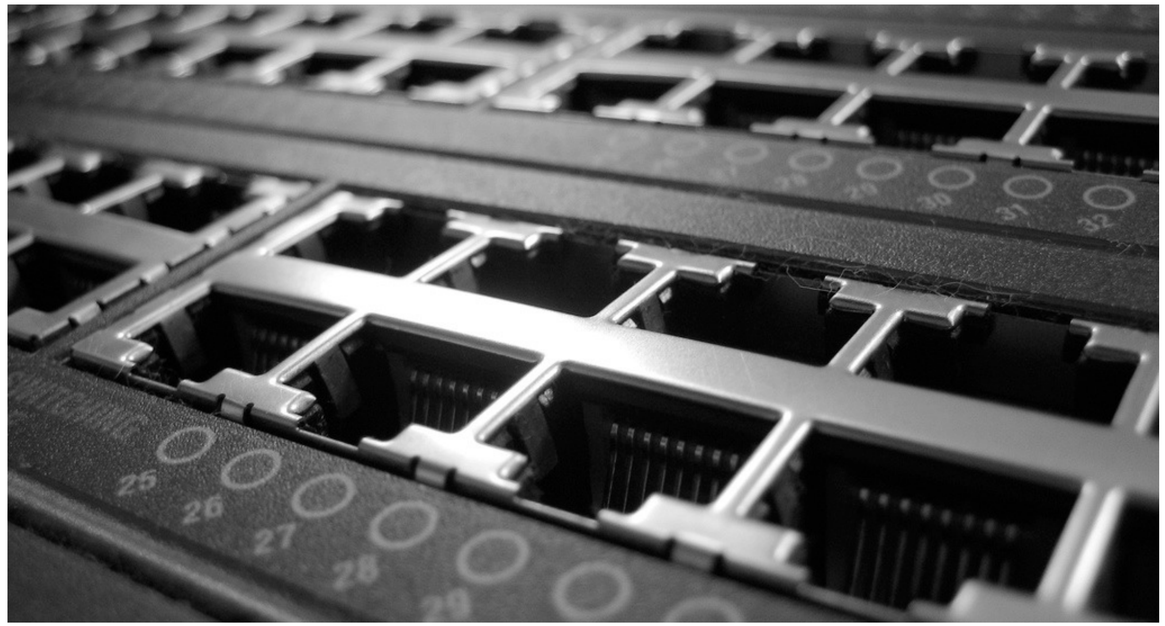


Bioinformatics in the Age of Big Data

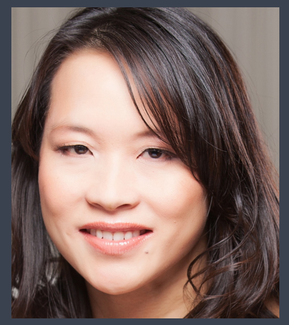
Shannan Ho Sui

Harvard T.H. Chan School of Public Health

April 26, 2017

<intro to core>

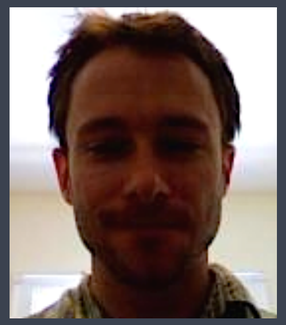
Harvard Chan Bioinformatics Core



Shannan Ho Sui

John Hutchinson

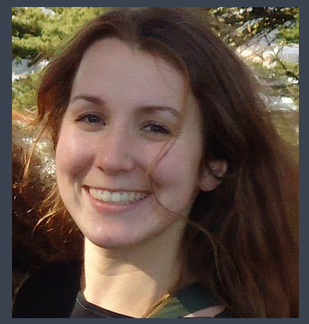
Brad Chapman



Rory Kirchner

Meeta Mistry

Radhika Khetani



Mary Piper

Lorena Pantano

Oliver Hofmann

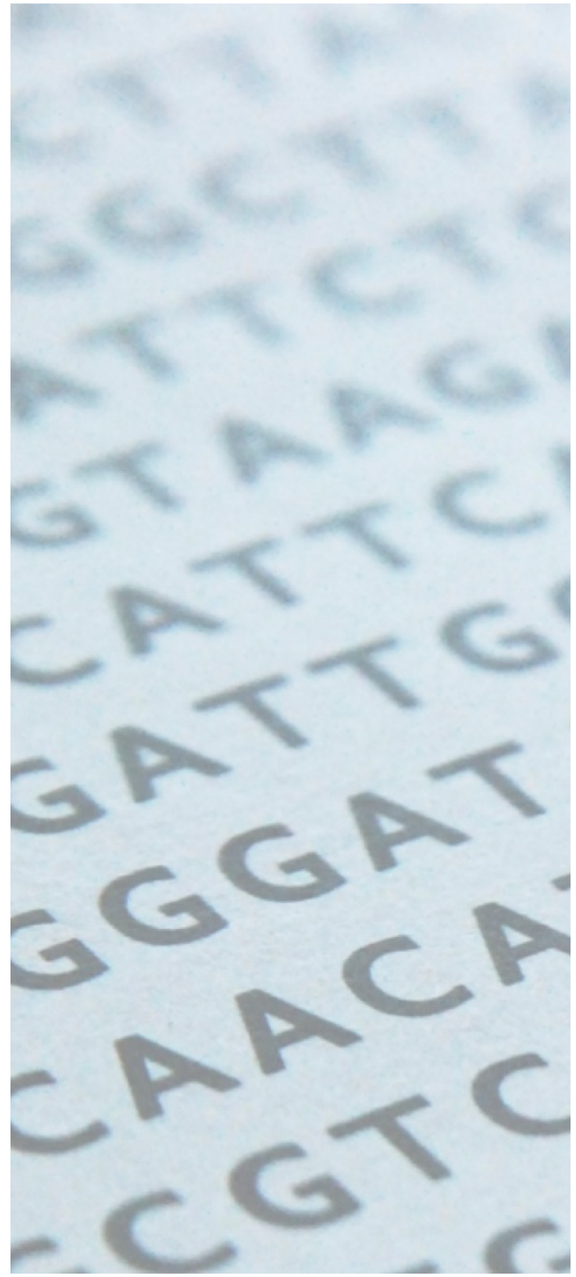
Peter Kraft

Harvard Chan Bioinformatics Core

Services offered by HBC

Consulting:

- RNA-seq, small RNA-seq and ChIP-seq
- Genome-wide methylation
- WGS, resequencing, exome-seq and structural variantion
- Gene expression arrays (microarrays)
- Functional enrichment
- Grant support



Services offered by HBC

Consulting:

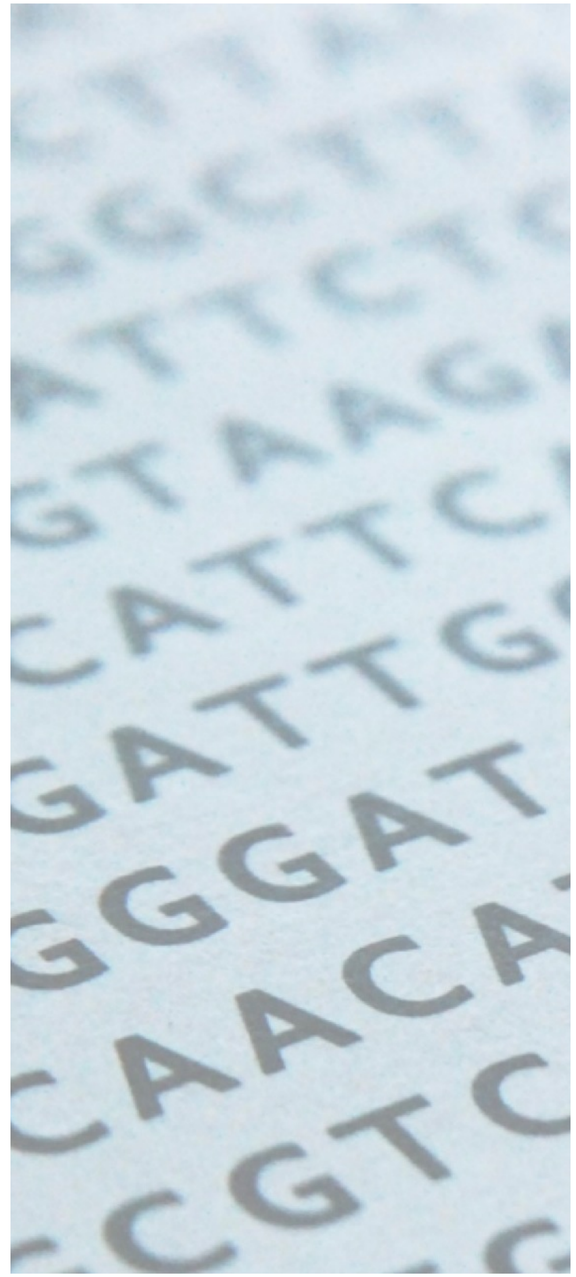
- RNA-seq, small RNA-seq and ChIP-seq
- Genome-wide methylation
- WGS, resequencing, exome-seq and structural variantion
- Gene expression arrays (microarrays)
- Functional enrichment
- Grant support

NGS-focused bioinformatics training:

Galaxy-based NGS analysis, Introductory and intermediate R,

Introductory Python, Introduction to Unix and HPC,

In-depth courses, and other.



HARVARD T.H. CHAN SCHOOL OF PUBLIC HEALTH

HARVARD STEM CELL INSTITUTE

HARVARD CATALYST
THE HARVARD CLINICAL AND TRANSLATIONAL SCIENCE CENTER

HARVARD MEDICAL SCHOOL

NIEHS / CFAR Bioinformatics Core

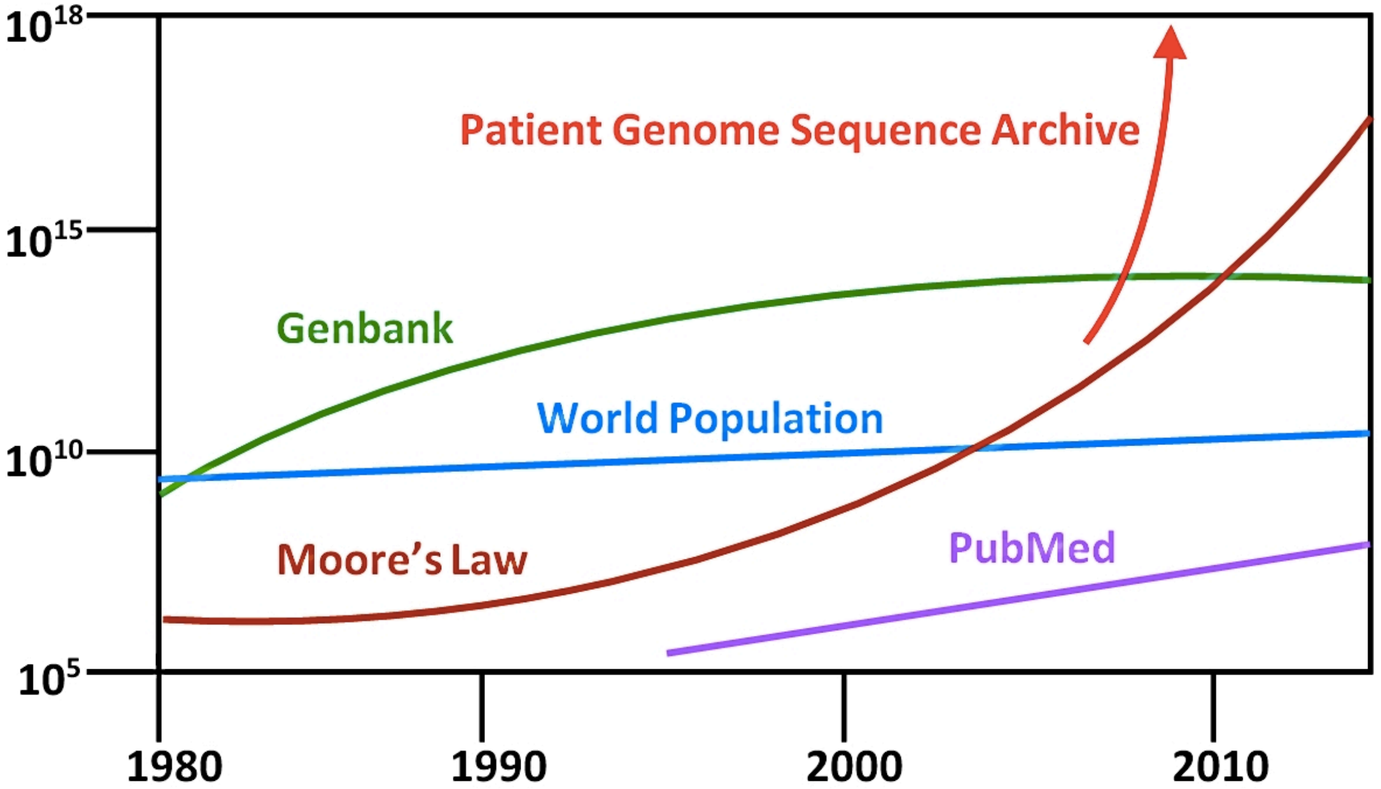
Center for Stem Cell Bioinformatics

Harvard Catalyst Bioinformatics Consulting

HMS Tools & Technology

Harvard NeuroDiscovery Center

</intro to core>



https://www.osehra.org

Figure 1: Approximate Growth of Different Data Populations

The pace of innovation in genomic data creation is much higher than the rate of innovation within genomic informatics

OPEN ACCESS

PERSPECTIVE

Big Data: Astronomical or Genomical?

Zachary D. Stephens, Skylar Y. Lee, Faraz Faghri, Roy H. Campbell, Chengxiang Zhai, Miles J. Efron, Ravishankar Iyer, Michael C. Schatz, Saurabh Sinha, Gene E. Robinson

Published: July 7, 2015 • https://doi.org/10.1371/journal.pbio.1002195

Article | Authors | Metrics | Comments | Related Content

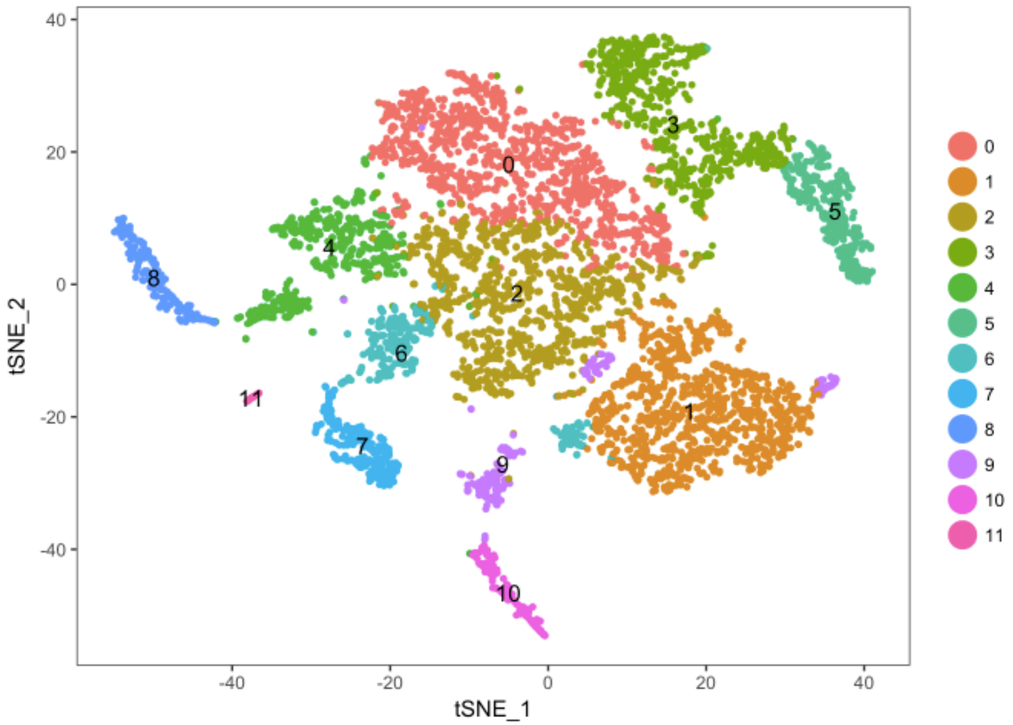
Abstract

Genomics is a Big Data science and is going to get much bigger, very soon, but it is not known whether the needs of genomics will exceed other Big Data domains. Projecting to the year 2025, we compared genomics with three other major generators of Big Data: astronomy, YouTube, and Twitter. Our estimates show that genomics is a "four-headed beast"—it is either on par with or the most demanding of the domains analyzed here in terms of data acquisition, storage, distribution, and analysis. We discuss aspects of new technologies that will need to be developed to rise up and meet the computational challenges that genomics poses for the near future. Now is the time for concerted, community-wide planning for the "genomical" challenges of the next decade.

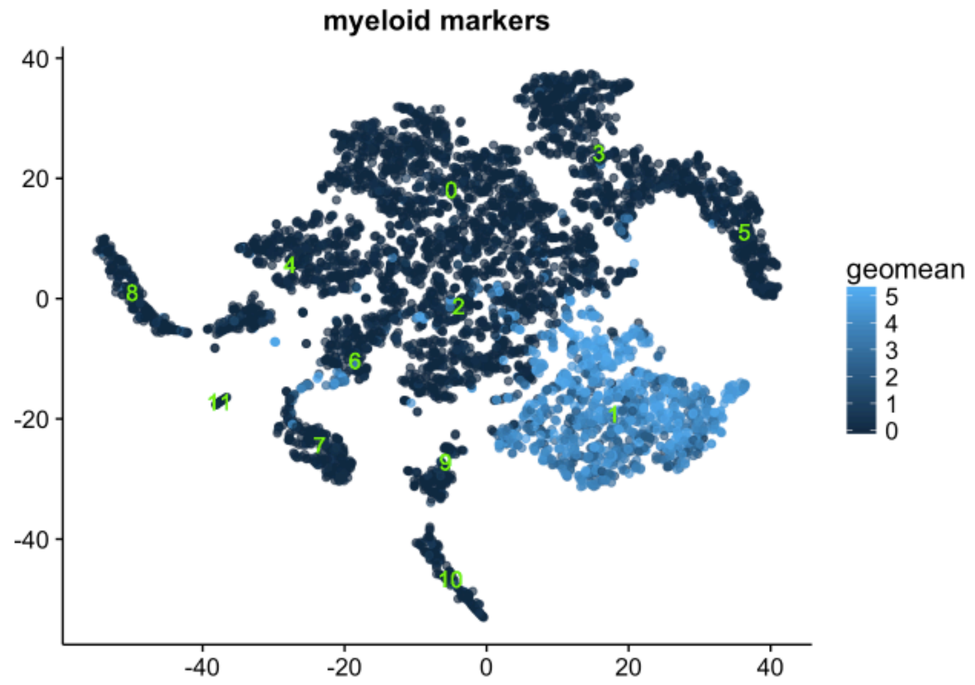
"A call to arms for big-data problems that span disciplines and that could benefit from a coordinated approach — such as the relative dearth of career paths for computational specialists in science, and the need for specialized types of storage and analysis capacity that will not necessarily be met by industrial providers."

- Gene Robinson (UIUC)

Data acquisition, distribution, storage, analysis



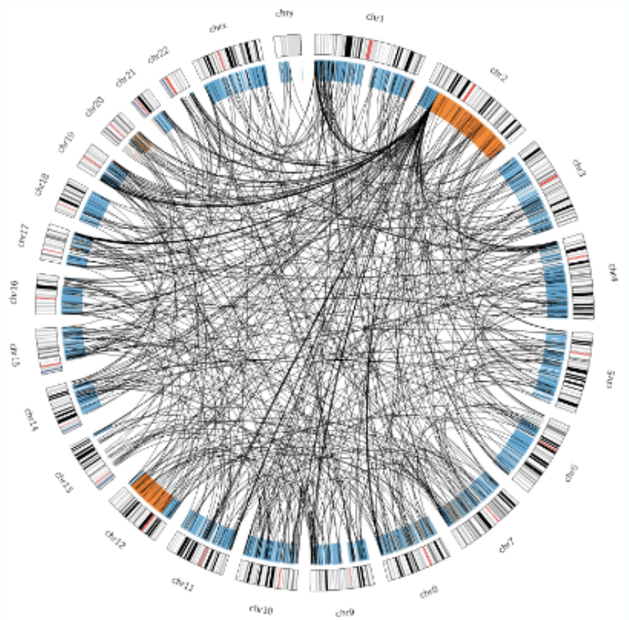
Myeloid markers



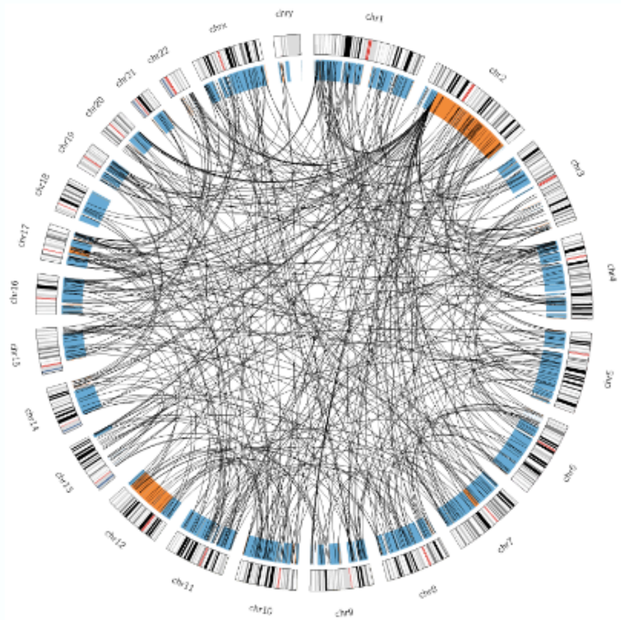
Single cell RNA sequencing

Subpopulations of cells in blood

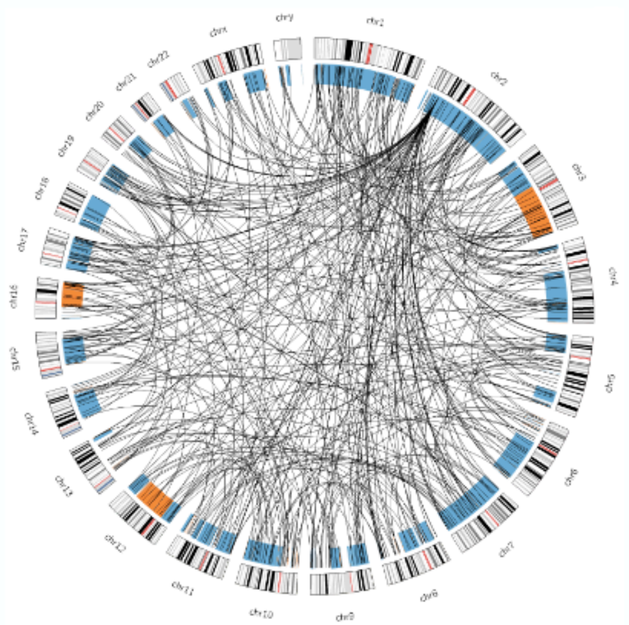
C5_KO2



C1C5_KO



C5_KO_P1C3C512



Variant discovery

Single cell RNA sequencing (scRNA-seq)

- Only detecting a small proportion of the overall number of genes expressed
- Lots of open problems like trying to model the dropout rate and visualizing this multidimensional data in ways we can easily interpret

Estimated Number of Cells
1,306,127

Post-Normalization Mean Reads per Cell
18,402

Median Genes per Cell
1,927

Sequencing

Pre-Normalization Number of Reads	30,282,984,60
Post-Normalization Number of Reads	24,036,055,22

1.3 Million Brain Cells from E18 Mice

Chromium Megacell Demonstration (v2 Chemistry) Dataset by Cell Ranger 1.3.0

Cells from cortex, hippocampus and subventricular zone of two E18 mice

- Combined cortex, hippocampus, and subventricular zone were purchased from BrainBits (C57EHCV). They were from 2 E18 C57BL/6 mice dissected on the same day, shipped overnight on ice, and stored at 4C until being prepared for scRNA-Seq.
- Brain tissues were dissociated following the Demonstrated Protocol for Mouse Embryonic Neural Tissue.
- 69 scRNA-Seq libraries were made from first mouse brain 2 days after the dissection. Another 64 scRNA-Seq libraries were made from second mouse brain 6 days after the dissection.
- 26bp Read 1 (16bp Chromium barcode and 10bp UMI), 98bp Read 2 (transcript), and 8bp i7 sample barcode for each scRNA-Seq library
- Sequenced on 11 Illumina Hiseq 4000 flow cells, and each sample was sequenced on multiple flow cells. Samples were downsampled and aggregated with approximately 18,500 reads per cell.
- 1,306,127 cells detected in total
- "aggr - Gene/cell matrix HDF5 (filtered)" contains the filtered matrix in HDF5 format. To load and process the matrix, please see our guided Python tutorial. Note that the file is too large to be loaded in R.
- "matrix of sampled 20k cells" contains the filtered gene-cell-barcode matrix of a randomly sampled 20k cells. This file can be opened in R using the function `get_matrix_from_h5` in the Cell Ranger: R Kit.
- Cell Ranger commands used to produce this aggregated dataset:

```
cellranger count --cells=10000 ...

cellranger aggr --id=neuron_aggregation --csv=aggregator.csv --nosecondary

cellranger reanalyze --id=neuron_reanalyze --matrix=filtered_matrix.h5 --params=reanalyze.csv
```

Not possible to work with this data in R

Cold Spring Harbor Laboratory

bioRxiv beta
THE PREPRINT SERVER FOR BIOLOGY

HOME | AI

Search

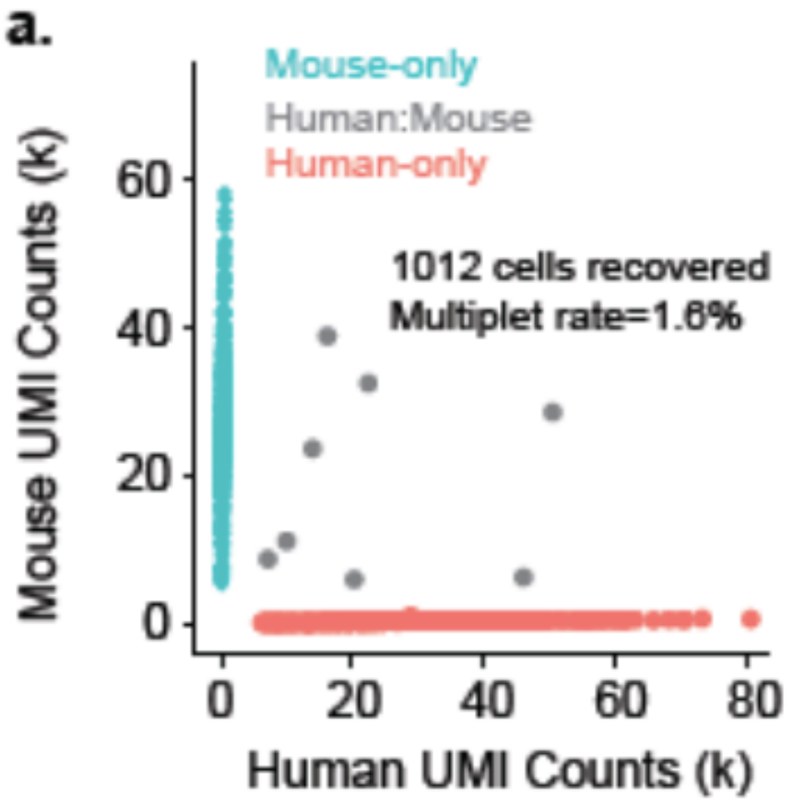
New Results

On the widespread and critical impact of systematic bias and batch effects in single-cell RNA-Seq data

Stephanie C Hicks, Mingxiang Teng, Rafael A Irizarry

doi: https://doi.org/10.1101/025528

Technical and experimental design problems



Single cell RNA-seq is not always single cell

Doublets are very difficult to detect computationally

Technical and experimental design problems

http://distill.pub/2016/misread-tsne/

illumina®

SIGN IN | VIEW CART | CONTACT US | SEARCH

ALL SYSTEMS | Sequencing Platforms | Microarray Scanners | NGS PLATFORM SELECTOR | QUESTIONS

Benchtop Sequencers | Production-Scale Sequencers




	NextSeq Series	HiSeq Series	HiSeq X Series†	NovaSeq Series
Popular Applications & Methods	Key Application	Key Application	Key Application	Key Application
Large Whole-Genome Sequencing (human, plant, animal)	●	●	●	●
Small Whole-Genome Sequencing (microbe, virus)	●	●		●
Exome Sequencing	●	●		●
Targeted Gene Sequencing (amplicon, gene panel)	●	●		●
Whole-Transcriptome Sequencing	●	●		●
Gene Expression Profiling with mRNA-Seq	●	●		●
miRNA & Small RNA Analysis	●	●		●
DNA-Protein Interaction Analysis	●	●		●
Methylation Sequencing	●	●		●

Cold Spring Harbor Laboratory

bioRxiv beta
THE PREPRINT SERVER FOR BIOLOGY

HOME | ABOUT | SUBMIT | ALERTS

Search

New Results

Previous

Index Switching Causes "Spreading-Of-Signal" Among Multiplexed Samples In Illumina HiSeq 4000 DNA Sequencing

Rahul Sinha, Geoff Stanley, Gunsagar Singh Gulati, Camille Ezran, Kyle Joseph Travaglini, Eric Wei, Charles Kwok Fai Chan, Ahmad N Nabhan, Tianying Su, Rachel Marie Morganti, Stephanie Diana Conley, Hassan Chaib, Kristy Red-Horse, Michael T Longaker, Michael P Snyder, Mark A Krasnow, Irving L Weissman

doi: https://doi.org/10.1101/125724

Posted April 9, 2017.

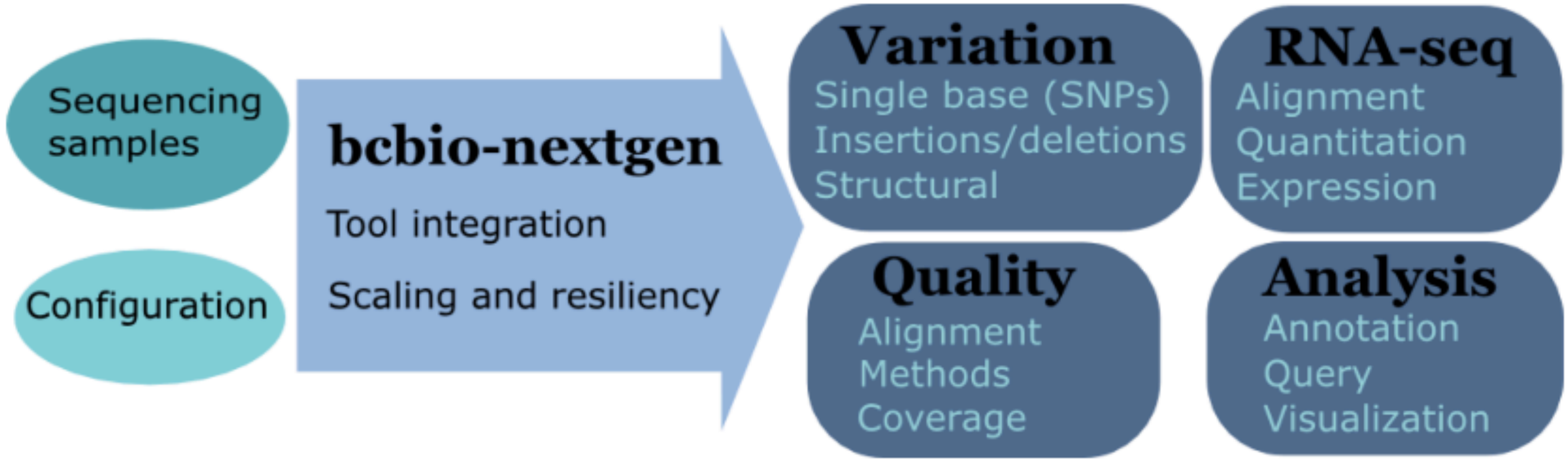
Download PDF

Email

How do we keep up?

- These are shared problems (academic, industry, startups)
- Build open source communities
- Community developed analyses
- Benchmarking and validation
- Interoperable infrastructure

White box software



https://github.com/chapmanb/bcbio-nextgen

Large scale infrastructure development

- Easy to configure
- “Best practice” methods
- Provide the community with a collected set of expertise, installation of tools and data, tool integration, validation, and scaling

- Aligners: bwa, novoalign, bowtie2, HISAT2
- Variantion: FreeBayes, GATK, VarDict, MuTecT2, Scalpel, SnpEff, VEP, GEMINI, Lumpy, Manta, CNVkit, WHAM
- RNA-seq: Tophat, STAR, Cufflinks, Sailfish
- Quality control: FastQC, samtools, Qualimap MultiQC
- Manipulation: bedtools, bcftools, biobambam picard, sambamba, samblaster, samtools, vcfl vt

Whole genome, deep coverage v1

Warning: the material on this page is considered out of date by the GSA team.

Best Practice Variant Detection with the GATK v2

Warning: the material on this page is considered out of date by the GSA team.

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
July 2012 edited February 4

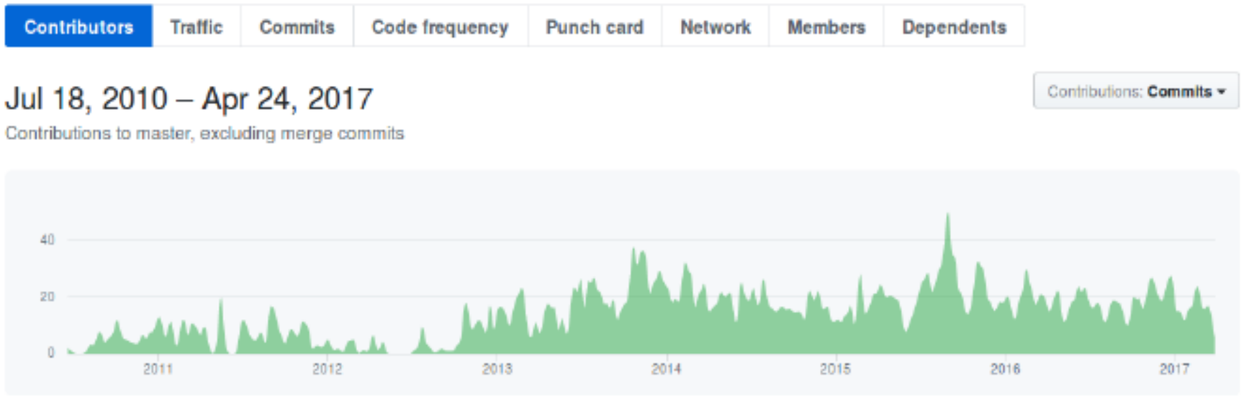
The Best Practices have been updated for GATK version 3. If you are running an older version, you should seriously consider upgrading. For more details

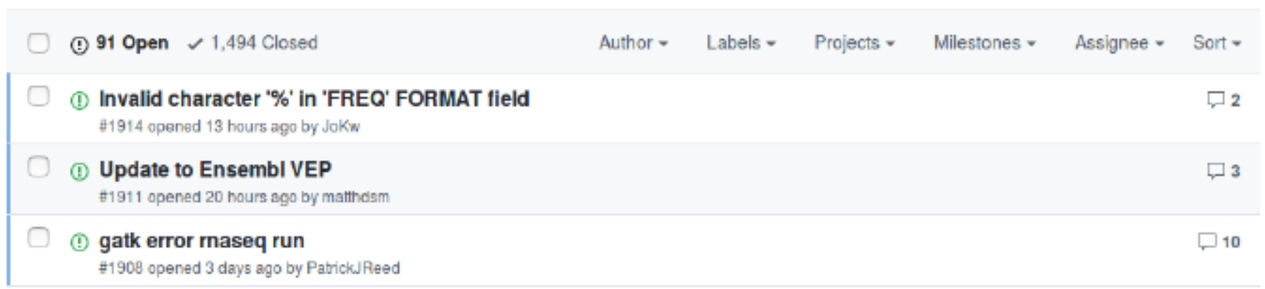
Complex, rapidly changing baseline functionality

A piece of software is being sustained if people are using it, fixing it, and improving it rather than replacing it.

http://software-carpentry.org/blog/2014/08/sustainability.html

It also needs to be sustained





https://github.com/chapmanb/bcbio-nextgen

The community provides sustainability and support

Validation

- Baseline for improving methods
- Unbiased algorithm comparisons

Genome in a Bottle Consortium

Global Alliance for Genomics & Healt

GC-TCGA DREAM Mutation Calling cha

//www.genomeinabottle.org/
//ga4gh.org/#/benchmarking-team
://www.synapse.org/#!Synapse:syn312572

Interoperability

- Bring compute and tools to the data using better abstractions

COMMON
WORKFLOW
LANGUAGE

- The future of genomic data is rich with promise and challenge
- Our goal is to build infrastructure and use state of the art methods to help answer these questions while keeping up with changing technologies and methods

These materials have been developed by members of the team at the Harvard Chan Bioinformatics Core (HBC). These are open access materials distributed under the terms of the Creative Commons Attribution license (CC BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.