

Gene expression analysis of duck liver steatosis from hybrid and parental species

Xi Liu, Frédéric Hérault, Christian Diot, Erwan Corre

▶ **To cite this version:**

Xi Liu, Frédéric Hérault, Christian Diot, Erwan Corre. Gene expression analysis of duck liver steatosis from hybrid and parental species. Journées Ouvertes Biologie, Informatique et Mathématiques (Jobim), Jul 2018, Marseille, France. , 2018, Abstracts Jobim 2018. hal-01842089

HAL Id: hal-01842089
https://hal.science/hal-01842089

Submitted on 17 Jul 2018

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

GENE EXPRESSION ANALYSIS OF DUCK LIVER STEATOSIS FROM HYBRID AND PARENTAL SPECIES

Xi Liu[1], Frédéric Hérault[2], Christian Diot[2], Erwan Corre[1]

[1] CNRS, Sorbonne Université, FR2424, ABiMS platform, Station Biologique, 29680, Roscoff, France - xi.liu@sb-roscoff.fr, erwan.corre@sb-roscoff.fr

[2] UMR 1348 PEGASE, Physiology, Environment and Genetics for the Animal and livestock Systems, INRA-Agrocampus Ouest, Saint-Gilles, France - frederic.herault@inra.fr, christian.diot@inra.fr

1. Background

Liver steatosis can occur spontaneously in wild waterfowls as a result of energy storage before migration. This ability is exploited since thousand years in domesticated birds to produce "foie gras" by overfeeding. However, different abilities for fatty liver production are known according to species. **To better describe the mechanisms involved in hepatic steatosis development and differences between species and hybrids**, next-generation sequencing and analyses was performed on RNAs extracted from the livers of ducks fed ad libitum or overfed.

Usually, such RNA-seq analyses involve the mapping of reads on a reference genome. However, when two different species are involved some mapping biases could be expected. Thus, alternative methods must be developed to be more appropriate for transcriptome analyses and comparisons between the two species and their hybrids. Analyses using both reference genome and *de novo* methods point out a good performance of the *de novo* method for the different species treatment revealing a new set of genes differentially expressed.

Keywords: RNA-seq, Inter-species Assembly, Gene expression, Liver steatosis

2. Samples

Genotypes



Anas *platyrhynchos* Common ducks "Pekin" — ♂♀ Hinny — ♀♂ Mulard —  Cairina *moschata* Muscovy ducks

"Foie gras" production: Mulard > Hinny; Muscovy > Pekin

Treatments

2 diets: Ad libitum & (14 days') Over feeding

Samples

4 genotypes x 2 treatments x 10 biological replicates

3. Assemblies

Sequencing

Illumina HiSeq 2000: 100bp paired-end x 80 libraries

Reference genome assembly

Available reference genome: *Anas platyrhynchos* *BGI_duck_1.0, INSDC Assembly GCA_000355885.1*

Adapter trimming ✚ Assembly: **TopHat2 + Cufflinks + Cuffmerge[1]** ⇨ Nb sequences 41K Nb residues 65M N50 2545

Reference genome is only available for Pekin and other species don't have a satisfying mapping rate on this reference, reference genome assembly is limited by the mapping biases.

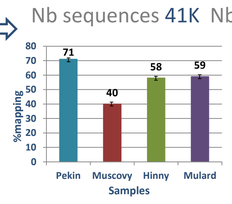


Figure 1 | Average of mapping rate on reference genome in function of the species by TopHat2. Similar result is shown by mapping reads back to reference genome assembly.

De novo assembly

Raw reads

	Pekin	Muscovy	Hinny	Mulard
Input read pairs	555M	514M	545M	639M
	⇩ Preprocessing: adapter removing, quality trimming and filtering: **Trimmomatic[2]**			
Both surviving	98,34%	98,24%	98,23%	98,22%
	⇩ Assemblies following by an in silico reads normalization: **Trinity[3] and Velvet-oases[4]**			
Normalization by set	28,70%	27,58%	31,53%	35,41%
Global normalization	6,68% ↔ 5,63%	7,17%	8,02% ↔ 5,32%	7,33%

Less divergence for parents than hybrids

	Pekin	Muscovy	Hinny	Mulard	Pekin+Muscovy	Hinny+Mulard	Hinny+Mulard
					Trinity		Velvet-oases
Nb sequences	491K	482K	631K	662K	808K	1389K	1329K
Nb residues	207M	546M	521M	554M	673M	870M	3365M
N50	2330	2694	1441	1464	1406	858	5014

✓ *Trinity has relatively a poorer performance on assembly with hybrid reads*

✓ *Velvet-oases generate very long sequences suspected to be chimeras*

⇩ Pseudo-mapping reads back to the assemblies: **Kallisto[5]**

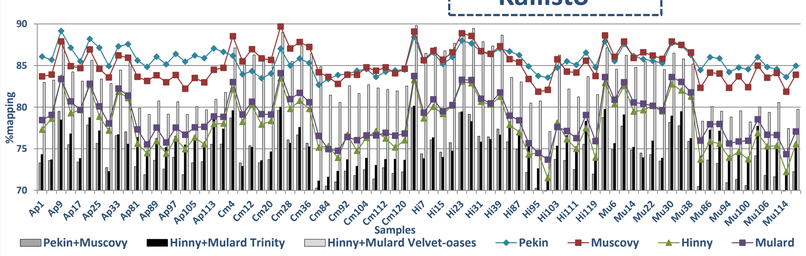


Figure 2 | Comparison of pseudoalignment rate in function of the samples between different assemblies. Kallisto analysis with no strand-specific paired-end pseudoalignment option. Pseudomapping rate on inter-species assemblies are represented as histogram. Pseudomapping rate on single species assemblies are represented as lines.

✓ *%mapping of single species assemblies are satisfied and globally better than that of inter-species assemblies, except velvet-oases assembly which displays a good rate*

✓ *For the single species assemblies, transcriptome of parents have a better rate than that of hybrids*

⇨Construct a meta-transcriptome using single species assemblies

⇩ Construction and evaluation of reference transcriptome: **DRAP[6]** ✚ **Kallisto + Busco[7]**

	Nb Seq	N50	Lg sum	Lg median
Input sequences information	2,265,631	1860	2,128,625,878	436
Coverage cleaning FPKM 1	175,572	2807	286,879,845	1000

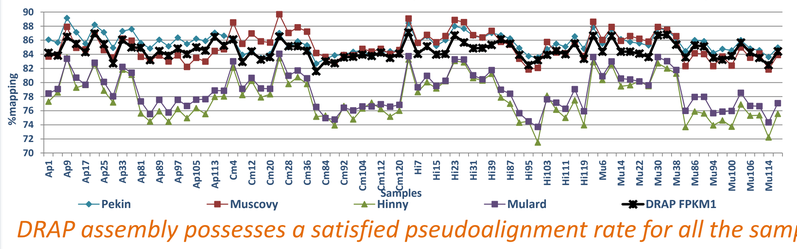


Figure 3 | Comparison of pseudoalignment rate in function of the samples between Trinity single species assemblies and DRAP meta-assembly (FPKM 1). Kallisto analysis with no strand-specific paired-end pseudoalignment option. Transrate v1.0.2 score of DRAP assembly is 0.2808.

DRAP assembly possesses a satisfied pseudoalignment rate for all the samples

BUSCO categories	Reference genome	Cufflinks assembly	DRAP FPKM1
Complete	82.9%	81.5%	97.1%
Complete and single-copy	81.2%	42.9%	59.1%
Complete and duplicated	1.7%	38.6%	38.0%
Fragmented	9.9%	11.2%	3.0%
Missing	7.2%	7.3%	0.0%

Table 1 | BUSCO notation assessment results. Total BUSCO groups searched is 303.

De novo approach discovers more new genes (97% completeness), provides a transcriptome more complete (0 missing) and less fragmented (only 3%).

⇨DRAP assembly was chosen as the reference transcriptome for further analyses.

4. Differential analysis

	DEG	Pekin	Muscovy	Mule	Hinny	common
DRAP transcriptome	up-regulated	2281	3450	4907	3901	539
	down-regulated	1468	2717	4013	3795	364
	all	3749	6167	8920	7696	906
Cufflinks transcriptome	up-regulated	1553	1371	1592	1314	520
	down-regulated	680	773	953	924	235
	all	2233	2144	2545	2238	758
Increase (%) DRAP vs Cufflinks	up-regulated	147	252	308	297	104
	down-regulated	216	351	421	411	155
	all	168	288	350	344	120

Table 2 | comparison of differentially expressed genes (DEG) of reference genome and *De novo* assemblies. For both transcriptomes, statistical tests are realized at gene level by edgeR[8] following a normalization step, then the results are filtered at p-value<0.05 and FC≥2. For DRAP assembly, methods DESeq2 and edgeR show similar results. As shown for Cufflinks assembly, 2281 and 1468 genes are found respectively up and down-regulated by overfeeding for Pekin, 758 DEG are in common for the 4 genotypes (also shown in Fig. 4C).

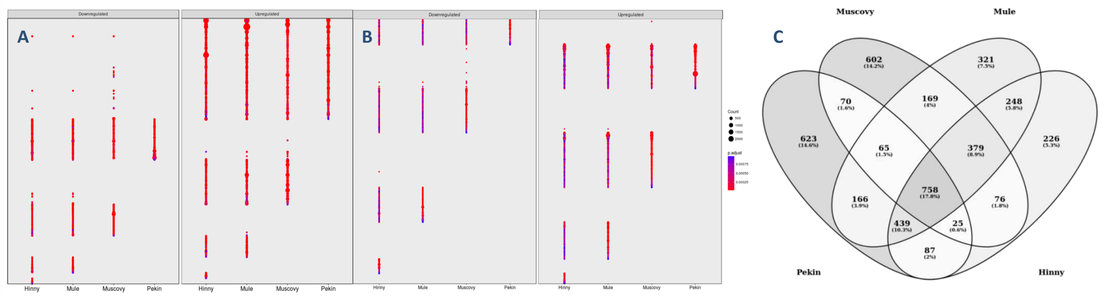
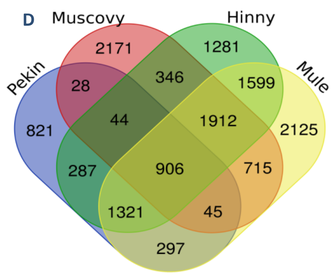


Figure 4 | Differentially expressed genes. Expression profiles (using clusterProfiler)[9] and Venn diagram[10] for reference genome approach (A and C) and for *de novo* assembly (B and D). A and B: left side down-regulated DEG by overfeeding, right side up-regulated DEG by overfeeding. Dot size corresponds to the count of FC. The p-values are shown in color bar, values increase from hot to cold color.



✓ *Globally more DEG of de novo transcriptome (see Tab. 2) but a similar expression profiles (Fig. 4a and Fig. 4b): A slight increase of DEG for Pekin (168%) and especially great increase for the other samples. ⇨ Reference genome approach seems pertinent with homologues species for DEG identification, but not with heterologs.*

✓ *More DEG in common of the 4 genotypes for de novo assembly (34% vs 24%)*

Comparison of Biological process GO terms occurrences between Cufflinks and DRAP DEG[11]

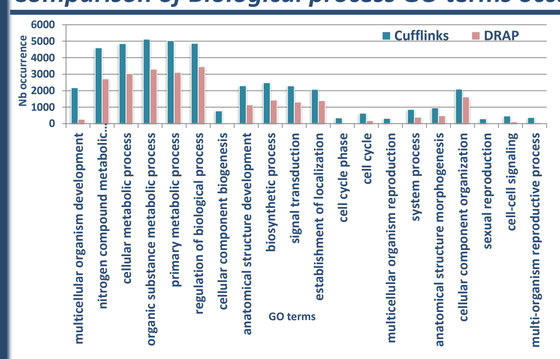


Figure 5 | comparison of the top 20 significant differences of Biological process (BP) GO terms .

Cufflinks 212 GO — **DRAP** 554 GO

⇩ χ^2 test, p-value<0.05

127GO terms have a significant difference of occurrences between Cufflinks and DRAP transcriptomes at p-value<0.05

Cufflinks specific (7 GO): cell cycle phase; multicellular organism reproduction; localization of cell; molting cycle; hair cycle phase; movement in environment of other organism involved in symbiotic interaction; menstrual cycle phase

DRAP specific (5 GO): execution phase of apoptosis; intermicrovillar adhesion; beta-amyloid clearance; dormancy process; innervation

Few method-specific BP GO terms

Significant BP GO terms occurrences using DRAP (annotated by Trinotate) are overestimated using Cufflinks. As there is more annotation with DRAP, our results suggest that there are more GO terms found overrepresented with DRAP, when excepted top 20 GO terms.

5. Functional annotation

✓ *Focus on the new species-specific DEG discovered by de novo approach and Go-enrichment analyses which may bring new functions.*

✓ *De novo approach bring us the opportunity to analyze genotype effect in addition to feeding effect. An ANOVA test will be realized to study the genotype and diet effect, also their interaction.*

6. Perspective

7. Bibliography

[1] Trapnell, Cole et al. Differential Gene and Transcript Expression Analysis of RNA-Seq Experiments with TopHat and Cufflinks: Nature Protocols 7.3 (2012): 562–578. PMC. Web. 26 June 2018.
[2] Bolger, A. M. et al. Trimmomatic: A flexible trimmer for Illumina Sequence Data. Bioinformatics, btu170 (2014).
[3] Grabherr, M. G. et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotech advance online publication, (2011).
[4] Schulz MH, Zerbino DR, Vingron M, Birney E. Oases: robust de novo RNA-seq assembly across the dynamic range of expression levels. Bioinformatics 2012 Apr 15;28(8):1086-92.
[5] Nicolas L Bray, Harold Pimentel, Páll Melsted and Lior Pachter, Near-optimal probabilistic RNA-seq quantification, Nature Biotechnology 34, 525–527 (2016), doi:10.1038/nbt.3519
[6] Cabau C, Escudié F, Djari A, Guiguen Y, Bobe J, Klopp C. (2017) Compacting and correcting Trinity and Oases RNA-Seq de novo assemblies. PeerJ 5:e2988.
[7] Felipe A. Simão, Robert M. Waterhouse, Panagiotis Ioannidis, Evgenia V. Kriventseva, Evgeny M. Zdobnov; BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs, Bioinformatics, Volume 31, Issue 19, 1 October 2015, Pages 3210–3212, https://doi.org/10.1093/bioinformatics/btv351
[8] Robinson MD, McCarthy DJ, Smyth GK (2010). "edgeR: a Bioconductor package for differential expression analysis of digital gene expression data." Bioinformatics, 26(1), 139-140.
[9] Yu G, Wang L, Han Y, He Q (2012). "clusterProfiler: an R package for comparing biological themes among gene clusters." OMICS: A Journal of Integrative Biology, 16(5), 284-287. doi: 10.1089/omi.2011.0118.
[10] http://bioinformatics.psb.ugent.be/webtools/Venn/
[11] https://github.com/anlebreton/GO_terms_FisherTest