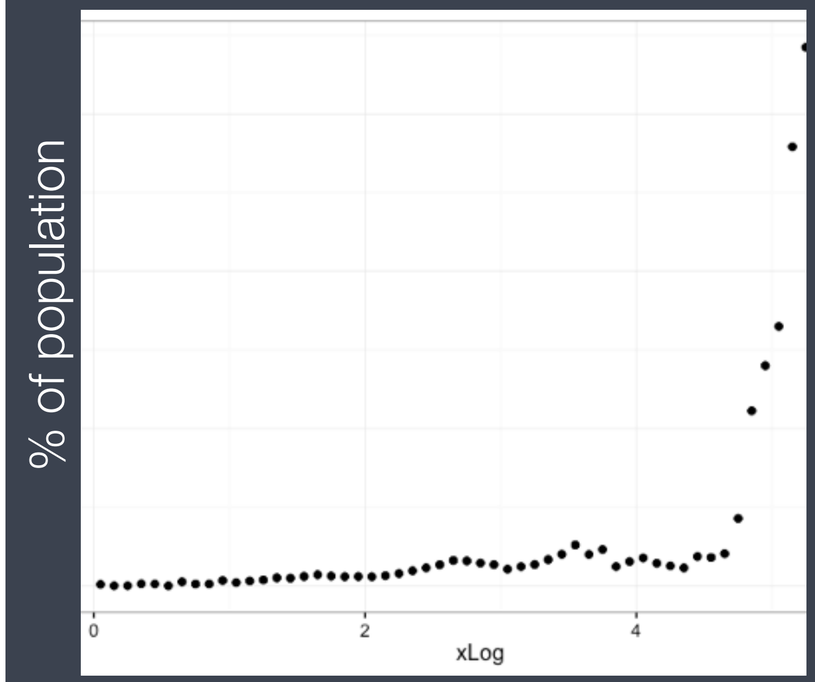


Single cell RNA-seq analysis pipeline

HARVARD T.H. CHAN
SCHOOL OF PUBLIC HEALTH

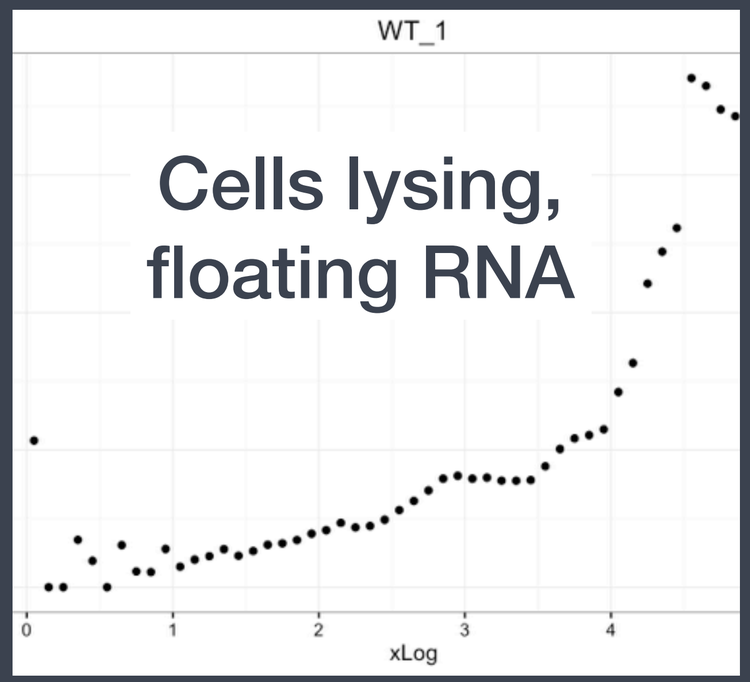
Cellular barcode distribution

perfect



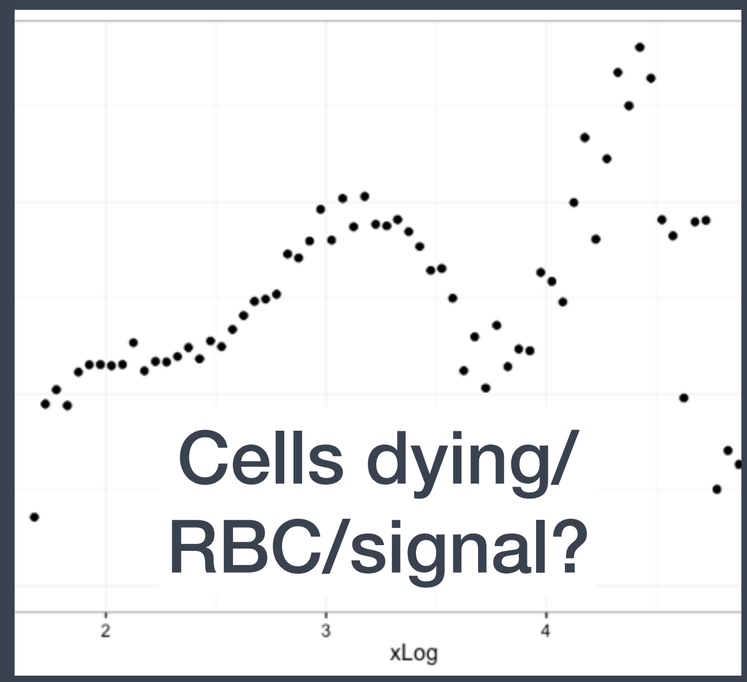
reads/cell barcode

ok



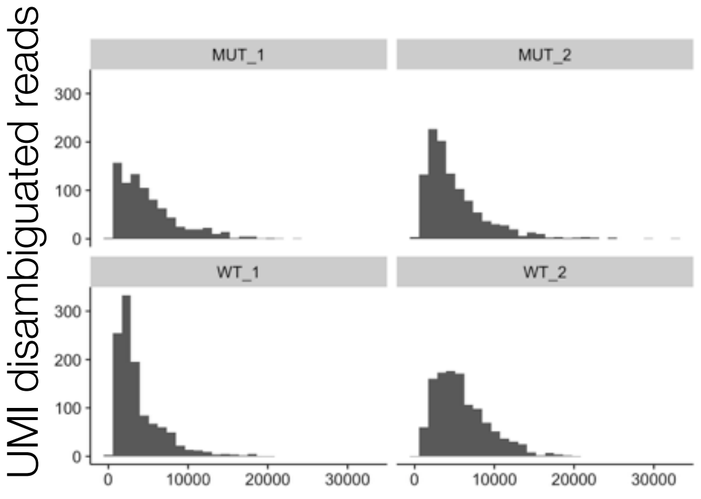
reads/cell barcode

poor



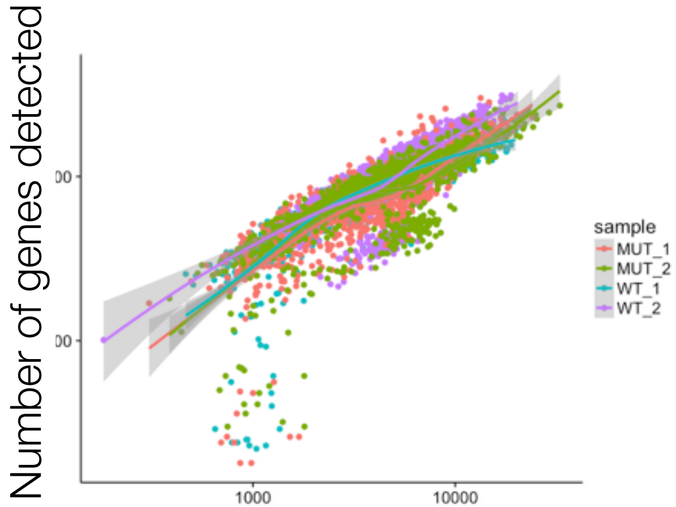
reads/cell barcode

HARVARD T.H. CHAN
SCHOOL OF PUBLIC HEALTH



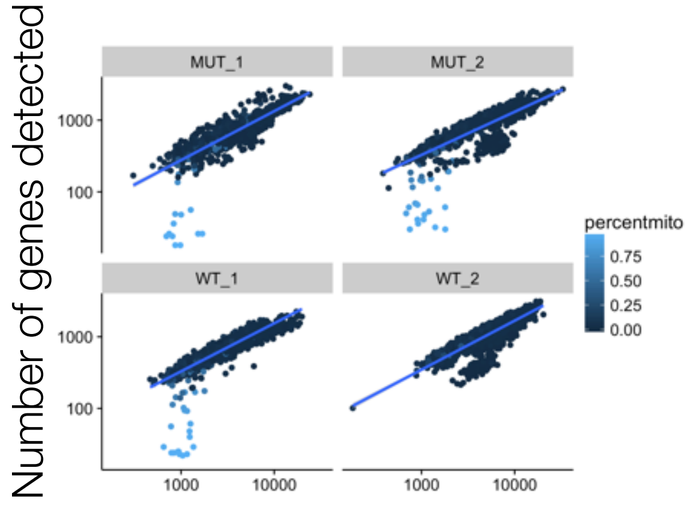
Total Counts

~5000 reads/cell
~1000 genes detected/cell



Total Counts

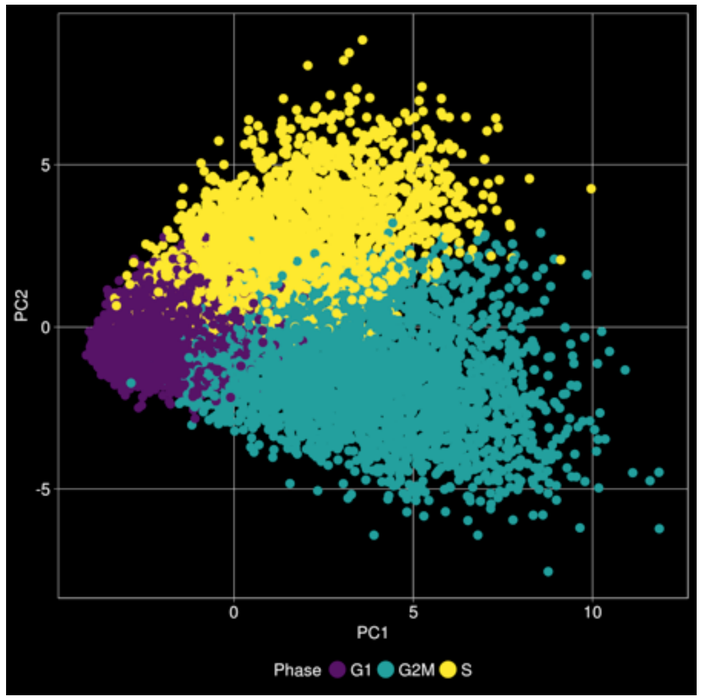
Cells with low complexity



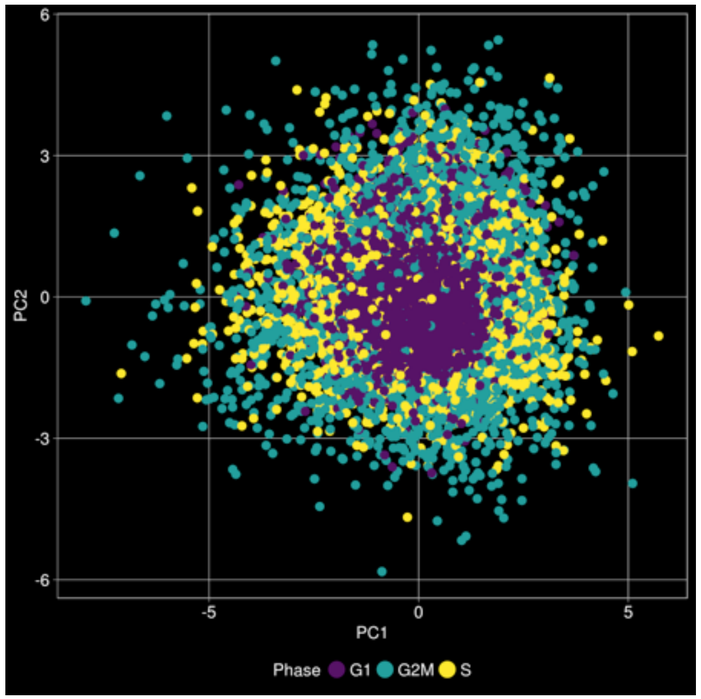
Total counts

Cells with mostly mitochondrial reads

Quality metrics



no correction

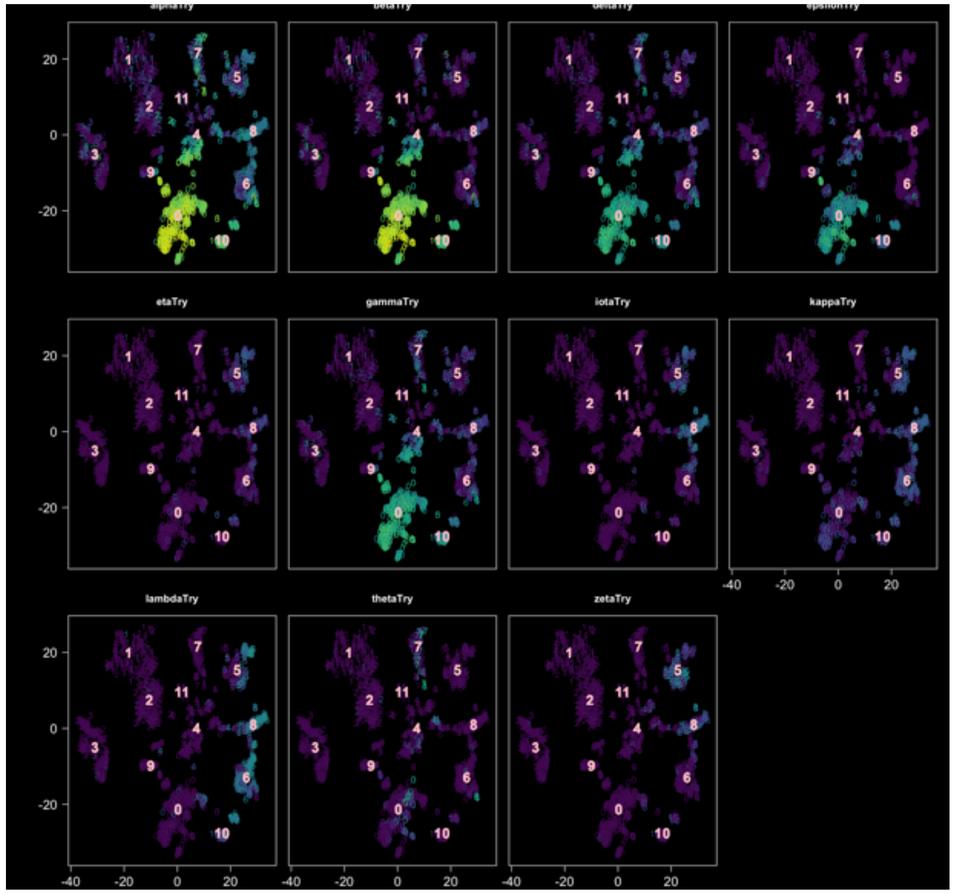


cycle regressed out

Cell cycle correction

bcbioSingleCell and bcbioRNASeq

- Bioconductor packages
- https://github.com/hbc/bcbioSingleCell
- https://github.com/hbc/bcbioRNASeq
- Templates for RNA-seq and single cell RNA-seq QC and analysis



F1000Research F1000Research 2017, 6:1976 Last updated: 20 NOV 2017

Check for updates

SOFTWARE TOOL ARTICLE

bcbioRNASeq: R package for bcbio RNA-seq analysis [version 1; referees: 1 approved with reservations]

Michael J. Steinbaugh [1*], Lorena Pantano [1*], Rory D. Kirchner [1], Victor Barrera [1], Brad A. Chapman [1], Mary E. Piper [1], Meeta Mistry [1], Radhika S. Khetani [1], Kayleigh D. Rutherford [1], Oliver Hofmann [2], John N. Hutchinson [1], Shannan Ho Sui [1]

[1] Harvard T.H. Chan School of Public Health, Boston, MA, 02115, USA
[2] University of Melbourne Center for Cancer Research, Melbourne, VIC, 3000, Australia

* Equal contributors

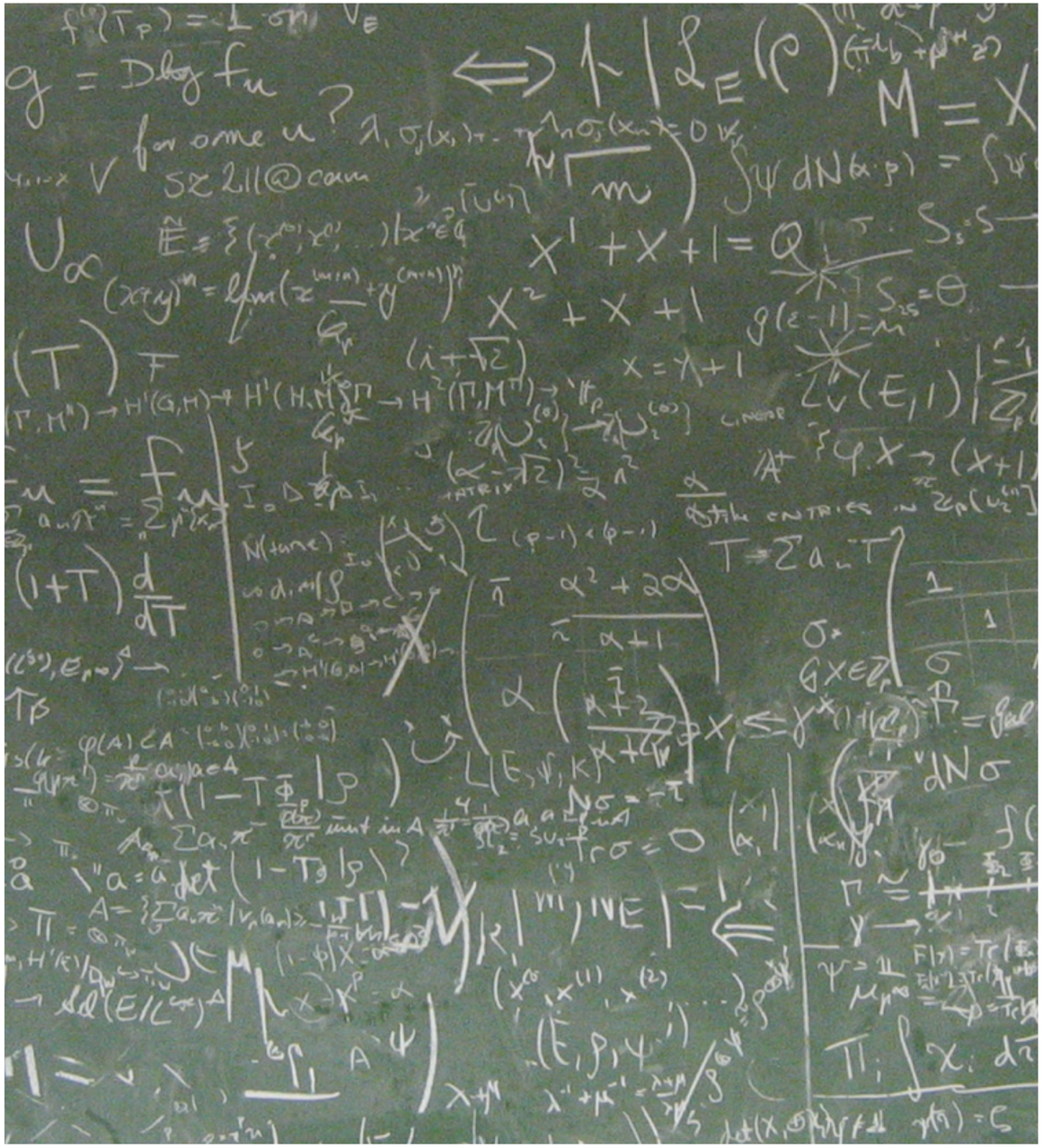
v1 **First published:** 08 Nov 2017, **6**:1976 (doi: 10.12688/f1000research.12093.1)
Latest published: 08 Nov 2017, **6**:1976 (doi: 10.12688/f1000research.12093.1)

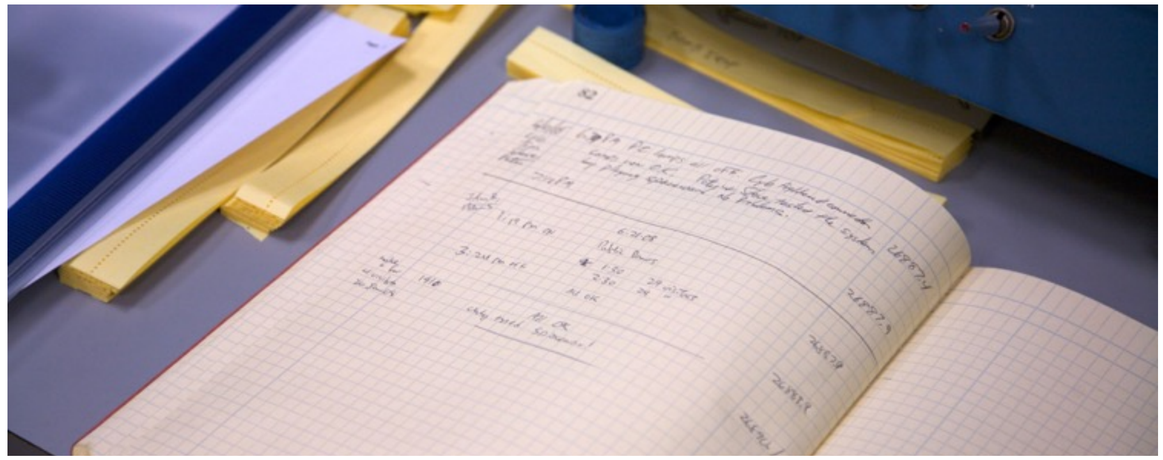
Open Peer Review

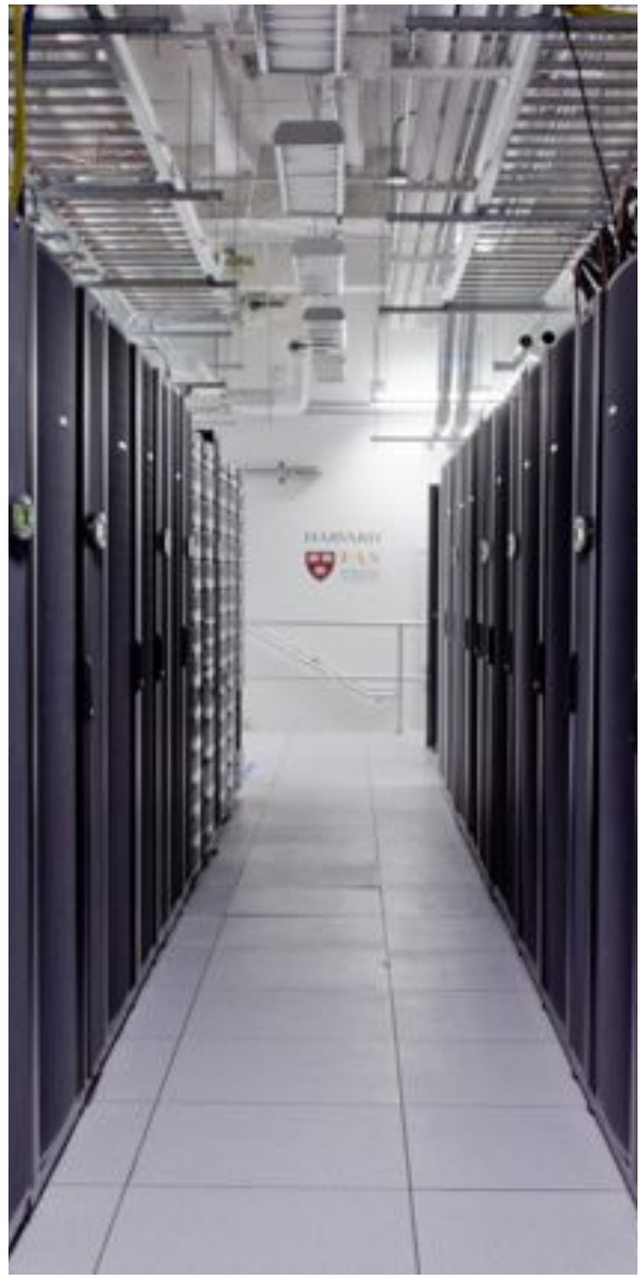
Consulting update

- 21 initial consults for grant support, experimental design, advice
- 22 projects directly supported with analysis
- Biggest users:
  - Perrimon Lab: 4 projects, 3 single cell
  - Church Lab: 5 projects
  - Brugge Lab: 1 single cell

<teaching>







Bioinformatics Training

Short courses:

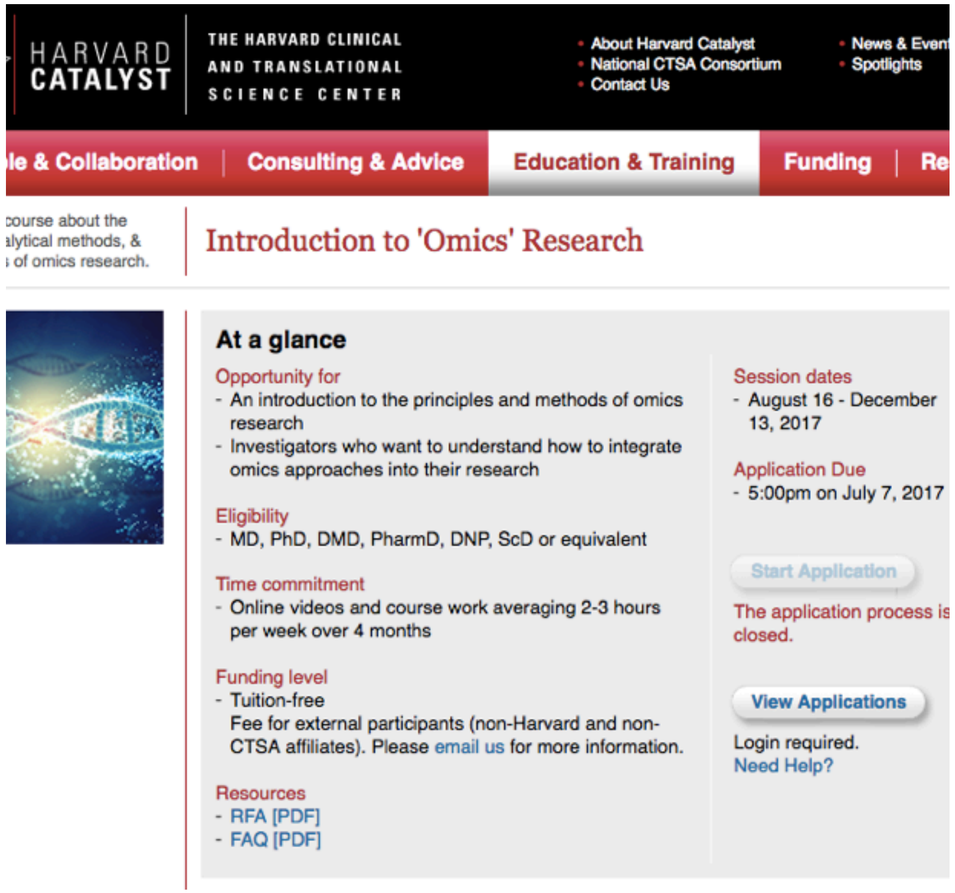
- NGS analysis
- UNIX and & compute clusters
- R
- Reproducible research (Github, Rmarkdown)

In-depth NGS course:

- UNIX & compute clusters
- NGS tools and analysis
- Functional analyses and plots with R

Training events

- May 31-Jul 7: In-depth NGS data analysis course
- Aug 23-24: Introduction to RNA-seq using high-performance computing
- Sep 13-14: Introduction to R
- Sep 25-27: Introduction to Unix (Catalyst)
- Oct 25- 26: Using DESeq2 for Differential Gene Expression
- Nov 1-3: Introduction to R: Basics, Plots, and RNA-seq Differential Expression Analysis (Catalyst)
- Nov 20-21: Introduction to RNA-seq using high-performance computing
- Dec 6-7: Introduction to R



Fall 2017 schedule:

Lessons	Date	Timing	Location	Prerequisites
Introduction to R & Visualizations with ggplot2	9/19/2017	2-4pm	HSPH, Building FXB, Room G11	None
Plotting and visualization in R using ggplot2 and other packages	10/17/2017	1-4pm	HSPH, Building FXB, Room G11	Beginner R or Intro R workshop
Functional analysis of gene lists	11/13/2017	1-4pm	TMEC, Room 328	Beginner R or Intro R workshop
Reproducible research using R (Rmarkdown: report generation)	12/11/2017	1-4pm	HSPH, Buliding FXB, Room G13	Beginner R or Intro R workshop

New workshop series: Current Topics in Bioinformatics

Trainings scheduled for 2018

- Jan 17-19: DataFest
- Jan 23-25: Introduction to Unix with RNA-seq
- Feb 27-Mar 1: Introduction to R with Differential Gene Expression (DGE)
- Apr 5-6: Introduction to R (Catalyst)
- Apr 12-13: Differential Gene Expression (Catalyst)

Summary

- Hired Michael Steinbaugh and Kayleigh Rutherford
- Increased our rate to $150/hour on July 1, 2017
- Transitioned to O2
- Began organizing our HMS data
- Continued to develop single cell RNA-seq QC and analysis pipeline
- Improvements to bcbio ChIP-seq and DGE pipelines
- Taught the in-depth NGS analysis course (5/31 - 7/7)
- New series of short modules
- Taught modules in BMI715 - Computing Skills for Biomedical Sciences (with Nils Gehlenborg)
- Templates to process screening data from ICCB-L
- ~20 labs supported this year

Contact us

Consults: **bioinformatics@hsph.harvard.edu**

Training: **hbctraining@hsph.harvard.edu**

Web: **bioinformatics.sph.harvard.edu**