CVA24v and receptors

04 September, 2020

Redmine issue:	4947
NBIS staff:	Nima Rafati (nima.rafati@nbis.se)
Principal investigator:	Niklas Arnberg(niklas.rnberg@umu.se)
Request by:	Nitesh Mistry(nitesh.mistry@umu.se)
Organisation:	Umea University
Estimated time:	60 h
Used time:	57 h

Contents

1 Work log

- **2020-03** Consultation
- **2020-05** Meeting and discussion on variant calls and ploidy of the samples
- **2020-06** Meeting and discussion on discovered SV
- **2020-08** Delivery of report and results

2 Practical information

2.1 Data responsibilities

Unfortunately, NBIS does not have resources to keep any files associated with the support request; we kindly suggest that you safely store the results delivered by us. In addition, we kindly ask that you remove the files from UPPMAX/UPPNEX. The main storage at UPPNEX is optimized for high-speed and parallel access, which makes it expensive and not the right place for long-term archiving. Please be considerate of your fellow researchers by not taking up this expensive space.

The responsibility for data archiving lies with universities and we recommend asking your local IT for support with long-term data storage. The Data Center at SciLifeLab may also be of help with discussing other options.

Please note that special considerations may apply to human-derived, sensitive personal data. This should be handled according to specific laws and regulations as outlined at the NBIS website.

2.2 Acknowledgements

If you are presenting the results in a paper, at a workshop or at a conference, we kindly remind you to acknowledge us according to the signed NBIS User Agreement:

NBIS staff should be included as co-authors if the support work leads to a publication and when this is merited in accordance to the ethical recommendations for authorship, *i.e.* the ICMJE recommendations. If applicable, please include *Nima Rafati, National Bioinformatics Infrastructure Sweden, Science for Life Laboratory, Stockholm University* as co-author. If the above is not applicable, please acknowledge NBIS like so: *Support by NBIS (National Bioinformatics Infrastructure Sweden) is gratefully acknowledged.*

In addition, Uppmax kindly asks you to acknowledge UPPMAX and SNIC. If applicable, please add: *The computations were performed on resources provided by SNIC through Uppsala Multidisciplinary Center for Advanced Computational Science (UPPMAX) under Project snic2020-15-10(computational)&snic2020-16-70(storage).*

In any and all publications based on data from NGI Sweden, the authors must acknowledge SciLifeLab, NGI and Uppmax, like so: *The authors would like to acknowledge support from Science for Life Laboratory, the National Genomics Infrastructure, NGI, and Uppmax for providing assistance in massive parallel sequencing and computational infrastructure.*

2.3 Closing procedures

You should soon be contacted by one of our managers, Jessica Lindvall *(jessica.lindvall@nbis.se)* or Henrik Lantz *(henrik.lantz@nbis.se)*, with a request to close down the project in our internal system and for invoicing matters. If we do not hear from you within **30 days** the project will be automatically closed and invoice sent. Again, we would like to remind you about data responsibility and acknowledgements, see the sections on data responsibilities and acknowledgments.

You are naturally more than welcome to come back to us with further data analysis request at any time via the support form. Thank you for using NBIS, we wish you the best of luck with your future research!

3 Summary

DSG2 is knocked out by CRISPR in HAP1 cells. This cell-line is known as haplpid cells; chr8 and 15 seem to have two copies. There are three samples:
- WT: wild type without any modification
- F9: KO DSG2
- H11: KO DSG2
Virus can infect both WT and H11 while F9 is resistant. Two different antibodies are used to detect expression of DSG2 protein in these samples. Figure 1 top panel shows detection of DSG2 protein using a antibody targeting domain 1 and 2 and lower panel by using a antibody targeting domain 3 and 4.

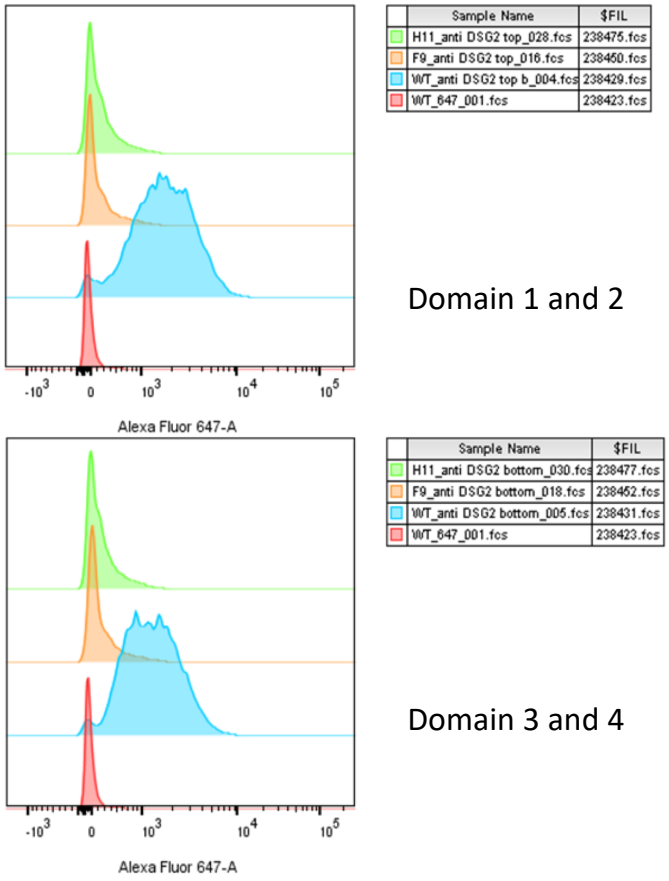

Figure 1: FACS sorting targeting domain 1 and 2 (top) and 3 and 4 (lower) panel.

To characterise the genetic difference between these samples whole-genome sequencing (WGS) was conducted. In this report we describe the analysis to identify genetic differentiation between samples which may be associated with response to infection. The initial variant calling was carried out by NGI based on GATK best practice. The WT sample did not have any calls as homozygous reference (0/0). Thus, we repeated the pipeline and compared the samples for small variants (SNP and INDEL). We also screened for structural variation (SV) and compared the samples.

4 Deliverables

- **Variant_stat:** Figures of variant calling statistics distribution.
- **VCF:** Final VCF of small variants by GATK and annotated for functional mutations by snpEff.
- **Freq:** A text file that reference allele frequency is calculated in all of the samples.
- **VCF:** Final VCF of structural variants by Manta and annotated for functional mutations.
- **Depth:** A text file consisting of raw depth of DSG2 region on chr18.
- **snpEff:** General statistics of functional annotation generated by snpEff.
- **Codes:** All the codes and part of results used in the report are located in github repository of NBISweden.
- **NBIS Report:** This report.

All the files are stored in */crex/proj/snic2020-16-70/private/SMS_4947_20_CVA24v_receptor/results/*. All the codes for analysis on Uppmax are sotred in */crex/proj/snic2020-16-70/private/SMS_4947_20_CVA24v_receptor/code/gene commands.sh* and all the codes for visualisation and report is stored in */crex/proj/snic2020-16-70/private/SMS_4947_20_CVA24v_receptor/SMS_4947_20_CVA24v_receptor.Rmd. The code to generate this report is stored in* /crex/proj/snic2020-16-70/private/SMS_4947_20_CVA24v_receptor/report/.

Data analysis
We used the alignments that was generated in pipeline used by NGI. Steps in this pipeline is as follows:
- FASTQC: Checking the quality of raw reads by FASTQC (Andrews, n.d.).
- Alignment: Aligning the reads by BWA (Li and Durbin 2010) on human reference genome (GRCh37/hg19).
- QC: Checking the quality of alignments by QualiMap (Okonechnikov, Conesa, and García-Alcalde 2016).
- Variant calling: Due to abovesaid reasons we repeated the variant calling by GATK (McKenna et al. 2010).

4.1 QC

We checked the quality of mapping for each sample by QualiMap (Okonechnikov, Conesa, and García-Alcalde 2016). Above 99% of reads were mapped to the genome. The average coverage across the genome was fairly variable among the samples (Table 1) but high enough to call variants.

Table 1: Mean coverage across the genome calculated by QualiMap.

Sample	Mean Coverage
F9	40.81
H11	56.20
WT	38.51

4.2 Variant calling

4.2.1 Small variant (SNP/INDEL)

For variant calling we followed GATK best practice (GATK4, version 4.1.1.0) (McKenna et al. 2010). We generated g.vcf files for each sample by GATK::HaplotypeCaller and genotyped the samples GATK::GenotypteGVCFs. To filter out low quality variants we extracted following statistics from vcf file for SNPs and INDELs separately and set a cutoff based on distributions of them. All the figures related to these statistics are located in _results/Variant_calling/Genotyping/*pdf_. Statistics and cut-off values for SNPs and INDELs are:

INDEL
- QUAL < 100.0
- QD < 10.0
- FS > 10.0
- ReadPosRankSum < -2.0
- MQRankSum < -2.0
- BaseQRankSum < -5.0
- SOR > 3.0
- genotype-filter DP < 10.0 || DP >200.0

SNP
- QUAL < 100.0
- QD < 10.0
- FS > 10.0
- MQ < 40.0
- MQRankSum < -8.0
- ReadPosRankSum < -5.0
- BaseQRankSum < -5.0
- SOR > 3.0
- genotype-filter DP < 10.0 || DP >400.0

Table 2: Frequency of small variants across the genome.

Variants	Frequency
SNP	2,643,008
INDEL	646,917

After filtering, we merged both vcf files (SNPs and INDELs) and kept only variants passed filtering. This resulted in **3,289,925** SNP/INDEL (Table 2):

We calculated reference allele frequency of each variant for each individual to identify differentiated sites between samples. The allele frequency distribution suggestes that all samples have very similar profile (Figure 2).

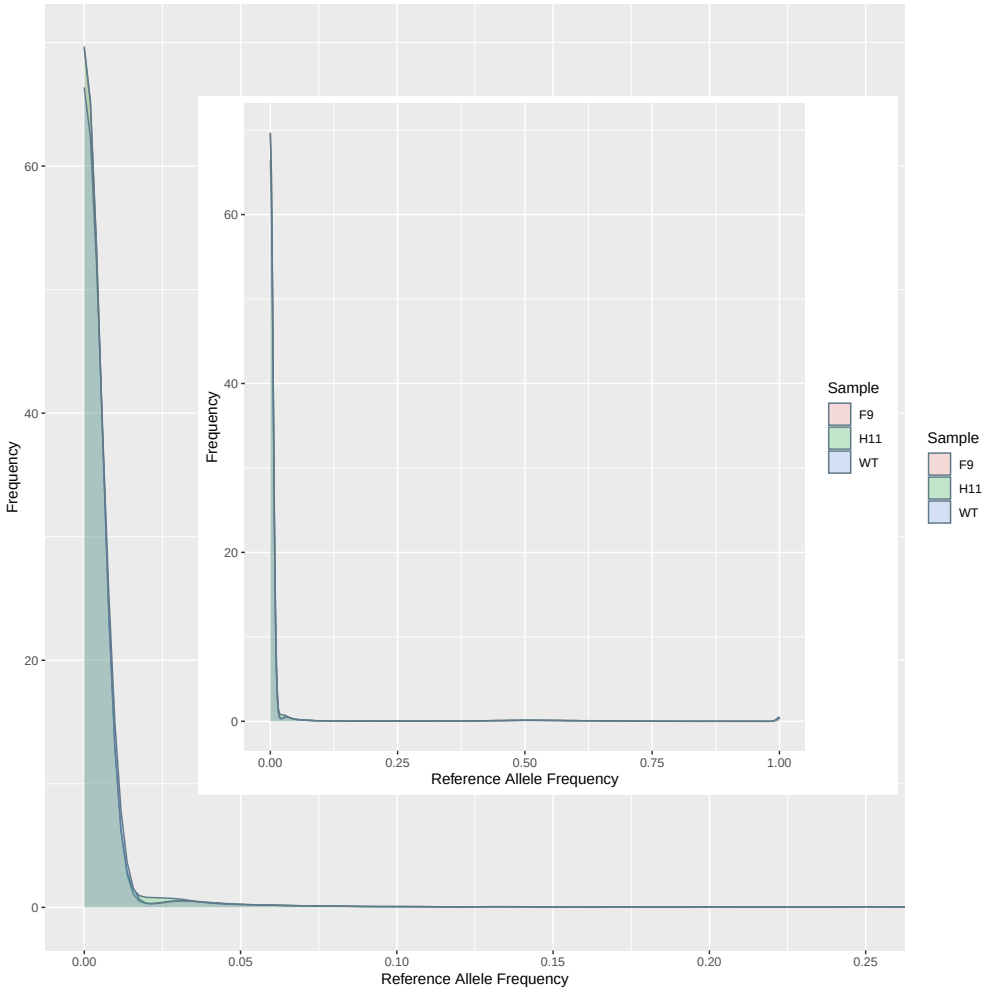

Figure 2: Reference allele frequency distribution across the genome. The inset figure shows full frequency spectrum (0-1).

4.2.2 Functional effect prediction

We used snpEff (Cingolani et al. 2012) and GRCh37.p13.RefSeq annotation to annotate the effect of identified variants on genes. Table 3 shows number of identified variants in different functional elements. Identified variants are mostly in non-coding regions (intergenic_region or intronic_region). Effect prediction

details is found here.

	Type	Count	Percent
1	3_prime_UTR_variant	21 430	0.58%
2	5_prime_UTR_premature_start_codon_gain_variant	459	0.01%
3	5_prime_UTR_variant	3 752	0.10%
4	bidirectional_gene_fusion	1	0.00%
5	conservative_inframe_deletion	45	0.00%
6	conservative_inframe_insertion	64	0.00%
7	disruptive_inframe_deletion	126	0.00%
8	disruptive_inframe_insertion	84	0.00%
9	downstream_gene_variant	181 580	4.90%
10	frameshift_variant	355	0.01%
11	gene_fusion	3	0.00%
12	initiator_codon_variant	1	0.00%
13	intergenic_region	1 852 514	49.95%
14	intragenic_variant	124 757	3.36%
15	intron_variant	1 308 566	35.29%
16	missense_variant	8 438	0.23%
17	non_coding_transcript_exon_variant	14 784	0.40%
18	non_coding_transcript_variant	14	0.00%
19	splice_acceptor_variant	87	0.00%
20	splice_donor_variant	65	0.00%
21	splice_region_variant	2 636	0.07%
22	start_lost	20	0.00%
23	stop_gained	85	0.00%
24	stop_lost	43	0.00%
25	stop_retained_variant	12	0.00%
26	synonymous_variant	8 555	0.23%
27	upstream_gene_variant	179 996	4.85%

Table 3: Functional annotation of identified variants.

In *DSG2* region there are 10 variants that listed in Table 4. Most of the variants are in intronic region but there are two frameshift mutations (exon 3 and 8) as well as a disruptive_inframe deletion (exon 5).

	Chr	Position	F9	H11	WT	Effect	Gene
1	18	29 086 973	0.00	0.00	0.00	intron_variant	DSG2
2	18	29 087 112	0.00	0.00	0.00	intron_variant	DSG2
3	18	29 089 395	0.00	0.00	0.00	intron_variant	DSG2
4	18	29 098 215	1.00	0.07	1.00	conservative_inframe_deletion	DSG2
5	18	29 098 226	1.00	0.17	1.00	splice_donor_variant	DSG2
5	18	29 098 226	1.00	0.17	1.00	conservative_inframe_deletion	DSG2
5	18	29 098 226	1.00	0.17	1.00	splice_region_variant	DSG2
5	18	29 098 226	1.00	0.17	1.00	intron_variant	DSG2
6	18	29 099 844	0.00	0.40	1.00	frameshift_variant	DSG2
7	18	29 101 063	0.65	1.00	1.00	disruptive_inframe_deletion	DSG2
8	18	29 104 820	1.00	0.44	1.00	frameshift_variant	DSG2
9	18	29 107 035	0.00	0.00	0.00	intron_variant	DSG2
10	18	29 132 175	0.00	0.00	0.00	downstream_gene_variant:intron_variant	DSG2:LOC100652770

Table 4: Identified mutations in DSG2 region and their functional impact. Values in columns F9, H11, and WT shows reference allele frequency

4.3 Large variant (Structural variant)

We screened *DSG2* region for structural variantion by Manta (Chen et al. 2016) for F9 and H11 samples. Similar to small variants we annotated identified variants by snpEff for functional changes on *DSG2* gene. This resulted in:

- F9: A deletion (~3 kb) chr18:29098216-29100979. This mutation completely remove exon 3 and 4 as well as part of exon 2. This deletion is not fixed in this sample (Figure 3)
- H11:
  i) An inversion (~2 kb) chr18:29098216-29099847. This inversion spans part of exon 2 and exon 3 and intronic region in between (Figure 3).
  ii) An insertion (225 bp) chr18:29104826. This mutation results in a stop_gain & disruptive_inframe_insertion on exon 8. The inserted sequence has high similarity to (MT270142.1)[https://www.ncbi.nlm.nih.gov/nucleotide/MT270142.1?report=genbank&log$=nuclalign&blast_rank=1&RID=KACJF7YS014] which is a cloning vector PAPs-CRISPR.

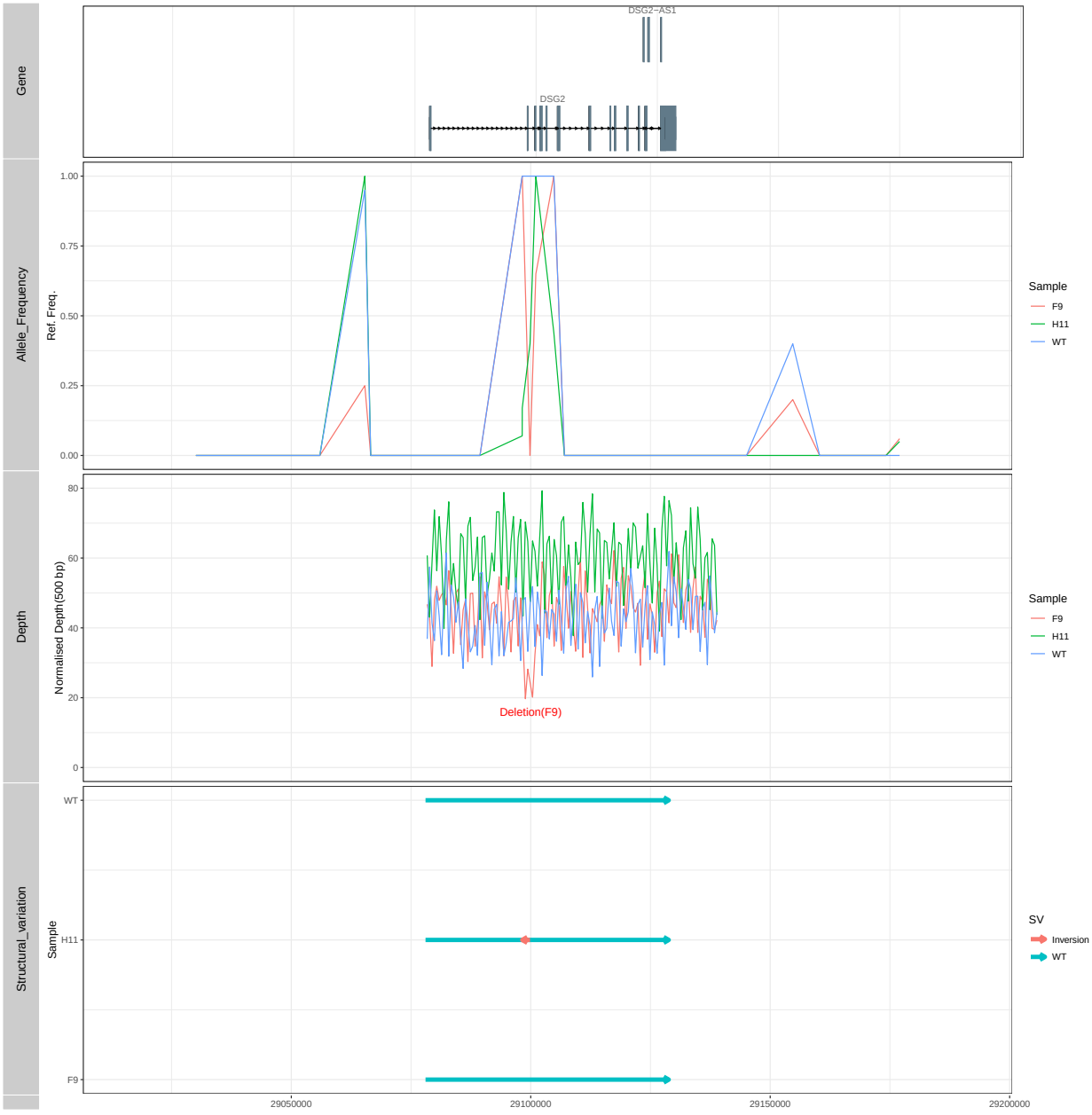

Figure 3: Summary of the identified variants. 1) gene model 2) reference allele frequency 3) normalised depth 4) inversion; inversion is shown by red arrow. 225 bp insertion was small to visualise here.

5 Concluding remarks

Results suggest that some of the identified variants have "HIGH" impact on functionality of *DSG2*. For instance deletion spanning exons in F9 and in particular inversion in H11 which introduces a stop codon. Identified small variants were mostly in intronic region as well as few frameshifts in coding regions. All in all, CRISPR has introduced different mutations in different samples which may be associated with different responses to infection. To explore this further we suggested additional experiments:

- Designing primers at breakpoints of identified inversion in H11 as well as deletion in F9 as discussed in the meeting on 4th of June.
- Designing primers to amplify cDNA.

The following figures show suggested PCR experiments in more details.

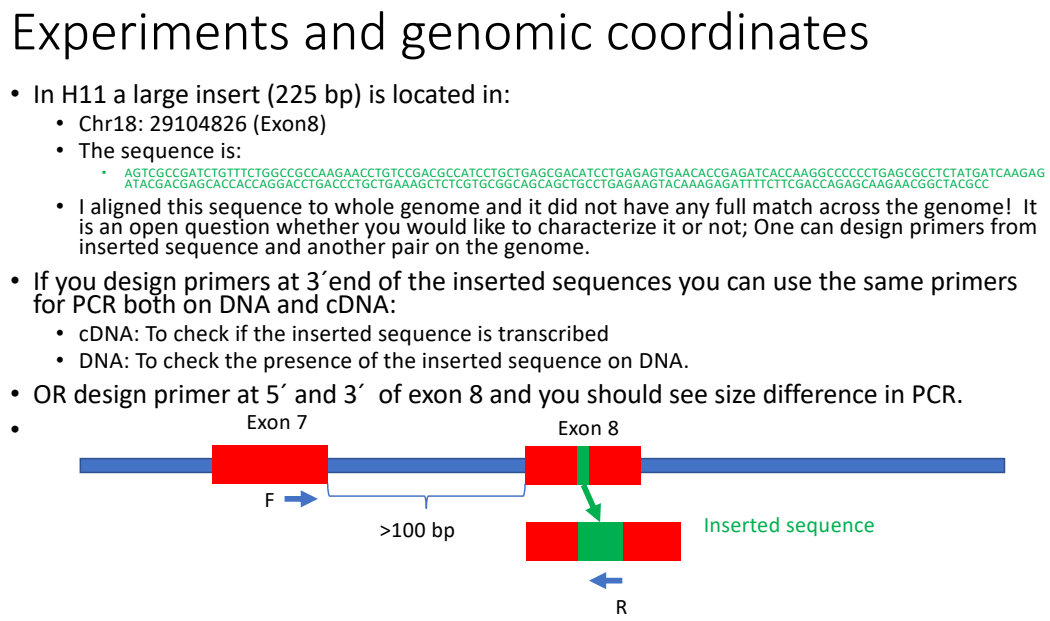

Figure 4: Suggested experiments discussed in the meeting.

Experiments and genomic coordinates

- **Please note that the STOP_GAIN is annotated as result of this insert.**
  - So, I suggest to design primers from 3´ site of this exon to amplify cDNA to capture potential new stop codon in the transcript.

Figure 5: Suggested experiments discussed in the meeting.

Experiments and genomic coordinates

- Inversion coordinate:
  - Chr18:29098216 – 29099847
- So you can design primers close to this coordinate but make sure that the distance to breakpoints should be different.

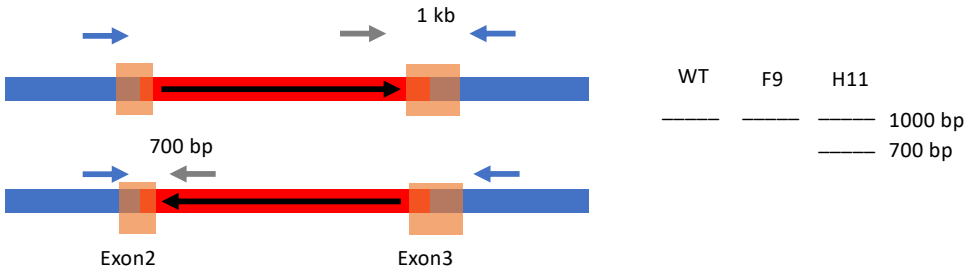

Figure 6: Suggested experiments discussed in the meeting.

It is also possible to look into other variants by allele frequency. Considering that you have a pool of cells in each sample, allele frequency will be a more accurate measure than genotypes. Thus, I filtered by reference allele frequency:
- F9 Ref freq <=0.2 while H11 & WT has ref freq >=0.8
- F9 Ref freq >=0.8 while H11 & WT has ref freq <=0.2
Figure 7 shows the distribution of these sites.

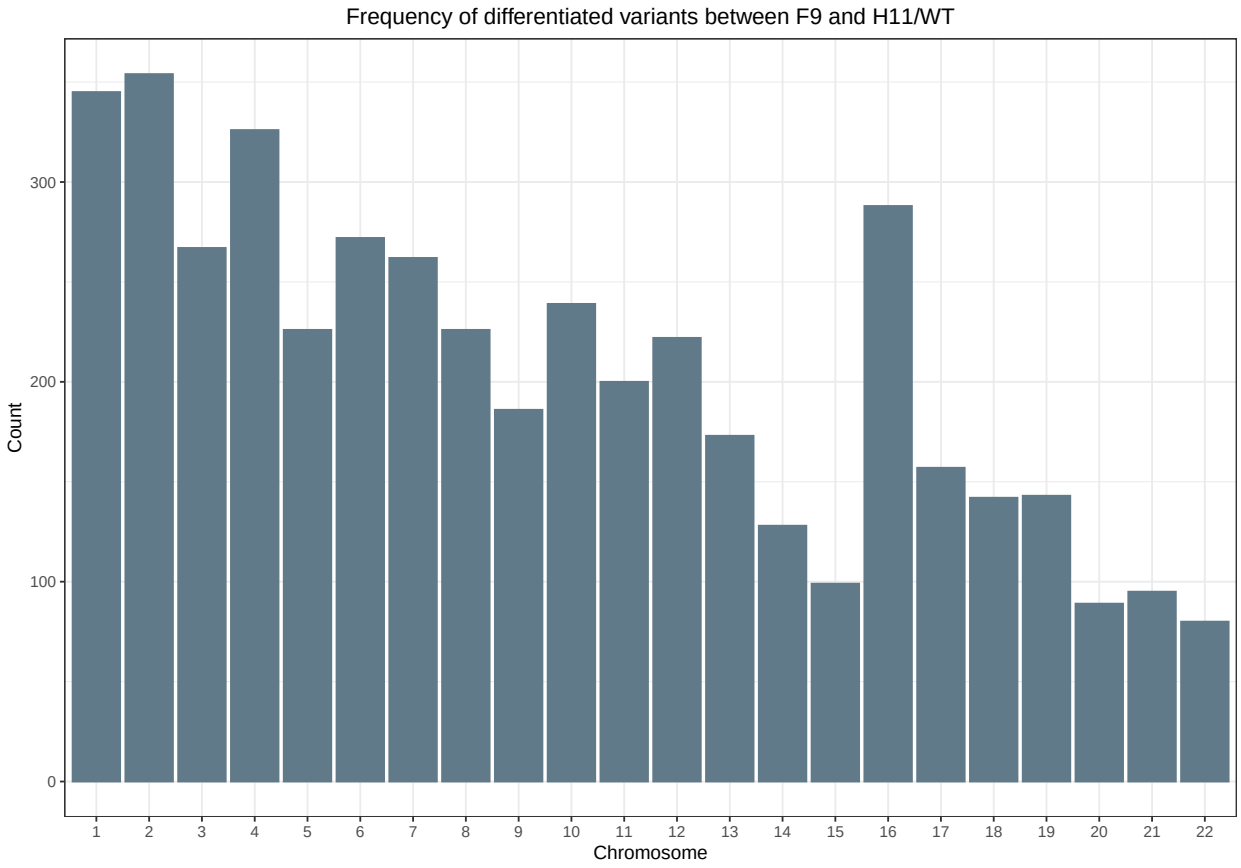

Figure 7: Suggested experiments discussed in the meeting.

There are 0.15% (~5000 SNP/INDEL) of total identified sites showing frequency difference between F9 and H11/WT. You can find list of these sites with allele frequency and their functional changes in *results/Variant_results/Variant_calling/Genotyping/Differentiated_sites_F9_H11_WT-snpEff.txt*. As an example there is a missense mutation (11:6588171) in *DNHD1* where reference allele frequency is *0* in F9 while H11 and WT have frequency of *1*. Or another missense mutation (19:9070194) in *MUC16* where F9 has frequency of *1* while H11 and WT has frequency of *0*.

Identified mutations can be correlated with exon maps 8 (from (Excoffon, Bowers, and Sharma 2014)) and anti body binding sites.

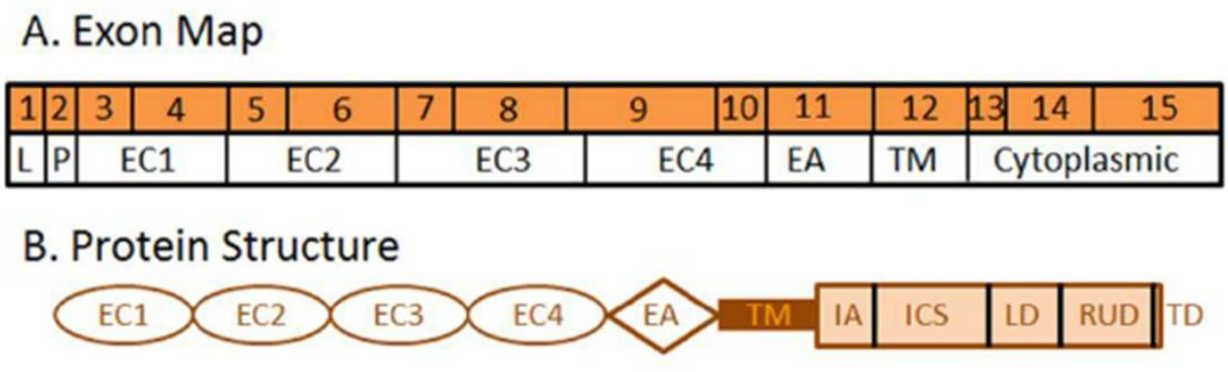

Figure 8: Exon map of DSG2 from Excoffon et al., 2014.

Reference

Andrews, S. n.d. "FastQC A Quality Control tool for High Throughput Sequence Data." citeulike-article-id:11583827%20http://www.bioinformatics.babraham.ac.uk/projects/fastqc/.

Chen, Xiaoyu, Ole Schulz-Trieglaff, Richard Shaw, Bret Barnes, Felix Schlesinger, Morten Källberg, Anthony J. Cox, Semyon Kruglyak, and Christopher T. Saunders. 2016. "Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications." *Bioinformatics* 32 (8): 1220–2. https://doi.org/10.1093/bioinformatics/btv710.

Cingolani, P., A. Platts, M. Coon, T. Nguyen, L. Wang, S. J. Land, X. Lu, and D. M. Ruden. 2012. "A Program for Annotating and Predicting the Effects of Single Nucleotide Polymorphisms, Snpeff: SNPs in the Genome of Drosophila Melanogaster Strain W1118; Iso-2; Iso-3." *Fly* 6 (2): 80–92.

Excoffon, Katherine J D A, Jonathan R Bowers, and Priyanka Sharma. 2014. "1. Alternative splicing of viral receptors: A review of the diverse morphologies and physiologies of adenoviral receptors." *Recent Research Developments in Virology* 9: 1–24. https://pubmed.ncbi.nlm.nih.gov/25621323%20https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4302334/.

Li, Heng, and Richard Durbin. 2010. "Fast and accurate long-read alignment with Burrows–Wheeler transform." *Bioinformatics* 26 (5): 589–95. https://doi.org/10.1093/bioinformatics/btp698.

McKenna, Aaron, Matthew Hanna, Eric Banks, Andrey Sivachenko, Kristian Cibulskis, Andrew Kernytsky, Kiran Garimella, et al. 2010. "The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data." *Genome Research* 20 (9): 1297–1303. https://doi.org/10.1101/gr.107524.110.

Okonechnikov, Konstantin, Ana Conesa, and Fernando García-Alcalde. 2016. "Qualimap 2: advanced multi-sample quality control for high-throughput sequencing data." *Bioinformatics (Oxford, England)* 32 (2): 292–94. https://doi.org/10.1093/bioinformatics/btv566.