# Structural differences among pig genomes illustrate genetic uniqueness of breeds

Zorc, M., Ogorevc, J., Škrlep, M., Bozzi, R., Petig, M., Fontanesi, L., Ovilo, C.,
Radović, Č., Kušec, G., Čandek-Potokar, M., **Dovč, P.**

University
of Ljubljana
Biotechnical
faculty

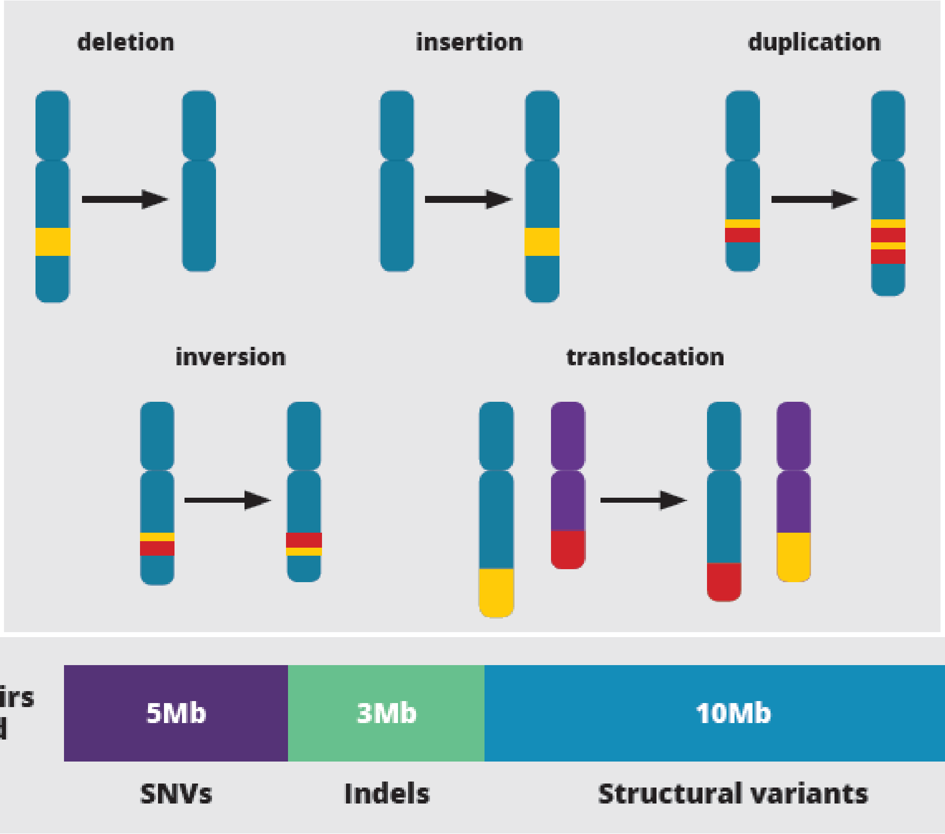
TREASURE

EAAP / 69th ANNUAL MEETING
Dubrovnik, Croatia
27th to 31st August 2018
Conventional and traditional livestock production systems – new challenges

# Outline

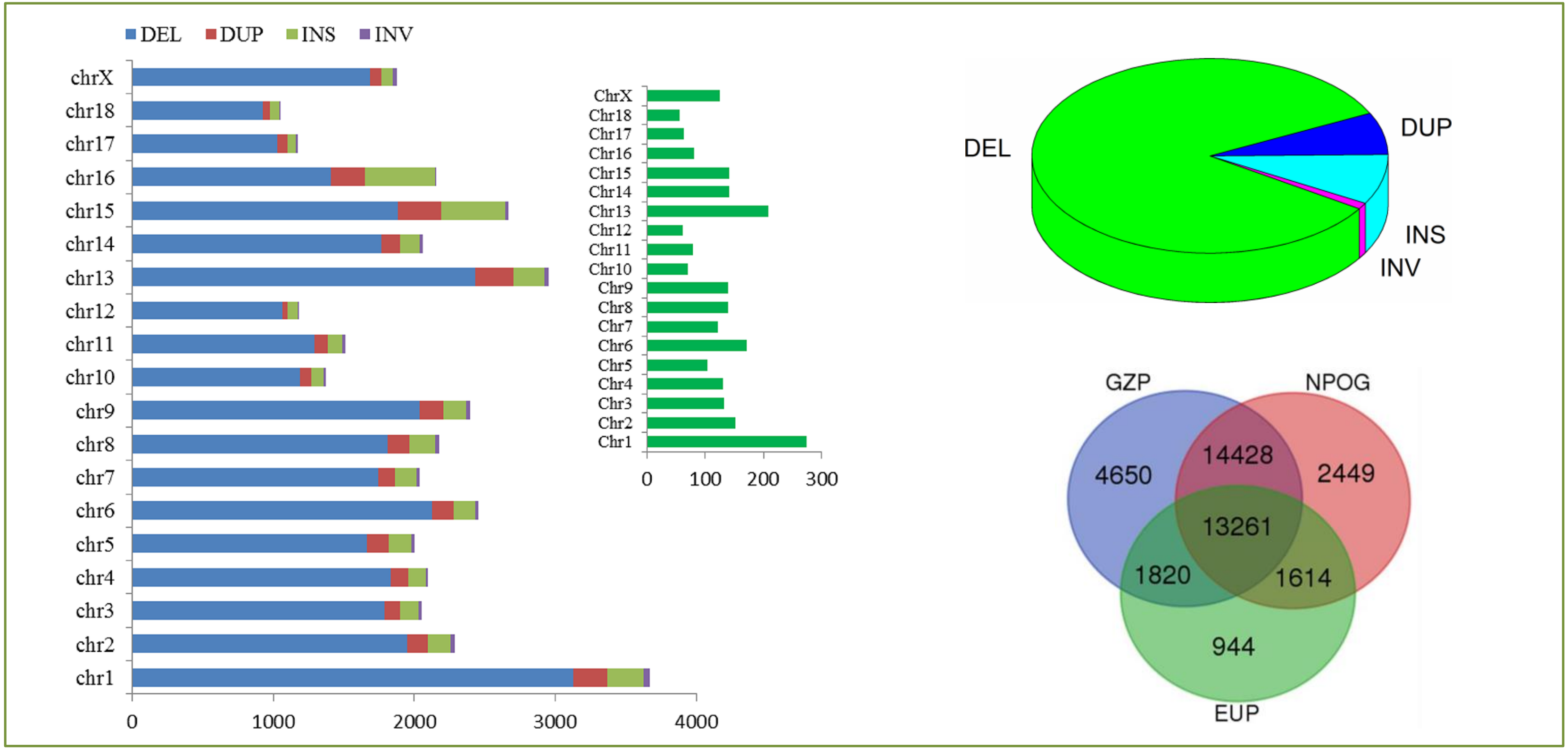
- Background
- Structural variants in pig
- Methodology
- Survey of structural variants in analysed breeds
- Some examples…
- Conclusions

# Relevance of SVs for human genetics



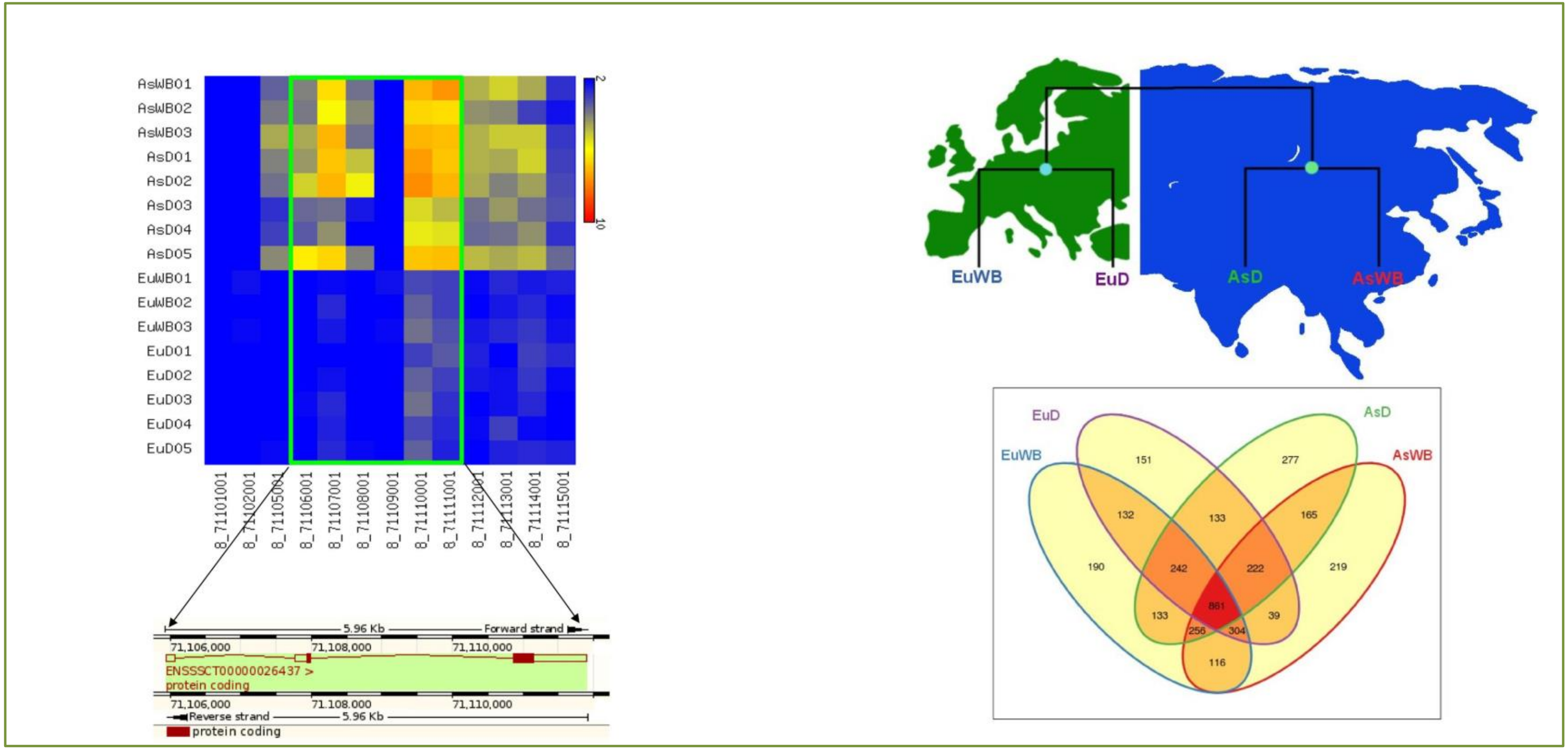
*Hickey L., 2017*

# Structural genome variation in pig



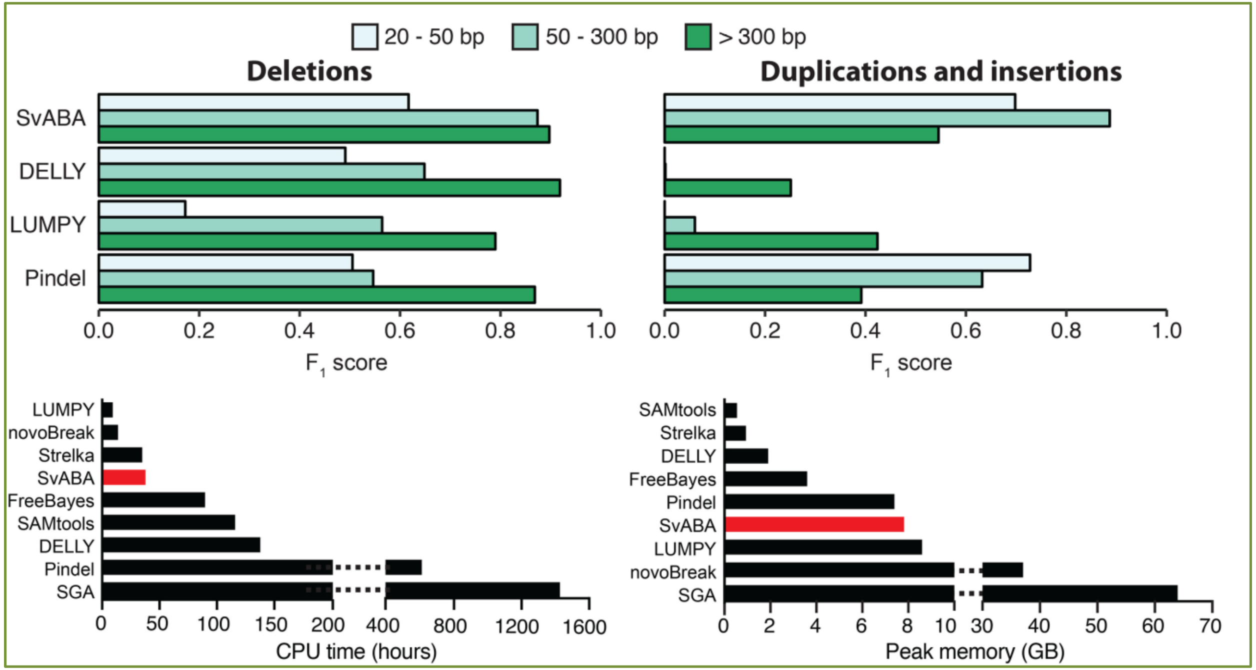
Chang Liu *et al.*, PLoSONE, 2018

# Evolutionary evidence from SVs in pig



Paudel *et al.*, BMC Genomics, 2013

# SV detection using NGS



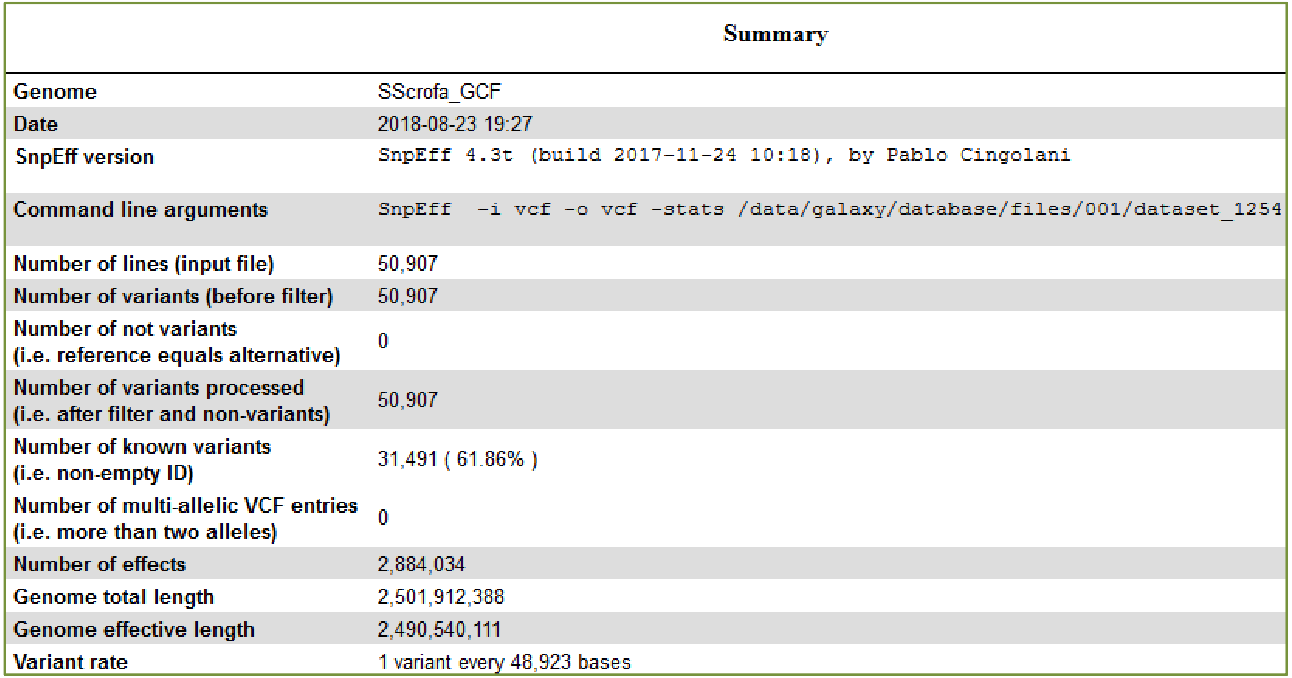
Wala *et. al*., Genome Research, 2018

# Material and methods

| Sequences | Quality control and mapping | SV calling | Functional annotation |
|---|---|---|---|
| • European Nucleotide Archive<br>• hi-seq, 150 bp | • FASTQC<br>• BWA-MEM<br>• Picard | • Lumpy | • SnpEff<br>• Ensembl-VEP<br>• STRING<br>• KOBAS 3.0 |
| • TREASURE selection of European pig breeds (20 genomes, coverage ~40)<br>• Publicly available WGS (30 genomes, coverage ~16) | • Reference genome: SScrofa 11.1 (NCBI) | • Number of reads to compute mean and stdev of read length: 10000 | • Sequence effect ontology<br>• KEGG pathways<br>• GO ontology |

# Summary of Lumpy tool

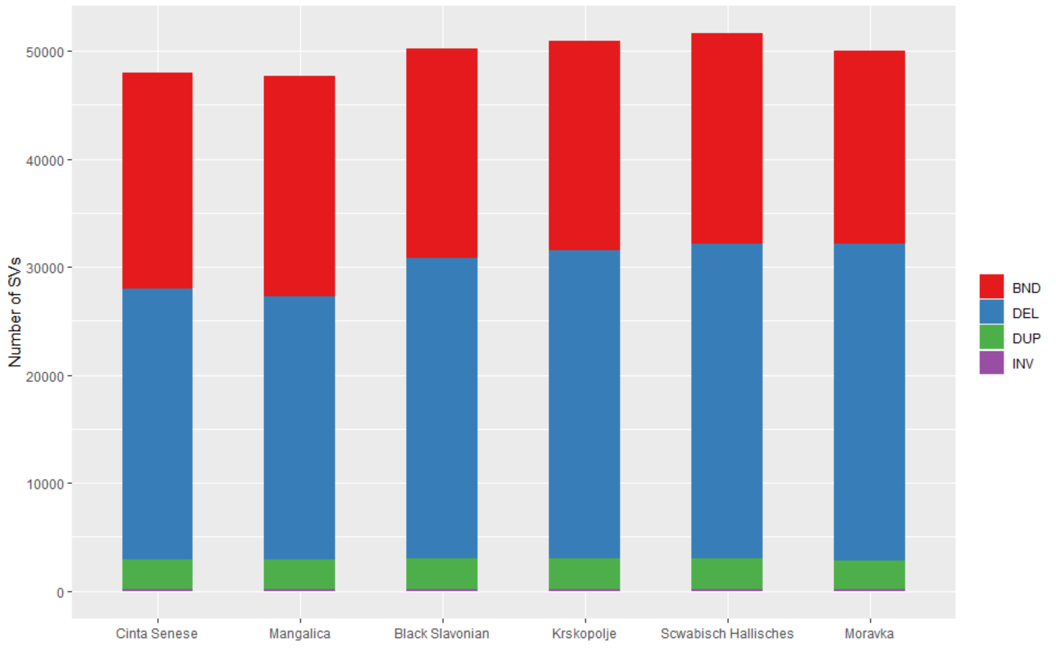
**SV calling**



| Summary | |
|---|---|
| Genome | SScrofa_GCF |
| Date | 2018-08-23 19:27 |
| SnpEff version | `SnpEff 4.3t (build 2017-11-24 10:18), by Pablo Cingolani` |
| Command line arguments | `SnpEff  -i vcf -o vcf -stats /data/galaxy/database/files/001/dataset_1254` |
| Number of lines (input file) | 50,907 |
| Number of variants (before filter) | 50,907 |
| Number of not variants (i.e. reference equals alternative) | 0 |
| Number of variants processed (i.e. after filter and non-variants) | 50,907 |
| Number of known variants (i.e. non-empty ID) | 31,491 ( 61.86% ) |
| Number of multi-allelic VCF entries (i.e. more than two alleles) | 0 |
| Number of effects | 2,884,034 |
| Genome total length | 2,501,912,388 |
| Genome effective length | 2,490,540,111 |
| Variant rate | 1 variant every 48,923 bases |

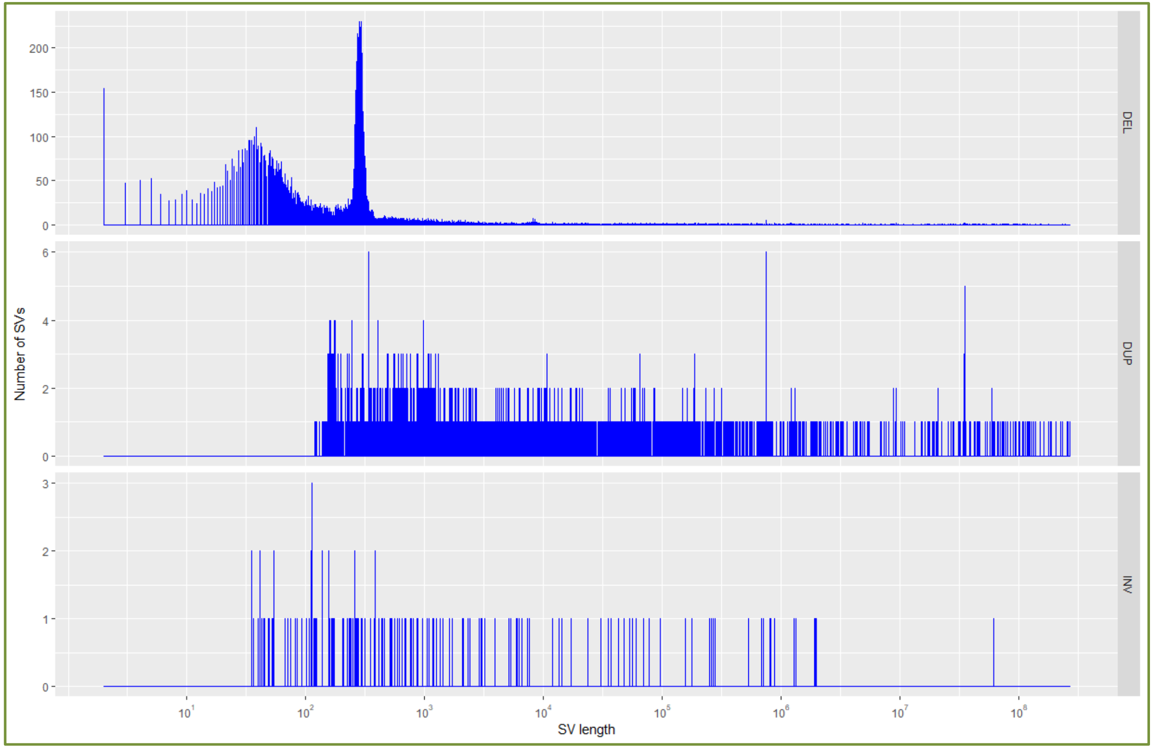
# SV types detected by Lumpy

**SV calling**



# Ranges of SVs detected by Lumpy

**SV calling**



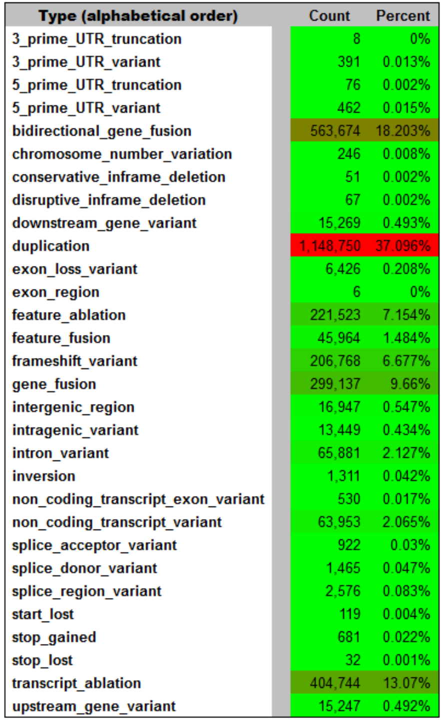
Krškopolje

EAAP, Dubrovnik, 30. Aug. 2018

# Possible consequences of SV

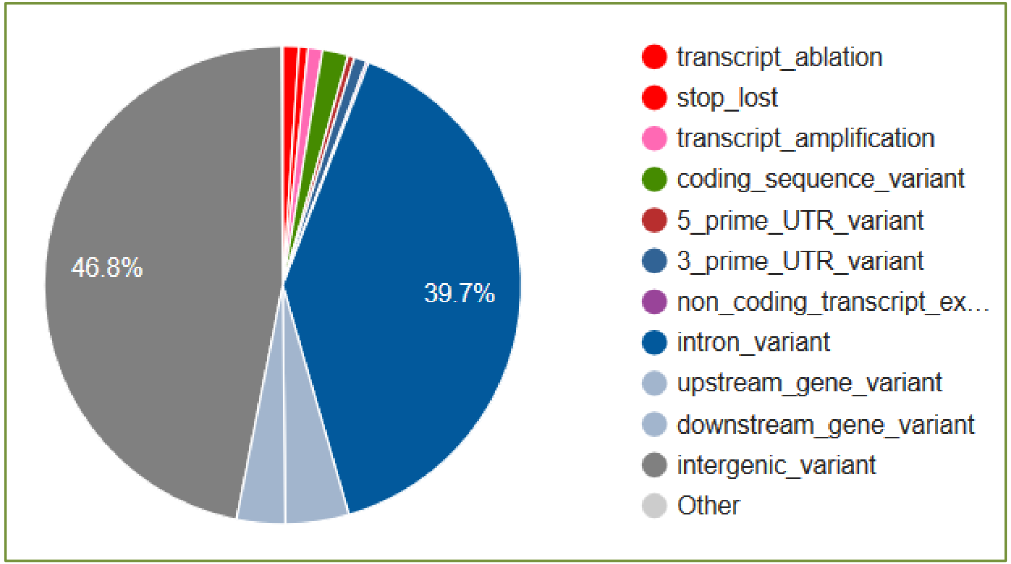
## SV annotation

SnpEff



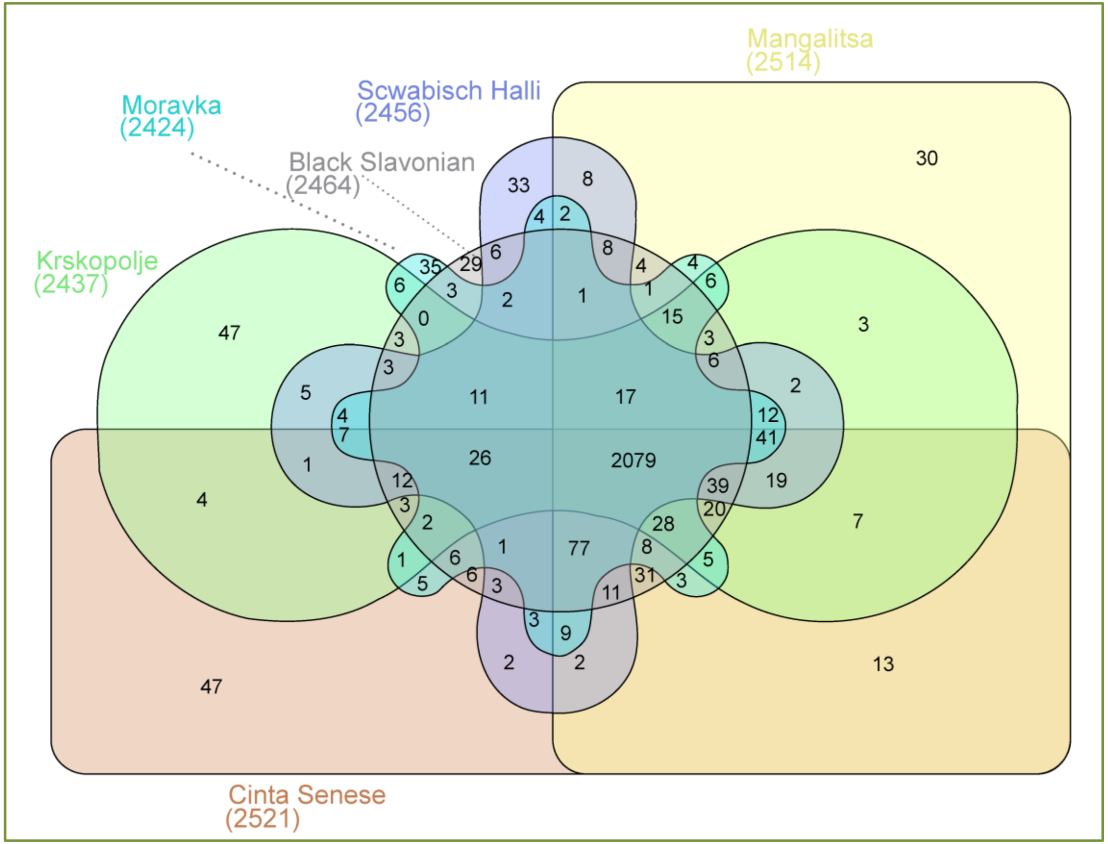
| Type (alphabetical order) | Count | Percent |
|---|---|---|
| 3_prime_UTR_truncation | 8 | 0% |
| 3_prime_UTR_variant | 391 | 0.013% |
| 5_prime_UTR_truncation | 76 | 0.002% |
| 5_prime_UTR_variant | 462 | 0.015% |
| bidirectional_gene_fusion | 563,674 | 18.203% |
| chromosome_number_variation | 246 | 0.008% |
| conservative_inframe_deletion | 51 | 0.002% |
| disruptive_inframe_deletion | 67 | 0.002% |
| downstream_gene_variant | 15,269 | 0.493% |
| duplication | 1,148,750 | 37.096% |
| exon_loss_variant | 6,426 | 0.208% |
| exon_region | 6 | 0% |
| feature_ablation | 221,523 | 7.154% |
| feature_fusion | 45,964 | 1.484% |
| frameshift_variant | 206,768 | 6.677% |
| gene_fusion | 299,137 | 9.66% |
| intergenic_region | 16,947 | 0.547% |
| intragenic_variant | 13,449 | 0.434% |
| intron_variant | 65,881 | 2.127% |
| inversion | 1,311 | 0.042% |
| non_coding_transcript_exon_variant | 530 | 0.017% |
| non_coding_transcript_variant | 63,953 | 2.065% |
| splice_acceptor_variant | 922 | 0.03% |
| splice_donor_variant | 1,465 | 0.047% |
| splice_region_variant | 2,576 | 0.083% |
| start_lost | 119 | 0.004% |
| stop_gained | 681 | 0.022% |
| stop_lost | 32 | 0.001% |
| transcript_ablation | 404,744 | 13.07% |
| upstream_gene_variant | 15,247 | 0.492% |

Ensembl VEP



# Genes affected by SVs

**SV annotation**



# Genes affected by SVs

**SV annotation**

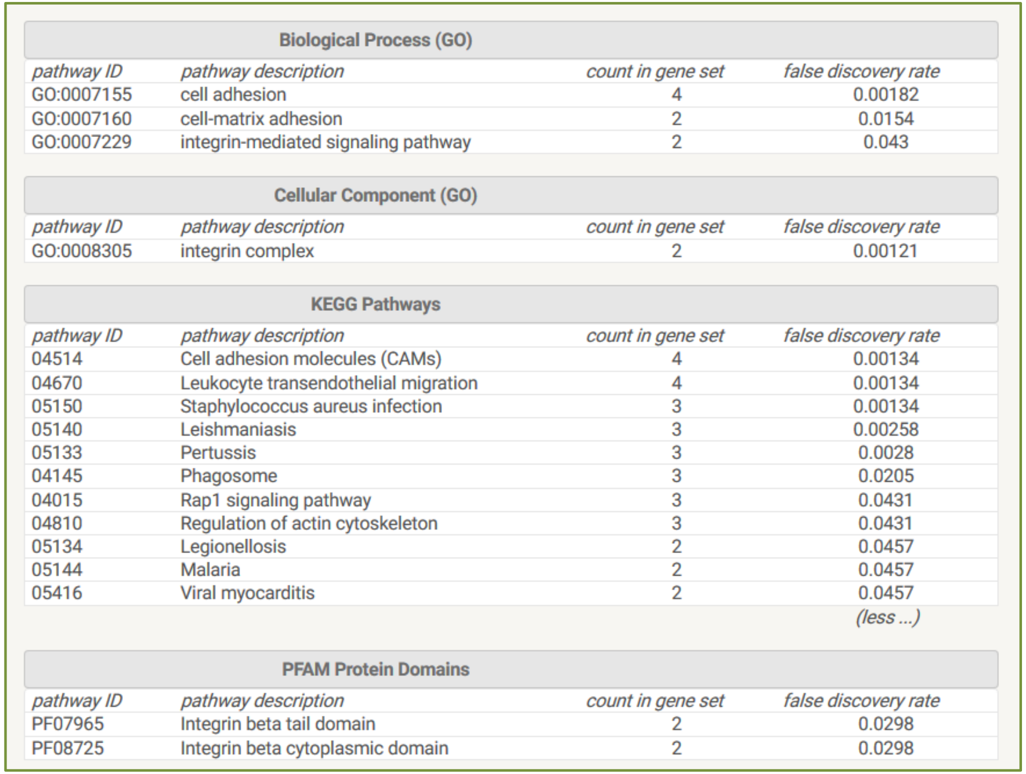
| | |
|---|---|
| Cinta Senese | *AK7, ALOX5AP, ARHGEF9, ARMC4, ATP10B, CCDC40, CTNND2, DDX55, EEA1, ELOVL2, FOXJ3, GPC6, GRIA3, ITGA2, LMO1, LTBP2, MEGF11, MGAT4C, MOCS2, MOXD1, NEBL, NSRP1, NUDCD3, ODF2L, PKHD1, PLCB1, PYGO1, RAPGEF6, SASH1, TMEM63A, ZBTB20, ZNF484* |
| Mangalitsa | *ABLIM1, ANO6, ATXN7L1, AUTS2, BIN3, COLEC10, DOK6, EPHA6, ERC2, FAM155A, FSTL5, KIAA1468, NKAIN2, SARM1,* ***TCF7L2****, TMF1, USP33* |
| Black Slavonian | *ABHD3, BBX, CDH8, CFAP70, DGKA, DOCK7, GHR, GLRA2, LRCH3, MDGA2, MGAT4D, PRIMPOL, SPRY1, TP63, TTC4, WDR93* |
| Krškopolje | *AKAP6, ARMC3, BEND7, CPEB3, CSMD3, DKK3, EHBP1, FHIT, ITGAM, KIAA0368, LRRIQ3, NDRG4, OPCML, PON3, PSD3, RTL4, SETX, SORBS2, SORCS2, STARD10, TBCK, TNC, TRPM8* |
| Scwabisch Hallisches | *LAGO4, ARID1B, C10H9orf3, COL23A1, CPED1, GGTA1P, GRIN2A, KLHDC4, MED1, NAV3, POT1, SLC1A3, SP4, TATDN1, VSNL1, ZNF366* |
| Moravka | *BCO2, DLG2, DMGDH, DNAH9, EFCAB11, FHOD3, FRMD4A, IFT57, KKLHL13, MAGI3, ME3, MID1, NCKAP1, PDLIM5, PDSS2, SEC23B, SLC44A5, STXBP5L, TAMM41, TGFBR3, TMEM260* |

*Table: Genes detected within SVs in one breed only*

- insulin measurement
- body mass index
- type II diabetes mellitus

# Functional annotation of genes affected by SVs

**SV annotation**



### Biological Process (GO)

| *pathway ID* | *pathway description* | *count in gene set* | *false discovery rate* |
|---|---|---|---|
| GO:0007155 | cell adhesion | 4 | 0.00182 |
| GO:0007160 | cell-matrix adhesion | 2 | 0.0154 |
| GO:0007229 | integrin-mediated signaling pathway | 2 | 0.043 |

### Cellular Component (GO)

| *pathway ID* | *pathway description* | *count in gene set* | *false discovery rate* |
|---|---|---|---|
| GO:0008305 | integrin complex | 2 | 0.00121 |

### KEGG Pathways

| *pathway ID* | *pathway description* | *count in gene set* | *false discovery rate* |
|---|---|---|---|
| 04514 | Cell adhesion molecules (CAMs) | 4 | 0.00134 |
| 04670 | Leukocyte transendothelial migration | 4 | 0.00134 |
| 05150 | Staphylococcus aureus infection | 3 | 0.00134 |
| 05140 | Leishmaniasis | 3 | 0.00258 |
| 05133 | Pertussis | 3 | 0.0028 |
| 04145 | Phagosome | 3 | 0.0205 |
| 04015 | Rap1 signaling pathway | 3 | 0.0431 |
| 04810 | Regulation of actin cytoskeleton | 3 | 0.0431 |
| 05134 | Legionellosis | 2 | 0.0457 |
| 05144 | Malaria | 2 | 0.0457 |
| 05416 | Viral myocarditis | 2 | 0.0457 |
| | | | *(less ...)* |

### PFAM Protein Domains

| *pathway ID* | *pathway description* | *count in gene set* | *false discovery rate* |
|---|---|---|---|
| PF07965 | Integrin beta tail domain | 2 | 0.0298 |
| PF08725 | Integrin beta cytoplasmic domain | 2 | 0.0298 |

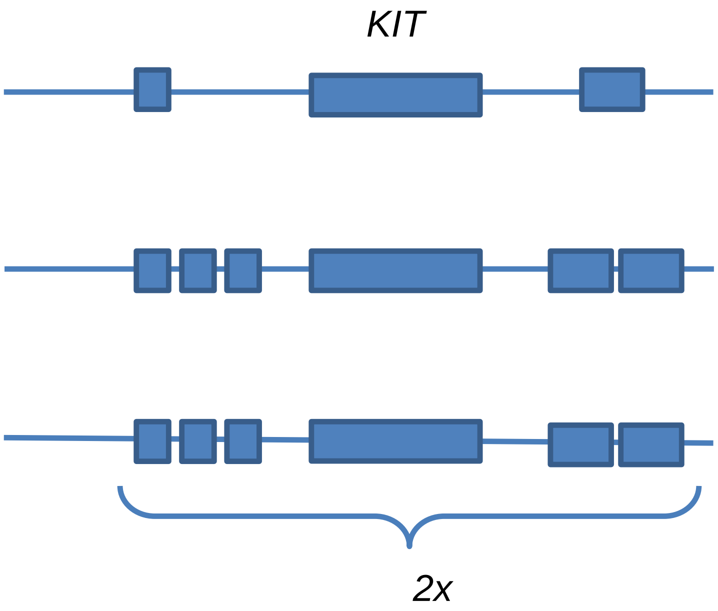
Krškopolje

# Examples

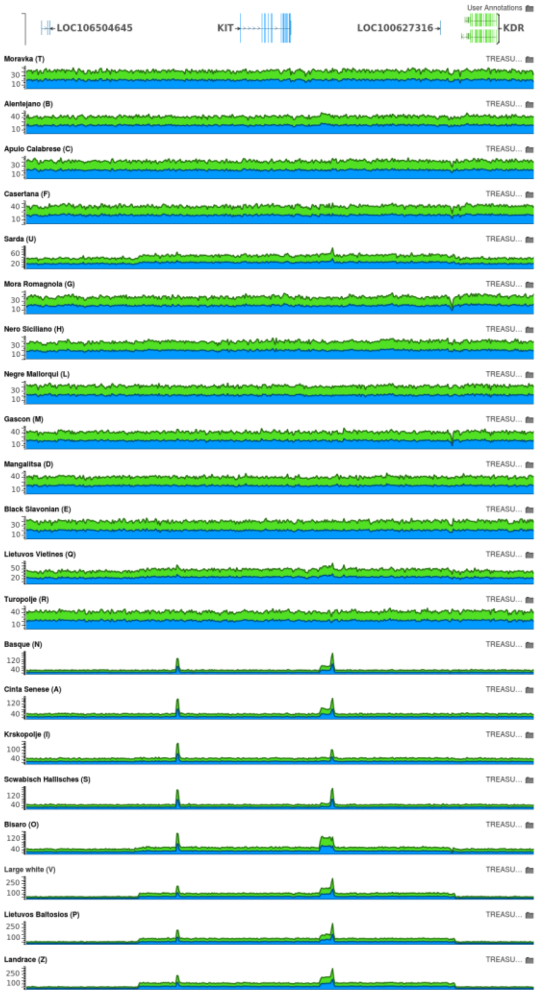
Deletions and duplications within genes:

- KIT, melanogenesis,
- UGT2B31, steroid hormone biosynthesis,
- SAL1, pheromone binding and in pre-mating behaviour,
- LEPR, appetite-regulating factor,
- MOGAT2 and miR-4335, fat digestion and absorption.

# Summary of SVs at KIT locus



# KIT

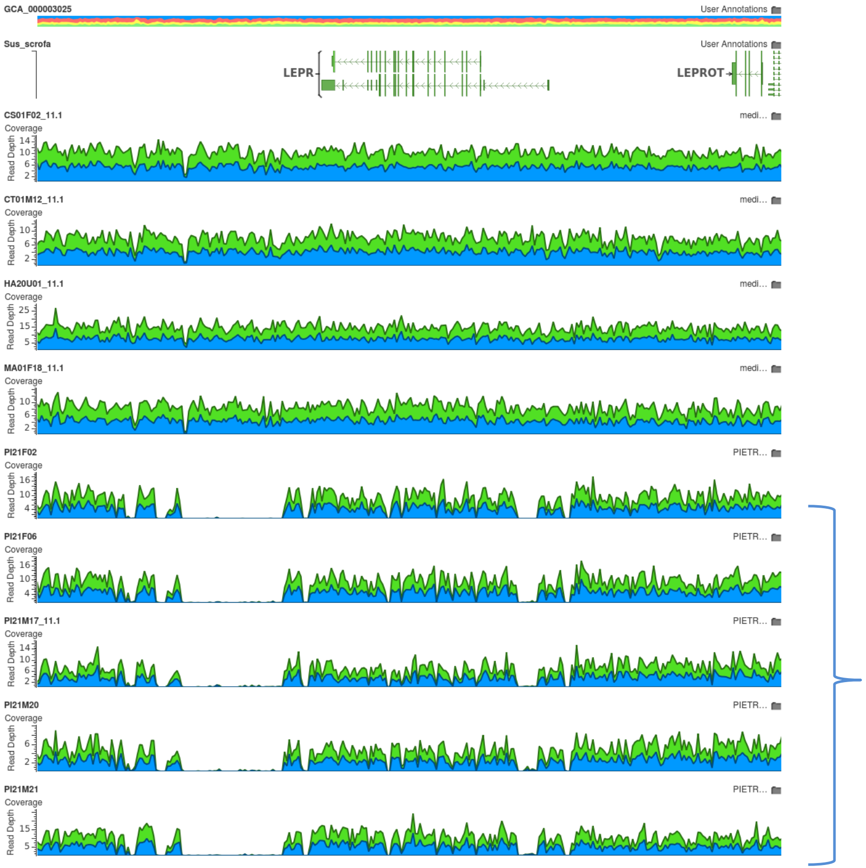


Moravka
Alentejano
Apulo Calabrese
Casertana
Sarda
Mora Romagnola
Nero Siciliano
Negre Mallorqui
Gascon
Mangalitsa
Black Slavonian
Lietuvos Vietines
Turopolje

Basque
Cinta Senese
Krskopolje
Scwabisch Hallisches
Bisaro

Large White
Lietuvos Baltosios
Landrace

# *LEPR*

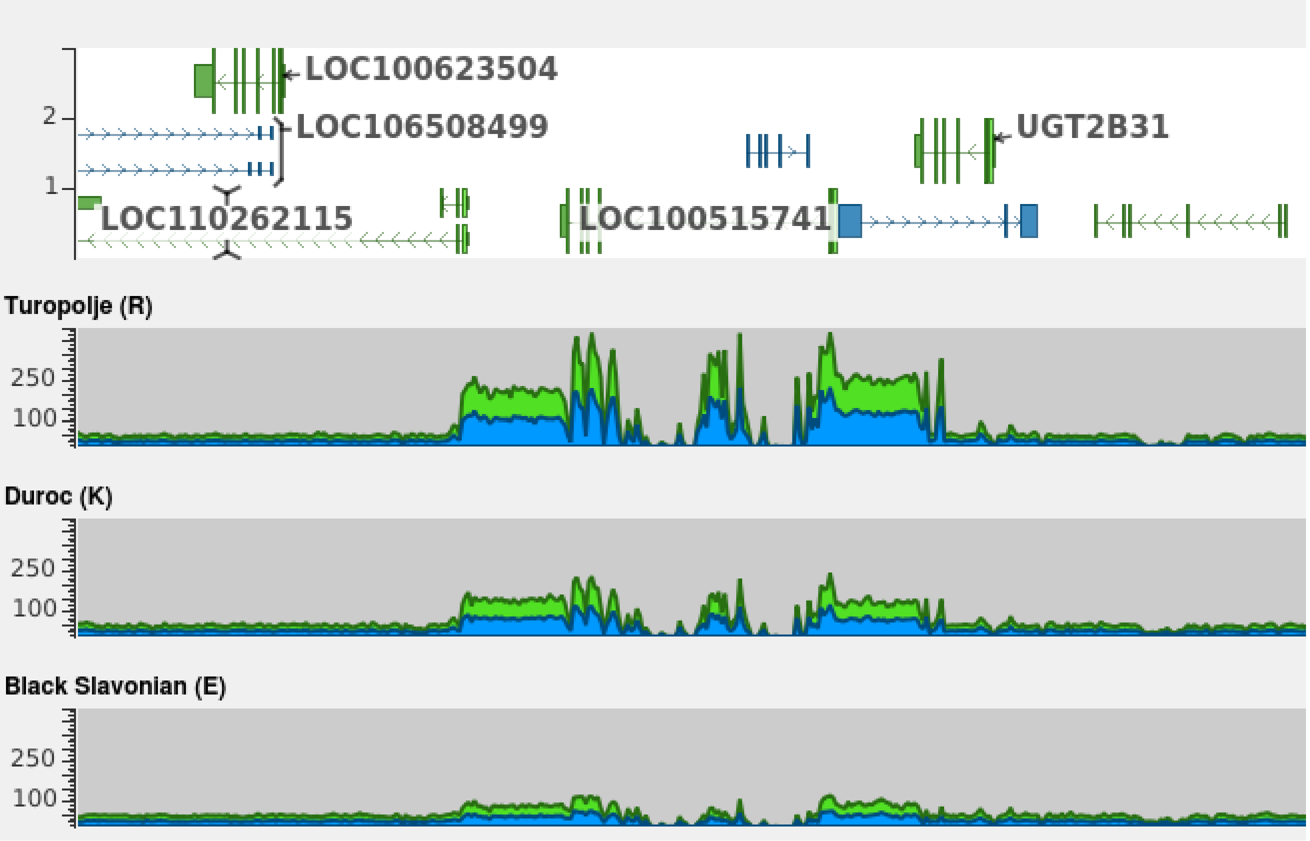


Moravka
Alentejano
Apulo Calabrese
Casertana
Sarda
Mora Romagnola
Nero Siciliano
Negre Mallorqui
Gascon
Mangalitsa
Black Slavonian
Lietuvos Vietines
Turopolje

Basque
Cinta Senese
Krskopolje
Scwabisch Hallisches
Bisaro
Large White
Lietuvos Baltosios
Landrace
Hampshire

Pietrain

# *UGT2B31*



Wild boar
Turopolje

Duroc
Hampshire
Mangalitsa
Moravka
Krskopolje
Alentejano
Lietuvos Vietines
Nero Siciliano
Negre Mallorqui
Large White
Apulo Calabrese
Sarda
Mora Romagnola
Scwabisch Hallisches

Black Slavonian
Cinta Senese
Bisaro
Casertana
Gascon
Basque
Landrace
Lietuvos Baltosios

# *SAL1*

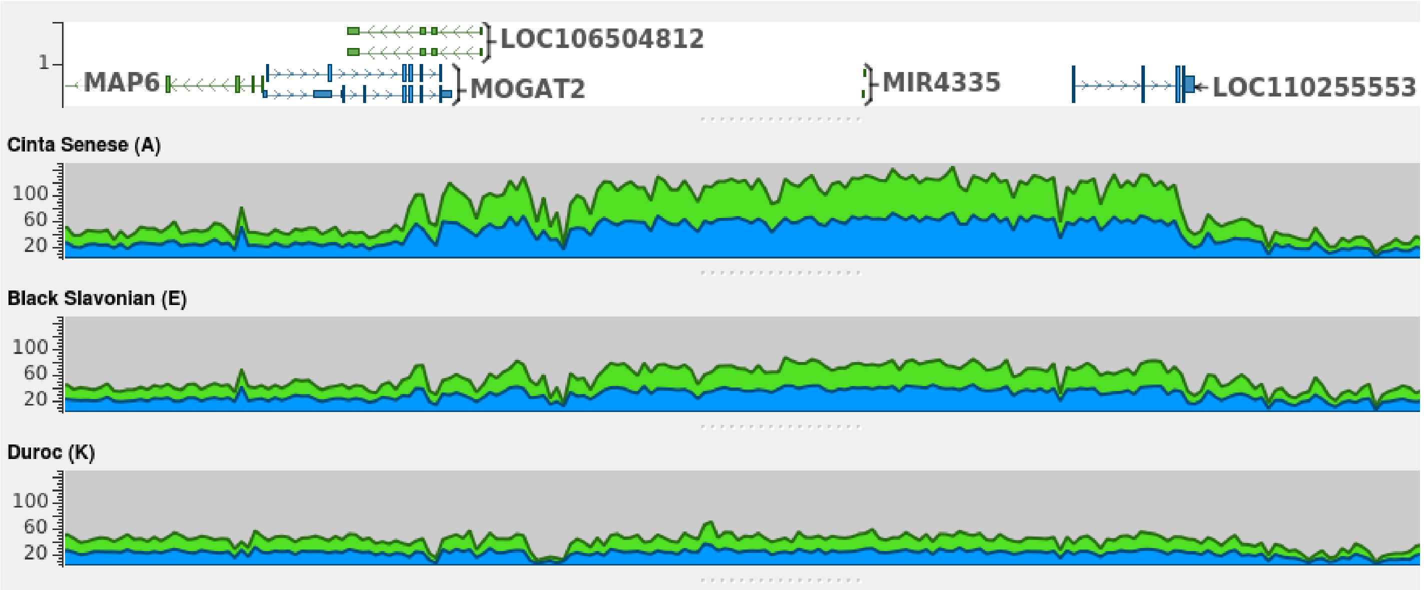
Duroc
Moravka
Hampshire

Krskopolje
Gascon
Large White
Landrace

Scwabisch Hallisches
Cinta Senese
Bisaro
Alentejano
Basque
Lietuvos Vietines
Nero Siciliano
Sarda

Turopolje
Black Slavonian
Casertana
Mangalitsa
Negre Mallorqui
Apulo Calabrese
Mora Romagnola
Lietuvos Baltosios

# *MOGAT2 and miR-4335*



Cinta Senese
Bisaro

Black Slavonian
Casertana
Turopolje
Moravka
Krskopolje
Alentejano
Lietuvos Vietines
Nero Siciliano
Large White
Apulo Calabrese
Sarda
Mora Romagnola
Landrace

Duroc
Gascon
Mangalitsa
Basque
Negre Mallorqui
Scwabisch Hallisches
Lietuvos Baltosios

# Conclusions

- Detection of SVs depends on sequence quality (coverage and length of reads) and bioinformatic tools used.
- Complete range of structural DNA variation cannot be investigated with a single tool.
- SVs affect a large of proportion of genomes (including coding and regulatory regions) and differ among breeds.
- SVs clearly have an impact on phenotypic diversity.

# Thank you for your attention!

Funded by European Union
Horizon 2020
Grant agreement No 634476