Connecting Galaxy to tools with alternative storage and compute models

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

1 July 2014

Tradeoff: Assessing code

Community > Implementation

Galaxy
Biopython: http://biopython.org
OpenBio: http://www.open-bio.org

Tradeoff: Reproducibility

Validation > Replication

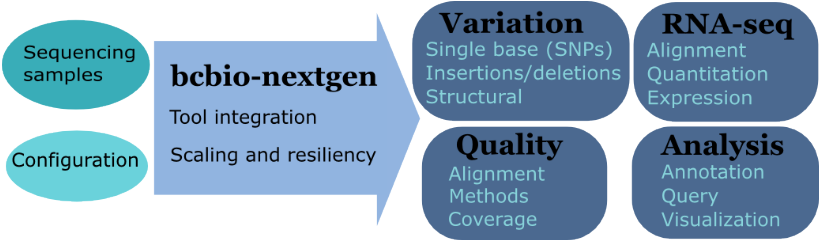
Genome in a Bottle: http://www.genomeinabottle.org/
ICGC-TCGA DREAM: https://www.synapse.org/#!Synapse:syn312572
SMaSH: http://smash.cs.berkeley.edu/

Tradeoff: Development focus

Scaling > Configurability

bcbio scaling: http://j.mp/bcbioscale

Overview



Sequencing samples

Configuration

bcbio-nextgen

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

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

Uses

- Aligners: bwa-mem, novoalign, bowtie2
- Variantion: FreeBayes, GATK, MuTecT, SnpEff, VEP, GEMINI, Lumpy, Delly
- RNA-seq: Tophat, STAR, cufflinks, HTSeq
- Quality control: fastqc, bamtools, RNA-SeQC
- Manipulation: bedtools, bcftools, biobambam, sambamba, samblaster, samtools, vcflib

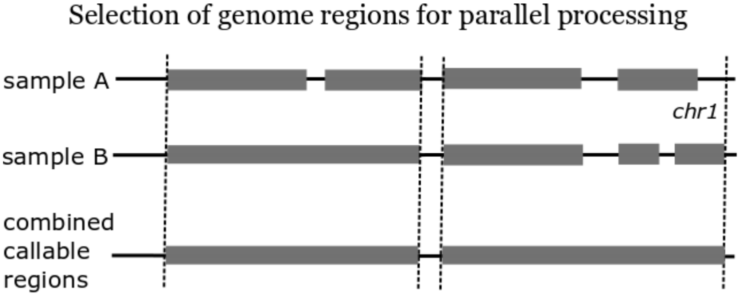
Provides

- Community – collected set of expertise
- Tool integration
- Validation – outputs + automated evaluation
- Installation of tools and data
- Scaling

Scaling: avoid intermediates

```
("{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")
```

Parallelization: split and batched jobs



Storage: extract outputs

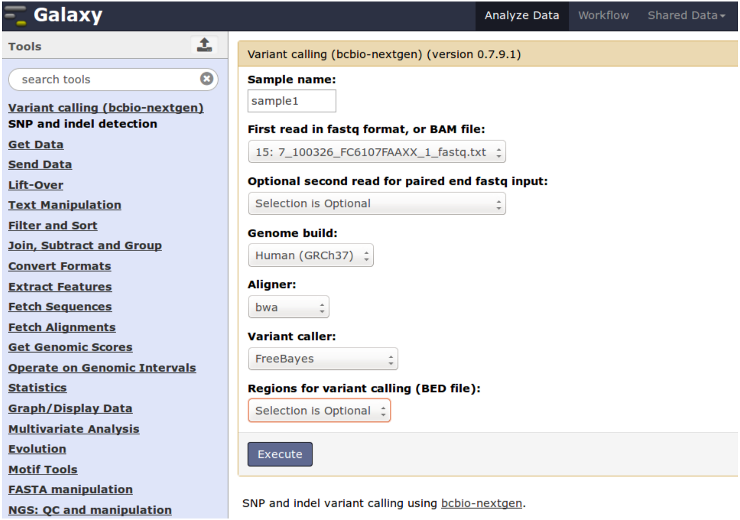
- Intermediates – 6x final

```
$ du -sh *
353G final
2.2T work
```

- 1500 whole genome scale – 110Tb

```
$ du -sh alz-p3f_2-g5/final
3.4T  alz-p3f_2-g5/final
$ ls -lhd *alz* | wc -l
31
```

bcbio as a Galaxy tool



Variant calling (bcbio-nextgen)
SNP and indel detection
Get Data
Send Data
Lift-Over
Text Manipulation
Filter and Sort
Join, Subtract and Group
Convert Formats
Extract Features
Fetch Sequences
Fetch Alignments
Get Genomic Scores
Operate on Genomic Intervals
Statistics
Graph/Display Data
Multivariate Analysis
Evolution
Motif Tools
FASTA manipulation
NGS: QC and manipulation

Variant calling (bcbio-nextgen) (version 0.7.9.1)

Sample name: sample1

First read in fastq format, or BAM file: 15: 7_100326_FC6107FAAXX_1_fastq.txt

Optional second read for paired end fastq input: Selection is Optional

Genome build: Human (GRCh37)

Aligner: bwa

Variant caller: FreeBayes

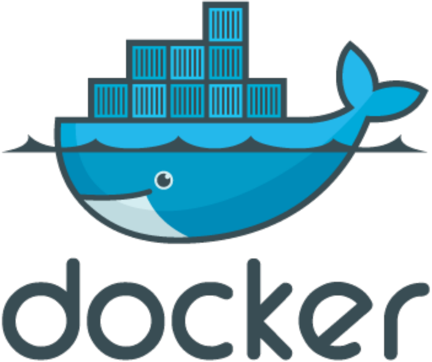
Regions for variant calling (BED file): Selection is Optional

Execute

SNP and indel variant calling using bcbio-nextgen.

https://github.com/chapmanb/bcbio-nextgen/tree/master/config/galaxy

Tool shed install: Docker



https://github.com/chapmanb/cloudbiolinux

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

Summary

- Focus: Community, Validation, Scaling
- bcbio-nextgen

  `https://github.com/chapmanb/bcbio-nextgen`
- Challenges: parallelization, scaling and storage
- Galaxy integration: Simple tool with Docker installation