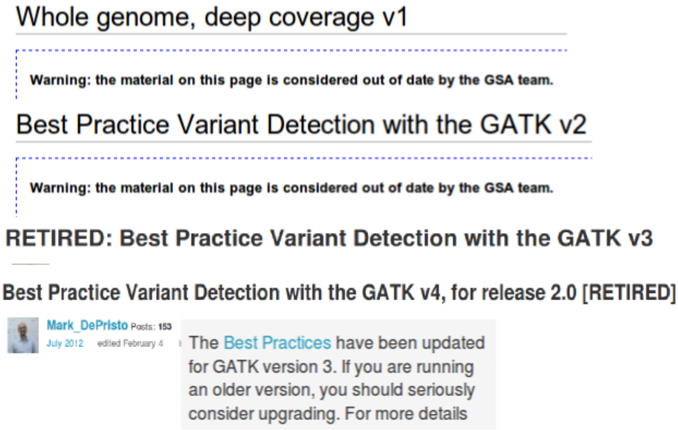
Supporting dynamic community developed biological pipelines

Brad Chapman
Bioinformatics Core, Harvard School of Public Health
https://github.com/chapmanb
http://j.mp/bcbiolinks

17 April 2014

Complex, rapidly changing pipelines



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Warning: the material on this page is considered out of date by the GSA team.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Warning: the material on this page is considered out of date by the GSA team.

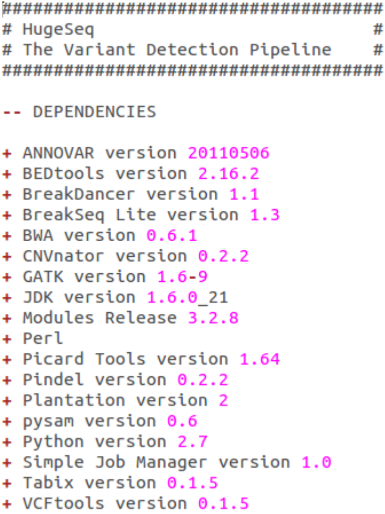
RETIRED: Best Practice Variant Detection with the GATK v3

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

Large number of specialized dependencies



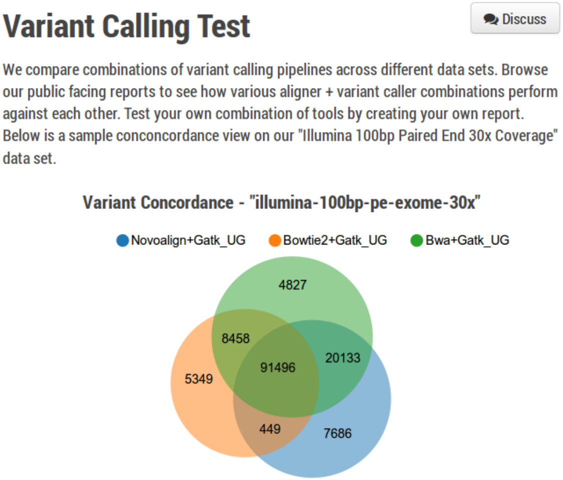
```
#####################################
# HugeSeq                           #
# The Variant Detection Pipeline    #
#####################################

-- 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
```

https://github.com/StanfordBioinformatics/HugeSeq

Quality differences between methods



Variant Calling Test

We compare combinations of variant calling pipelines across different data sets. Browse our public facing reports to see how various aligner + variant caller combinations perform against each other. Test your own combination of tools by creating your own report. Below is a sample conconcordance view on our "Illumina 100bp Paired End 30x Coverage" data set.

Variant Concordance - "illumina-100bp-pe-exome-30x"

http://www.bioplanet.com/gcat

Scaling on full ecosystem of clusters

DRMAA
Distributed Resource Management
Application API — www.drmaa.org

openlava
Open Source Workload Management

GRID ENGINE

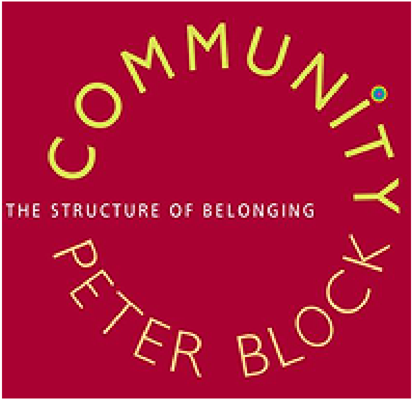
slurm
workload manager

amazon
web services™

Platform LSF

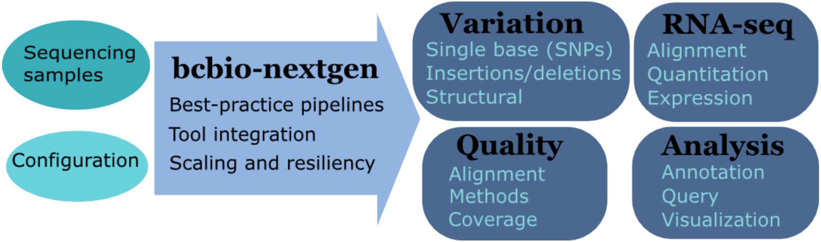
Torque

Solution



http://www.amazon.com/Community-Structure-Belonging-Peter-Block/dp/1605092770

Overview



Sequencing samples

Configuration

bcbio-nextgen

Best-practice pipelines

Tool integration

Scaling and resiliency

Variation

Single base (SNPs)

Insertions/deletions

Structural

RNA-seq

Alignment

Quantitation

Expression

Quality

Alignment

Methods

Coverage

Analysis

Annotation

Query

Visualization

Uses

- Aligners: bwa, novoalign, bowtie2
- Variantion: FreeBayes, GATK, VarScan, MuTecT, SnpEff
- RNA-seq: tophat, STAR, cufflinks, HTSeq
- Quality control: fastqc, bamtools, RNA-SeQC
- Manipulation: bedtools, bcftools, biobambam, sambamba, samblaster, samtools, vcflib

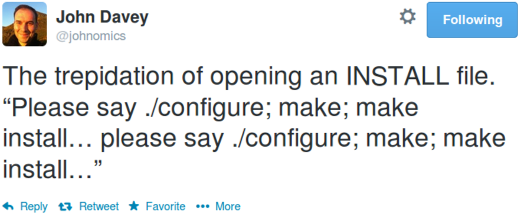
Provides

- Best practice analysis pipelines
- Tool integration
- Multi-platform support
- Scaling

Development goals

- Community developed
- Quantifiable
- Scalable
- Reproducible

Community: installation



John Davey
@johnomics

The trepidation of opening an INSTALL file. “Please say ./configure; make; make install… please say ./configure; make; make install…”

Automated Install

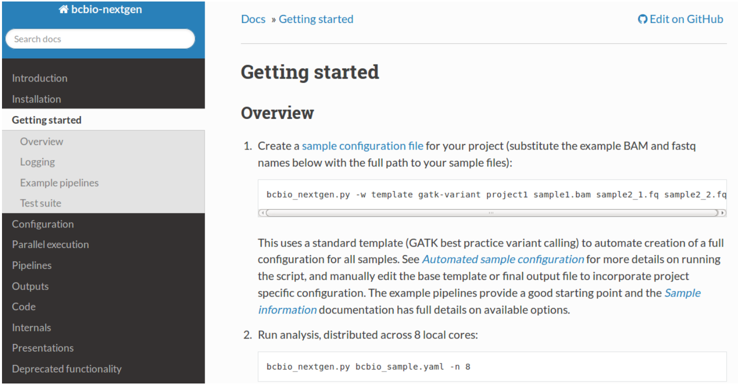
Bare machine to ready-to-run pipeline, tools and data

- CloudBioLinux: http://cloudbiolinux.org
- Homebrew: https://github.com/Homebrew/homebrew-science
- Conda: http://j.mp/py-conda

Easier install

Docker

Community: documentation



Getting started

Overview

1. Create a sample configuration file for your project (substitute the example BAM and fastq names below with the full path to your sample files):

```
bcbio_nextgen.py -w template gatk-variant project1 sample1.bam sample2_1.fq sample2_2.fq
```

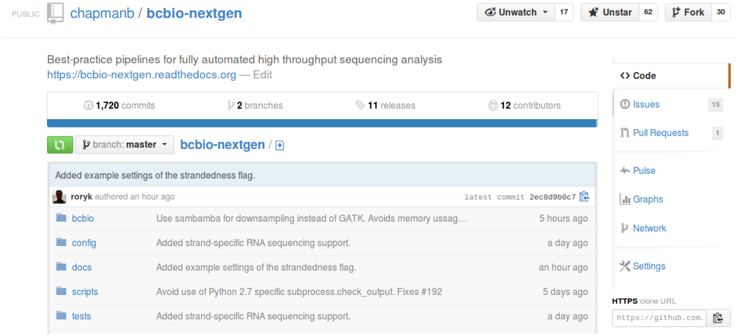
This uses a standard template (GATK best practice variant calling) to automate creation of a full configuration for all samples. See *Automated sample configuration* for more details on running the script, and manually edit the base template or final output file to incorporate project specific configuration. The example pipelines provide a good starting point and the *Sample information* documentation has full details on available options.

2. Run analysis, distributed across 8 local cores:

```
bcbio_nextgen.py bcbio_sample.yaml -n 8
```

https://bcbio-nextgen.readthedocs.org

Community: contribution



PUBLIC chapmanb / bcbio-nextgen

Best-practice pipelines for fully automated high throughput sequencing analysis
https://bcbio-nextgen.readthedocs.org — Edit

1,720 commits · 2 branches · 11 releases · 12 contributors

Added example settings of the strandedness flag.

roryk authored an hour ago · latest commit 2ec8d9b0c7

bcbio	Use sambamba for downsampling instead of GATK. Avoids memory ussag…	5 hours ago
config	Added strand-specific RNA sequencing support.	a day ago
docs	Added example settings of the strandedness flag.	an hour ago
scripts	Avoid use of Python 2.7 specific subprocess.check_output. Fixes #192	5 days ago
tests	Added strand-specific RNA sequencing support.	a day ago

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

Validation

Tests for implementation and methods

- Currently:
  - Family/population calling
  - RNA-seq differential expression
  - Structural variations
- Expand to:
  - Cancer tumor/normal

    http://j.mp/cancer-var-chal

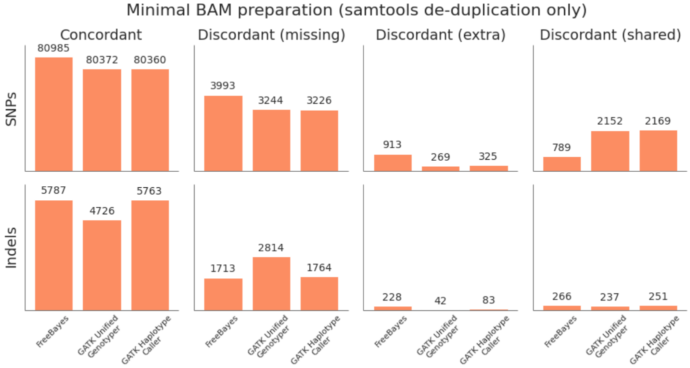
Example evaluation

- Variant calling
  - GATK UnifiedGenotyper
  - GATK HaplotypeCaller
  - FreeBayes
- Two preparation methods
  - Full (de-duplication, recalibration, realignment)
  - Minimal (only de-duplication)

Reference materials

http://www.genomeinabottle.org/

Quantify quality

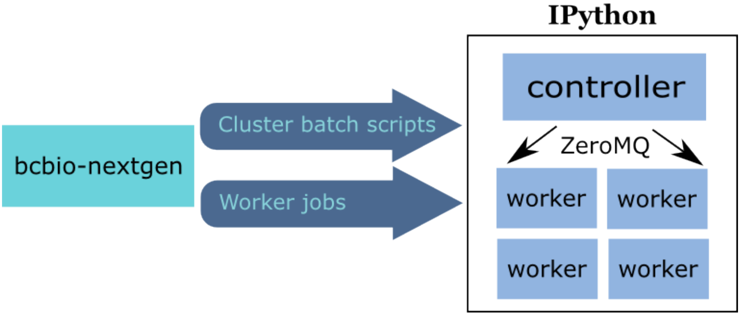


- Quantification details: `http://j.mp/bcbioeval2`

Validation enables scaling

- Little value in realignment when using haplotype aware caller
- Little value in recalibration when using high quality reads
- Streaming de-duplication approaches provide same quality without disk IO

Scaling overview



- Infrastructure details: http://j.mp/bcbioscale
- IPython: http://ipython.org/ipython-doc/dev/parallel/index.html

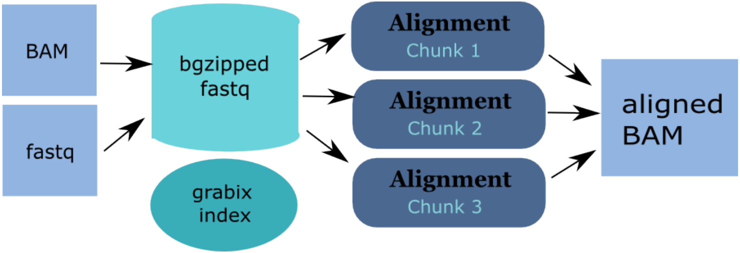
Current target environment

- Cluster scheduler
  - SLURM
  - Torque
  - SGE
  - LSF
- Shared filesystem
  - NFS
  - Lustre
- Local temporary disk
  - SSD

Scaling improvements

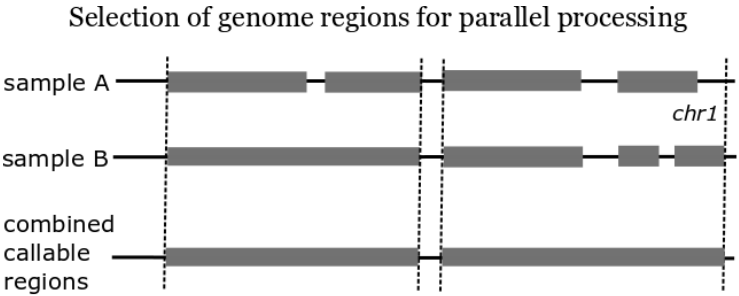
- Split alignments
- Split by genome regions
- Manage memory
- Avoid IO

Alignment parallelization



`https://github.com/arq5x/grabix`

Variant calling parallelization



Selection of genome regions for parallel processing
sample A
chr1
sample B
combined callable regions

Memory usage

Configuration

```
bwa:
  cmd: bwa
  cores: 16
samtools:
  cores: 16
  memory: 2G
gatk:
  jvm_opts: ["-Xms750m", "-Xmx2750m"]
```

Batch file

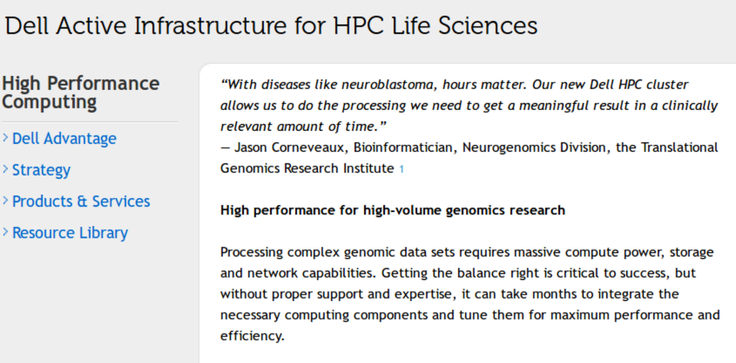
```
#PBS -l nodes=1:ppn=16
#PBS -l mem=45260mb
```

Avoid filesystem IO

Pipes and streaming algorithms

```
("{bwa} mem -M -t {num_cores} -R ’{rg_info}’ -v 1 "
 "  {ref_file} {fastq_file} {pair_file} "
 "| {samblaster} "
 "| {samtools} view -S -u /dev/stdin "
 "| {sambamba} sort -t {cores} -m {mem} --tmpdir {tmpdir}"
 "   -o {tx_out_file} /dev/stdin")
```

Dell System



Dell Active Infrastructure for HPC Life Sciences

High Performance Computing

- Dell Advantage
- Strategy
- Products & Services
- Resource Library

"With diseases like neuroblastoma, hours matter. Our new Dell HPC cluster allows us to do the processing we need to get a meaningful result in a clinically relevant amount of time."
— Jason Corneveaux, Bioinformatician, Neurogenomics Division, the Translational Genomics Research Institute 1

High performance for high-volume genomics research

Processing complex genomic data sets requires massive compute power, storage and network capabilities. Getting the balance right is critical to success, but without proper support and expertise, it can take months to integrate the necessary computing components and tune them for maximum performance and efficiency.

Glen Otero, Will Cottay

http://dell.com/ai-hpc-lifesciences

Evaluation details

System

- 400 cores
- 3Gb RAM/core
- Lustre filesystem
- Infiniband network

Samples

- 60 samples
- 30x whole genome (100Gb)
- Illumina
- Family-based calling

Timing: Alignment

Step	Time	Processes
Alignment preparation	13 hours	BAM to fastq; bgzip; grabix index
Alignment	30 hours	bwa-mem alignment
BAM merge	7 hours	Merge alignment parts
Alignment post-processing	6 hours	Calculate callable regions

Timing: Variant calling

Step	Time	Processes
Post-alignment BAM preparation	6 hours	De-duplication
Variant calling	18 hours	FreeBayes
Variant post-processing	2 hours	Combine variant files; annotate: GATK and snpEff

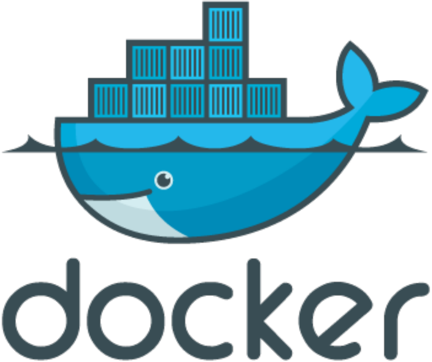
Timing: Analysis and QC

Step	Time	Processes
BAM merging	6 hours	Combine post-processed BAM file sections
GEMINI	3 hours	Create GEMINI SQLite database
Quality Control	5 hours	FastQC, alignment and variant statistics

Timing: Overall

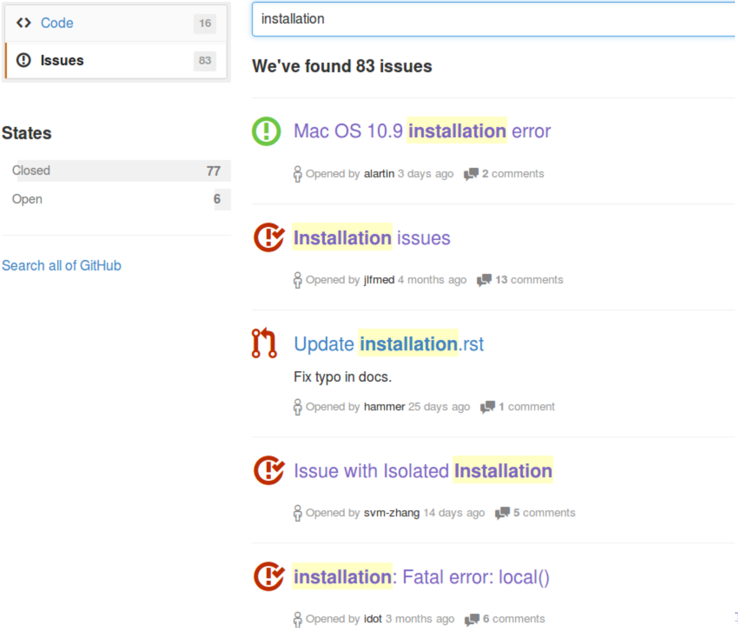
- 4 days for 60 samples
- ~2 hours per sample at 400 cores
- In progress: optimize for single samples

Reproducible environment



http://docker.io

Consistent support environment



installation

We've found 83 issues

States

Closed	77
Open	6

Search all of GitHub

Mac OS 10.9 **installation** error

Opened by alartin 3 days ago 2 comments

Installation issues

Opened by jlfmed 4 months ago 13 comments

Update **installation**.rst

Fix typo in docs.

Opened by hammer 25 days ago 1 comment

Issue with Isolated **Installation**

Opened by svm-zhang 14 days ago 5 comments

installation: Fatal error: local()

Opened by idot 3 months ago 6 comments

Docker benefits

- Fully isolated
- Reproducible – store full environment with analysis (~1Gb)
- Improved installation – single download + data

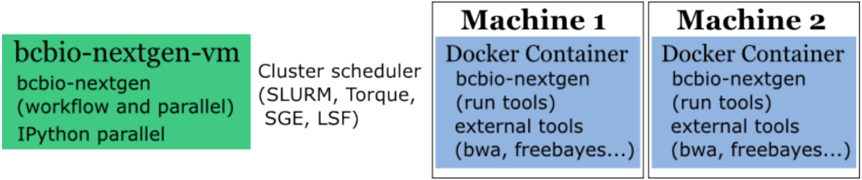
bcbio with Docker

- External Python wrapper
  - Installation
  - Start and run containers
  - Mount external data into containers
  - Parallelize
- All analysis tools inside Docker

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

http://j.mp/bcbiodocker

Docker HPC parallelization



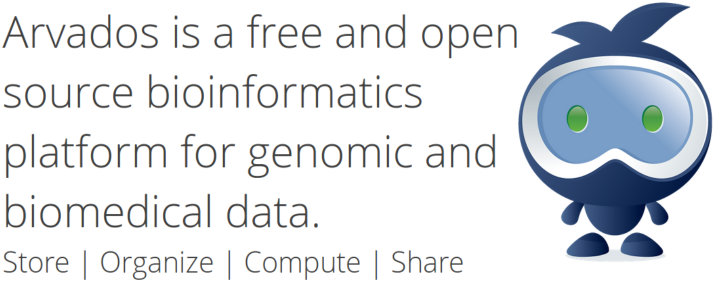
Consistent scaling environment

Amazon challenges

- Cost – spot instances
- Disk – local scratch, no EBS
- Organization – no shared filesystems, S3 push/pull
- Data – reconstitute on minimal machines
- Security – encryption at rest

Clusterk http://clusterk.com/

Program provenance



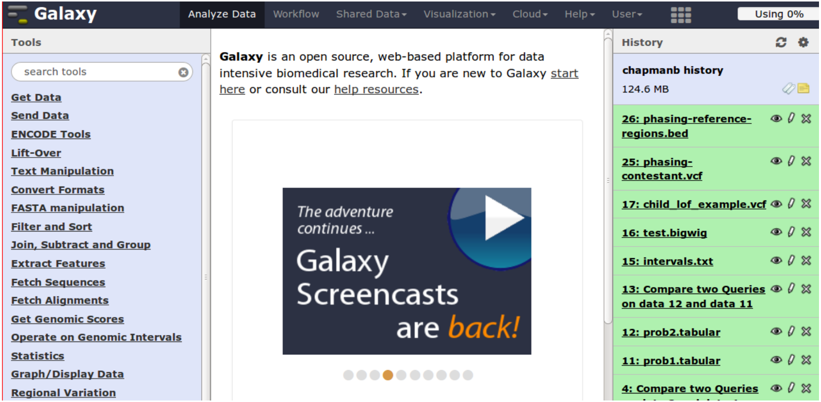
Arvados is a free and open source bioinformatics platform for genomic and biomedical data.

Store | Organize | Compute | Share

https://arvados.org/

https://curoverse.com/

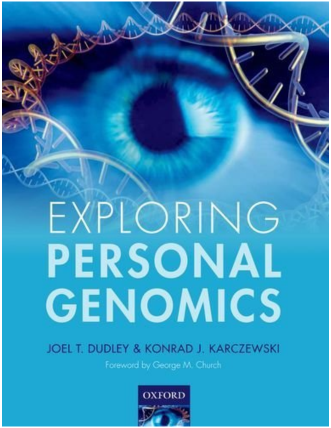
Integrated



Galaxy is an open source, web-based platform for data intensive biomedical research. If you are new to Galaxy start here or consult our help resources.

https://usegalaxy.org/

Accessible


EXPLORING PERSONAL GENOMICS
JOEL T. DUDLEY & KONRAD J. KARCZEWSKI
Foreword by George M. Church
OXFORD

http://exploringpersonalgenomics.org/

Summary

- Community developed pipelines > challenges
- Focus
  - Community: easy to install and contribute
  - Validation of methods
  - Scalability
  - Reproducibility and virtualization
- Widely accessible

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

http://j.mp/bcbiolinks