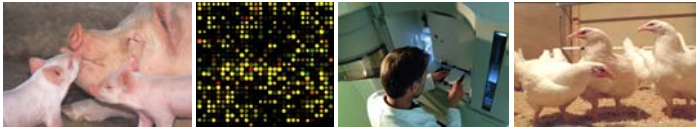




EADGENE and SABRE Post-Analyses Workshop

12-14th November 2008, Animal Sciences Group, Wageningen UR, Lelystad



Annotation results comparison



This publication represents the views of the Authors, not the EC. The EC is not liable for any use that may be made of the information.



The aims of the workshop

- Presenting the different annotation strategies and pipelines used by the teams collaborating in EADGENE.
- Comparing the pipelines on one or two data sets in order to present the impact on the annotation files provided to the biologists.
- Discussing the possible new needs, options and evolutions of these tools with other/new users.



EADGENE and SABRE Post-Analyses Workshop
 12-14th Nov 2008, Animal Sciences Group, Wageningen UR, Lelystad

2



Why should we compare?

- Several teams work on the annotation within work-package 1.3.
- They share the general ideas but work for different users which have different needs.
- They use different databases and different tools.
- The pipeline and individual results have been presented during yesterday's session.
- What about the correspondence between the results?
- What is the best solution?

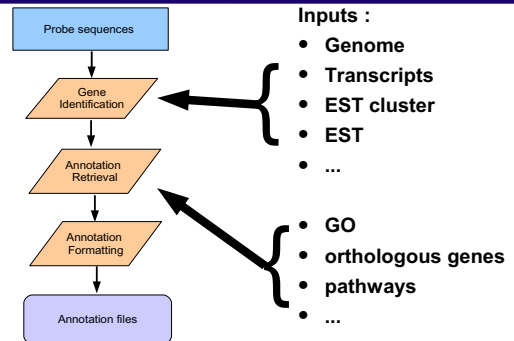


EADGENE and SABRE Post-Analyses Workshop
 12-14th Nov 2008, Animal Sciences Group, Wageningen UR, Lelystad

3



The probe re-annotation process



EADGENE and SABRE Post-Analyses Workshop
 12-14th Nov 2008, Animal Sciences Group, Wageningen UR, Lelystad

4



Some take home messages from yesterday

- Re-annotation is important because the knowledge of genome and the genes (transcripts) evolves.
- The criteria to link a probe to a gene are not fixed.
- Is possible get more links but the quality of the links will decrease (user decision).
- A criteria linked to the probability of expression could be added.



EADGENE and SABRE Post-Analyses Workshop
 12-14th Nov 2008, Animal Sciences Group, Wageningen UR, Lelystad

5



The comparison data-sets

- Chicken oligos :

Species	Origin	Nb oligos	Design year	Number of slides
Chicken	ARK-Genomics	20,460	2005	RI 445 / INRA 200
Cattle	BOMC / ARK Genomics	24,000	2003	RI 85
Pig	DIAS	25 210	2003	0

- Subset of the oligo-set : 791 oligos
- Up and down regulated from the "MM8_MM24.txt" file
- ~~Fig cDNA :~~
- ~~cDNA present in the "Supplemental_data_table1.xls" file,~~
- ~~Subset of 241 probes~~



EADGENE and SABRE Post-Analyses Workshop
 12-14th Nov 2008, Animal Sciences Group, Wageningen UR, Lelystad

6

The reference data

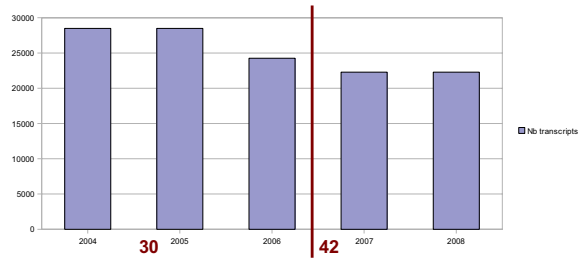
- Provided by Roslin
- Two columns linking oligo to Ensembl gene name
- New reference column => roslin gene

	A	B	C	D	E
1	Reporter_id	Reporter Name	Reporter BioSequence DatabaseEntry [ensembl]	Gene ID	
286	RIGG09418	ENSGALG00000015169	ENSGALG00000015169	ENSGALG00000015169	
287	RIGG09463	ENSGALG00000015551	ENSGALG00000015551	ENSGALG00000015551	
288	RIGG09566	ENSGALG00000016434	ENSGALG00000016434	...	
289	RIGG09584	ENSGALG00000016580	ENSGALG00000016580	...	
290	RIGG09586	ENSGALG00000016605	ENSGALG00000016605	ENSGALG00000016605	
291	RIGG09644	ENSGALG00000017038	ENSGALG00000017038	ENSGALG00000017038	
292	RIGG09662	ENSGALG00000017170	ENSGALG00000017170	ENSGALG00000017170	
293	RIGG09684	ENSGALG00000017281	ENSGALG00000017281	ENSGALG00000017281	

30 42

Evolution of the number of chicken transcripts

- Two genome builds
- Three gene builds

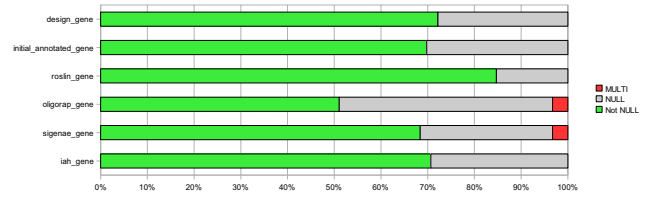


The files provided by different groups

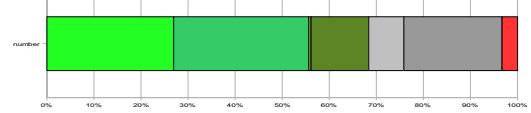
- WU
- IAH
- INRA

oligo_name	design_gene	initial_annotated_gene	roslin_gene	oligozap_gene	signeae_gene	iah_gene
RIGG09001	NULL	NULL	NULL	NULL	NULL	NULL
RIGG09018	ENSGALG00000016333	NULL	ENSGALG00000016333	ENSGALG00000016333	ENSGALG00000016333	ENSGALG00000016333
RIGG09028	NULL	ENSGALG00000006385	ENSGALG00000006385	ENSGALG00000006385	ENSGALG00000006385	ENSGALG00000006385
RIGG09063	ENSGALG00000005781	ENSGALG00000005781	ENSGALG00000005781	ENSGALG00000005781	ENSGALG00000005781	ENSGALG00000005781
RIGG09158	NULL	NULL	NULL	NULL	NULL	NULL
RIGG09252	ENSGALG0000002272	ENSGALG0000002272	ENSGALG0000002272	ENSGALG0000002272	ENSGALG0000002272	ENSGALG0000002272
RIGG09266	NULL	NULL	NULL	NULL	NULL	NULL
RIGG09278	NULL	ENSGALG00000020528	ENSGALG00000020528	ENSGALG00000020528	ENSGALG00000020528	ENSGALG00000020528
RIGG09286	NULL	ENSGALG00000012412	ENSGALG00000012412	ENSGALG00000012412	ENSGALG00000012412	ENSGALG00000012412
RIGG09293	ENSGALG00000018023	ENSGALG00000018023	ENSGALG00000018023	ENSGALG00000018023	ENSGALG00000018023	ENSGALG00000018023
RIGG09279	NULL	ENSGALG00000032790	ENSGALG00000032790	ENSGALG00000032790	ENSGALG00000032790	ENSGALG00000032790
RIGG09300	NULL	ENSGALG00000018409	ENSGALG00000018409	ENSGALG00000018409	ENSGALG00000018409	ENSGALG00000018409
RIGG09383	NULL	ENSGALG00000000346	ENSGALG00000000346	ENSGALG00000000346	ENSGALG00000000346	ENSGALG00000000346
RIGG09385	NULL	ENSGALG00000000611	ENSGALG00000000611	ENSGALG00000000611	ENSGALG00000000611	ENSGALG00000000611
RIGG09412	NULL	ENSGALG00000006524	ENSGALG00000006524	NULL	ENSGALG00000006524	ENSGALG00000006524

Annotated oligos

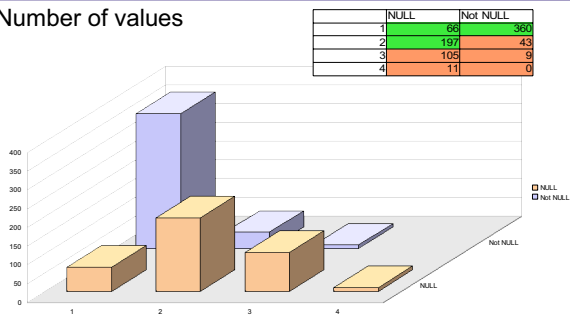


Signeae : more information about the quality of the link



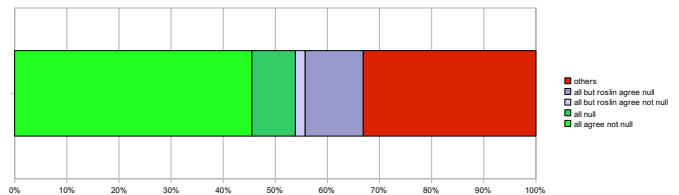
Different cases

- Number of values



Comparison to the reference

- How many of the pipeline produced annotations are corresponding to the reference?
- How many of the pipeline produced annotations are homogeneous and different from the reference?



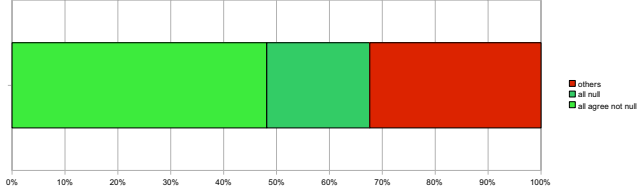
Pipeline agreement versus initial annotation

oligo_name1	roslin_gene	oligorap_gene	sigenae_gene	iah_gene
RI6600731	ENSGAL00000000744	NULL	NULL	NULL
RI6601135	ENSGAL000000009544	NULL	NULL	NULL
RI6601383	ENSGAL000000005040	NULL	NULL	NULL
RI6601390	ENSGAL000000122551	NULL	NULL	NULL
RI6601513	ENSGAL000000012233	NULL	NULL	NULL
RI6601569	ENSGAL000000007154	NULL	NULL	NULL
RI6601623	ENSGAL000000124007	NULL	NULL	NULL
RI6602264	ENSGAL000000017008	NULL	NULL	NULL
RI6602418	ENSGAL00000015544	NULL	NULL	NULL
RI6603205	ENSGAL000000011221	NULL	NULL	NULL
RI6603743	ENSGAL000000013920	NULL	NULL	NULL

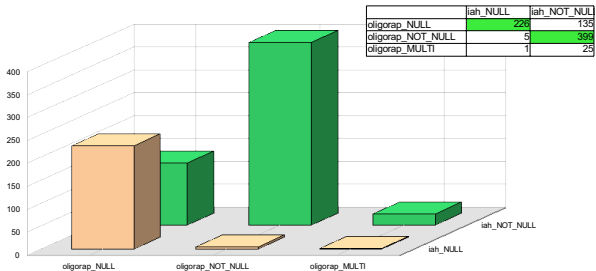
oligo_name1	roslin_gene	oligorap_gene	sigenae_gene	iah_gene
RI6618189	ENSGAL000000014728	NULL	ENSGAL00000002778	ENSGAL00000002778
RI6618405	ENSGAL00000015214	NULL	ENSGAL00000002971	ENSGAL00000002971
RI6618406	ENSGAL00000015217	NULL	ENSGAL00000012106	ENSGAL00000012106
RI6618730	ENSGAL00000015752	NULL	ENSGAL000000090740	ENSGAL000000090740
RI6619513	ENSGAL00000017060	NULL	ENSGAL00000016431	ENSGAL00000016431
RI6619628	ENSGAL00000017257	NULL	ENSGAL0000002518	ENSGAL0000002518
RI6619753	ENSGAL00000024059	NULL	ENSGAL00000023973	ENSGAL00000023973
RI662024	ENSGAL00000011434	NULL	ENSGAL00000007930	ENSGAL00000007930
RI664196	ENSGAL00000007916	ENSGAL00000009056	ENSGAL00000012616	ENSGAL00000012616
RI664307	ENSGAL00000009056	ENSGAL00000009054	ENSGAL00000021211	ENSGAL00000021211
RI664852	ENSGAL00000012599	ENSGAL00000012616	ENSGAL00000021211	ENSGAL00000021211
RI664859	ENSGAL00000012566	ENSGAL00000012616	ENSGAL00000021211	ENSGAL00000021211
RI664858	ENSGAL00000014577	ENSGAL00000021211	ENSGAL00000021211	ENSGAL00000021211
RI664919	ENSGAL00000014057	ENSGAL00000022870	ENSGAL00000022870	ENSGAL00000022870
RI664979	ENSGAL00000017124	ENSGAL00000006230	ENSGAL00000006230	ENSGAL00000006230
RI6620451	ENSGAL00000019971	ENSGAL00000006028	ENSGAL00000006028	ENSGAL00000006028

Comparison between the pipelines

- How many of the pipeline produced annotations are equal?

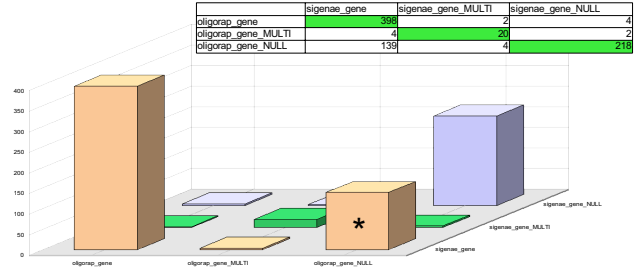


OligoRAP versus IAH



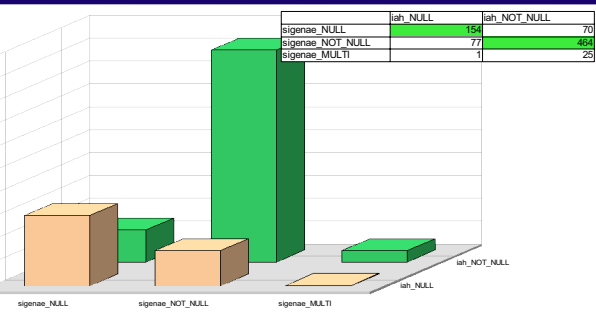
- 387 genes out of 399 are corresponding

OligoRAP versus SIGENAE



- 397 genes out of 398 are corresponding
- 68% of * are from category 3 or 4.

IAH versus SIGENAE



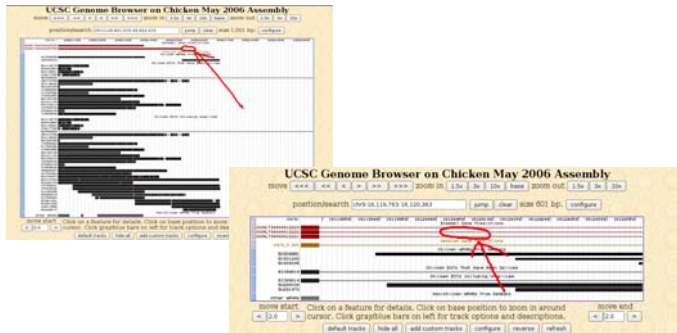
- 420 genes out of 464 are corresponding

Case analysis

- It is work in progress.
- We have produced lists of cases to analyze
- The first analysis show :
 - that most of the differences are linked to the use or not of loose links,
 - that some differences come from the thresholds used,
 - that some differences come from the use of the surrounding information (Unigene UTR links).

probe_id_key	seq_id	start	end	strand	maxBlock	percent_id	gene	seq_is_chr	category	origin
936	ENSGAL70000003358	1407	1422	1	16	22.2222	ENSGAL00000002143	0	5	exon
936	ENSGAL70000017109	349	406	1	52	72.2222	ENSGAL00000001051	0	5	exon

Case analysis 2



Conclusions

- ✦ All annotation pipelines are not equivalent.
- ✦ The key element is the link between the oligo and the gene.
- ✦ If we want to add links the quality of the link has to be taken into account by the biologists (if not it can lead to a wrong paths)
- ✦ Adding information about the compliance of the oligo on technical criteria (single strand folding,...).
- ✦ Using array experiment results to get more information on annotation (expression rate, RT PCR results,...)
- ✦ Using new sequencing technologies instead of micro-arrays

Acknowledgments

EADGENE Management team :

- ✦ Caroline Channing
- ✦ Sandrine Ayuso
- ✦ Thu Bizat

Wageningen University :

- ✦ Pieter Neerincx
- ✦ Haisheng Nie
- ✦ Martien Groenen
- ✦ Jack Leunissen

IAH :

- ✦ Dennis Prickett
- ✦ Michael Watson

INRA :

- ✦ Pierrot Casel
- ✦ Sandrine Laguarrigue
- ✦ Frederic Lecerf
- ✦ Li Jiang
- ✦ Gwénola Tosser
- ✦ Yannick Faulconnier