‘To sequence or not to sequence’ is not a question anymore. BUT...

Vladimír Beneš
21 June 2011

http://www.genecore.embl.de
More data on their way to you!

v3 flowcell
imaged area larger by 50 %!
v3 sequencing chemistry

Old Cluster Amplification
Some GC-rich clusters poorly resolved/not detected at very high densities

New Cluster Amplification
Larger, brighter GC-rich clusters are well resolved and detected at very high densities
Increasing yield

Sequence output from Illumina reads per lane

- **Millions of reads PF**

- **Illumina Arrives**
  - 0.3
  - 1.0
  - 1.3.2
  - 1.3.4
  - 1.4

- **GAIIx**
  - 1.6

- **SCS2.6, v4 kits**

- **SCS 2.8, cBot, v5 kits**

- **GA**

- **HiSeq2000**
Improving quality of called bases

Average quality per cycle, RNA-Seq

- 36 Base mRNAsSeq
- 105 Base V4 mRNAsSeq
- s_6
- TruSeq 150 R1
- 76 Base mRNAsSeq
- 76 Base V5 mRNAsSeq
- s7 V5
- TruSeq 150 R2
- 76 Base V4 mRNAsSeq
- v5105
- s8 v5
Challenges

• The only thing constant in life is change…
• Distorted expectations of users
• Data (‘massive’ amounts, formats…)
• Interpretation of results (suboptimal experimental design; is everything relevant?)
• Incomplete understanding of sources of error and bias in MPS data
Bias is never good...
Comparison of Sequence Reads Obtained from Three Next-Generation Sequencing Platforms

Shingo Suzuki¹, Naoaki Ono¹, Chikara Furusawa¹, Bei-Wen Ying¹, Tetsuya Yomo¹,²,³*

¹ Department of Bioinformatics Engineering, Graduate School of Information Science and Technology, Suita, Osaka, Japan, ² Exploratory Research for Advanced Technology (ERATO), Japan Science and Technology Agency, Suita, Osaka, Japan, ³ Graduate School of Frontier Biosciences, Osaka University, Suita, Osaka, Japan

RNA-seq: technical variability and sampling


Lauren M McIntyre (mcintyre@ufl.edu)
Kenneth K Lopiano (Klopiano@ufl.edu)

Sequencing

METHOD

Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries

Geng Tian¹,²

Kensuke Nakamura¹
Hirofumi Yoshikawa¹
Hiroki Takahashi¹

Daniel Aird¹, Michael G Ross¹, Wei-Sheng Chen², Maxwell Danielsson², Timothy Fennell³, Carsten Russ¹, David B Jaffe¹, Chad Nusbaum¹, Andreas Gnirke¹
Hype/hope curve

- Peak of Inflated Expectations
- Plateau of Productivity
- Slope of Enlightenment
- Trough of Disillusionment
- Technology Trigger

TIME

VISIBILITY
Massively parallel sequencing for monitoring genetic consistency and quality control of live viral vaccines

Alexander Neverov and Konstantin Chumakov

Center for Biologics Evaluation and Research, Food and Drug Administration, Rockville, MD 20852

PNAS (2010)

Kahvejian, Nat Biotech (2008)
### Available MPS applications

<table>
<thead>
<tr>
<th>Application</th>
<th>Supported</th>
</tr>
</thead>
<tbody>
<tr>
<td>transcriptome</td>
<td>yes</td>
</tr>
<tr>
<td>RNA-Seq, Tag-Seq</td>
<td></td>
</tr>
<tr>
<td>miRnome</td>
<td>yes</td>
</tr>
<tr>
<td>smallRNA-Seq</td>
<td></td>
</tr>
<tr>
<td>protein-NA interactions</td>
<td>yes</td>
</tr>
<tr>
<td>ChIP-Seq, CLIP-Seq</td>
<td></td>
</tr>
<tr>
<td>epigenome</td>
<td>yes</td>
</tr>
<tr>
<td>Methyl-Seq</td>
<td></td>
</tr>
<tr>
<td>de novo &amp; re-sequencing</td>
<td>yes</td>
</tr>
<tr>
<td>Metagenomics</td>
<td>yes</td>
</tr>
<tr>
<td>Genome capture, multiplexing</td>
<td>yes</td>
</tr>
</tbody>
</table>
## MPS methods used in epigenomics

<table>
<thead>
<tr>
<th>Epigenetic modification</th>
<th>Method</th>
</tr>
</thead>
<tbody>
<tr>
<td>DNA methylation</td>
<td>MethylC-seq</td>
</tr>
<tr>
<td></td>
<td>BS-seq</td>
</tr>
<tr>
<td></td>
<td>MeDIP-seq</td>
</tr>
<tr>
<td></td>
<td>MRE-seq</td>
</tr>
<tr>
<td></td>
<td>MethylCap-seq</td>
</tr>
<tr>
<td></td>
<td>RRBS</td>
</tr>
<tr>
<td>Histone post-translational modifications</td>
<td>ChIP-seq</td>
</tr>
<tr>
<td>Histone variants</td>
<td>ChIP-seq</td>
</tr>
<tr>
<td>Chromatin modifiers and remodelers</td>
<td>ChIP-seq</td>
</tr>
<tr>
<td>Chromatin accessibility</td>
<td>DNaseI-seq</td>
</tr>
<tr>
<td></td>
<td>FAIRE-seq</td>
</tr>
<tr>
<td></td>
<td>Sono-seq</td>
</tr>
<tr>
<td>Nucleosome positioning and turnover</td>
<td>MNase-seq</td>
</tr>
<tr>
<td></td>
<td>CATCH-IT</td>
</tr>
<tr>
<td>Long-range chromatin interactions</td>
<td>Hi-C</td>
</tr>
<tr>
<td></td>
<td>ChIA-PET</td>
</tr>
<tr>
<td>Allele-specific chromatin signatures</td>
<td>haploChIP</td>
</tr>
</tbody>
</table>

Importance of experimental design

What do I want to study?

"Would you tell me, please, which way I ought to go from here?"
"That depends a good deal on where you want to get to," said the Cat.
"I don’t much care where--" said Alice.
"Then it doesn’t matter which way you go," said the Cat.
"--so long as I get SOMEWHERE," Alice added as an explanation.
"Oh, you’re sure to do that," said the Cat, "if you only walk long enough."

Lewis Carroll’s Alice in Wonderland
### Which sequencing mode to use?

<table>
<thead>
<tr>
<th>Sequencing type</th>
<th>Recommendation</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Exon capture</strong></td>
<td>50Mb Kit, human: 105b SR – to get sufficient coverage</td>
</tr>
<tr>
<td><strong>Whole genome sequencing</strong></td>
<td><em>Large rearrangements</em>: Mate-pairs large insert</td>
</tr>
<tr>
<td></td>
<td><em>Resequencing</em>: SNPs/indels: Coverage is good 100+ PE.</td>
</tr>
<tr>
<td></td>
<td>If you don’t get the coverage at the start you’ll regret it 😞.</td>
</tr>
<tr>
<td><strong>RNA-Seq</strong></td>
<td><strong>Coverage is the key!</strong></td>
</tr>
<tr>
<td></td>
<td><em>Tag counting</em>: large number of mappable tags 36-50b SR should suffice.</td>
</tr>
<tr>
<td></td>
<td><em>Transcriptome assembly or exon usage</em>: 75+ single or 75+ PE</td>
</tr>
<tr>
<td></td>
<td>depending on a de novo/spliced read mapping approach or map pairs to detect also alternative splicing.</td>
</tr>
<tr>
<td></td>
<td><em>Strand-specific libraries</em>: complex insight into transcriptome</td>
</tr>
<tr>
<td><strong>Chip-Seq</strong></td>
<td>36b SR unless you have real concerns about ‘alignability’ of your target (i.e. some strange looking enhancer region)</td>
</tr>
<tr>
<td><strong>Multiplexing</strong></td>
<td>Coverage is the key!</td>
</tr>
</tbody>
</table>
MPS workflow

1) Library preparation

2) Cluster generation on a flow cell

SE, PE reads, 36-150 bases

3) Sequencing & imaging

4) Data processing & analysis
MPS library preparation

5′AATGATACGGCGACCACCGA-ACACTCTTTCCCTACACGACGCTCTTTCCGATCT--INSERT--TCGTATGCCGTCTTCTGCTTGTTACTATGCGCTGGTGGCT-TGTGAGAAGGGATGTGCTGAGAAGGCTAGA--INSERT--AGCATACGGCGAGACGACGA5′

where

- 5′AATGATACGGCGACCACCGA is the P5 attachment/amplification primer sequence
- 5′CAAGCAGAAGACGGCATACGA is the P7 attachment/amplification primer sequence
- 5′ACACTCTTTCCCTACACGACGCTCTTTCCGATCT is the SBS3 sequencing primer sequence
- INSERT is a complex mix of DNA fragments
Forked adapters

Adapter

ACACTTTTCCCTACACGAC

GCTTTCCGAT

CGAAGGCTA

GTTCGTCTTCTGCCGTATGCT

G TC P-

GATCGGAAGAGC

CTAGCCTTCTCG

TCGTATGCGTGTTTTCTGCTTG

P-GATCGGAAGAGC

T CAGCCTTCTCG

CAGCACATCCCTTTCACACA

Adapter
Library preparation II.

(a) A conventional protocol

(b) A “nextera” protocol

at least 500 ng of gDNA

<50 ng of gDNA!

Adey et al., Genome Biology (2010)
Library preparation III.

- Strict QC of starting material (GiGo)
  - Qubit quantification
  - gel images, bioanalyzer/experion traces
Library preparation IV.

- Bioruptor, probe (ChIP-Seq)
- Covaris vs nebulization
- Kits (proprietary, home-brewed, NEB!)
- Size selection using gel extractor, E-gel, Pippin prep, SPRIworks…,
- Lo-bind tubes!
RNA-Seq libraries

- rRNA depletion (oligo-dT beads, Ribo-Minus, Ribo-Zero…) BUT mitochondria-derived rRNA mostly ignored!!

- strand-specific library, Levine et al., Nat Meth (2010)
The rocks and shallows of deep RNA sequencing: Examples in the *Vibrio cholerae* RNome

CARSTEN A. RAABE,1 CHEE HOCK HOE,2 GERRIT RANDAU,1 JUERGEN BROSIOUS,1,3 THEAN HOCK TANG,2 and TIMOFEY S. ROZHDESTVENSKY1,3

1Institute of Experimental Pathology, University of Muenster, 48149 Muenster, Germany
2Infectious Diseases Cluster, Advanced Medical and Dental Institute (AMDI), Universiti Sains Malaysia, 13200 Penang, Malaysia

ABSTRACT

New deep RNA sequencing methodologies in transcriptome analyses identified a wealth of novel nonprotein-coding RNAs (npcRNAs). Recently, deep sequencing was used to delineate the small npcRNA transcriptome of the human pathogen *Vibrio cholerae* and 627 novel npcRNA candidates were identified. Here, we report the detection of 223 npcRNA candidates in *V. cholerae* by different cDNA library construction and conventional sequencing methods. Remarkably, only 39 of the candidates were common to both surveys. We therefore examined possible biasing influences in the transcriptome analyses. Key steps, including tailing and adapter ligations for generating cDNA, contribute qualitatively and quantitatively to the discrepancies between data sets. In addition, the state of 5'-end phosphorylation influences the efficiency of adapter ligation and C-tailing at the 3'-end of the RNA. Finally, our data indicate that the inclusion of sample-specific molecular identifier sequences during ligation steps also leads to biases in cDNA representation. In summary, even deep sequencing is unlikely to identify all RNA species, and caution should be used for meta-analyses among alternatively generated data sets.
<table>
<thead>
<tr>
<th>No.</th>
<th>Step</th>
<th>Substeps leading to possible bias</th>
</tr>
</thead>
</table>
| 1   | Preparation and (counter) selection of RNA starting material         | Selection of subcellular organelles or fractions  
Selection as polysomal mRNAs or other ribonucleoprotein complexes (RNPs)  
Method of RNA preparation (e.g., loss of small RNAs after LiCl precipitation)  
Size selection on columns or gels  
Selection of polyadenylated or capped RNAs  
Counter-selection of undesired rRNAs, tRNAs, etc. by affinity methods |
| 2   | Removal or addition of RNA modifications                             | Mostly at the termini, such as decapping, dephosphorylation, or phosphorylation                                                                                 |
| 3   | Extension of RNA 3’-ends                                             | Tailing with oligo(A) or oligo(C) using poly(A) polymerase  
Ligation of oligonucleotide adapters  
Ligation of oligonucleotide adapters |
| 4   | Extension of RNA 5’-ends                                             |                                                                                                                                                               |
| 5   | Reverse transcription                                               | RNA modifications and secondary structures can lead to premature stop of extension                                                                           |
| 6   | Adapter ligation to 3’-ends of (first) single-stranded cDNA          |                                                                                                                                                               |
|     | (in applicable protocol variants)                                   |                                                                                                                                                               |
| 7   | PCR amplification                                                    | Can lead to bias in amplicon representation due to template size, base composition, repeat content, hairpin structures, etc.                                  |
| 8   | Cloning efficiency                                                  | In protocols where cDNA is cloned in, e.g., plasmid vectors prior to sequencing  
Adapter restriction sites for cloning might also be present on cDNA |
| 9   | Computational analyses                                              | Different filters and stringency                                                                          |

Shading indicates steps examined in this study.
Library quantification & QC

- Qubit
- Bioanalyzer
  - HS DNA Chip
  - DNA 1000 Chip

![Image of Qubit instrument]

**Overall Results for sample 8**:

- Number of peaks found: 2

**Peak table for sample 8**:

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>15</td>
<td>4.20</td>
<td>424.2</td>
<td>Lower Marker</td>
</tr>
<tr>
<td>2</td>
<td>81</td>
<td>0.93</td>
<td>17.4</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>358</td>
<td>4.57</td>
<td>19.3</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>1,500</td>
<td>2.10</td>
<td>2.1</td>
<td>Upper Marker</td>
</tr>
</tbody>
</table>
Methyl-Seq

Pacific Biosciences

Detection volume

Modified bases

By courtesy of Pacific Bioscience
Direct detection of methylated bases

Data integration

ChIP-seq  | RNA-seq quantification  | RNA-seq discovery
---|---|---
Integrate: RNA-seq, ChIP-seq and external data
Analyze:
- Associated genes
- Differential expression
- Novel splice isoforms
- Motif finding
- Expression levels
- Novel gene models
Aggregate and identify:
- Binding sources
- Novel transfrag
- Enriched regions
- Density on known exons
Maps read:
- Contiguous reads
- Splice-crossing reads

De novo transcript assembly

Pepke et al., Nature Methods (2009)
Where are we heading?
Nucleic acids detection and sequencing techniques

Kahvejian et al., Nat Biotech (2008)
We are drowning in information and starving for knowledge.
Rutherford D. Roger
MPS features

- Unprecedented discovery power
- Hypothesis-free
- Almost unbiased results
- Sensitivity & specificity
- For tag-counting applications truly whole-genome, -transcriptome, -methylome… view
- Only one source of ‘technology’ noise
Acknowledgments

Bettina Haase
Dinko Pavlinic
Jens Stolte
Jonathon Blake
Tobias Rausch
Jürgen Zimmermann
All our users and former colleagues

Science is built with facts as a house is with stones, but a collection of facts is no more a science than a heap of stones is a house.
Jules Henri Poincare

Have a nice day!