DOI: 10.21451/1984-3143-AR866

Anim. Reprod., v.13, n.3, p.153-159, Jul./Sept. 2016

Strategies to avoid drowning in the deep sequencing data flood

S. Bauersachs[1]

ETH Zurich, Animal Physiology, Institute of Agricultural Sciences, Zurich, Switzerland.

Abstract

The enormous technological progress in the field of functional genomics during the last 15 years had a significant impact on animal sciences. With the development of Next Generation Sequencing it became feasible to analyze genomes and transcriptomes within short time frames and affordable costs. One major challenge of this rapid development is to manage the data flood and to perform data analysis and integration in an optimal manner. This review provides some information about a typical analysis pipeline for RNA-Sequencing (RNA-Seq) data and a strategy for the analysis of small RNA-Seq data derived from species with poor annotation for non-coding RNA genes. Furthermore, problems regarding gene annotation in livestock species and their possible implications for data analysis and interpretation are discussed. Despite of not yet solved problems and challenges with respect to data analysis and integration the approaches in the field of functional genome analysis opened up new ways to try to understand the complex trait fertility.

Keywords: animal breeding, bioinformatics tools, biology of reproduction, deep sequencing, Galaxy project.

The impact of functional genomics on life science research

The tremendous technological advances in the field of functional genomics during the last decades had a strong impact on research in animal sciences. This is reflected, e.g., by a dramatic increase of the number of publications containing respective keywords (Fig. 1). The first wave started with the broad application of DNA microarrays end of the nineties, and a similar rise is observed for studies using RNA sequencing (RNA-Seq). With respect to livestock the increase of the number of published transcriptome studies showed a shift of two to three years.

The development of Next Generation Sequencing (NGS) facilitated the analysis of genomes and transcriptomes in an extremely short time at affordable costs (Goodwin *et al*., 2016). With the newest instruments for the generation of so-called "short reads" (up to 2X 150 bp, Illumina HiSeq 4000) it is currently possible to obtain up to 1.5 Tera bases corresponding to 5 billion reads per run or 12 genomes or 100 transcriptomes or 180 exomes per instrument run which takes 3.5 days. Furthermore, Third Generation sequencers deliver extremely long reads and can be used to sequence full-length RNA molecules (messenger as well as long non-coding RNAs) or to bridge longer repetitive genomic sequences to fill the gaps of the current versions of genome sequence assemblies (Goodwin *et al*., 2016). But also in the field of proteome analysis, the techniques have advanced, particularly mass spectrometric methods. Improvement has been achieved mainly with respect to sensitivity and quantification (Zhang *et al*., 2014a, b). Furthermore, NGS techniques have been refined in order to analyze tiny amounts of RNA or DNA. Whereas early RNA-Seq library preparation protocols needed starting material (total RNA) in the microgram range, modern standard protocols start from 100 ng of total RNA. Special protocols were developed to perform RNA-Seq even for a few or single cells such as oocytes and early embryos but also with parts of neuronal cells (Liu *et al*., 2014; Hrdlickova *et al*., 2016; Marr *et al*., 2016).

With this rapid development, particularly for NGS, a big challenge came up with respect to data analysis, interpretation, and integration (Rajasundaram and Selbig, 2016; Sun and Hu, 2016; Suravajhala *et al*., 2016). More and more data sets are generated for the analysis of gene expression at the level of RNA and proteins as well as for the genome-wide identification of sequence variants correlating with the trait fertility (Bauersachs, 2014; Bauersachs and Wolf, 2015). The combination of data from genome-wide association studies (GWAS) or quantitative trait locus (QTL) studies with corresponding data derived from gene expression analyses has a great potential to improve the understanding of the trait fertility with respect to the effects of sequence variations on gene expression regulation. A number of attempts to integrate these data have been performed for cattle (Pimentel *et al*., 2011; Minten *et al*., 2013; Moore *et al*., 2016).

In addition to the classical gene products mRNA and protein also non-coding RNA molecules are investigated which mainly have a role in regulation of gene expression (Bidarimath *et al*., 2014; Kotaja, 2014). Particularly, microRNAs (miRNAs), short non-coding regulatory RNAs, play a major role in the regulation of gene expression mainly at the level of repression of translation of specific target mRNAs as well as mRNA degradation (Krol *et al*., 2010). The expression of miRNAs in endometrium and in the embryo/conceptus has already been investigated in a number of studies (Ponsuksili *et al*., 2014; Krawczynski *et al*., 2015a, b).

For various reasons, such as not well standardized data analysis pipelines, incomplete genome sequence assemblies for livestock species, and incomplete gene annotation the analysis and the comparability of different data sets is complicated. This

[1]Corresponding author: stefan.bauersachs@usys.ethz.ch
Phone: +41(44)632-2631
Received: June 16, 2016
Accepted: July 7, 2016

is even more complicated if omics data has been generated in different labs using various technological platforms (Bauersachs, 2014). To solve these problems will be one of the main tasks for future research if the scientific community is interested in exploiting the potential of omics studies and in a real progress in the field, i.e., the understanding of fertility as a complex trait.

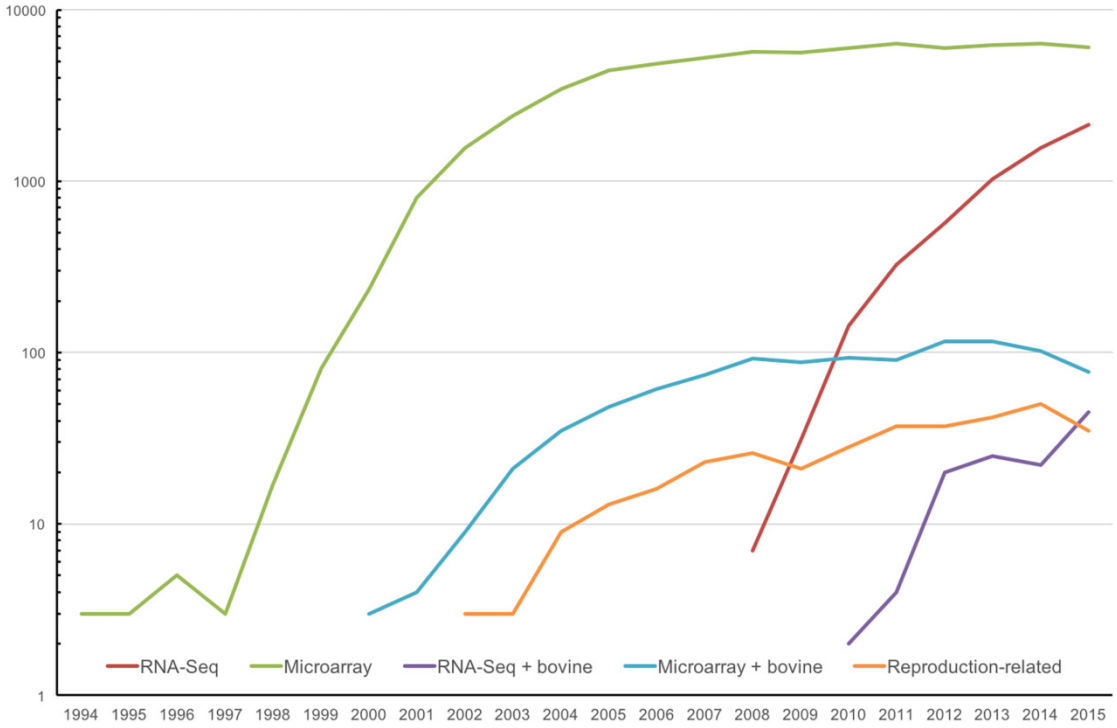

Figure 1. Pubmed search for abstracts containing keywords related to transcriptome analyses. RNA-Seq: keyword "RNA-Seq"; Microarray: keyword "microarray"; Microarray+bovine: keywords "microarray", "bovine", "*Bos taurus*", "cattle"; RNA-Seq+bovine: keywords "RNA-Seq", "bovine", "*Bos taurus*", "cattle"; Reproduction-related: using a combination of keywords for transcriptome analysis, livestock species, and reproductive organs.

Typical data analysis pipeline and statistical analysis

A typical data analysis pipeline for RNA-Seq data comprises several steps starting from the obtained sequence reads (Fastq files). Usually, the sequence reads are first trimmed based on quality scores (e.g. with Trimmomatic), i.e., bases with low quality at the ends (mainly found at the 3' end) are removed. Since RNA-Seq libraries often contain a certain percentage of cDNA inserts shorter than the read length, some reads run into the adapter sequence which has to be removed using a respective tool. To get information for the quality of the sequence data, Fastq files are checked before and after processing steps (e.g. FastQC) to ensure that all files have a comparable quality and to identify potential sequencing artifacts. After these data processing steps, the remaining reads are usually mapped to a reference genome or a transcriptome. The first is usually performed by the use of a spliced read mapper, e.g., Tophat2 (Kim *et al*., 2013) or HISAT (Kim *et al*., 2015). After assigning the sequence reads to a specific location in the genome the reads are counted for each exon, transcript or each gene. This can be performed on the basis of available gene annotation from NCBI or Ensembl. Alternatively, the data itself can be used to complement existing gene annotation using tools like Cufflinks or StringTie (Trapnell *et al*., 2012; Pertea *et al*., 2015). In the first years of RNA-Seq data analysis most of the tools were only available in command line mode running on Linux systems. With the integration into the Galaxy platform, a web browser-based genome analysis tool (Blankenberg *et al*., 2010), complex large-scale analyses can be performed without informatics or programming expertise (Giardine *et al*., 2005). Finally, these steps result in a read count table that is used for analysis of differential gene expression. Widely used tools for the analysis of read count data and the identification of differentially expressed genes (DEG) are the BioConductor R packages EdgeR (Robinson *et al*., 2010) and DESeq2 (Love *et al*., 2014). Since a local installation of Galaxy on a LINUX server is necessary to analyze bigger data sets such as RNA-Seq data an alternative way is to do the complete analysis of RNA-Seq data by the use of R and BioConductor on a desktop computer (Anders *et al*., 2013).

Analysis of small RNA-Seq data sets with special adaptation to poorly annotated species

For the analysis of small RNA-Seq data sets a modified analysis pipeline is needed compared to the basic analysis pipeline for RNA-Seq data since the resulting reads represent, at least in theory, the entire

sequence of a small ncRNA. The typical processing steps of the FastQ files starting with quality control up to adapter clipping are similar. However, the use of spliced mappers like TopHat2 (Kim *et al*., 2013) or HISAT (Kim *et al*., 2015) for the analysis of smallRNA-Seq data sets is not appropriate because small RNAs are usually neither spliced nor found in coding regions of annotated genes. This leads to the necessity of a different mapping and sequence annotation strategy. For example, mapping to a reference genome using the Burrows-Wheeler Alignment tool (BWA; Li and Durbin, 2009) to map against a reference genome or NCBI BLAST (Altschul *et al*., 1997) for short sequences which is also available in Galaxy (Blankenberg *et al*., 2010) are suitable options. The BWA aligner works best for well annotated genomes where almost all short ncRNAs are known. So, the obtained sequences are just mapped to the corresponding genes and miRNA sequences (canonical and isomiRs) can be easily analyzed with tools like miRDeep2 (Friedlander *et al*., 2012). Because the BLAST algorithm is too slow for the analysis of too high numbers of sequence comparisons the number of unique sequences found in smallRNA-Seq libraries have to be appropriately filtered, e.g., based on a counts per million (CPM) cut-off, to reduce the number of sequences from hundreds of thousands or even millions to several thousand. This filtering removes at the same time sequences without biological relevance or sequences which are very likely to be the result of sequencing artifacts. An example for this data analysis strategy is shown in Fig. 2.

A challenging problem for livestock species including pig and cattle is the rather low number of annotated small ncRNAs, which complicates the use of BWA for mapping and miRDeep2 for identification of miRNAs. Furthermore, small RNA libraries usually contain also many other small RNAs in addition to miRNAs, such as fragments of ribosomal RNAs (rRNA), transfer RNAs (tRNA), small nucleolar RNAs (snoRNA), small nuclear RNAs (snRNA), and Piwi-associated small RNAs (piRNAs) in case of germline cells (Cole *et al*., 2009). Although the prediction of novel miRNAs can be performed by the use of miRDeep2 (Friedlander *et al*., 2012), the annotation of sequence fragments derived from other RNA molecules is more difficult. In contrast in humans, a great variety of ncRNAs is known compared to other mammalian species. This information can be used to improve annotation of small RNA data from other species since many of these RNAs are highly conserved. The use of BLASTn-short (local installation in Galaxy) for sequence comparison to all available sequences for RNA molecules of the target species and the inclusion of ortholog information derived from well annotated species from different annotation sources significantly improves the annotation of identified sequences found in the small RNA-Seq results to 80-90%, depending on the species and the sample type (Jochen Bick, 2016; ETH Zurich; personal communication). The consideration of the frequent occurrence of sequences representing isoforms of miRNAs (isomiRs; Krawczynski *et al*., 2015a; Zhang *et al*., 2016) can further improve sequence annotation. IsomiRs result from imprecise and alternative cleavage during the pre-miRNA processing and post-transcriptional modifications. The isomiRs show different miRNA stability, sub-cellular localization, and target selection (Zhang *et al*., 2016). Since post-transcriptional modification during miRNA processing also leads to the addition of nucleotides not matching to the genome sequence those isomiRs cannot be easily mapped using BWA and/or miRDeep2. Using the annotation strategy based on BLASTn searches following statistical data analyses can be performed including various types of small ncRNAs or miRNAs only. Furthermore, based on the attempt to annotate as much as possible of the obtained sequences, percentages of read counts in relation to the total number of read counts can be calculated for individual types of ncRNAs. This can also help to identify technical or biological outliers in a data set.

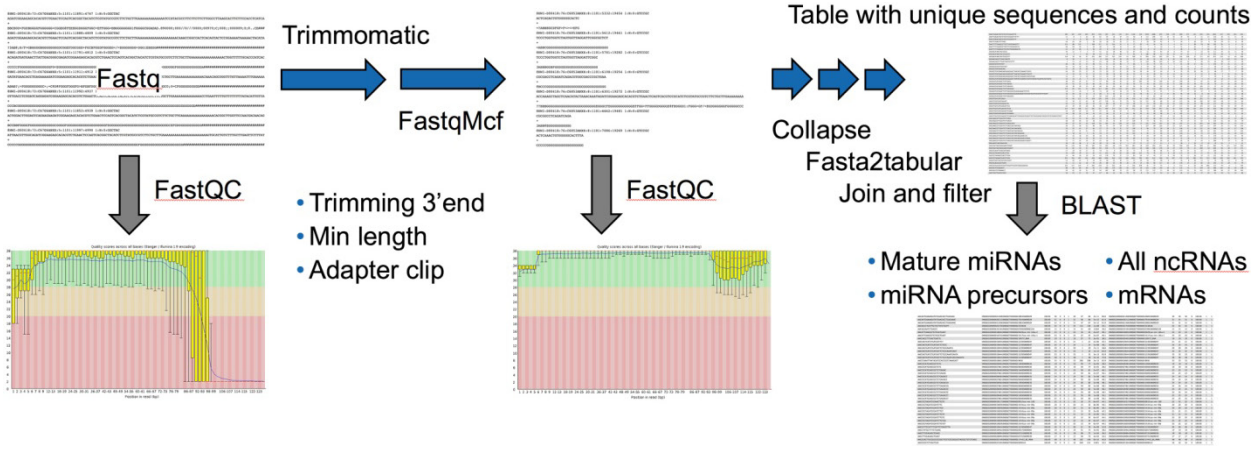

Figure 2. Workflow for data analysis of small RNA-Seq data performed by the use of Galaxy tools. The workflow goes from left to the right and includes processing of sequence files, generation of a read count table, and annotation of the obtained sequences. Fastq: sequence files derived from Illumina sequencer; FastQC: tool for quality control of fastq files; Trimmomatic and FastqMcf: tools for processing fastq files; BLAST: Basic Local Alignment Search Tool.

Gene annotation in livestock

The efforts to sequence the human genome (Lander *et al*., 2001) were extremely high and cost alone the US tax payer almost three billion dollars. In addition to the genome sequence itself large projects were performed to sequence full-length cDNAs from mRNA derived from almost all human tissues (Wiemann *et al*., 2001) to obtain information about transcribed regions in the genome, gene structures, and transcript isoforms. Meanwhile, genomes have been sequenced also for livestock species (Elsik *et al*., 2009; Wade *et al*., 2009; Groenen *et al*., 2012). However, gene annotation for these species is still based in large part on the comparison to human or mouse orthologous genes. In addition, different annotation pipelines, e.g., NCBI and Ensembl, provide gene models which show sometimes substantial differences making the correct assignment of genes annotated with different pipelines at the same genomic locus difficult. The corresponding information for the assignment of genes in Entrez Gene to genes in Ensembl is incomplete (NCBI->Ensembl) or incorrect (Ensembl->Entrez Gene). This is a serious problem if useful information such as ortholog annotation found in one database should be assigned to gene IDs of the other database.

For the functional annotation and downstream bioinformatics analysis the use of gene IDs of livestock species is not optimal and leads to information loss. The reason for that is incomplete annotation, i.e., many genes still do not have the official gene symbol and are not assigned to functional annotation databases such as Gene Ontologies (Ashburner *et al*., 2000) and KEGG pathway database (Kanehisa and Goto, 2000). To avoid this loss of information the putative human ortholog information can be used. One resource for ortholog information is for example EnsemblCompara (Vilella *et al*., 2009; Pignatelli *et al*., 2016). In order to combine information derived from different databases provided by the NCBI and Ensembl we are developing a Mammalian Ortholog and Annotation database (MOA-Db) integrated in the Galaxy platform in our group (Jochen Bick, 2016; ETH Zurich; unpublished results). A schematic overview of the MOA-Db is shown in Fig. 3.

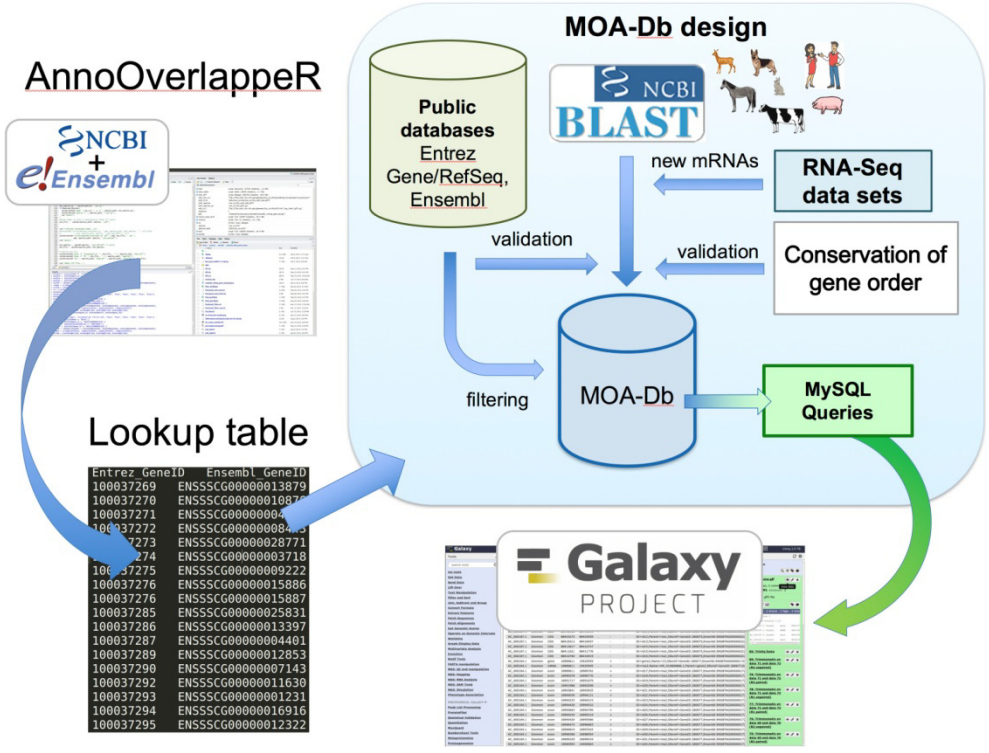

Figure 3. Development of a Mammalian Ortholog and Annotation database (MOA-Db). Based on information derived from public databases and available RNA-Seq data sets an annotation database is built for a number of mammalian species including gene annotation from NCBI and Ensembl as well as ortholog information. The ortholog relationships are based on information extracted from databases such as EnsemblCompara as well as on crosswise global BLAST comparisons of all transcripts annotated at NCBI for each species.

Conclusions

The development of functional genomics approaches opened new ways to improve our understanding of the complex trait fertility. After the first wave of enthusiasm it is becoming more and more evident that there is a number of big challenges in the context of data analysis and integration. The main problems are inherent in missing standards for data analysis pipelines, integration of different kinds of data sets, bias in data sets related to different laboratories, protocols, and the use of different platforms. Furthermore, a particular challenge is the integration of results from different omics approaches, such as genome, transcriptome, proteome, and metabolome analysis. A major obstacle for the integration of omics

data sets is the existence of a plethora of different databases and corresponding identifiers as well as incomplete, inconsistent, and not coordinated gene annotations, e.g., when comparing genome annotation at NCBI and Ensembl. In addition, an insufficient and/or erroneous gene or protein annotation leads to a significant loss of information and in the worst case to wrong data interpretation. Despite of all these problems and challenges, the development of the new sequencing technologies and the foreseeable even more exciting developments with respect to functional genomics technologies promise a new era of research in the animal sciences.

References

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*, 25:3389-3402.

Anders S, McCarthy DJ , Chen Y, Okoniewski M, Smyth GK, Huber W, Robinson MD. 2013. Count-based differential expression analysis of RNA sequencing data using R and bioconductor. *Nat Protoc*, 8:1765-1786.

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G. 2000. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet,* 25:25-29.

Bauersachs S. 2014. Combined analysis of transcriptome studies of bovine endometrium during the preimplantation phase and comparison to results from ovine and porcine preimplantation endometrium. *In*: Juengel JL, Miyamoto A, Price C, Reynolds LP, Smith MF, Webb R (Ed.). *Reproduction in Domestic Ruminants VIII*: *Proceedings of the Ninth International Symposium on Reproduction in Domestic Ruminants*. Leicestershire, UK: Context Products Ltd.. pp. 167-177.

Bauersachs S, Wolf E. 2015. Uterine responses to the preattachment embryo in domestic ungulates: recognition of pregnancy and preparation for implantation. *Annu Rev Anim Biosci*, 3:489-511.

Bidarimath M, Khalaj K, Wessels JM, Tayade C. 2014. MicroRNAs, immune cells and pregnancy. *Cell Mol Immunol*, 11:538-547.

Blankenberg D, Von Kuster G, Coraor N, Ananda G, Lazarus R, Mangan M, Nekrutenko A, Taylor J. 2010. Galaxy: a web-based genome analysis tool for experimentalists. *Curr Protoc Mol Biol*, Chapter 19, Unit 19.10.1-21.

Cole C, Sobala A, Lu C, Thatcher SR, Bowman A, Brown JW, Green PJ, Barton GJ, Hutvagner G. 2009. Filtering of deep sequencing data reveals the existence of abundant Dicer-dependent small RNAs derived from tRNAs. *RNA*, 15:2147-2160.

Elsik CG, Tellam RL, Worley KC, Gibbs RA, Muzny DM, Weinstock GM, Adelson DL, Eichler EE,Elnitski L, Guigó R, Hamernik DL, Kappes SM, Lewin HA, Lynn DJ, Nicholas FW, Reymond A, Rijnkels M, Skow LC, Zdobnov EM, Schook L, Womack J, Alioto T, Antonarakis SE, Astashyn A, Chapple CE, Chen HC, Chrast J, Câmara F, Ermolaeva O, Henrichsen CN, Hlavina W, Kapustin Y, Kiryutin B, Kitts P,Kokocinski F, Landrum M, Maglott D, Pruitt K, Sapojnikov V, Searle SM, Solovyev V, Souvorov A, Ucla C, Wyss C, Anzola JM, Gerlach D, Elhaik E, Graur D, Reese JT, Edgar RC, McEwan JC, Payne GM, Raison JM, Junier T, Kriventseva EV, Eyras E, Plass M, Donthu R, Larkin DM, Reecy J, Yang MQ, Chen L,Cheng Z, Chitko-McKown CG, Liu GE, Matukumalli LK, Song J, Zhu B, Bradley DG, Brinkman FS, Lau LP, Whiteside MD, Walker A, Wheeler TT, Casey T,German JB, Lemay DG, Maqbool NJ, Molenaar AJ, Seo S, Stothard P, Baldwin CL, Baxter R, Brinkmeyer-Langford CL, Brown WC, Childers CP, Connelley T, Ellis SA, Fritz K, Glass EJ, Herzig CT, Iivanainen A, Lahmers KK, Bennett AK, Dickens CM, Gilbert JG, Hagen DE, Salih H, Aerts J, Caetano AR, Dalrymple B, Garcia JF, Gill CA, Hiendleder SG, Memili E, Spurlock D, Williams JL, Alexander L, Brownstein MJ, Guan L, Holt RA, Jones SJ, Marra MA, Moore R, Moore SS, Roberts A, Taniguchi M, Waterman RC, Chacko J, Chandrabose MM, Cree A, Dao MD, Dinh HH, Gabisi RA, Hines S, Hume J, Jhangiani SN, Joshi V,Kovar CL, Lewis LR, Liu YS, Lopez J, Morgan MB, Nguyen NB, Okwuonu GO, Ruiz SJ, Santibanez J, Wright RA, Buhay C, Ding Y, Dugan-Rocha S, Herdandez J, Holder M, Sabo A, Egan A, Goodell J, Wilczek-Boney K, Fowler GR, Hitchens ME, Lozado RJ, Moen C, Steffen D, Warren JT, Zhang J, Chiu R,Schein JE, Durbin KJ, Havlak P, Jiang H, Liu Y, Qin X, Ren Y, Shen Y, Song H, Bell SN, Davis C, Johnson AJ, Lee S, Nazareth LV, Patel BM, Pu LL, Vattathil S, Williams RL Jr, Curry S, Hamilton C, Sodergren E, Wheeler DA, Barris W, Bennett GL, Eggen A, Green RD, Harhay GP, Hobbs M, Jann O, Keele JW, Kent MP, Lien S, McKay SD, McWilliam S, Ratnakumar A, Schnabel RD, Smith T, Snelling WM, Sonstegard TS, Stone RT, Sugimoto Y, Takasuga A, Taylor JF, Van Tassell CP, Macneil MD, Abatepaulo AR, Abbey CA, Ahola V, Almeida IG, Amadio AF, Anatriello E, Bahadue SM, Biase FH, Boldt CR, Carroll JA, Carvalho WA, Cervelatti EP, Chacko E, Chapin JE, Cheng Y, Choi J, Colley AJ, de Campos TA, De Donato M, Santos IK, de Oliveira CJ, Deobald H, Devinoy E,Donohue KE, Dovc P, Eberlein A, Fitzsimmons CJ, Franzin AM, Garcia GR, Genini S, Gladney CJ, Grant JR, Greaser ML, Green JA, Hadsell DL, Hakimov HA, Halgren R, Harrow JL, Hart EA, Hastings N, Hernandez M, Hu ZL, Ingham A, Iso-Touru T, Jamis C, Jensen K, Kapetis D, Kerr T, Khalil SS, Khatib H,Kolbehdari D, Kumar CG, Kumar D, Leach R, Lee JC, Li C, Logan KM, Malinverni R, Marques E, Martin WF, Martins NF, Maruyama SR, Mazza R, McLean KL, Medrano JF, Moreno BT, Moré DD, Muntean CT, Nandakumar HP, Nogueira MF, Olsaker I, Pant SD, Panzitta F, Pastor RC, Poli MA, Poslusny N,Rachagani S, Ranganathan S, Razpet A, Riggs PK, Rincon G, Rodriguez-Osorio N, Rodriguez-Zas SL, Romero NE, Rosenwald A, Sando L, Schmutz SM,Shen L, Sherman L, Southey BR, Lutzow YS, Sweedler JV, Tammen I, Telugu

BP, Urbanski JM, Utsunomiya YT, Verschoor CP, Waardenberg AJ, Wang Z,Ward R, Weikard R, Welsh TH Jr, White SN, Wilming LG, Wunderlich KR, Yang J, Zhao FQ. 2009. The genome sequence of taurine cattle: a window to ruminant biology and evolution. *Science*, 324:522-528.

Friedlander MR , Mackowiak SD, Li N, Chen W, Rajewsky N. 2012. miRDeep2 accurately identifies known and hundreds of novel microRNA genes in seven animal clades. *Nucleic Acids Res*, 40:37-52.

Giardine B, Riemer C, Hardison RC, Burhans R, Elnitski L, Shah P, Zhang Y, Blankenberg D, Albert I, Taylor J, Miller W, Kent WJ, Nekrutenko A. 2005. Galaxy: a platform for interactive large-scale genome analysis. *Genome Res*, 15:1451-1455.

Goodwin S, McPherson JD, McCombie WR. 2016. Coming of age: ten years of next-generation sequencing technologies. *Nat Rev Genet*, 17:333-351.

Groenen MA, Archibald AL, Uenishi H, Tuggle CK, Takeuchi Y, Rothschild MF, Rogel-Gaillard C, Park C, Milan D, Megens HJ, Li S, Larkin DM, Kim H, Frantz LA, Caccamo M, Ahn H, Aken BL, Anselmo A, Anthon C, Auvil L, Badaoui B, Beattie CW, Bendixen C, Berman D, Blecha F, Blomberg J, Bolund L, Bosse M,Botti S, Bujie Z, Bystrom M, Capitanu B, Carvalho-Silva D, Chardon P, Chen C, Cheng R, Choi SH, Chow W, Clark RC, Clee C, Crooijmans RP, Dawson HD, Dehais P, De Sapio F, Dibbits B, Drou N, Du ZQ, Eversole K, Fadista J, Fairley S, Faraut T, Faulkner GJ, Fowler KE, Fredholm M, Fritz E, Gilbert JG, Giuffra E,Gorodkin J, Griffin DK, Harrow JL, Hayward A, Howe K, Hu ZL, Humphray SJ, Hunt T, Hornshøj H, Jeon JT, Jern P, Jones M, Jurka J, Kanamori H, Kapetanovic R, Kim J, Kim JH, Kim KW, Kim TH, Larson G, Lee K, Lee KT, Leggett R, Lewin HA, Li Y, Liu W, Loveland JE, Lu Y, Lunney JK, Ma J, Madsen O, Mann K, Matthews L, McLaren S, Morozumi T, Murtaugh MP, Narayan J, Nguyen DT, Ni P, Oh SJ, Onteru S, Panitz F, Park EW, Park HS, Pascal G, Paudel Y,Perez-Enciso M, Ramirez-Gonzalez R, Reecy JM, Rodriguez-Zas S, Rohrer GA, Rund L, Sang Y, Schachtschneider K, Schraiber JG, Schwartz J, Scobie L, Scott C, Searle S, Servin B, Southey BR, Sperber G, Stadler P, Sweedler JV, Tafer H, Thomsen B, Wali R, Wang J, Wang J, White S, Xu X, Yerle M, Zhang G, Zhang J, Zhang J, Zhao S, Rogers J, Churcher C, Schook LB. 2012. Analyses of pig genomes provide insight into porcine demography and evolution. *Nature*, 491:393-398.

Hrdlickova R, Toloue M, Tian B. 2016. RNA-Seq methods for transcriptome analysis. *Wiley Interdiscip Rev RNA*. doi: 10.1002/wrna.1364.

Kanehisa M, Goto S. 2000. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res*, 28:27-30.

Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL. 2013. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol*, 14:R36.

Kim D, Langmead B, Salzberg SL. 2015. HISAT: a fast spliced aligner with low memory requirements. *Nat Methods*, 12:357-360.

Kotaja N. 2014. MicroRNAs and spermatogenesis. *Fertil Steril*, 101:1552-1562.

Krawczynski K, Bauersachs S, Reliszko ZP, Graf A, Kaczmarek MM. 2015a. Expression of microRNAs and isomiRs in the porcine endometrium: implications for gene regulation at the maternal-conceptus interface. BMC *Genomics*, 16:906.

Krawczynski K, Najmula J, Bauersachs S, Kaczmarek MM. 2015b. MicroRNAome of porcine conceptuses and trophoblasts: expression profile of micrornas and their potential to regulate genes crucial for establishment of pregnancy. *Biol Reprod*, 92:21.

Krol J, Loedige I, Filipowicz W. 2010. The widespread regulation of microRNA biogenesis, function and decay. *Nat Rev Genet*, 11:597-610.

Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, Funke R, Gage D, Harris K, Heaford A, Howland J, Kann L, Lehoczky J, LeVine R, McEwan P, McKernan K, Meldrim J, Mesirov JP, Miranda C, Morris W, Naylor J, Raymond C, Rosetti M, Santos R, Sheridan A, Sougnez C, Stange-Thomann Y, Stojanovic N, Subramanian A, Wyman D, Rogers J, Sulston J, Ainscough R, Beck S, Bentley D, Burton J, Clee C, Carter N, Coulson A, Deadman R, Deloukas P, Dunham A, Dunham I, Durbin R, French L, Grafham D, Gregory S, Hubbard T, Humphray S, Hunt A, Jones M, Lloyd C, McMurray A, Matthews L, Mercer S, Milne S, Mullikin JC, Mungall A, Plumb R, Ross M, Shownkeen R, Sims S, Waterston RH, Wilson RK, Hillier LW, McPherson JD, Marra MA, Mardis ER, Fulton LA, Chinwalla AT, Pepin KH, Gish WR, Chissoe SL, Wendl MC, Delehaunty KD, Miner TL, Delehaunty A, Kramer JB, Cook LL, Fulton RS, Johnson DL, Minx PJ, Clifton SW, Hawkins T, Branscomb E, Predki P, Richardson P, Wenning S, Slezak T, Doggett N, Cheng JF, Olsen A, Lucas S, Elkin C, Uberbacher E, Frazier M, Gibbs RA, Muzny DM, Scherer SE, Bouck JB, Sodergren EJ, Worley KC, Rives CM, Gorrell JH, Metzker ML, Naylor SL, Kucherlapati RS, Nelson DL, Weinstock GM, Sakaki Y, Fujiyama A, Hattori M, Yada T, Toyoda A, Itoh T, Kawagoe C, Watanabe H, Totoki Y, Taylor T, Weissenbach J, Heilig R, Saurin W, Artiguenave F, Brottier P, Bruls T, Pelletier E, Robert C, Wincker P, Smith DR, Doucette-Stamm L, Rubenfield M, Weinstock K, Lee HM, Dubois J, Rosenthal A, Platzer M, Nyakatura G, Taudien S, Rump A, Yang H, Yu J, Wang J, Huang G, Gu J, Hood L, Rowen L, Madan A, Qin S, Davis RW, Federspiel NA, Abola AP, Proctor MJ, Myers RM, Schmutz J, Dickson M, Grimwood J, Cox DR, Olson MV, Kaul R, Raymond C, Shimizu N, Kawasaki K, Minoshima S, Evans GA, Athanasiou M, Schultz R, Roe BA, Chen F, Pan H, Ramser J, Lehrach H, Reinhardt R, McCombie WR, de la Bastide M, Dedhia N, Blöcker H, Hornischer K, Nordsiek G, Agarwala R, Aravind L, Bailey JA, Bateman A, Batzoglou S, Birney E, Bork P, Brown DG, Burge CB, Cerutti L, Chen HC, Church D, Clamp M, Copley RR, Doerks T, Eddy SR, Eichler EE, Furey TS, Galagan J, Gilbert JG, Harmon C, Hayashizaki Y, Haussler D, Hermjakob H, Hokamp K, Jang W, Johnson LS, Jones TA, Kasif S, Kaspryzk A, Kennedy S, Kent WJ, Kitts P, Koonin EV, Korf I,Kulp D, Lancet D, Lowe TM, McLysaght

A, Mikkelsen T, Moran JV, Mulder N, Pollara VJ, Ponting CP, Schuler G, Schultz J, Slater G, Smit AF, Stupka E, Szustakowki J, Thierry-Mieg D, Thierry-Mieg J, Wagner L, Wallis J, Wheeler R, Williams A, Wolf YI, Wolfe KH, Yang SP, Yeh RF, Collins F, Guyer MS,Peterson J, Felsenfeld A, Wetterstrand KA, Patrinos A, Morgan MJ, de Jong P, Catanese JJ, Osoegawa K, Shizuya H, Choi S, Chen YJ, Szustakowki J. **International Human Genome Sequencing Consortium**. 2001. Initial sequencing and analysis of the human genome. *Nature*, 409:860-921.

Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, 25:1754-1760.

Liu N, Liu L, Pan X. 2014. Single-cell analysis of the transcriptome and its application in the characterization of stem cells and early embryos. *Cell Mol Life Sci*, 71:2707-2715.

Love MI, Huber W, Anders S. 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol*, 15:550.

Marr C, Zhou JX, Huang S. 2016. Single-cell gene expression profiling and cell state dynamics: collecting data, correlating data points and connecting the dots. *Curr Opin Biotechnol*, 39:207-214.

Minten MA, Bilby TR, Bruno RG, Allen CC, Madsen CA, Wang Z, Sawyer JE, Tibary A. Neibergs HL, Geary TW, Bauersachs S, Spencer TE. 2013. Effects of fertility on gene expression and function of the bovine endometrium. *PLoS One*, 8:e69444.

Moore SG, Pryce JE, Hayes BJ, Chamberlain AJ, Kemper KE, Berry DP, McCabe M, Cormican P, Lonergan P, Fair T, Butler ST. 2016. Differentially expressed genes in endometrium and corpus luteum of holstein cows selected for high and low fertility are enriched for sequence variants associated with fertility. *Biol Reprod*, 94:19.

Pertea M, Pertea GM, Antonescu CM, Chang TC, Mendell JT, Salzberg SL. 2015. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat Biotechnol*, 33:290-295.

Pignatelli M, Vilella AJ, Muffato M, Gordon L, White S, Flicek P, Herrero J. 2016. ncRNA orthologies in the vertebrate lineage. *Database (Oxford)*, 2016:bav127.

Pimentel EC, Bauersachs S, Tietze M, Simianer H, Tetens J, Thaller G, Reinhardt F, Wolf E, Konig S. 2011. Exploration of relationships between production and fertility traits in dairy cattle via association studies of SNPs within candidate genes derived by expression profiling. *Anim Genet*, 42:251-262.

Ponsuksili S, Tesfaye D, Schellander K, Hoelker M, Hadlich F, Schwerin M, Wimmers K. 2014. Differential expression of miRNAs and their target mRNAs in endometria prior to maternal recognition of pregnancy associates with endometrial receptivity for in vivo- and in vitro-produced bovine embryos. *Biol Reprod*, 91:135.

Rajasundaram D, Selbig J. 2016. More effort - more results: recent advances in integrative 'omics' data analysis. *Curr Opin Plant Biol,* 30:57-61.

Robinson MD, McCarthy DJ, Smyth GK. 2010. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26:139-140.

Sun YV, Hu YJ. 2016. Integrative analysis of multi-omics data for discovery and functional studies of complex human diseases. *Adv Genet*, 93:147-190.

Suravajhala P, Kogelman LJ, Kadarmideen HN. 2016. Multi-omic data integration and analysis using systems genomics approaches: methods and applications in animal production, health and welfare. *Genet Sel Evol,* 48:38.

Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, Pimentel H, Salzberg SL, Rinn JL, Pachter,L. 2012. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat Protoc*, 7:562-578.

Vilella AJ, Severin J, Ureta-Vidal A, Heng L, Durbin R, Birney E. 2009. EnsemblCompara GeneTrees: Complete, duplication-aware phylogenetic trees in vertebrates. *Genome Res*, 19:327-335.

Wade CM, Giulotto E, Sigurdsson S, Zoli M, Gnerre S, Imsland F, Lear TL, Adelson DL, Bailey E, Bellone RR, Blöcker H, Distl O, Edgar RC, Garber M, Leeb T, Mauceli E, MacLeod JN, Penedo MC, Raison JM, Sharpe T, Vogel J, Andersson L, Antczak DF, Biagi T, Binns MM, Chowdhary BP, Coleman SJ, Della Valle G, Fryc S, Guérin G, Hasegawa T, Hill EW, Jurka J, Kiialainen A, Lindgren G, Liu J, Magnani E, Mickelson JR, Murray J, Nergadze SG, Onofrio R, Pedroni S, Piras MF, Raudsepp T, Rocchi M, Røed KH, Ryder OA, Searle S, Skow L, Swinburne JE, Syvänen AC, Tozaki T, Valberg SJ, Vaudin M, White JR, Zody MC; Broad Institute Genome Sequencing Platform; Broad Institute Whole Genome Assembly Team, Lander ES, Lindblad-Toh K. 2009. Genome sequence, comparative analysis, and population genetics of the domestic horse. *Science*, 326:865-867.

Wiemann S, Weil B, Wellenreuther R, Gassenhuber J, Glassl S, Ansorge W, Böcher M, Blöcker H, Bauersachs S, Blum H, Lauber J, Düsterhöft A, Beyer A,Köhrer K, Strack N, Mewes HW, Ottenwälder B, Obermaier B, Tampe J, Heubner D, Wambutt R, Korn B, Klein M, Poustka A. 2001. Toward a catalog of human genes and proteins: sequencing and analysis of 500 novel complete protein coding human cDNAs. *Genome Res*, 11:422-435.

Zhang G, Annan RS, Carr SA, Neubert TA. 2014a. Overview of peptide and protein analysis by mass spectrometry. *Curr Protoc Mol Biol*, 108:10.21.1-10.21.30.

Zhang Z, Wu S, Stenoien DL, Pasa-Tolic L. 2014b. High-throughput proteomics. *Annu Rev Anal Chem*, 7:427-454.

Zhang Y, Zang Q, Xu B, Zheng W, Ban R, Zhang H, Yang Y, Hao Q, Iqbal F, Li A, Shi Q. 2016. IsomiR Bank: a research resource for tracking IsomiRs. *Bioinformatics*.