THETA
Statistical Genetics Group
*Institute of Animal Genetics*
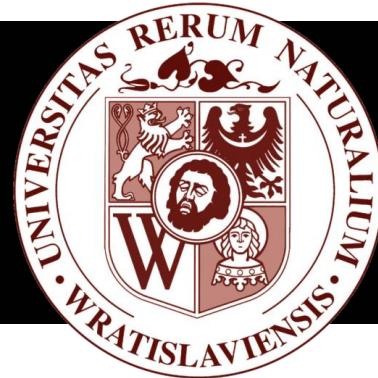
Impact of copy number polymorphisms on gene expression

M. Mielczarek, M. Frąszczak, J.Szyda

Background & Objectives

- Genome-wide CNVs detection
- **CNV impact on gene expression**

source: https://illustrated-glossary.nejm.org
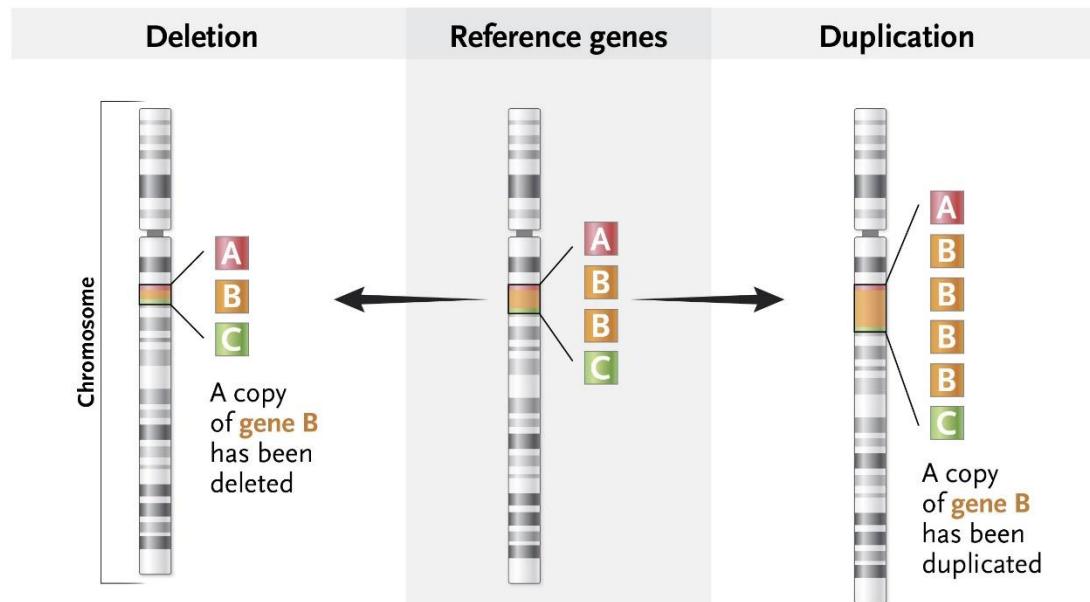
Dataset

Accession → NCBI
BioProject PRJNA354435
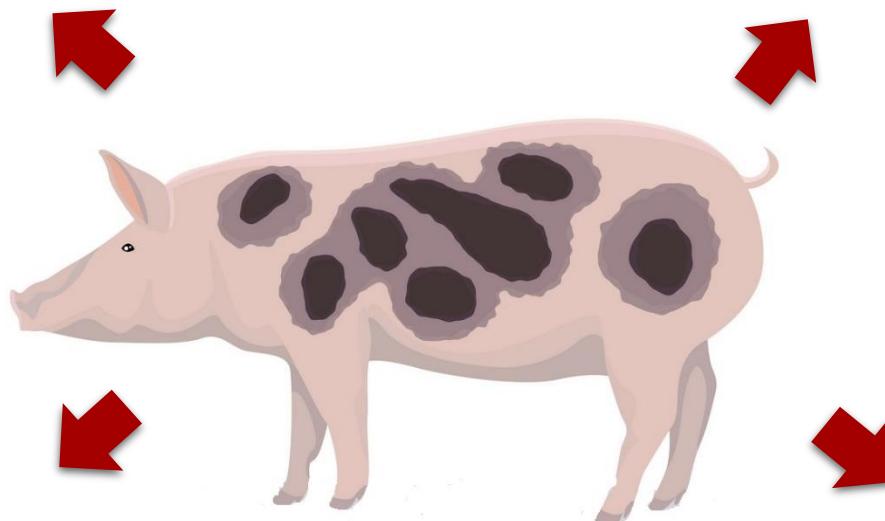
one Duroc x Pietrain
crosbreed pig ♀

DNA-seq:
Illumina HiSeq 2500
150×2 PE
average coverage: 15

RNA-seq:
Illumina HiSeq 2500
150×2 PE
fat: 2 x 93 725 452
liver: 2 x 78 943 463
muscle: 2 x 85 905 149

Methods

1. CNVs detection (DNA)

2. Transcript quantification (RNA)

3. CNV impact on gene expression investigation

Methods

1. CNVs detection (DNA)

quality control & data filtering	**FASTQC & Trimm.**
↓	
alignment to Sscrofa11.1	**BWA**
↓	
quality control & data processing after alignment	**Picard & Samtools**
↓	
CNV detection & validation	**CNVnator & Pindel**
↓	
CNV annotation	**VEP**

2. Transcript quantification (RNA)

quality control & data filtering	**FASTQC & Trimm.**
↓	
transcript quantification & normalization	**kallisto**

3. CNV impact on gene expression investigation

|	
Permutation tests (Wilcoxon test) & Correlation	**R**

Results

CNVs characterization → **the numer of CNVs**

deletions = 420

transcripts · intergenic

148 (35%)
272 (65%)

duplications = 304

transcripts · intergenic

148 (49%)
156 (51%)
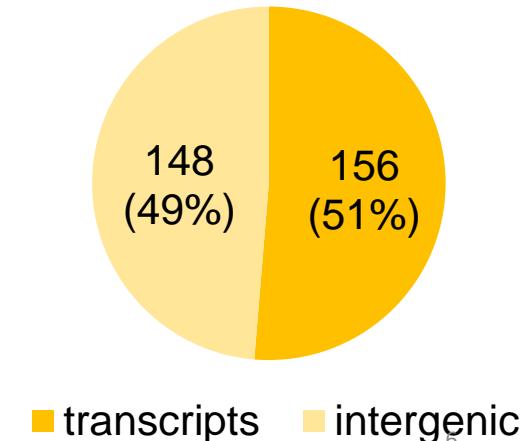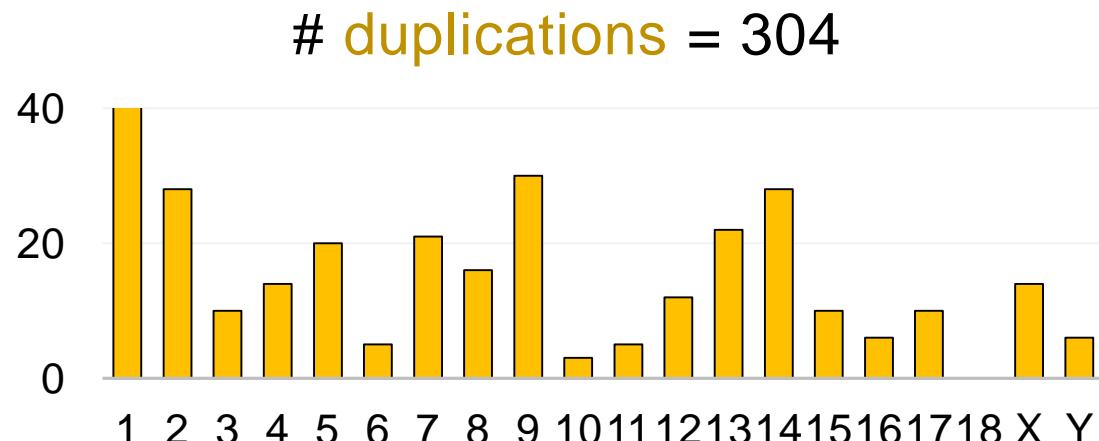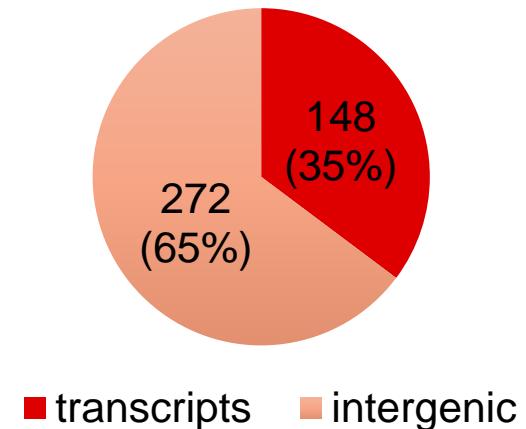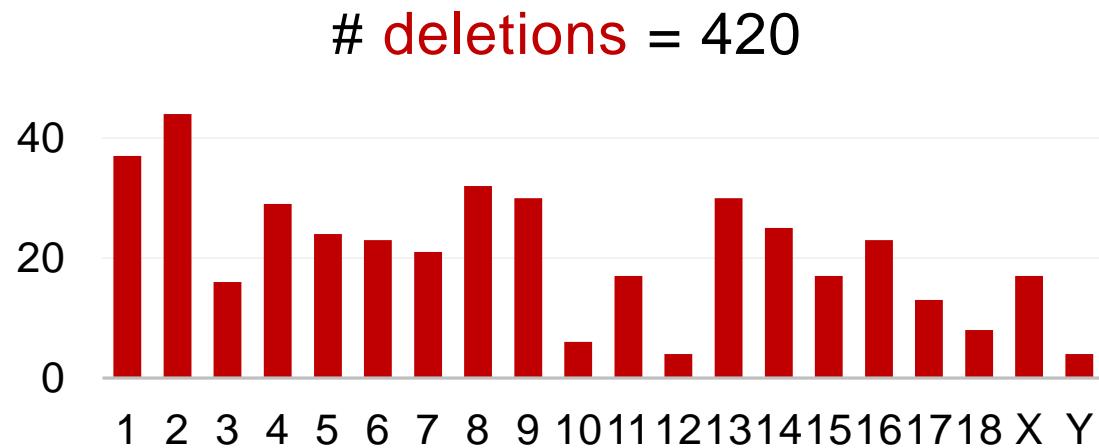
Results

CNVs characterization → **CNVs length (bp)**

deletions
min:	500
max:	133 700
mean:	7 047
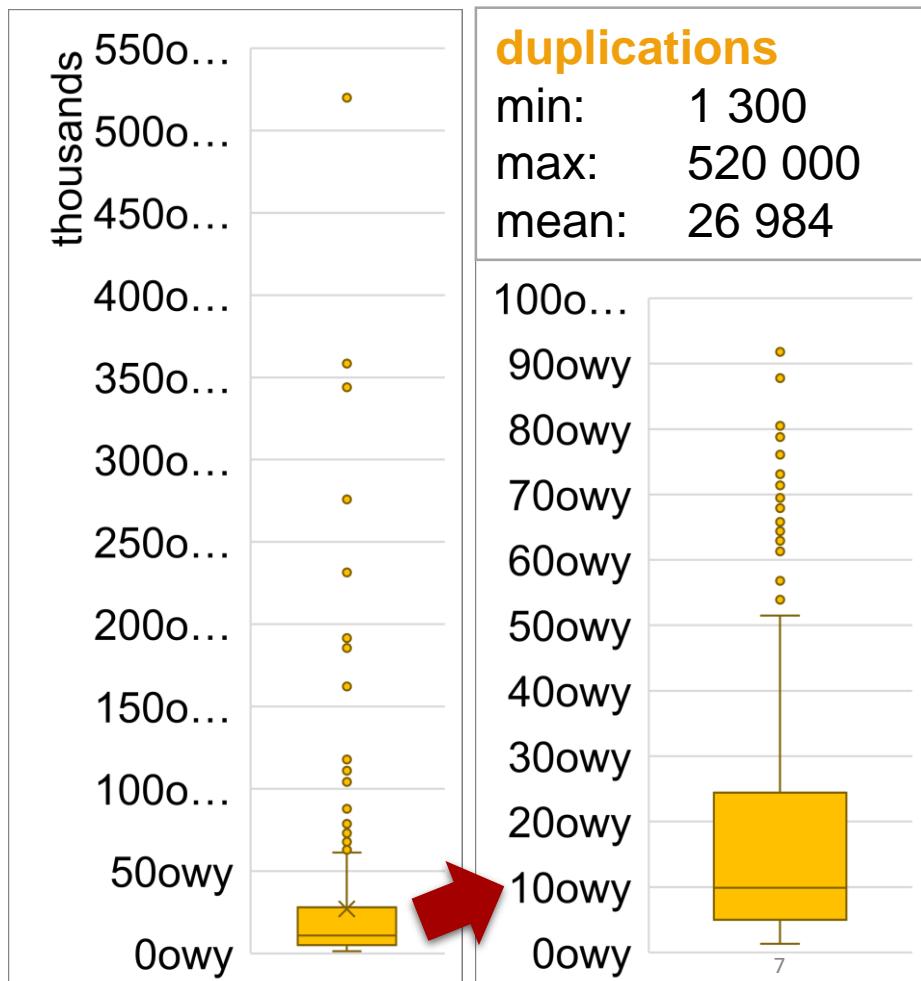
duplications
min:	1 300
max:	520 000
mean:	26 984
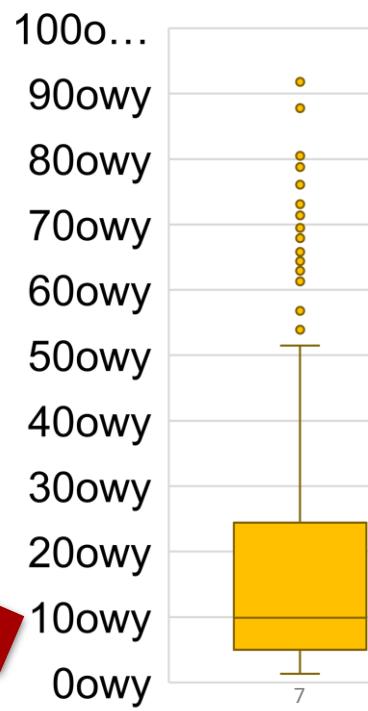
Results

CNVs characterization → **CNVs functional annotation**

Deletions

GO	KEGG	QTL
-	-	-

Duplications

GO (+/-)	KEGG (+)	QTL
● sensory perception of smell (+) ● G-protein coupled receptor (+) ● cellular metabolic processes (-) ● regulation of metabolic processes (-)	● pathways of lipid metabolism class ● olfactory transduction ● inflammatory mediator regulation ● carcinogenesis	● production traits and the meat quality ● skeletal system and bone structure ● immune system ● …and others

Results

CNVs impact → **CNVs in transcripts**

Permutation Test

H_0: $E_{CNV} \geq E_{random}$ E_{CNV} - expression of a transcript with CNV

H_1: $E_{CNV} < E_{random}$ E_{random} - expression of a random transcript

	Deletions	Duplications
fat	p=0.0008	p<0.0001
muscle	p=0.0010	p<0.0001
liver	p=0.1418	p=0.5908

Results

CNVs impact →

size of a deleted transcript

liver — $r = 4.2654 \cdot 10^{-5}$, $p = 0.4956$

fat — $r = 0.0001$, $p = 0.4866$

muscle — $r = -0.002$, $p = 0.6937$

expression vs % of deleted transcript
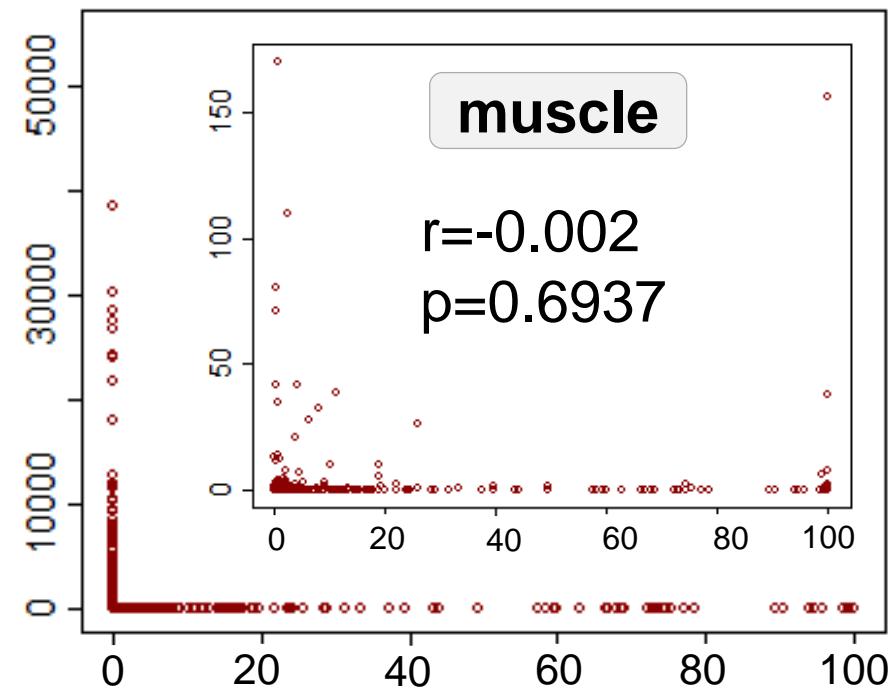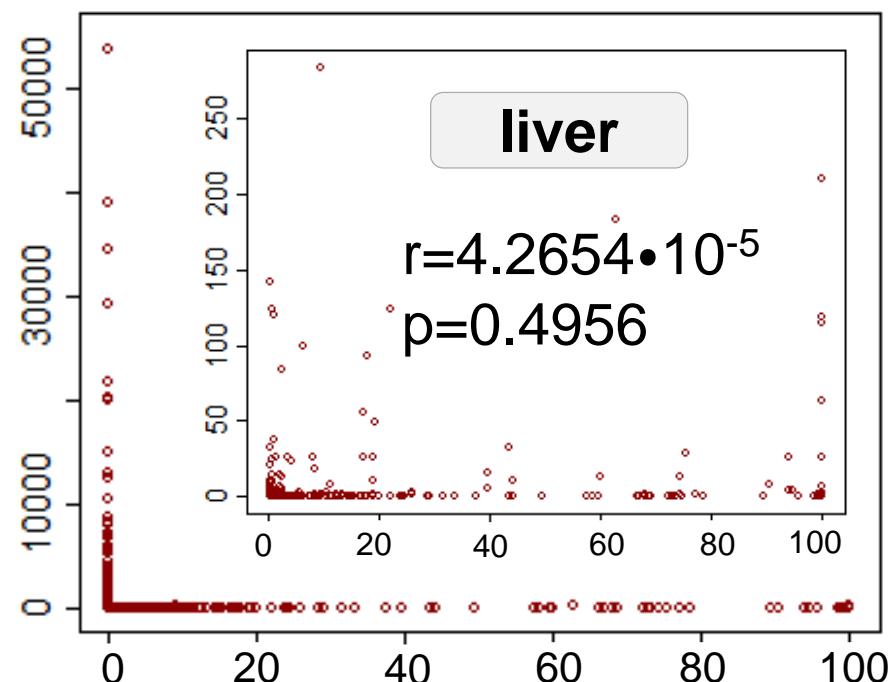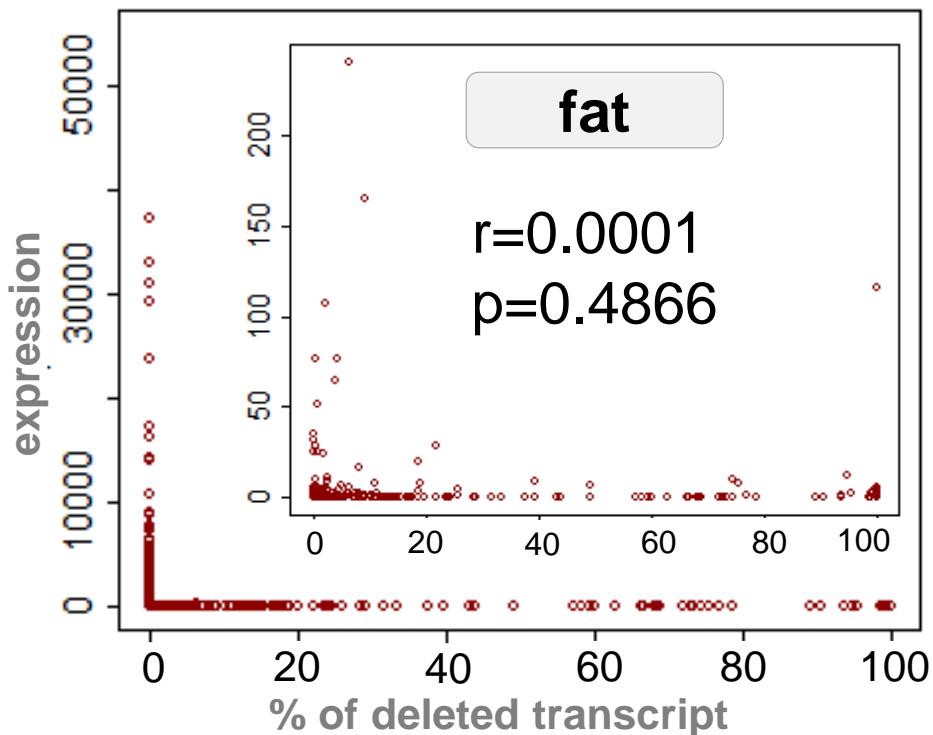
Results

CNVs impact →

size of duplicated transcript

liver

r=-0.0024
p=0.6894

fat

r=0.0055
p=0.1296

muscle

r=0.0054
p=0.1347

expression

% of duplicated transcript
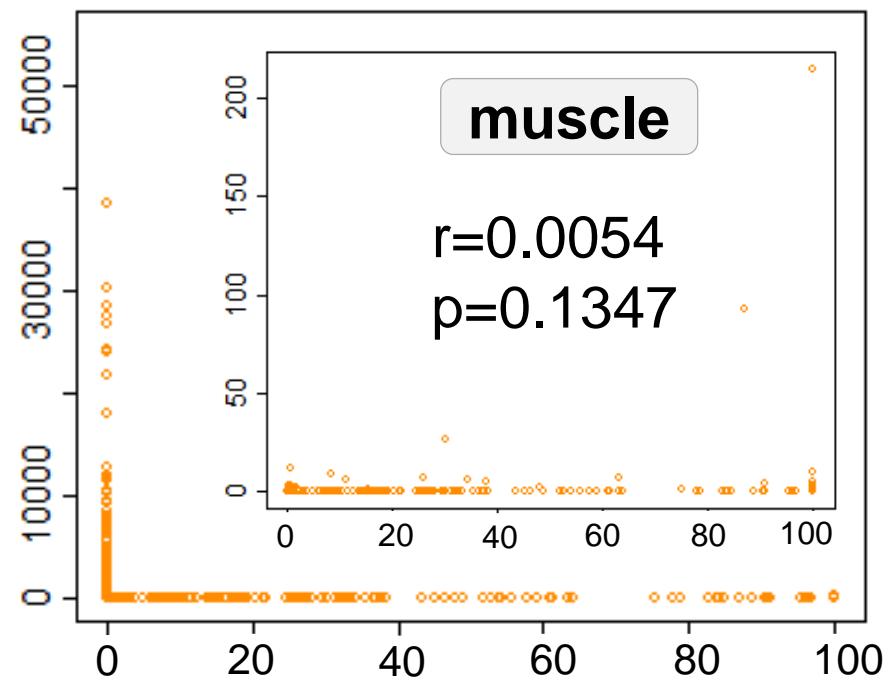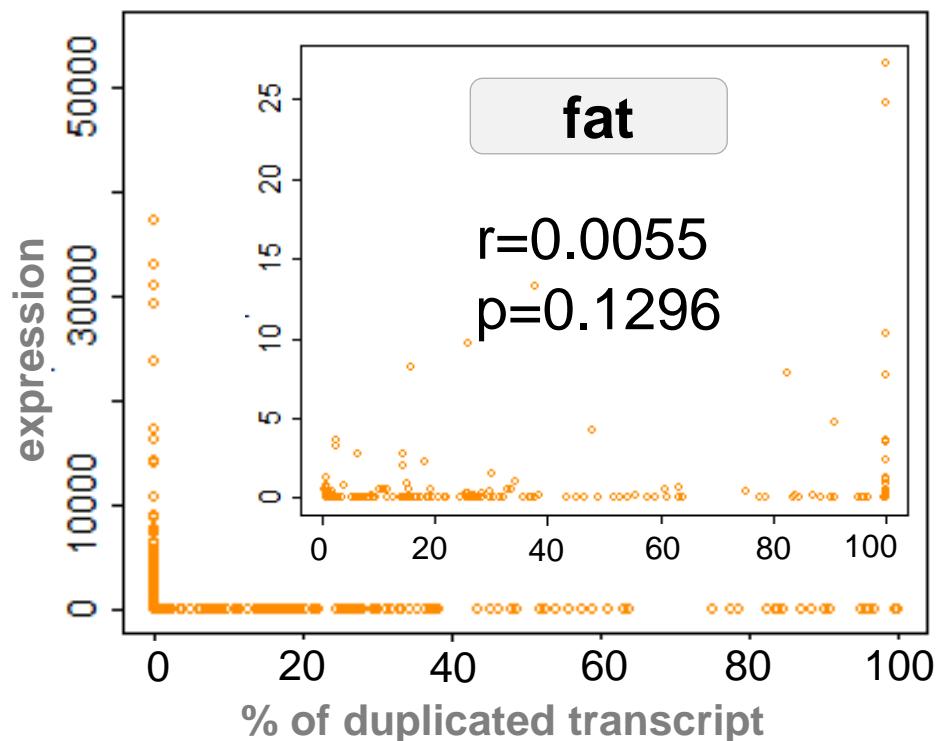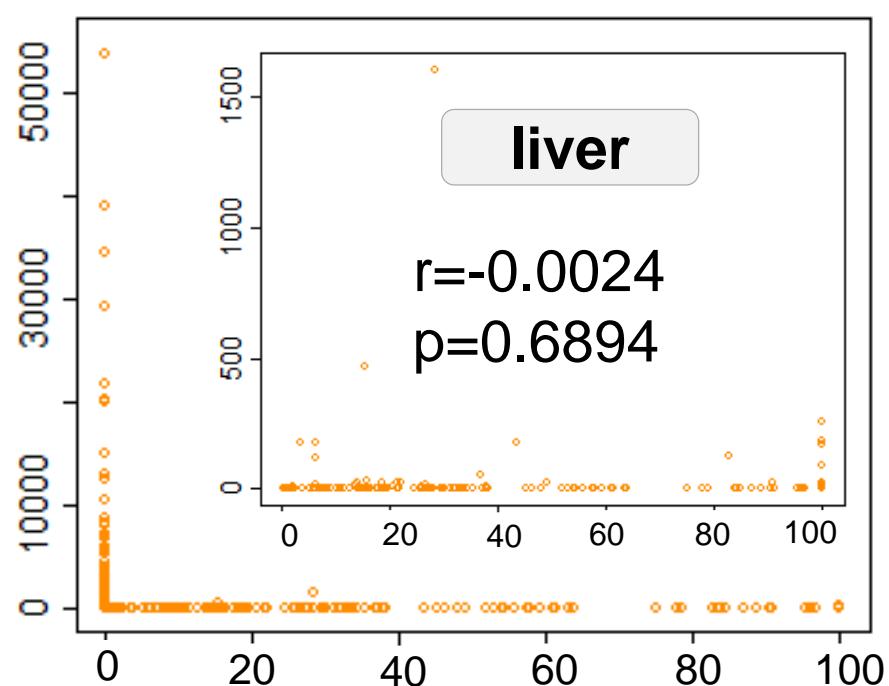
Summary & conclusions

- Genome: more deletions than duplications
                        deletions shorter than duplications

- Transcriptome: less deletions than duplications,
                        deletions located in „neutral" regions
                        „important" regions duplicated

- Deletions and duplications → lower expression in fat and muscle

- No correlation between the size of deletions/duplications and expression level

Future work

- DNA-seq and RNA-seq of 20 male Polish Large White pigs
- Matched by age, sex, breed and environmental factors (diet, housing, slaughter conditions)
- Additionall CNVs validation by PCR
- Statistical analysis to provide population-wide inferences

Acknowledgements

Leading National Research
Centre in Poland

Poznan Supercomputing
and Networking Center

Polish National
Science Centre

…thank YOU!

CNVs impact → **occurrence of CNVs in transcripts**

Permutation Test
- H_0: $E_{CNV} \geq E_{random}$ E_{CNV} - expression of a transcript with CNV
- H_1: $E_{CNV} < E_{random}$ E_{random}- expression of a random transcript

$$p = \frac{1}{k} \sum_{i=1}^{k} \mathbb{I}(W_{h_i}(Z) \geq W(X,Y))$$

where **k**=50,000 denoted the total number of permutations in the vector containing expression of all transcripts, **W** denoted the Wilcoxon rank sum test statistics, **X** denoted the vector of expression values for transcripts overlapping with del/dup, **Y** denoted the vector of expression values for transcripts which not contain del/dup, **Z** represented the vector of all investigated transcript expression values (the combined samples X and Y) and **h** represented permuted vector of Z.