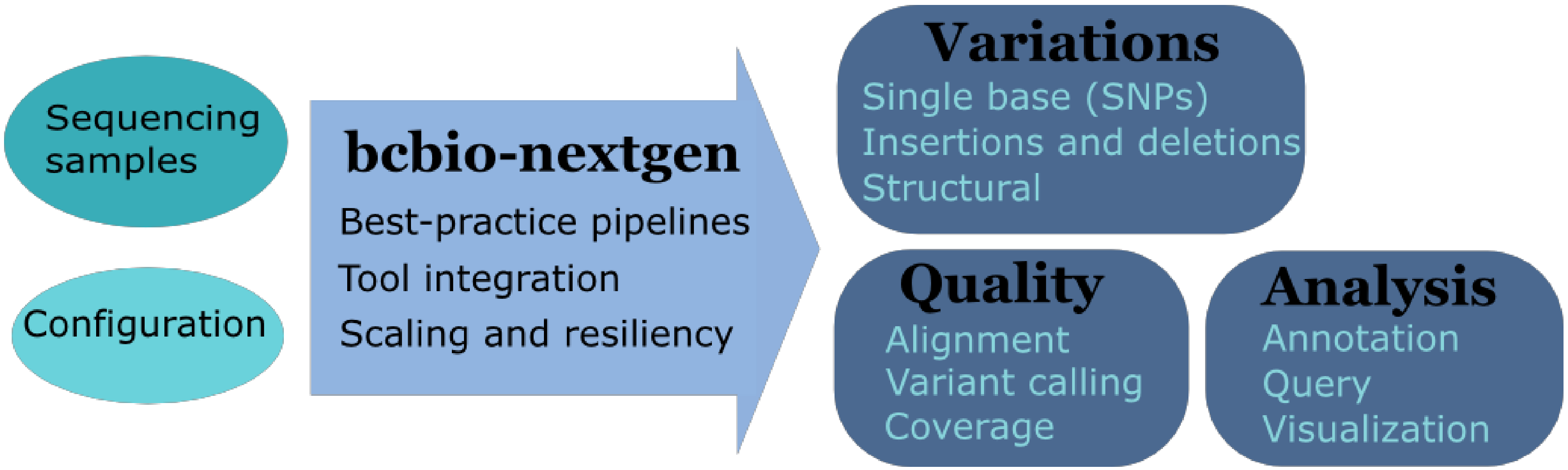
Quantifiable and scalable detection of genomic variants

Brad Chapman

Bioinformatics Core, Harvard School of Public Health

@chapmanb

26 June 2013



Sequencing samples

Configuration

bcbio-nextgen

Best-practice pipelines

Tool integration

Scaling and resiliency

Variations

Single base (SNPs)

Insertions and deletions

Structural

Quality

Alignment

Variant calling

Coverage

Analysis

Annotation

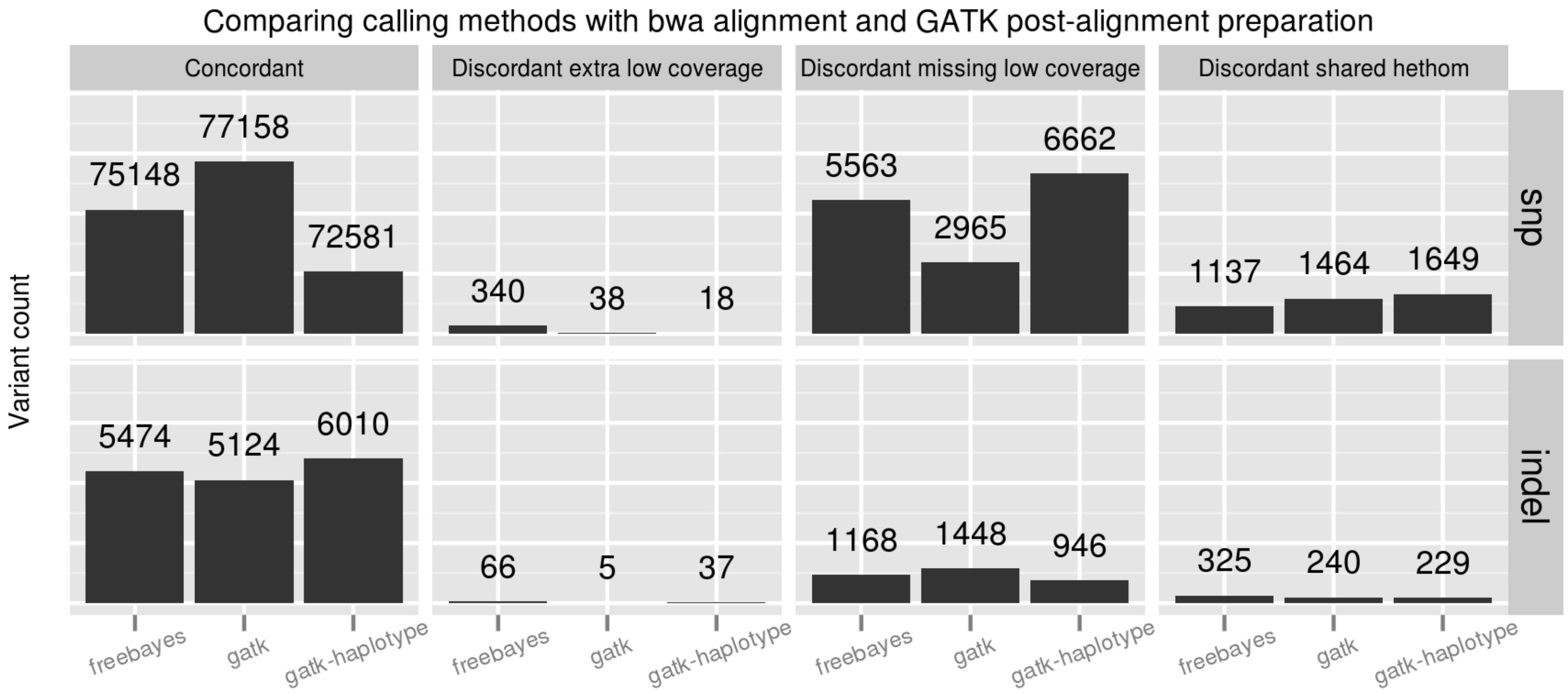
Query

Visualization

Development goals

- Quantifiable: assess variant quality
- Scalable: 1500 whole genome samples
- Reproducible: text-configurable, provenance, version tracking
- Community developed: open source, documented and widely deployable

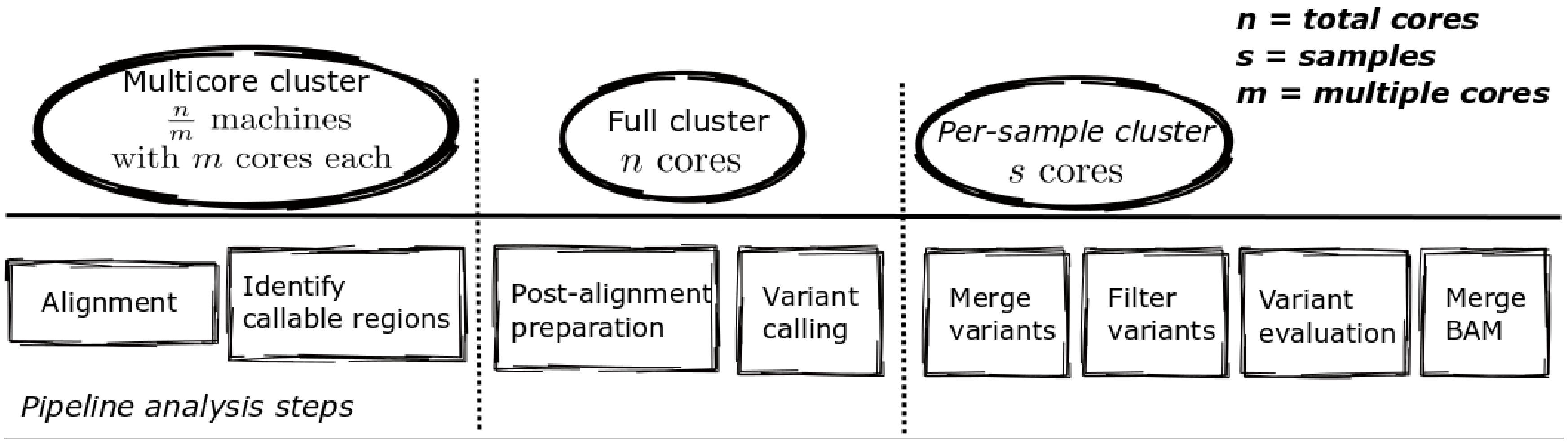
Quantify quality



Reference materials:
http://www.genomeinabottle.org/

Parallel scaling

Heterogeneous cluster creation during variant calling



Infrastructure:
http://ipython.org/ipython-doc/dev/parallel/index.html

Reproducible configuration

```
- files: [NA12878-NGv3-LAB1360-A_1.fastq.gz, NA12878-NGv3-LAB1
360-A_2.fastq.gz]
  description: NA12878
  analysis: variant2
  genome_build: GRCh37
  algorithm:
    aligner: bwa
    recalibrate: gatk
    realign: gatk
    variantcaller: [gatk, freebayes, gatk-haplotype]
    coverage_interval: exome
    coverage_depth: high
    platform: illumina
    quality_format: Standard
    validate: NA12878-nist-v2_13-NGv3-pass.vcf
```

Community developed

- Fully automated installation: CloudBioLinux
- Deployable on multiple clusters (LSF, SGE, Torque)
- Integrated with web platforms (Galaxy, STORMSeq)
- Open source and documented

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