

INTEGRATION
AND MODELING
for PREDICTIVE
BIOLOGY



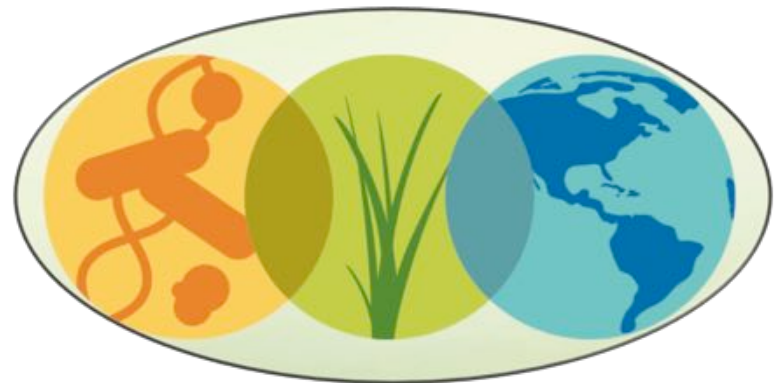
Variation & RNA-seq Services

Michael Schatz

Cold Spring Harbor Laboratory

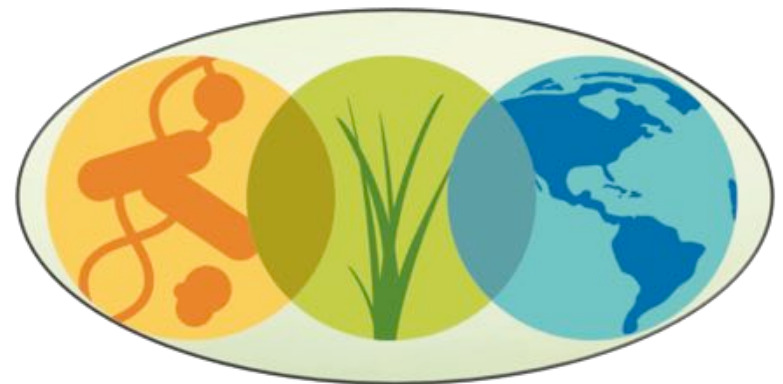
Agenda

1. Getting Started
2. Variation services
3. RNA-seq services



Agenda

1. Getting Started
2. Variation services
3. RNA-seq services



Samples to discoveries



Logging In

- IRIS: command-line access to KBase functionality inside web browser
- Download and install tools on Mac or Linux
- KBase Labs offer early access to web-based interfaces
- Narrative user interface

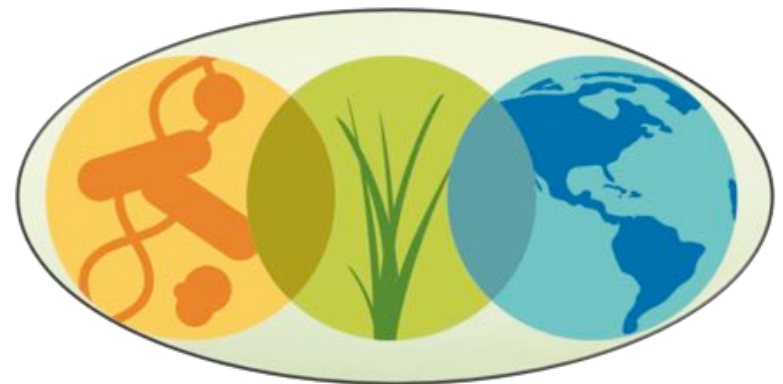
The screenshot shows the KBase website homepage. At the top, there is a navigation bar with links for 'About', 'News', 'User Zone', 'Developer Zone', 'KBase Labs', and 'Contact Us'. A search bar is located in the top right corner. The main content area features a large heading: 'The Department of Energy Systems Biology Knowledgebase (KBase) is an emerging software and data environment designed to enable researchers to collaboratively generate, test and share new hypotheses about gene and protein functions, perform large-scale analyses on a scalable computing infrastructure, and model interactions in microbes, plants, and their communities. KBase provides an open, extensible framework for secure sharing of data, tools, and scientific conclusions in predictive and systems biology.' Below this, there are three circular call-to-action buttons: 'Try KBase Now' (with a blue flower icon), 'Download the Tools' (with a red toolbox icon), and 'Visit KBase Labs' (with a gear icon). To the right, there are sections for 'Latest News' and 'Upcoming Events'. At the bottom, there is a 'Glimpse the future' button and a footer with 'Sign up for a KBase account', 'KBase is sponsored by the U.S. Department of Energy's Office of Biological and Environmental Research', and 'Acknowledgements Privacy and Security'.

<http://kbase.us/>

<https://gologin.kbase.us/SignUp>

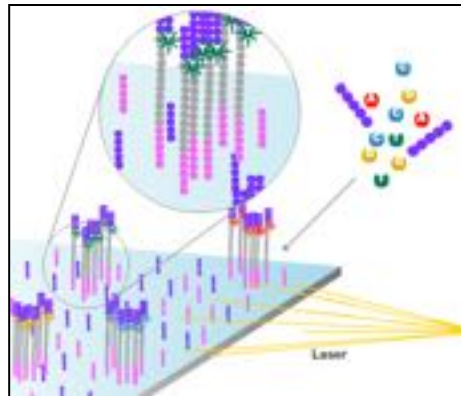
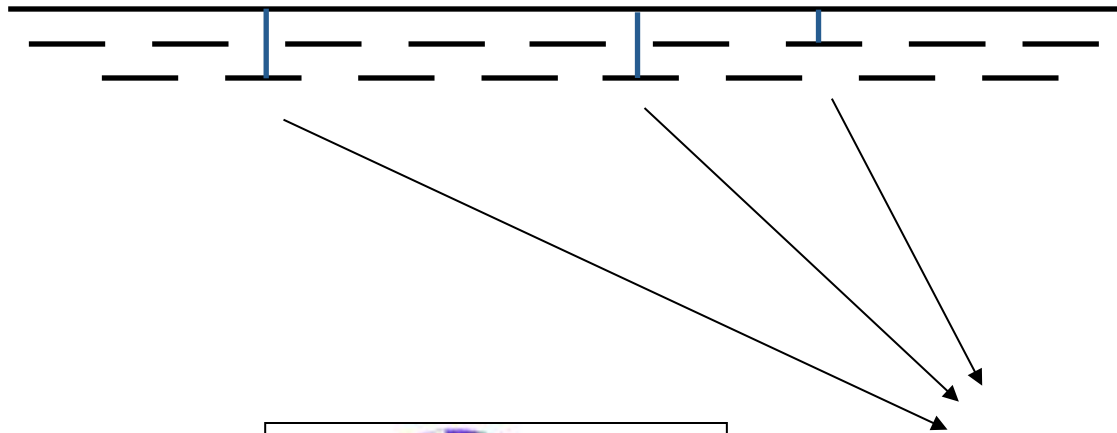
Agenda

1. Getting Started
2. Variation services
3. RNA-seq services



Resequencing & Variations

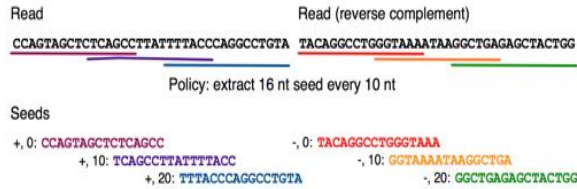
How does your sample compare to the reference?



- Plant Height
- Drought Resistance
- Biomass production

Algorithms for Mapping & Genotyping

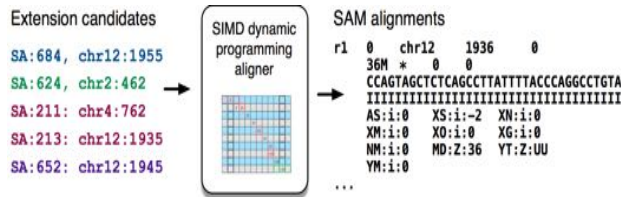
1. Split read into segments



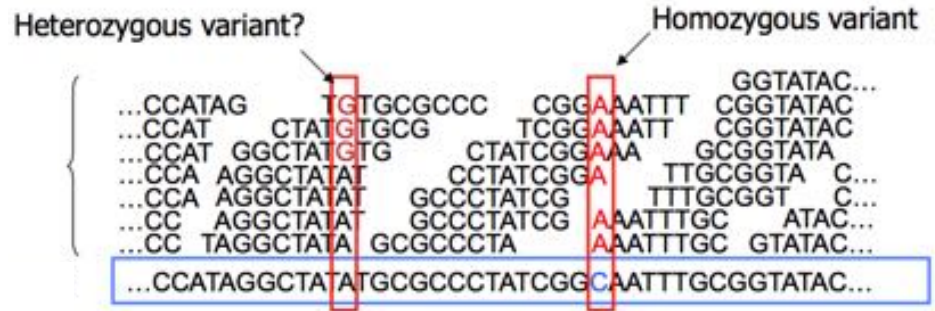
2. Lookup each segment and prioritize



3. Evaluate end-to-end match



Fast gapped-read alignment with Bowtie 2
Langmead & Salzberg. (2012) Nature Methods. 9:357-359.



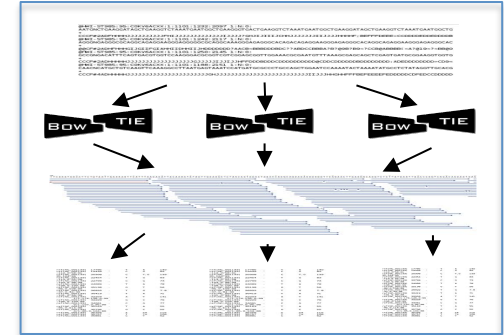
- Distinguishing SNPs from sequencing error typically a likelihood test of the coverage
 - Hardest to distinguish between errors and heterozygous SNP.
 - Coverage is the most important factor!
 - Target at least 10x, 30x more reliable

The Sequence Alignment/Map format and SAMtools
Li H et al. (2009) Bioinformatics. 25:16 2078-9

Variation Services API 1.0

Genotyping API

- **Bowtie:** Launch alignment task with Bowtie
- **BWA:** Launch alignment task with BWA
- **SNPCalling:** Launch SNPcalling task with SAMTools
- **SortAlignments:** Launch task to sort by chromosome



Data API

- **List:** List files in a directory
- **Fetch:** Fetch files from HDFS
- **Put:** Put files into HDFS
- **RM:** Delete files on HDFS
- **FetchBAM:** On-the-fly conversion to BAM
- **PutFastq:** Put reads into HDFS with conversion

Job API

- **ClusterStatus:** return basic status of cluster (jobs running, nodes available, etc)
- **JobStatus:** Given a JobID, returns current status
- **ListJobs:** List JobID running with a given username
- **KillJob:** Kills a given JobID

Notes:

- All calls are authenticated with KBase username/password

Reads to SNPs in 5 easy steps

1. Identify reference genome

```
$ all_entities_Genome -f scientific_name | grep -i 'Populus'
```

2. Upload Reads to KBase cloud

```
$ jk_fs_put_pe populus.1.fq.gz populus.2.fq.gz populus
```

3. Align Reads with Bowtie2

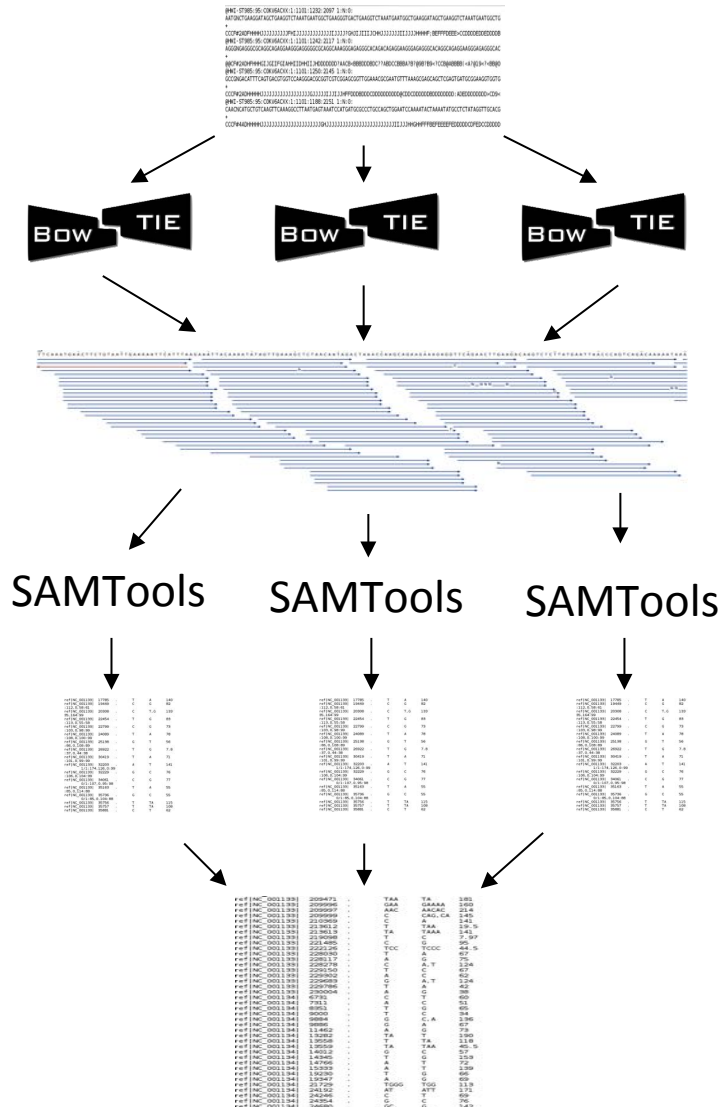
```
$ jk_compute_bowtie -in=populus.pe \
  -org=populus -out=populus_align
```

4. Call SNPs with SAMTools

```
$ jk_compute_samtools_snp -in=populus_align \
  -org=populus -out=populus_snps
```

5. Merge and Download VCF files

```
$ jk_compute_vcf_merge -in=populus_snps \
  --alignments=populus_align -out=populus.vcf
$ jk_fs_get populus.vcf
```



Maize population analysis

Align & call SNPs from 131 maize samples

1TB fastq / 408Gbp input data

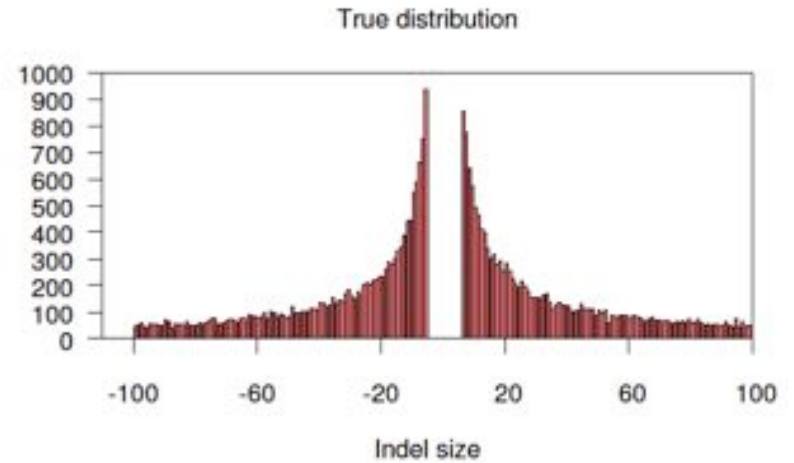
	Serial	KBase cloud (small)	KBase Cloud (large)
Config	1 core (1 node)	210 cores (15 nodes)	854 cores (61 nodes)
Bowtie2	1311 hr*	19.5 hr	5 hr
Sort	58 hr*	N/A	N/A
Samtools	58 hr*	3.5 hr	1.5 hr
End-to-End	1427 hr*	23 hr	6.5 hr
Speedup	1x	62x	219x

*estimated time

Variation Services 2.0 Sneak peak

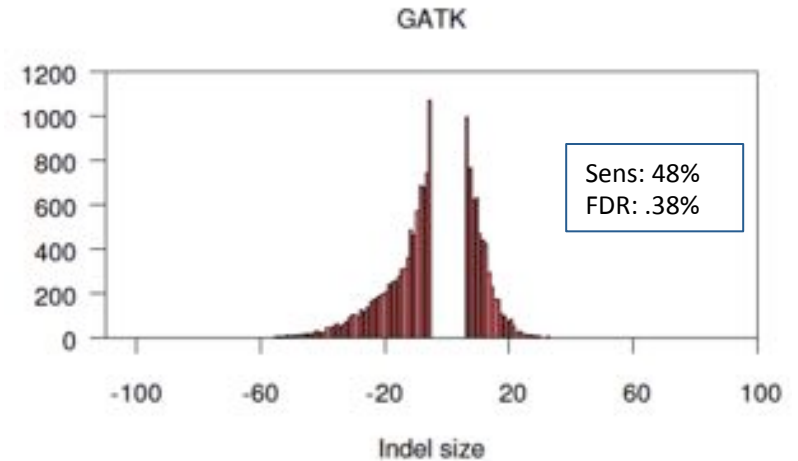
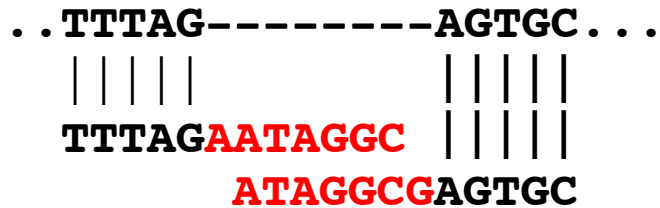
SNPs + Short Indels

High precision and sensitivity



“Long” Indels (>5bp)

Reduced precision and sensitivity



Analysis confounded by sequencing errors, localized repeats, allele biases, and mismapped reads

Scalpel: Haplotype microassembly

DNA sequence **micro-assembly** pipeline for accurate detection and validation of *transmitted* and *de novo* mutations

Features

1. Combine **mapping** and **assembly**
2. Exhaustive search of **haplotypes**
3. **De novo mutations**



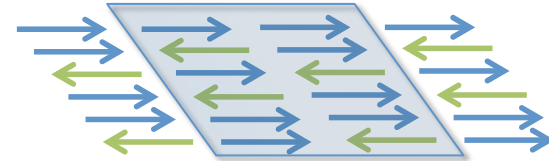
NRXN1 *de novo* SNP
(auSSC12501 chr2:50724605)

Accurate detection of de novo and transmitted INDELs within exome-capture data using micro-assembly

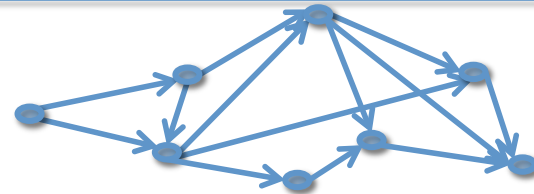
Narzisi, G et al (2014) *bioRxiv* doi: 10.1101/001370

Algorithm Overview

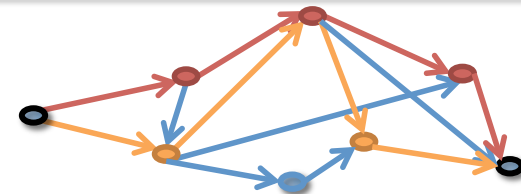
Extract reads mapping within the exon including (1) well-mapped reads, (2) soft-clipped reads, and (3) anchored pairs



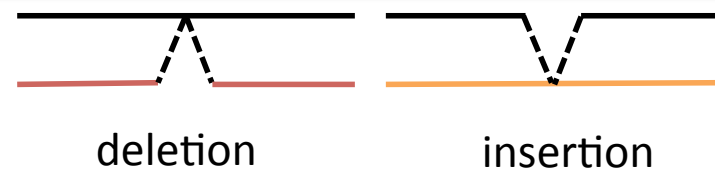
Decompose reads into overlapping k -mers and construct de Bruijn graph from the reads



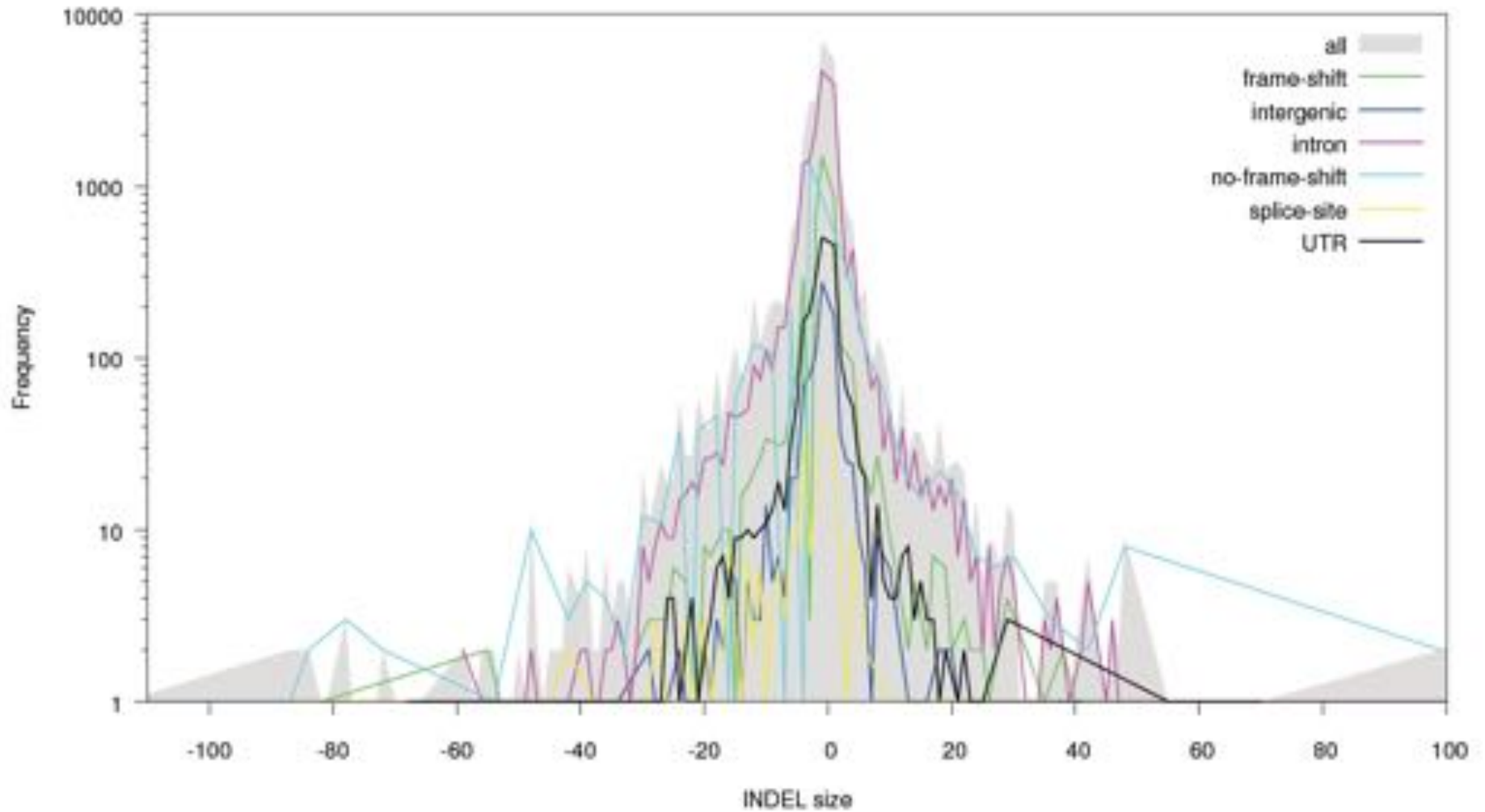
Find end-to-end haplotype paths spanning the region



Align assembled sequences to reference to detect mutations



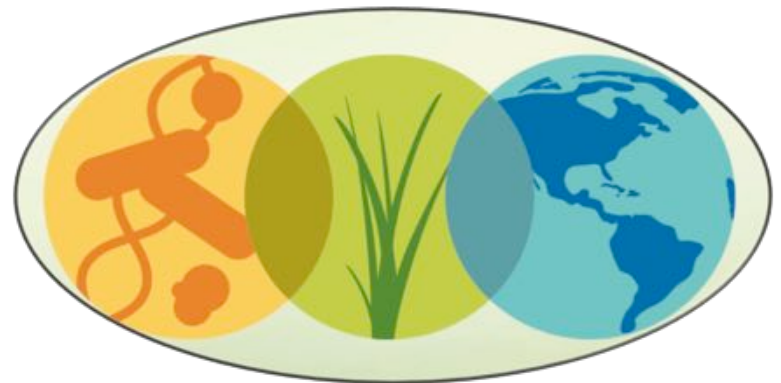
Population Analysis



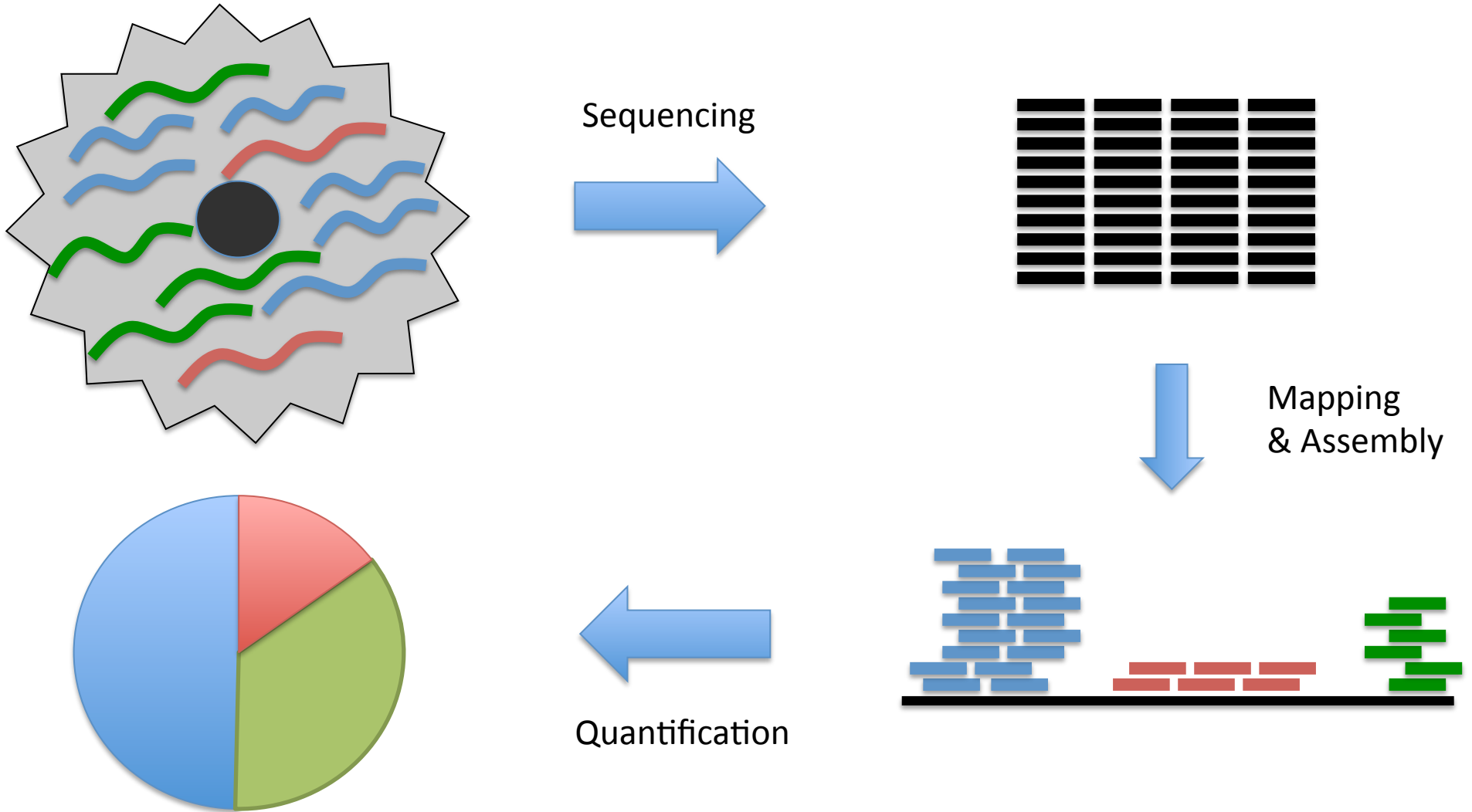
See Ranjan's Talk at 5:40

Agenda

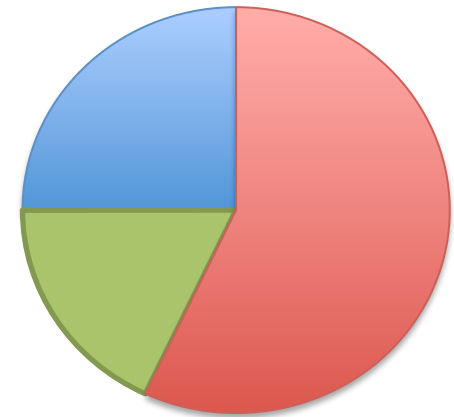
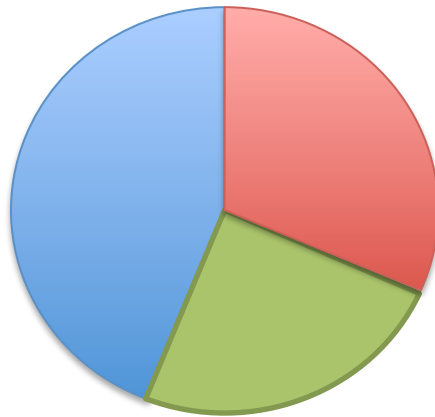
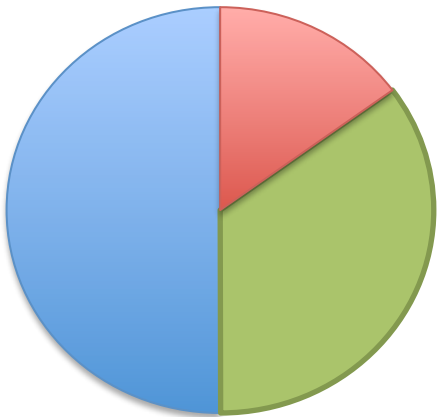
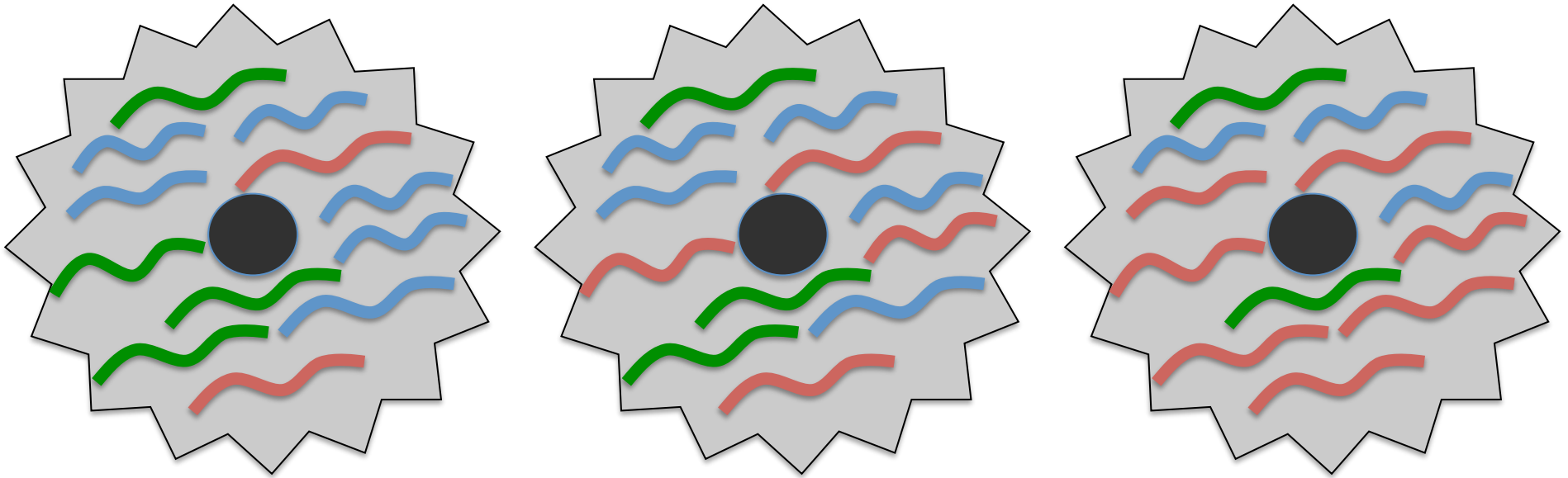
1. Getting Started
2. Variation services
3. RNA-seq services



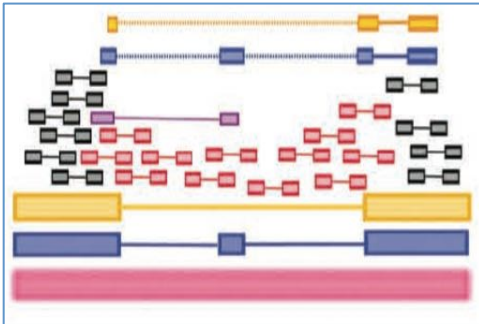
RNA-seq Overview



RNA-seq Overview



RNA-seq Challenges

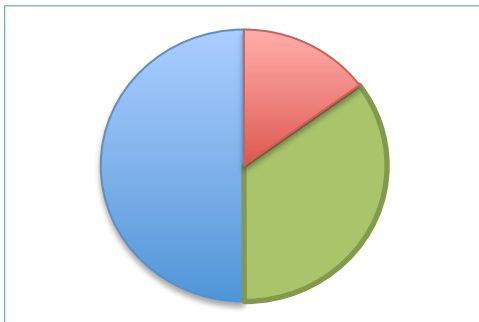


Challenge 1: Eukaryotic genes are spliced

Solution: Use a spliced aligner, and assemble isoforms

TopHat: discovering spliced junctions with RNA-Seq.

Trapnell et al (2009) *Bioinformatics*. 25:0 1105-1111

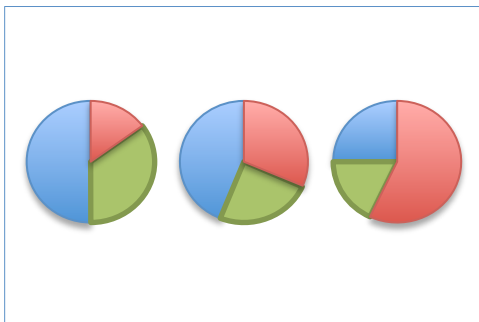


Challenge 2: Read Count != Transcript abundance

Solution: Infer underlying abundances (e.g. FPKM)

Transcript assembly and quantification by RNA-seq

Trapnell et al (2010) *Nat. Biotech.* 25(5): 511-515



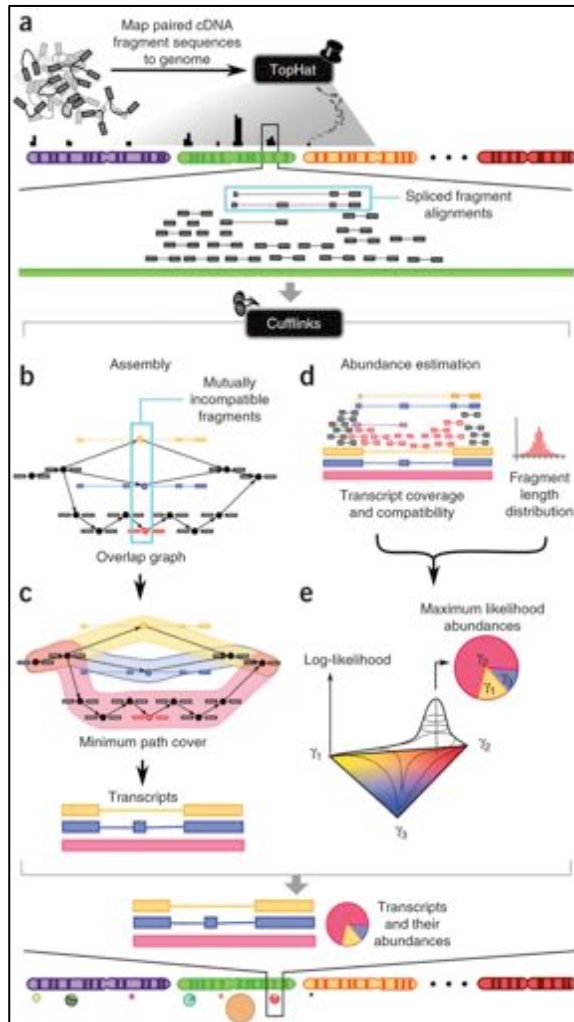
Challenge 3: Transcript abundances are stochastic

Solution: Replicates, replicates, and more replicates

RNA-seq differential expression studies: more sequence or more replication?

Liu et al (2013) *Bioinformatics*. doi:10.1093/bioinformatics/btt688

Identifying Differentially Expressed Genes



1. Spliced alignment with TopHat

```
$ jk-compute-tophat -in=t1.1.fq.gz,t1.2.fq.gz -ref=ecoli -out=t1-tophat -align_opts=-p8
```

2. Assemble and quantify expression with Cufflinks

```
$ jk-compute-cufflinks -in=t1-tophat/accepted_hits.bam -out=t1-cufflinks \
  -assembly_opts=-p8
```

3. Merge samples

```
$ jk-compute-cuffmerge -in=t1-cufflinks/transcripts.gtf,t2-cufflinks/transcripts.gtf \
  -ref=ecoli -out=cuffmerge-out -assembly_opts=-p8
```

4a. Identify DE genes

```
$ jk-compute-cuffdiff -in=t1-tophat/accepted_hits.bam,t2-tophat/accepted_hits.bam \
  -out=cuffdiff-out -ref=ecoli -assembly_opts=-p8 -condn_labels=T1,T2 \
  -merged_gtf=cuffmerge-out/merged.gtf
```

4b. Discover novel genes & isoforms

```
$ jk-compute-cuffcompare -in=t1-cufflinks/transcripts.gtf,t2-cufflinks/transcripts.gtf \
  -out=cuffcompare-out -ref_gtf=cuffmerge-out/merged.gtf
```

RNA-seq 2.0: Long Read Analysis



- Long-read single-molecule sequencing has potential to directly sequence full length transcripts
 - Error corrected reads almost perfectly match the genome, pinpointing splice sites, identifying alternative splicing

Hybrid error correction and de novo assembly of single-molecule sequencing reads.
Koren, S, Schatz, MC, et al. (2012) *Nature Biotechnology*. doi:10.1038/nbt.2280

Additional Resources

Resource	URL
KBase	http://kbase.us/
Getting Started	http://kbase.us/for-users/user-home/
Variation Services	http://kbase.us/for-users/tutorials/analyzing-data/variation-service/
RNA-seq Services	http://kbase.us/for-users/tutorials/analyzing-data/plant-genome-analysis/plant-expression-service/
Bowtie2	http://bowtie-bio.sourceforge.net/bowtie2/index.shtml
BWA	http://bio-bwa.sourceforge.net/
SAMTools	http://samtools.sourceforge.net/
Cufflinks	http://cufflinks.cbc.umd.edu/
KBase Contact	http://kbase.us/contact-us/
Survey	https://www.surveymonkey.com/s/KB-user-info

Thank you!

<http://schatzlab.cshl.edu>
@mike_schatz / @DOEKBase

