Nuevas herramientas de secuenciación (RNA seq) para el análisis de características complejas

Juan F. Medrano

Dept. of Animal Science
University of California, Davis

INIA, Madrid, España Octubre 14, 2011

J.F. Medrano / U.C. Davis
Temas

Análisis del transcriptoma con RNA sequencing

Transcriptoma de la leche a diferentes etapas lactancia
  Lactación temprana: proteínas de la leche
  Lactación tardía: enzimas proteolíticas

Características complejas/validación de reguladores
  Oligosacáridos de la leche
  Contenido de citrato en la leche
  Nutrigenómica, estudio en pez zebra
RNA sequencing procedure

Sample collection  RNA preparation and sequencing

RNA extraction

Extract mRNA from tissue
mRNA

m^G

AAAAA(C)_4

Fragmented mRNA

Random prime to convert to cDNA

Add adapters, PCR

Multiplex indexing adapter ligation

~220 M

 Millions of reads

<table>
<thead>
<tr>
<th>Tissue</th>
<th>Lane 1</th>
<th>Lane 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Brainstem</td>
<td>20.6</td>
<td>20.9</td>
</tr>
<tr>
<td>Cerebral Cortex</td>
<td>17.1</td>
<td>20.7</td>
</tr>
<tr>
<td>Hypothalamas</td>
<td>17.3</td>
<td>18.7</td>
</tr>
<tr>
<td>Gonadal fat</td>
<td>15.5</td>
<td>14.1</td>
</tr>
<tr>
<td>Pituitary</td>
<td>17.7</td>
<td>19.7</td>
</tr>
<tr>
<td>Liver</td>
<td>14.4</td>
<td>16.3</td>
</tr>
</tbody>
</table>

Total reads 102.6, 110.4

J.F. Medrano / U.C. Davis
Mapping sequencing reads to exons

Reference genome sequence

Sequencing reads

Digital expression read-out

Position in the genome

Morozova et al. 2009,

Gene expression

Measured by counting sequence reads

RPKM value = Reads per kilobase of exon per million mapped reads

Gene structure

SNP discovery

Software used:

J.F. Medrano / U.C. Davis
RNA-Sequence Analysis Workflow

I. Sequence analysis
   - Importing sequence reads and QC
   -Assembly to Reference Genome
       - De novo assembly
       - SNP detection
       - DIP detection
       - New transcripts

II. RNA-Seq analysis
   - Transcriptome (RPKM)
   - Exons/genes discovery
   - Experimental comparison
       - Compare multiple samples
       - Transformation and normalization
       - Statistical analysis

III. Functional annotation, Blast2Go
    - Pathway Analysis, IPA

J.F. Medrano / U.C. Davis
## RNA seq vs. Microarray

<table>
<thead>
<tr>
<th>Gene</th>
<th>Mammary RPKM</th>
<th>RNA-Seq Reads</th>
<th>Affymetrix Expression values</th>
</tr>
</thead>
<tbody>
<tr>
<td>CSN2</td>
<td>174686</td>
<td>1351852</td>
<td>14.31</td>
</tr>
<tr>
<td>LGB</td>
<td>129059</td>
<td>737858</td>
<td>14.11</td>
</tr>
<tr>
<td>CSN3</td>
<td>44255</td>
<td>271151</td>
<td>14.24</td>
</tr>
<tr>
<td>LALBA</td>
<td>34007</td>
<td>177313</td>
<td>14.08</td>
</tr>
<tr>
<td>CSN1S1</td>
<td>32345</td>
<td>277713</td>
<td>13.12</td>
</tr>
<tr>
<td>GLYCAM1</td>
<td>22015</td>
<td>102009</td>
<td>13.92</td>
</tr>
<tr>
<td>CSN1S2</td>
<td>14670</td>
<td>120333</td>
<td>13.83</td>
</tr>
<tr>
<td>MFGE8</td>
<td>4398</td>
<td>42588</td>
<td>12.55</td>
</tr>
<tr>
<td>FASN</td>
<td>2332</td>
<td>130664</td>
<td>12.67</td>
</tr>
<tr>
<td>LTF</td>
<td>817</td>
<td>14570</td>
<td>13.33</td>
</tr>
<tr>
<td>AGPAT6</td>
<td>425</td>
<td>7269</td>
<td>9.47</td>
</tr>
<tr>
<td>MUC1</td>
<td>411</td>
<td>4510</td>
<td>10.32</td>
</tr>
<tr>
<td>SLC29A1</td>
<td>253</td>
<td>4480</td>
<td>8.49</td>
</tr>
<tr>
<td>CIDEA</td>
<td>185</td>
<td>5352</td>
<td>11.12</td>
</tr>
<tr>
<td>PTGDS</td>
<td>147</td>
<td>877</td>
<td>10.86</td>
</tr>
<tr>
<td>TSTA3</td>
<td>125</td>
<td>1356</td>
<td>8.66</td>
</tr>
<tr>
<td>FOLR1</td>
<td>113</td>
<td>1571</td>
<td>9.66</td>
</tr>
<tr>
<td>BANF1</td>
<td>108</td>
<td>1400</td>
<td>8.90</td>
</tr>
<tr>
<td>VAT1</td>
<td>93</td>
<td>1852</td>
<td>9.56</td>
</tr>
<tr>
<td>DAP</td>
<td>86</td>
<td>1522</td>
<td>10.62</td>
</tr>
<tr>
<td>MST1</td>
<td>1.18</td>
<td>21</td>
<td>3.20</td>
</tr>
<tr>
<td>FGD1</td>
<td>1.17</td>
<td>29</td>
<td>4.10</td>
</tr>
<tr>
<td>PTGS1</td>
<td>1.16</td>
<td>25</td>
<td>5.41</td>
</tr>
<tr>
<td>MORC4</td>
<td>1.16</td>
<td>19</td>
<td>3.56</td>
</tr>
<tr>
<td>TOR1AIP2</td>
<td>1.14</td>
<td>18</td>
<td>3.01</td>
</tr>
<tr>
<td>CHRND</td>
<td>1.14</td>
<td>17</td>
<td>4.55</td>
</tr>
<tr>
<td>ARID3B</td>
<td>1.14</td>
<td>20</td>
<td>3.78</td>
</tr>
<tr>
<td>TMEM59L</td>
<td>1.13</td>
<td>13</td>
<td>3.58</td>
</tr>
<tr>
<td>RAMP1</td>
<td>1.12</td>
<td>15</td>
<td>4.17</td>
</tr>
<tr>
<td>FUT1</td>
<td>1.10</td>
<td>19</td>
<td>3.21</td>
</tr>
</tbody>
</table>

### Highly expressed genes (~180 genes)

**Dynamic Range**
- RPKM: 817 - 174,686
- Affy 12.5 – 14.3

### Medium expressed genes (~6,026)

**Dynamic Range**
- RPKM: 86 - 425
- Affy 8.5 – 11.1

### Low expressed genes (~11,024)

**Dynamic Range**
- RPKM: 1.10 – 1.18
- Affy 3.01 – 5.41

RPKM: Reads per kilo base of exon length / million reads

J.F. Medrano / U.C. Davis
~40% of the variance in protein level is explained by mRNA levels. Most of these 40% is due to differences in transcription rate.
Milk transcriptome at different stages of lactation

~18,000 of 26,000 genes are expressed
~9,000 genes are ubiquitously expressed at all stages

<table>
<thead>
<tr>
<th></th>
<th>D15</th>
<th>D90</th>
<th>D250</th>
</tr>
</thead>
<tbody>
<tr>
<td>Highly expressed &gt;500 RPKM</td>
<td>86</td>
<td>140</td>
<td>150</td>
</tr>
<tr>
<td>10 genes represent this % of reads</td>
<td>61%</td>
<td>11%</td>
<td>19%</td>
</tr>
</tbody>
</table>

D15

Molecular and Cellular Functions

- Lipid Metabolism
- Molecular Transport
- Small Molecule Biochemistry
- Cellular Growth and Proliferation
- Cell Death

Milk components
- Casein/whey proteins
- Glycam1-mucin

D250

Molecular and Cellular Functions

- Cell Cycle
- DNA Replication, Recombination, and Repair
- Cell Death
- Post-Translational Modification
- RNA Post-Transcriptional Modification

- antiapoptotic
- immune system
- Proteolytic enzymes

J.F. Medrano / U.C. Davis
Gene expression pattern of highly expressed genes at day 15 representing 61% of all sequence reads.

Protein in cow milk remains fairly constant

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Protein %</td>
<td>3.13±0.2</td>
</tr>
<tr>
<td>Casein %</td>
<td>2.38±0.21</td>
</tr>
</tbody>
</table>
Milk transcriptome at different stages of lactation

~18,000 of 26,000 genes are expressed
~9,000 genes are ubiquitously expressed at all stages

<table>
<thead>
<tr>
<th></th>
<th>D15</th>
<th>D90</th>
<th>D250</th>
</tr>
</thead>
<tbody>
<tr>
<td>Highly expressed &gt;500 RPKM</td>
<td>86</td>
<td>140</td>
<td>150</td>
</tr>
<tr>
<td>10 genes represent this % of reads</td>
<td>61%</td>
<td>11%</td>
<td>19%</td>
</tr>
</tbody>
</table>

D15

**Molecular and Cellular Functions**

- Lipid Metabolism
- Molecular Transport
- Small Molecule Biochemistry
- Cellular Growth and Proliferation
- Cell Death

Milk components
- Casein/whey proteins
- Glycam1-mucin

D250

**Molecular and Cellular Functions**

- Cell Cycle
- DNA Replication, Recombination, and Repair
- Cell Death
- Post-Translational Modification
- RNA Post-Transcriptional Modification
- apoptotic
- immune system
- Proteolytic enzymes

J.F. Medrano / U.C. Davis
Proteolytic enzymes in milk:

Plasmin (alkaline serum protease)

Cathespins (lysosomal proteases)

**Role:**
- Mammary development
- Microbial interactions
- Effect on fermented products and cheese
- Sensory quality of milk
- Potential neutraceticals

---

*J.F. Medrano / U.C. Davis*
SNP discovery in 14 Holstein cows
107,639 SNP in coding regions

SNP validation with dbSNP

Criteria
- Quality score
- >10 reads
- Min 2 reads/SNP
- No SNP on read ends

(Canovas et al
MammGen 2010)
Milk oligosaccharide structures

Lacto-N-Tetraose
[M+ Na]$^+$=732.3

Lacto-N-Neohexose
[M+ Na]$^+$=1097.4

Isomeric fucosylated Lacto-N-Hexose
[M+ Na]$^+$=1243.4

Difucosyllacto-N-Hexaose
[M+ Na]$^+$=1389.5

Zivkovic A M, Barile D Adv Nutr 2011
Syalic Acid Metabolism
genes in milk

Wickramsinghe et al PloSONE 2011
128 genes from 10 functional oligosaccharide metabolism categories in mammals

502 SNP in coding regions
- Directly genotyped by RNAseq
- Genotyping array
- Association study

Wickramsinghe et al PloSONE 2011
Non-synonymous SNP in glycosylation-related genes that showed a damaging effect in the encoded protein (Polyphen analysis)

<table>
<thead>
<tr>
<th>Gene</th>
<th>Breed</th>
<th>Reference position</th>
<th>Reference</th>
<th>Allele variations</th>
<th>Amino acid change</th>
<th>PolyPhen prediction</th>
</tr>
</thead>
<tbody>
<tr>
<td>FUT2</td>
<td>Holstein</td>
<td>917</td>
<td>A</td>
<td>G/A</td>
<td>Tyr306Cys damaging</td>
<td>(score 0.999)</td>
</tr>
<tr>
<td>FUCA1</td>
<td>Holstein</td>
<td>15,334</td>
<td>C</td>
<td>C/T</td>
<td>Thr246Ile damaging</td>
<td>(score 0.567)</td>
</tr>
<tr>
<td>NEU3</td>
<td>Holstein</td>
<td>12,263</td>
<td>T</td>
<td>T/G</td>
<td>Phe159Cys damaging</td>
<td>(score 0.997)</td>
</tr>
<tr>
<td>MGAT4A</td>
<td>Holstein, Jersey</td>
<td>64,673</td>
<td>G</td>
<td>A/G</td>
<td>Gly467Arg damaging</td>
<td>(score 0.972)</td>
</tr>
<tr>
<td>MGAT4A</td>
<td>Holstein, Jersey</td>
<td>45,746</td>
<td>C</td>
<td>T/C</td>
<td>Arg to STOP N/A</td>
<td></td>
</tr>
<tr>
<td>NEU3</td>
<td>Brown Swiss</td>
<td>12,064</td>
<td>T</td>
<td>C/T</td>
<td>Trp93Arg damaging</td>
<td>(score 0.942)</td>
</tr>
<tr>
<td>MGAT4B</td>
<td>Brown Swiss</td>
<td>2,434</td>
<td>C</td>
<td>C/T</td>
<td>Ser299Phe damaging</td>
<td>(score 0.991)</td>
</tr>
<tr>
<td>MGAT5</td>
<td>Brown Swiss</td>
<td>75,165</td>
<td>C</td>
<td>C/T</td>
<td>Ser182Phe damaging</td>
<td>(score 0.69)</td>
</tr>
<tr>
<td>GALNT1</td>
<td>Brown Swiss</td>
<td>46,723</td>
<td>T</td>
<td>T/G</td>
<td>Phe259Val damaging</td>
<td>(score 0.995)</td>
</tr>
<tr>
<td>GNE</td>
<td>Shorthorn</td>
<td>18,142</td>
<td>G</td>
<td>G/T</td>
<td>Asp338Tyr damaging</td>
<td>(score 0.952)</td>
</tr>
</tbody>
</table>
**Target Validation**

Pathway analysis

→ SNP selection  *(Canovas et al Mamm Genome, 2010)*

→ Marker-trait association studies

→ Definition of regulators

**Example**: genes responsible for variation of CITRATE content in cow milk (130-160mg/100ml).

**Citrate in milk**

- Involved in Ca and P balance
- Heat Stability
- Aids in protein coagulation, flavor and aroma
- Provides protein stability
- Primary buffer in milk

*J.F. Medrano / U.C. Davis*
Pathway of fatty acid synthesis in ruminant mammary tissue

Numbers in parenthesis correspond to average expression values (RPKM) measured by RNA-seq in milk samples.

J.F. Medrano / U.C. Davis
Zebrfish muscle tissue response to a plant protein diet

$n=440$

Average weight = 52 mg
Muscle from 8 males
pool RNA (4 fish/pool)
2 RNA-seq libraries
54 differentially expressed genes

Average weight = 228 mg
Muscle from 8 males
pool RNA (4 fish/pool)
2 RNA-seq libraries
70 differentially expressed genes

Low growth fish: protein synthesis, cellular morphology, skeletal and muscle system development, and tissue morphology.

High growth fish: lipid metabolism, vitamin and mineral metabolism and oxidation reduction.
Population fish (24 families)

Parents (48 fish)

Four low growth fish/ family
N= 96

Four high growth fish/ family
N= 96

165 SNP / 240 samples

<table>
<thead>
<tr>
<th>ID Gen</th>
<th>Gene</th>
<th>SNP</th>
<th>Minor allele</th>
<th>Minor allele frequency</th>
<th>p-value</th>
<th>FDR</th>
<th>slope</th>
<th>Amino acids</th>
</tr>
</thead>
<tbody>
<tr>
<td>ENSDARG00000000</td>
<td>N</td>
<td>A/T</td>
<td>T</td>
<td>0.129</td>
<td>1.60E-05</td>
<td>0.001233</td>
<td>-110.1670988</td>
<td>Synonym</td>
</tr>
<tr>
<td>ENSDARG00000000</td>
<td>A</td>
<td>T/C</td>
<td>T</td>
<td>0.200</td>
<td>0.0033</td>
<td>0.172945</td>
<td>12.7210075</td>
<td>Synonym</td>
</tr>
<tr>
<td>ENSDARG00000000</td>
<td>P</td>
<td>T/A</td>
<td>A</td>
<td>0.132</td>
<td>0.0050</td>
<td>0.195037</td>
<td>39.98901273</td>
<td>Synonym</td>
</tr>
<tr>
<td>ENSDARG00000000</td>
<td>C</td>
<td>A/C</td>
<td>C</td>
<td>0.031</td>
<td>0.0056</td>
<td>0.17339</td>
<td>-134.1560644</td>
<td>Ile500Leu</td>
</tr>
<tr>
<td>ENSDARG000000045864</td>
<td>Tmod1</td>
<td>G/C</td>
<td>C</td>
<td>0.223</td>
<td>0.0061</td>
<td>0.158305</td>
<td>61.38335784</td>
<td>Ser141Thr</td>
</tr>
</tbody>
</table>

J.F. Medrano / U.C. Davis
Conclusiones

• El workflow analítico de RNAseq aplicado a caracteres complejos es una robusta herramienta para incrementar el conocimiento biológico de los mismos.
  - cuantificación precisa del nivel de expresión génica con una alta correlación a los niveles de proteína.
  - el descubrimiento de nuevos tránscritos
  - la identificación de nuevos SNP y otras variantes a través de un completo genotipado del exoma del organismo
  - permitiendo la identificación de otros organismos presentes en el material biológico

• La combinación de RNAseq en el análisis de vías metabólicas e identificación de SNP con estudios de asociación es una forma experimental para definir módulos reguladores clave de caracteres complejos.

J.F. Medrano / U.C. Davis
Acknowledgements

Medrano Lab
Alma Islas, Gonzalo Rincon, Saumya Wickramasinghe, Pilar Ulloa, Angela Canovas (IRTA, Spain)

UCDavis Genome Center

Colaboradores
Carlito Lebrilla (UCDavis)
Bruce German (UCDavis)
Rafael Jimenez-Flores (CalPoly, SLO)
Armand Sanchez (UAB)

Financiamiento

[Logos and affiliations]
Genetic Principles Governing the Rate of Progress of Livestock Breeding, JAS 1939

“As a starting point suppose that we were given a reasonably complete map of all of the chromosomes, showing the location of all important genes affecting the character in question as well as of convenient marker genes. What could we do with it?”

Sewall Wright 1939