Scalable Software for Analyzing Large Collections of RNA Sequencing Data

Ben Langmead
Assistant Professor, JHU Computer Science
langmea@cs.jhu.edu, langmead-lab.org, @BenLangmead

IDIES Symposium
Johns Hopkins University, October 17, 2014
Improving sequencers

Source: www.illumina.com

2. **This Bizarre Organism Builds Itself a New Genome Every Time It Has Sex**, by Greg Miller. Wired, September 17, 2014.


## Big projects

<table>
<thead>
<tr>
<th>Study</th>
<th>Approx # samples</th>
</tr>
</thead>
<tbody>
<tr>
<td>ENCODE</td>
<td>100</td>
</tr>
<tr>
<td>GEUVADIS</td>
<td>465</td>
</tr>
<tr>
<td>Depression Genes Network</td>
<td>950</td>
</tr>
<tr>
<td>TCGA</td>
<td>&gt;2,000</td>
</tr>
<tr>
<td>GTEx</td>
<td>&gt;10,000</td>
</tr>
</tbody>
</table>
Why study big public datasets?

- To make discoveries missed by original authors
- To combine datasets in new ways
- To add power to a smaller experiment
- As proving ground for new methods
MapReduce: aggregate, compute, repeat

- **Aggregate**
  - **Group** and **order** data
- **Compute**
  - **Run** a simple program concurrently over all the ordered groups
Cloud-scale RNA-sequencing differential expression analysis with Myrna.
Myrna design

RNA reads, group A

Align

Group by region, order along ref

Overlap & count

Group by sample, order by count

Normalize

Group by gene

DE Test

Order by p value

Results

## Myrna results

Amazon Elastic MapReduce, c1.xlarge instances

<table>
<thead>
<tr>
<th></th>
<th># Samples</th>
<th>Input size (GB, gzipped)</th>
<th># CPUS</th>
<th>Wall clock time</th>
<th>Analysis cost</th>
<th>Cost per input GB</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pickrell et al</td>
<td>69</td>
<td>42</td>
<td>320</td>
<td>1h:38m</td>
<td>$65.60</td>
<td>$1.56</td>
</tr>
<tr>
<td>GEUVADIS</td>
<td>465</td>
<td>1,068*</td>
<td>600</td>
<td>19h:08m</td>
<td>$364.50</td>
<td>$0.34</td>
</tr>
</tbody>
</table>

Pickrell et al data transfer: ~$12, ~1hr

GEUVADIS data transfer: ~$34, ~14hr

* analyzing mate 1


ReCount: digested RNA-seq using Myrna

- Normalized gene-count tables encompassing 18 different published studies, 475 samples, >8 billion RNA-seq reads
Rail-RNA

- Bring scalability closer to frontier of RNA-seq analysis
- Benefit maximally from analyzing many samples at once

Abhinav Nellore  Jacob Pritt  Jeff Leek
On 40 c3.2xlarge EC2 machines

Dashed lines: linear extrapolations

Costs per sample
N=20: $4.20
N=40: $3.15
N=80: $2.63

Number $N$ of c3.2xlarge nodes
Recall of instances where intron is overlapped by read*; no annotation provided

*sim of 40 mil 76-bp reads w/ coverage distribution derived from YRI GEUVADIS sample
Thank you

Abhinav Nellore  
Rail

Jacob Pritt  
Rail

Kasper Hansen  
Myrna

Jeff Leek  
Rail
Myrna

Contact:
Email: langmea@cs.jhu.edu
Web: www.langmead-lab.org
Twitter: @BenLangmead