Sensing foul *AIRE*:

Investigating possible reporter genes for AIRE mutations

Amund Holte Berger

This thesis is submitted in partial fulfilment of the requirements for the degree of
Master of Science in Biomedical Sciences

Department of Biomedicine / Department of Clinical Science (K2)
Faculty of Medicine
University of Bergen
Spring 2018

Acknowledgements

I would first like to thank my supervisors, Stefan Johansson, Eirik Bratland and Per Knappskog for great supervision. Stefan Johansson for his excellent organisational skills, theoretical, genetic and bioinformatical help. Eirik Bratland for his help in hands-on lab training and knowledge of all things AIRE and immunology. Per Knappskog for his troubleshooting skills, and his methodological help.

Thanks go out to all colleagues at the K.G. Jebsen Center for Autoimmune Disorders, most especially Haydee Artaza Alvarez and Alexander Hellesen. Haydee for all the help with bioinformatics, particularly the development of our pipeline using Kallisto/DESeq2, and for patiently answering all my questions. Alexander for all his help with immunology, AIRE and our research approach.

I would also like to thank the people at the Research and Development (FoU) group in the medical genetics lab (MGM). Most especially Jorunn Skeie Bringsli for help in navigating the process of Sanger sequencing, and Guri Elisabeth Matre for patiently teaching the theory and process of quantitative polymerase chain reaction. Also, I would like to thank Hilde Eldevik Rusaas for performing the RNAseq library preparation.

From the Genomics Core Facility (GCF) of the MGM, I would like to thank Rita Holdhus for the RNAseq sequencing and Tomasz Stokowy for help with initial bioinformatics with the first RNAseq, using the traditional hisat2/featureCounts/DESeq2 pipeline.

Finally, I would like to thank Nicolas Delhomme and Bastian Schiffthaler from the University of Umeå, Charlotte Soneson from the University of Zürich, Matthew Macmanes from the University of New Hampshire, Aaron Lun and everyone else at the EMBL-EBI in Hinxton for an excellent course on Advanced RNAseq analysis. Thanks for giving me many pointers and ideas to try out.

Table of contents

Abbreviations

21OH	*21-hydroxylase*, used as positive FLAG control
AIRE	Autoimmune regulator, a transcriptional regulator
APECED	Autoimmune polyendocrinopathy candidiasis ectodermal dystrophy, an alternative name for APS-1
APS-1	Autoimmune polyendocrine syndrome type 1
ARMC5	Armadillo Repeat-Containing Protein 5
Brd4	Bromodomain-containing protein 4, transcriptional inhibitor
BSA	Bovine serum albumin, albumin protein extract
C311Y	AIRE variant with the C311Y mutation, a dominant missense mutation
CARD	Caspase recruitment domain, a protein interaction domain
CBP	CREB-binding protein, a transcriptional activator
CCNH	Cyclin H
CD	Cluster of differentiation, surface molecule or protein used to differentiate between cell types
cDNA	Complementary DNA, DNA reverse transcribed from mRNA, lacks introns
CHAC1	ChaC Glutathione Specific Gamma-Glutamylcyclotransferase 1
CNS1	Conserved non-coding sequence 1, an enhancer sequence
CTCF	CCCTC-Binding Factor, a transcriptional inhibitor
Ctcfl	CCCTC-Binding Factor Like, a transcriptional enhancer, works by replacing CTCF
cTEC	Cortical thymic epithelial cells
DN	Double negative T-Cells, does not have the surface markers CD4 or CD8
DP	Double positive T-Cells, has both the surface markers CD4 and CD8
ERMAP	Erythroblast Membrane Associated Protein
FACS	Fluorescence-activated cell sorting, cell sorting method using flow-cytometry
FC	Fold change, a number describing the change in quantity between two states in a comparison
FLAG	Protein tag with the DYKDDDDK sequence
GAPDH	Glyceraldehyde-3-Phosphate Dehydrogenase
GOrilla	Gene Ontology enRIchment anaLysis and visuaLizAtion, a gene ontology enrichment tool
GRCh38	Genome Reference Consortium Human Build 38, reference genome
H3K27ac	Acetylated Lysine 27 on the histone tail of histone H3
H3K4me0	Unmethylated Lysine 4 on the histone tail of histone H3
H3K4me1	Mono-methylated Lysine 4 on the histone tail of histone H3
H3K4me3	Tri-methylated Lysine 4 on the histone tail of histone H3
Hnrnpl	Heterogeneous Nuclear Ribonucleoprotein L, a transporter of P-TEFb
HSR	Homogenously staining region, the previous name for the CARD domain in AIRE
IGFL1	Insulin Growth Factor-Like Family Member 1
IL10RA	Interleukin 10 Receptor Subunit Alpha
INHBE	Inhibin Beta E Subunit
Irf4	Interferon Regulatory Factor 4, a transcriptional regulator
Irf8	Interferon Regulatory Factor 8, a transcriptional regulator
JMJD6	Arginine Demethylase and Lysine Hydroxylase, a pre-mRNA splicing factor
KRT14	Keratin 14

LDS-PAGE Lithium dodecyl sulphate polyacrylamide gel electrophoresis, a protein electrophoresis method
MHC Major histocompatibility complex, antigen presenting, also known as HLA
MITE Mutagenesis by integrative tiles, a multiplexed controlled mutagenesis method
mTEC Medullary thymic epithelial cells
Myc Protein tag with the EQKLISEEDL sequence
NFκB Nuclear Factor κB, a signalling complex
NHEJ Non-homologous end joining, a DNA repair pathway
NLS Nuclear localisation sequence, a peptide sequence in a protein necessary for nucleus entry
NMD Nonsense mediated decay, an RNA degradation pathway triggered by premature stop codons
P300 E1A binding protein, a transcriptional coactivator
padj FDR adjusted p-value, or Benjamini-Hochberg (BH) adjusted p-value
PBS Phosphate buffered saline, a buffer solution
PHD Plant homeodomain, a zinc finger structure
PTA Peripheral tissue antigens
P-TEFb Positive Transcription Elongation Factor b, protein kinase and transcriptional activator
PTH Parathyroid hormone
PVDF A polyvinylidene difluoride membrane
qPCR Quantitative polymerase chain reaction, gene expression analysis method, also known as Real-Time PCR
R257X AIRE variant with the R257X mutation, recessive nonsense mutation
RANK Receptor Activator of Nuclear Factor κB, member of the TNF family
RANKL Receptor Activator of Nuclear Factor κB ligand
RelA the NFκB subunit p65
RIN RNA integrity number, a measure of RNA quality, a number between 1-10 where 10 is the highest quality
RNAseq RNA sequencing, a method for transcriptome analysis
rRNA Ribosomal RNA, RNA used directly in ribosomal structures necessary for protein synthesis
S100A8 S100 Calcium Binding Protein A8
SAND DNA recognition domain, named for the proteins: Sp100, AIRE, NucP41/75, and DEAF-1
SLC3A2 Solute Carrier Family 3 Member 2
SLC7A11 Solute Carrier Family 7 Member 11
SP Single positive T-cells, has either CD4 or CD8
TBS Tris buffered saline, a buffer
TBS-T A TBS buffer with added Tween20
Tbx21 T-Box 21, a transcriptional regulator
Tcf7 Transcription factor 7, a transcriptional regulator
TCR T-cell receptor, antigen sensing
TNF Tumour necrosis factor, a transmembrane cell signalling protein family
TOP1 Topoisomerase I, responsible for single stranded breaks in DNA
TOP2 Topoisomerase II, part of the NHEJ DNA repair pathway, responsible for double stranded breaks in DNA
TPM Transcripts per million, absolute expression value normalised against gene length and sequencing depth
TRA Tissue-restricted self-antigens
Tregs Regulatory T-cells, positive for CD4, CD25 and FOXP3
TSA Tissue-specific antigens
UPR Unfolded protein response, a cell response to misfolded proteins
Wt Wildtype, the predominant gene variant

Summary

The autoimmune regulator protein, known as AIRE, is a potent transcriptional regulator active in medullary thymic epithelial cells (mTECs) of the thymus, where it is able to switch on the expression of thousands of genes commonly only expressed in specialised peripheral tissues. This ability of AIRE makes it a crucial component in the immune system, specifically for the process of negative selection, in which T-cells are evaluated in their ability to recognise the body's own proteins. This works as a check-point to avoid autoimmunity, and T-cells that bind to AIRE induced proteins are terminated as they are considered dangerous for the organism. Disruption of AIRE function by mutations leads to the disease autoimmune polyendocrine syndrome type 1 (APS-1), in which autoimmune T-cells initiate destructive processes affecting a variety of functions in the body. Clinically, APS-1 is defined as the presence of at least two out of three major manifestations: Addison's disease (adrenal insufficiency), hypoparathyroidism and chronic mucocutaneous candidosis. However, patients may not necessarily present the major manifestations, and may also exhibit a variety of other manifestations, both of which may be related to the severity of the underlying mutation.

AIRE consists of a number of functional domains; a CARD sequence used for AIRE dimerisation, a SAND domain for general DNA interaction, and two PHD zinc fingers, one used for histone interaction, and the other for protein recruitment and interaction. AIRE induction works in a stochastic manner, targeting genes that are passively downregulated by methylated histone marks, or that are otherwise actively repressed. Therefore AIRE induced genes differ between cells because of the sheer amount of inducible genes, as well as between different cell types, because different cell types will repress different genes.

In order to study AIRE function and inform larger sequencing and GWAS studies, we are aiming to develop a functional screening assay for *AIRE* mutations, using a deep mutational scanning approach. This would attempt to characterise the functional effect of any hypothetical mutation within *AIRE* and would require a robust reporter gene. This would be a gene with high expression in *AIRE* wildtype, but low expression in an *AIRE* mutant, while preferably encoding a cell surface protein for easier FACS sorting. To identify these reporter genes we aimed to develop a robust cell system with AIRE. This cell system needed to be

amenable to large-scale transfection and FACS sorting, and be robustly expressing functional AIRE proteins. We, therefore, investigated the expression of known AIRE reporter genes, curated from the literature, and evaluated the usability of these genes as possible reporters. Furthermore, we developed a protocol for AIRE inducible gene discovery using RNA sequencing and evaluated different methodological approaches to this.

We successfully established a robust cell system based on *AIRE* transfected HEK293FT cells, which exhibited substantial AIRE expression. By using qPCR probes for known AIRE induced genes, we also confirmed that AIRE was functionally active.

We found that the previously reported AIRE regulated genes *KRT14* and *S100A8* could be used as reporter genes based on lower expression in selected *AIRE* mutants. We found that RNAseq is highly consistent between experiments and across methodological approaches when it comes to library preparation, and correlates well with results using qPCR. However, we were unable to identify new reporter genes fitting our criteria, and the reporter gene candidates *KRT14* and *S100A8* were too weakly expressed to be detected by RNAseq. Comparing the AIRE wildtype with untransfected cells yielded substantial transcriptome differences, consistent with the literature, yet did not yield usable reporter genes. Comparing the AIRE mutants R257X and C311Y with the wildtype, in order to find downregulated genes in the mutants, we found a large population of upregulated genes in the R257X mutant and little difference between C311Y and wildtype. Neither of these mutants has previously been investigated using transcriptome analysis, and so it is uncertain how representative these results are, but they are consistent across our experiments. Western blot analysis showed some degradation in all transfected populations, yet substantial degradation of the C311Y mutant, suggesting a possible instability in this variant.

AIRE is a fascinating transcriptional regulator able to induce the expression of repressed genes, but the knowledge of AIRE and its function is still incomplete. The failure of our RNAseq approach to detect AIRE reporter genes indicates that changes in methodology are required. Such changes may still render a deep mutational scanning approach a viable option for the purpose of studying AIRE.

1. Introduction

AIRE or the *Autoimmune Regulator* is a gene located on the long arm of chromosome 21 (21q22.3) that encodes a protein that is 545 amino acids long and 57.7 kDa in size.[1, 2, 3] AIRE works as a transcriptional regulator, and is able to induce the expression of thousands of genes coding for proteins known as tissue-restricted self-antigens (TRA), tissue-specific antigens (TSA), or peripheral tissue antigens (PTA).[4] These are proteins that are usually expressed only in particular tissues. This ability of AIRE makes it a crucial part of the negative selection of T-cells in the thymus as part of the process known as central tolerance.[5] Mutations in *AIRE* may disrupt this process, leading to the development of the disease autoimmune polyendocrine syndrome type 1 (APS-1).

AIRE expression

Because of AIRE's ability to induce expression of such a broad spectrum of proteins, it is crucial that its expression is tightly controlled. AIRE is predominantly expressed in a subset of thymus cells called medullary thymic epithelial cells (mTEC), although a few B-cells, a few cells in the spleen and lymph nodes, and testicular germ cells are also able to express AIRE.[6] While AIRE is expressed principally in mTEC, it is confined to a mTEC population characterised by high MHC class II expression on its surface in addition to expression of the cluster of differentiation (CD) costimulatory proteins CD80 and CD86.[7, 8] Expression of these proteins in addition to the numerous TRA induced by AIRE expression make these $mTEC^{Hi}$ cells excellent self-antigen presenting cells. In mTEC AIRE localises to the nucleus where it multimerises and forms nuclear bodies, structures composed of chromatin and protein in the nucleus, also known as nuclear dots, and in the cytoplasm where it co-localises with cytoskeletal filaments.[9, 10, 11] In order to be restricted to such a small number of cells, AIRE expression is controlled in a variety of ways, some of which have been delineated. In mTEC H3K4me3 methylation of the histones in the promoter area of *AIRE* relaxes the chromatin structure and allows access to the promoter sequence.[12] In addition, a signalling pathway necessary for AIRE expression is the tumour necrosis factor (TNF) receptor family member Receptor Activator of Nuclear Factor κB (RANK) signalling pathway.[8] RANK Ligand (RANKL) on the surface of other thymic cells activate RANK signalling in a few mTEC, which leads to the activation of the canonical NFκB signalling pathway. The release of the NFκB signalling

complex allows movement of this complex consisting of the NFκB subunits p65 (RelA) and p50 into the nucleus where it binds to the CNS1 enhancer sequence located around 3kbp upstream of *AIRE* and which contains two 10bp RelA binding motifs.[13] There is also evidence of efficient AIRE expression being dependent on a complex of the transcriptional regulators Interferon Regulatory Factor 4 (Irf4), Interferon Regulatory Factor 8 (Irf8), T-Box 21 (Tbx21), and Transcription Factor 7 (Tcf7), in addition to the replacement of the inhibitor CCCTC-Binding Factor (CTCF) with CCCTC-Binding Factor Like (Ctcfl).[14] Another control of AIRE concerns the proper splicing of the *AIRE* pre-mRNA, where the splicing regulator Arginine Demethylase And Lysine Hydroxylase (JMJD6) is necessary for the proper splicing of *AIRE.*[15] Thus, if JMJD6 is not present the *AIRE* mRNA will retain intron 2 with a premature stop codon leading to activation of the nonsense mediated decay (NMD) pathway, and degradation of the mRNA before translation.

AIRE protein

The protein structure of AIRE is composed of a series of protein domains (**Fig. 1**[16, 17]), as well as other motifs of known function. These consist of a Caspase recruitment domain (CARD), a SAND domain (named for the proteins Sp100, AIRE, NucP41/75, and DEAF-1), two plant homeodomain (PHD) zinc fingers, a Nuclear Localisation Sequence (NLS), and four LXXLL sequences.[18, 19, 20] The CARD, previously known in AIRE as the homogeneously staining region (HSR), is important for AIRE oligomerisation.[20, 21] In addition, posttranslational modification of the CARD sequence has been linked to interaction with the inhibitor Bromodomain-containing protein 4 (Brd4) attracting the Positive Transcription Elongation Factor (P-TEFb) that is needed for successful transcription of AIRE regulated genes.[22] The SAND domain is a well-conserved ~80 amino acid sequence found in a range of DNA interacting nuclear proteins.[19] The DNA binding properties of the SAND domain are thought to be linked to a positively charged surface patch containing the conserved amino acid sequence KDWK; however, AIRE's SAND domain instead contains the modified amino acid sequence NKAR. Also, AIRE does not have the Zinc-binding sequence that is present in other SAND domains.[23] The AIRE sequence consists of PHD zinc fingers; however, it is likely that these two PHD domains have different functions within AIRE.[24] PHD1 has been linked to interaction with the tail of histone H3 (**Fig. 2**), specifically H3K4 when this lysine is unmethylated (H3K4me0).[25, 26, 27]

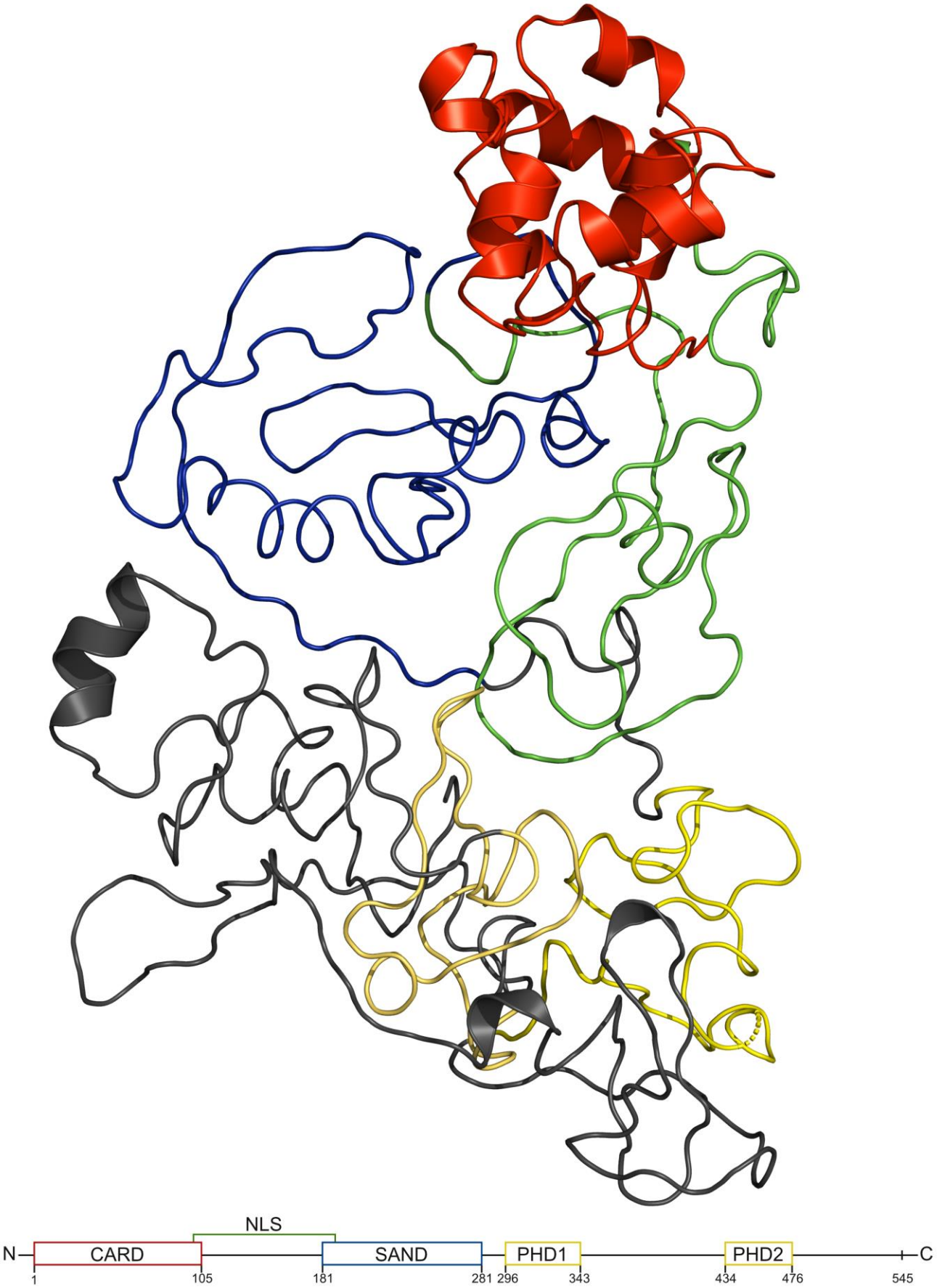


Figure 1 Estimated structure and overview of the AIRE protein with its functional domains. AIRE structure model estimated from homologous sequences using Phyre2[16] (64% of residues with >90% confidence). Model annotated and rendered in PyMOL[17]. CARD domain in red, Nuclear localisation sequence (NLS) in green, SAND domain in blue, and two PHD zinc fingers in yellow. Overview created with Scribus.

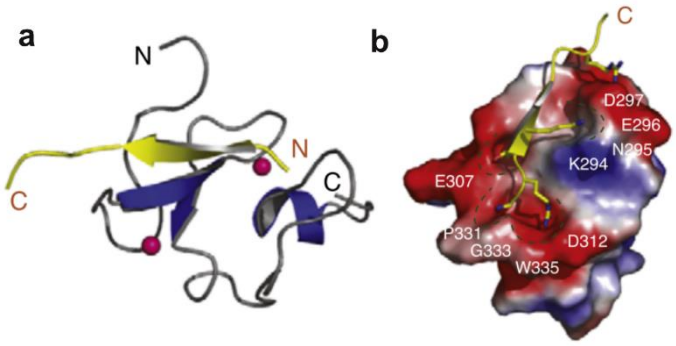


Figure 2 3D representation of NMR structure of AIRE PHD1-Histone H3 interaction. Cartoon representation in **a**, surface representation in **b**, histone sequence in yellow. Surface charge in **b** visualised as negative in red and positive in blue. Image reproduced in part from Chakravarty et. al. 2009 with permission.[27]

Trimethylation of lysine 4 of histone H3 (H3K4me3) is characteristic of transcriptional activation, and so PHD1 recognition of a nonmethylated H3K4 is indicative of PHD1 interaction with inactive genes.[28] While the surface of PHD1 is negatively charged, facilitating interaction with the positively charged histone tail, the PHD2 surface is positively charged, indicating that PHD2 is not involved in histone interactions.[29] Instead, mounting evidence supports the importance of the PHD2 domain in AIRE partner interaction.[24, 29] Deletion of the PHD2 domain has been shown to stop AIRE interaction with known partners with functions related to transcription, chromatin binding and nuclear transport.[24] The NLS is a signal required for the proper transport of large proteins from the cytosol to the nucleus through the nuclear pore complex.[30] The NLS in AIRE is characterised as a monopartite NLS that allows AIRE to be transported after interaction with the adaptor protein importin α.[31] AIRE also contains four distinct sequences with the amino acid pattern LXXLL, which are known nuclear receptor interaction domains.[32]

AIRE protein interactions

AIRE interacts with a variety of proteins, with the majority of its functional domains. AIRE's nuclear receptor interaction domains may attract CREB-binding protein (CBP) and the E1A binding protein p300 (P300), which has been found to induce acetylation of AIRE, reducing transcription of AIRE target genes.[18, 33] This process can be reversed by the deacetylase Sirtuin-1 which seems to counteract the CBP/P300 inhibition of AIRE, and thus be necessary for AIRE activity.[34] While AIRE has been found to target inactive genes marked by H3K4me0, AIRE also interacts with the repressive complex ATF7ip-MBD1, thereby targeting genes actively repressed in the cell.[35] There are many other AIRE interacting proteins important in AIRE target gene transcription, consistent with AIRE interaction in large multimeric protein complexes, and these can be categorised (**Fig. 3**) into proteins that are linked to transcription, pre-mRNA processing, chromatin binding and structure, and nuclear transport.[24, 36] After

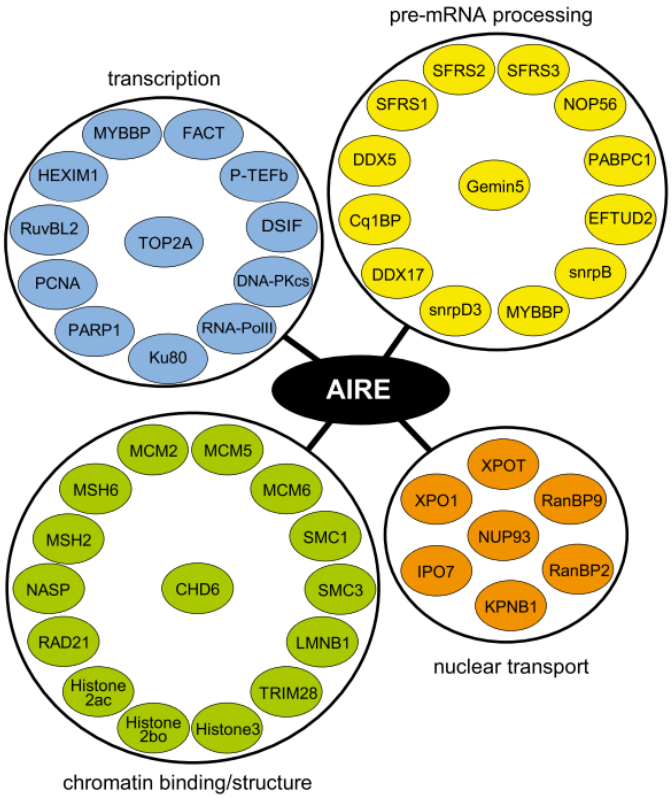


Figure 3 Overview of some AIRE interacting proteins categorised into functional groups. Image taken from Yang et. al. 2013.[24]

AIRE is recruited to inactive or repressed genes marked by H3K4me0 or ATF7ip-MBD1, AIRE is found to migrate over and to co-localise to super-enhancers.[37] Super-enhancers are long chromatin sections marked with H3K27ac and H3K4me1 histone marks that hosts a high number of transcription factors. These sections are possibly working as depots for efficient transcription, looping around to active transcription sites to interact with RNA polymerase II and its preinitiation complex. One of the proteins associated with super-enhancers is the protein Brd4 which binds to AIRE.[37] Brd4 binding leads to recruitment of the positive transcription elongation factor P-TEFb, which is transported to AIRE by the Heterogeneous Nuclear Ribonucleoprotein L (Hnrnpl). [22, 38, 39, 40] P-TEFb is a transcriptional elongation factor that works by phosphorylating stalled RNA-polymerase II, thereby releasing it to continue gene transcription of AIRE target genes.[41, 42] In addition to inducing gene transcription, AIRE association to super-enhancer regions leads to interaction of AIRE with the DNA topoisomerase TOP1.[37] TOP1 introduces single-strand nicks in the DNA, where only one of the DNA strands are cut, and it is possible that these nicks recruit the non-homologues end joining (NHEJ) DNA repair complex DNA-PK, consisting of the protein kinase DNA-PKcs, the Ku80 -Ku70 heterodimer, PARP-1, FACT, and the topoisomerase TOP2. While it is unclear exactly how this complex works together with AIRE, it does introduce single nicks (TOP1) or double-stranded breaks (TOP2) that relaxes the chromatin structure allowing for efficient transcription. Furthermore, it is possible that this works as a histone eviction complex that removes histones in front of the transcribing RNA polymerase II, keeping the transcription complex from slowing down.[37] In addition to initiating and ensuring efficient transcription of AIRE target genes, other partners are important in the efficient processing of

the resulting pre-mRNA. AIRE interacts with a variety of these genes (**Fig. 3**) exemplified by the small nuclear ribonucleoprotein EFTUD2, that has been shown to localise to AIRE containing nuclear bodies.[43] This pre-mRNA processing complex ensures that the pre-mRNA is spliced into mRNA as fast as possible, consistent with the fact that AIRE greatly increases the mRNA levels of its target genes, yet the pre-mRNA levels of those same genes remain low.[43] In addition, the mRNA processing might affect the stability of the transcripts, as the AIRE induced genes have a relatively long half-life compared to other genes.

AIRE induced gene expression

AIRE is responsible for the promiscuous expression of TRA in mTEC of the thymus, giving a very small subset of cells in the thymus one of the broadest gene expression profiles of any cell.[44, 45, 46, 47] These TRA genes are usually specific to distinct tissues, and have a low expression in mTEC when AIRE is inactive.[48] Of the up to 19,293 protein-coding genes expressed in AIRE positive mTEC, AIRE induction is responsible for increased expression of up to 3980.[47] Interestingly, overrepresented among the genes not induced in these cells are the functional gene ontology categories of the olfactory and vomeronasal receptors. The expression profile of mTEC, consisting of induced and repressed genes, is not dependent solely on AIRE, as the introduction of AIRE into cells and cell systems other than mTEC results in different AIRE induced expression profiles.[48, 49] This is consistent with the way AIRE targets inactive, or actively repressed genes, genes that will necessarily differ between cell types.[26, 35, 48] Single cell studies of individual mTEC show that the AIRE induced gene expression differs significantly from cell to cell, in a stochastic manner, meaning that each cell will have different gene expression from each other, with no single cell able to express the full AIRE induced transcriptome.[50] However, the gene expression in a single cell is not entirely stochastic, as genes located close to each other in the genome have a higher chance of being expressed together.[51] The expression of AIRE induced genes while high in each single cell, are often low on a population level, because of their infrequent expression.[47]

Central tolerance

The expression of TRA by the function of AIRE is crucial to the process of central tolerance of thymocytes (developing T-cells) in the adaptive immune system. Central tolerance is the process in which the developing lymphocytes are checked for their ability to bind to the body's own proteins, called self-antigens.[52] The mammalian immune system is divided into two different systems, the passive fast acting innate immune system, and the active slower acting adaptive immune system. The adaptive immune system requires the rearrangement of the genes encoding B- and T-cell receptors in order for the adaptive immune system to recognise unknown pathogens and foreign molecules. However, this extraordinary ability means that a few of these rearrangements might lead to autoimmune cells that recognise the body's own proteins. These cells must, therefore, be identified and eliminated in order to ensure no autoimmune diseases arise. Both developing B-cells and T-cells must undergo central tolerance; however, while B-cells undergo this process in the bone marrow, immature T-cells are transported to the thymus. Immature T-cells undergo a variety of tests in the thymus, where they move gradually from the outer cortex into the medulla (**Fig. 4**).[53] After entering the thymic cortex through blood vessels from the bone marrow, immature thymocytes are double negative (DN) for the molecules CD4 and CD8. In their journey through the cortex, these cells go through multiple stages (DN1-4) in which their T-cell receptors (TCR) are rearranged, and they end up expressing both the CD4 and

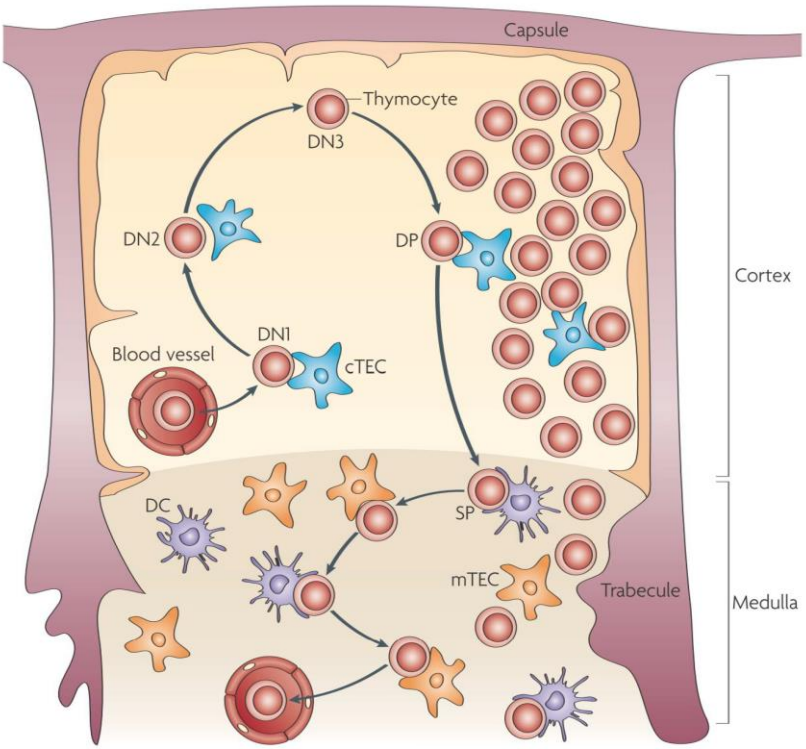


Figure 4 The path taken by an immature thymocyte through the thymus. Immature thymocytes from the bone marrow enter the cortex through blood vessels, then mature from CD8/CD4 double negative (DN) to double positive (DP) T-cells. The double positive T-cells get checked for their ability to bind MHC molecules in positive selection by cTEC, then their inability to bind the body's own proteins in negative selection by AIRE expressing mTEC. Image reproduced from Klein et. al (2009) with permission.[53]

CD8 molecules as double positive (DP) T-cells. First, these thymocytes are checked for their ability to present TCR on their surface, signifying a successful receptor rearrangement. Secondly, the thymocytes undergo the process of positive T-cell selection, in which their ability to bind to and recognise the antigen presenting major histocompatibility complex (MHC) class I or II receptors on the surface of antigen-presenting cells are checked by binding MHC molecules on the surface of cortical thymic epithelial cells (cTEC). The majority of thymocytes are eliminated in this step, where they die by neglect after not receiving certain survival signals. In the process of positive selection, surviving cells also differentiate into single positive (SP) T-cells by becoming either CD4 positive T-helper cells or CD8 positive cytotoxic T-cells depending on increased binding affinity to MHC class II and I respectively. Finally, if the thymocytes survive the positive T-cell selection, they migrate into the medulla of the thymus where they undergo the process of negative T-cell selection. This process is where AIRE is a crucial component, and consists of checking the thymocytes in their ability to bind MHC receptors presenting peptides from the body's own TRA on the surface of mTEC. T-cells that bind too strongly to the MHC presenting a self-peptide undergo apoptotic deletion, or gets turned into the immunosuppressive CD4, FOXP3 and CD25 positive regulatory T-cells (Tregs)[54], while the T-cells with a low or intermediate affinity for the MHC/TRA complex are allowed to migrate out of the thymus into peripheral tissues.

APS-1

Mutations in *AIRE* leading to a failure of negative T-cell selection is the cause of the autoimmune disorder APS-1 also known as autoimmune polyendocrinopathy candidiasis ectodermal dystrophy (APECED).[2] APS-1 (OMIM 240300) is a monogenic disease, with predominantly autosomal recessive inheritance, although dominant forms have also been reported.[55, 56, 57] The prevalence of APS-1 is estimated to be 1:100 000 on a worldwide basis, while the prevalence in Norway is around 1:90 000. The prevalence peaks in certain populations such as amongst Persian Jews (1:9 000), Sardinians (1:14 000) or Finns (1:25 000).[58, 59] APS-1 is characterised by the three major manifestations of Addison's disease, hypoparathyroidism and chronic mucocutaneous candidosis, although other symptoms vary significantly amongst patients.[60] Addison's disease is caused by low or absent production of the steroid hormones cortisol and aldosterone, causing patients to become fatigued, causing the skin to darken, leading to weight loss, and the desire to consume salt.[59]

Hypoparathyroidism is defined by low production of parathyroid hormone (PTH), low calcium but high phosphate in the blood, leading to muscle cramps, grand mal seizures and clumsiness.[59] Finally, chronic mucocutaneous candidosis is chronic infection with the yeast *Candida albicans* typically as an infection of the mouth, but it can also spread to the throat, intestines and fingernails.[59] Symptoms include soreness of the corner of the mouth, problems in consuming acidic or spicy food, trouble swallowing if the infection spreads to the throat, abdominal pain, diarrhoea and flatulence if it spreads to the intestines. Candidosis is usually the first symptom to appear, followed by hypoparathyroidism and then Addison's disease, although not all patients will contract all three disease aspects.[59, 61] Addison's disease and hypoparathyroidism can both prove fatal if not diagnosed and treated in time, such treatment mainly consisting of hormone replacement therapy. Disease onset is usually during childhood and adolescence, though milder dominant forms may have a later onset.[57, 59] Other common manifestations are type-1 diabetes mellitus, ovarian insufficiency, enamel hypoplasia (deficient enamel of the teeth), alopecia (loss of hair), keratitis (inflammation of the eye), and vitiligo (loss of pigment in the skin).[59, 62]

AIRE mutations

Over 100 mutations in *AIRE* have been reported to cause APS-1 (www.hgmd.cf.ac.uk), with the major Finnish mutation R257X and the 13 base-pair deletion C322del13 being the most prevalent. R257X is a recessive nonsense mutation changing from an arginine to a premature stop codon within the SAND domain, leading to a loss of both PHD zinc fingers (**Fig. 5**).[63] The 13 base pair deletion is located within the first PHD zinc finger, disrupting it and introducing a premature stop codon leading to a truncated protein without the second PHD zinc finger.[63] Some examples of rare AIRE mutations are the dominant negative mutation C311Y, leading to a disruption of the zinc-binding structure of the first PHD zinc finger, and the splice mutation c.879+1G>A, causing a truncation of the SAND domain.[57, 58] *AIRE* mutations found

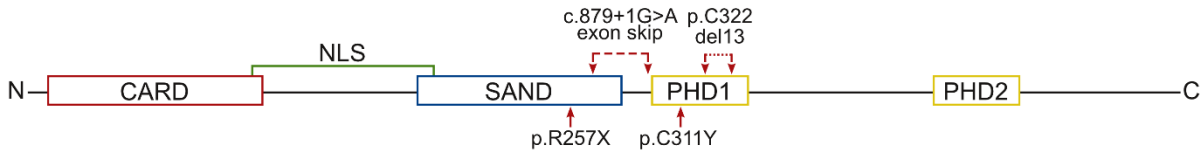


Figure 5 Overview of the AIRE protein with its functional domains and some known mutations. Mutations shown are the recessive R257X major Finnish mutation, the exon skipping 879+1G>A, the dominant C311Y, and the 13 base-pair deletion C322del13. Graphic created with Scribus.

in patients are spread throughout the gene sequence, although the majority of them are located within the CARD, and the two PHD zinc-finger sequences.[64] Curiously, most dominant mutations have so far been located in exon 8, at the start of the first PHD domain.

Deep mutational scanning assay

While the traditional way to discover disease-causing mutations is by sequencing of affected patients after clinical diagnosis has been suspected or confirmed, new techniques make it possible to investigate the negative aspects of any possible mutation in vitro. Using a technique called deep mutational scanning, a library of gene variants can be synthesised, transfected into a cell population, which is then screened for functional effect and sequenced (**Fig. 6**).[65, 66] The first step of this process is the creation of a library of gene variants containing single base substitutions for any locus in the gene using saturation mutagenesis. One method of saturation mutagenesis is to use a programmable microarray with subsequent PCR amplification such as in the technique mutagenesis by integrated tiles (MITE).[67] MITE consists of the creation of a library of short sequence tiles flanked by adaptor sequences, which are synthesised with one single base difference from the consensus sequence. These tiles are subsequently inserted into a plasmid containing the rest of the investigated gene. The plasmids are then transfected into a cell line, in a concentration ensuring that the majority of cells will only contain a single plasmid. The next step of the deep mutational scanning process is the classification of cell populations depending on the gene activity. Genes coding for enzymes with clear delineated substrates are the easiest to investigate in this manner; however, many genes will need to be investigated using indirect reporter genes.[67] Using fluorescence-activated cell sorting (FACS) with fluorescent antibodies targeting the proteins of these reporter genes, populations with lower target gene activity can be isolated, and subsequently sequenced to identify the underlying mutation.[66, 68] Because of the high volume of sequencing, only a massively parallel high-throughput sequencing platform is sufficient. Sequencing accuracy can be mitigated with a high number of reads such that individual read errors are eliminated as a consequence of the high volume of the consensus sequence. [69]

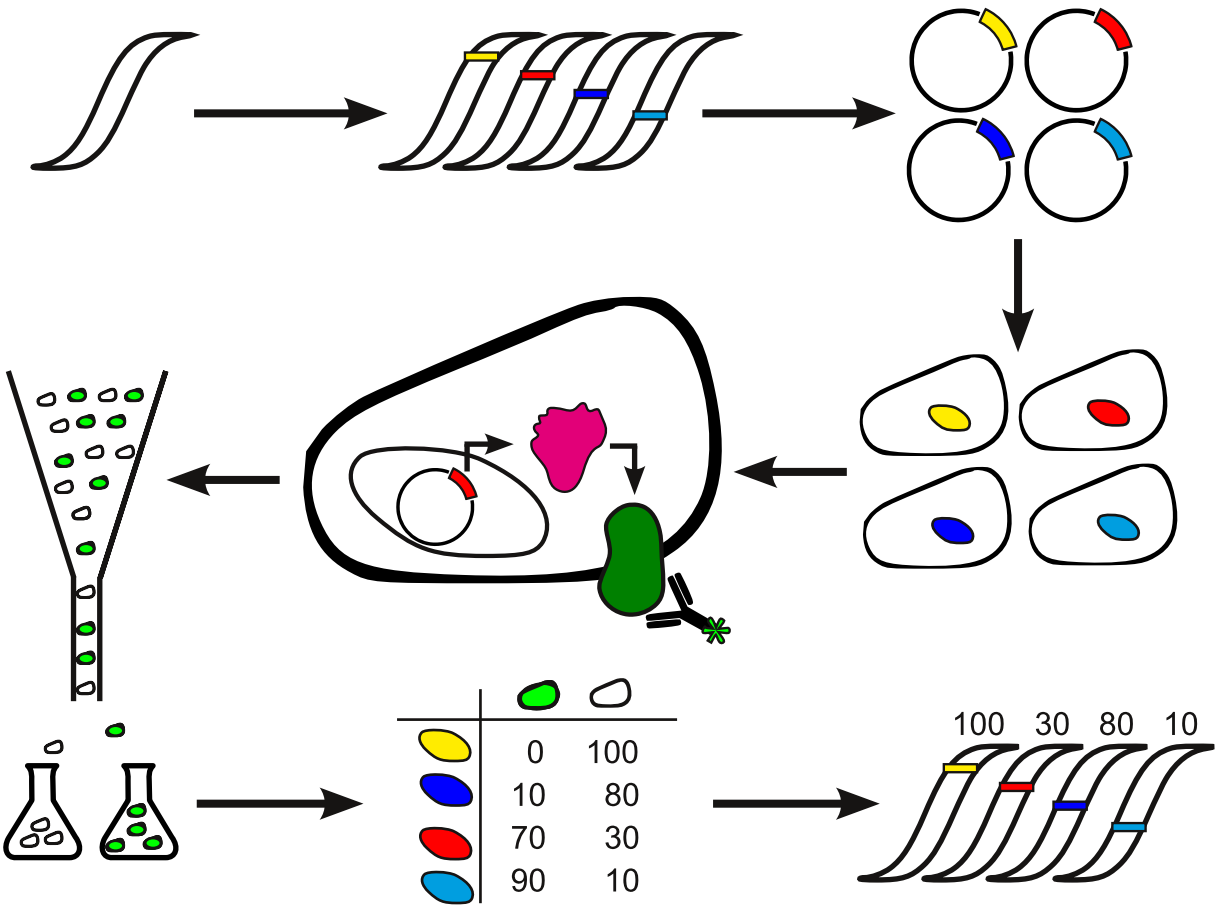


Figure 6 Overview of a general approach to deep mutational scanning. First a library is created of all possible mutations in the gene of interest. These mutated genes are inserted into plasmids and transfected into cells that are subsequently grown in culture. The mutated genes will express the protein of interest and the protein will induce expression of a reporter protein, depending on the severity of the mutation in the gene. A fluorescent antibody detects the presence of the reporter protein and FACS sorts the populations. The cells are then sequenced and the mutations are ranked in their severity on the basis of the number of cells expressing the reporter protein. Graphic created with Scribus, Inkscape and GIMP.

Methods of reporter gene discovery

In order to determine possible reporter genes, genes reported previously in the literature can be evaluated using real-time quantitative polymerase chain reaction (qPCR), or transcriptomes of cells can be investigated using RNA sequencing (RNAseq) to identify new reporter genes. qPCR is a technique that uses complementary DNA (cDNA) synthesised from messenger RNA (mRNA) in a PCR reaction that can be tracked using fluorescent probes that are specific to individual genes.[70] As the PCR reaction proceeds, the number of cycles a gene probe needs to reach a particular threshold level of fluorescence can quantify the original sample concentration. qPCR is a highly sensitive technique; unfortunately, because it requires specific probes, this technique cannot be used to find reporter genes without prior

knowledge and is limited to a small number of gene probes for each experiment. In attempting to discover new reporter genes, RNAseq can be used to investigate differences in gene expression between cell populations.[71] Compared to older methods like microarray, RNAseq has the advantage of discovering genes in an unbiased manner without preconceptions, although it is not as sensitive to weakly expressed genes as qPCR because of limited read depth.[71, 72] RNA samples used in RNAseq is first isolated from a cell population, and then either depleted according to some criteria (rRNA depletion or mRNA selection) or used with all RNA available. The resulting RNA is then converted into cDNA and fragmented into small sequences with flanking adaptor sequences.[72] These fragments are then sequenced according to the sequencing technology of choice. A good reporter gene in a multiplexed reporter assay is a gene with a substantial difference between populations as quantified by fold change (FC), but also high expression in absolute terms. Also, a gene that codes for a secreted protein does not work well when using FACS, where a membrane-bound protein with extracellular epitopes would be preferable. Selected candidate reporter genes can be evaluated using qPCR with populations transfected with either the wildtype or mutants with known disease-causing effect. If the possible reporter genes are downregulated in the populations transfected with the mutated gene, it may be used as a reporter gene if its protein is also detectable. Evaluation of the reporter proteins can be performed using Western blot and flow cytometry. While Western blot is able to accurately detect the presence of the reporter from a protein lysate, regardless of its intracellular localisation, it can only indirectly be used to quantify the expression of the reporter protein. Flow cytometry, on the other hand, can accurately quantify the number of cells that express the reporter but require different protocols depending on the cellular localisation.[73] The ability of flow cytometry to also combine different antibodies enables it to find subpopulations, such as cells that have been successfully transfected and that express one or more reporter proteins. In addition, flow cytometry uses the same underlying technology as FACS, a lamellar flow where cells are transported one by one past a detector sensing fluorescence, size or complexity.[73] The main difference between FACS and flow cytometry is the ability of FACS to separate the cells into different containers depending on their fluorescence after detection.[68]

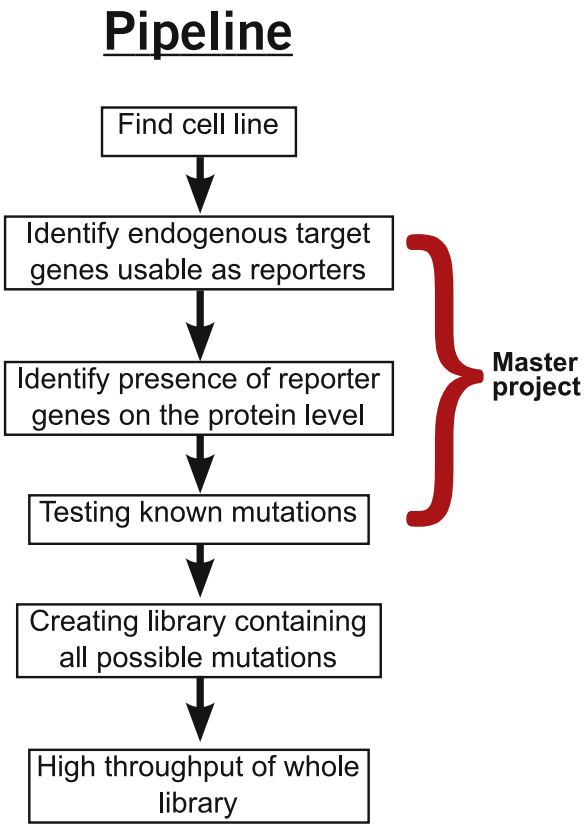


Figure 7 Pipeline for the functional screening assay using deep mutational scanning. Scope of master project outlined in red. Graphic created with Scribus.

Study approach

As a means to create a deep mutational scanning assay for all possible mutations in *AIRE,* we have searched for possible reporter genes that can be used to detect the functional effect of *AIRE* mutations. The deep mutational scanning pipeline is outlined in **Figure 7**, with the steps within the scope of this project outlined in red. HEK293FT embryonic kidney cells were transfected with a FLAG and MYC tagged *AIRE*, RNA was isolated, and the transcriptome investigated using qPCR and RNAseq. Genes previously mentioned as AIRE dependent in the literature and new reporter gene candidates identified using RNAseq were investigated using qPCR. As a selection strategy, the clinically relevant *AIRE* mutants R257X and C311Y were generated and used to check for downregulation of possible reporter genes. In addition, we evaluated the viability of RNAseq in identifying weakly expressed genes, compared RNAseq and qPCR sensitivity and compared various RNAseq library preparation methods.

2. Aims

Mutations in AIRE, a unique transcriptional regulator, lead to the rare autoimmune disorder APS-1. In order to better our understanding of AIRE, and inform larger sequencing and GWAS efforts, our group is developing a multiplexed screening assay to characterise the functional effect of any mutation in *AIRE*. This screening assay uses a deep mutational scanning approach, which has previously been used in our group to investigate monogenic diabetes.

In the process of developing this functional screening assay, a method to check for any functional impact of mutations in *AIRE* needs to be developed. Because of the nature of AIRE, it cannot directly be investigated using its binding to a discrete promoter sequence or one clear downstream effector. Instead, AIRE induces thousands of genes, some of which may be usable as indicators of AIRE function.

The overall aim of this project has been to develop methods to identify possible reporter genes amongst the AIRE induced genes that would be usable in a multiplexed functional screening assay. To that effect our main aim can be divided into multiple objectives:

- Developing a robust cell system with AIRE expression. This cell system needs to be amenable to large-scale transfection and FACS sorting, and be robustly expressing functional AIRE proteins.

- Evaluate the cell system for AIRE functionality, investigate the activity of known AIRE reporter genes, and evaluate the usability of these genes as possible reporters.

- Developing a protocol for AIRE inducible gene discovery using RNA sequencing, and evaluate different methodological RNAseq approaches.

- Comparing the performance of different RNAseq approaches to each other and compare them to the established method of qPCR.

3. Materials and Methods

Reagents

All reagents were purchased from Thermo Fisher if not otherwise stated.

Plasmid amplification and mutagenesis

In order to express AIRE and AIRE mutants in cell lines, a variety of plasmids were used in this work. A pCMV plasmid containing Flag and Myc tagged *AIRE* wildtype was purchased from OriGene (OriGene Cat#: RC213497). This plasmid was used for all *AIRE* Wt transfections and used as a base for creating the *AIRE* mutations R257X and C311Y as described below. To have a positive control of the transfection a plasmid containing Flag and Myc tagged *21-Hydroxylase* (Origene Cat#: RC216416) was used. Plasmids were amplified using TOP10 competent *Escherichia coli* cells from Thermo Fisher and purified using a QIAprep Spin Miniprep Kit from Qiagen. Mutations were then created using the QuikChange II Site-Directed Mutagenesis kit from Agilent Technologies (Cat#: 200524-5). The R257X mutation was created using mutagenesis primers from Eurogentec with the forward sequence GAA-GCC-TCT-GGT-TTG-AGC-CAA-GGG-AG and the reverse sequence CTC-CCT-TGG-CTC-AAA-CCA-GAG-GCT-TC. Similarly, the C311Y mutation was created using mutagenesis primers from Eurogentec with the forward sequence GAG-CTC-ATC-TGC-TAT-GAC-GGC-TGC-CC and the reverse sequence GGG-CAG-CCG-TCA-TAG-CAG-ATG-AGC-TC. The *AIRE* and mutated *AIRE* containing plasmids were confirmed to be accurate using Sanger sequencing with the Applied Biosystems 3730 DNA analyser.

Cell culture and transfection

HEK293FT human embryonic kidney cells (RRID: CVCL_6911) were grown in a medium consisting of Dulbecco's Modified Eagle Medium (Cat#: 31966-021) supplemented with 4.5 g/l D-Glucose, Pyruvate, 10% (v/v) Fetal Bovine Serum (FBS), and 1% (v/v) Penicillin-Streptomycin. The cells were incubated at 37°C in a 5% CO2 humidified incubator until reaching 80-100% confluency. The cells were subsequently counted using a Countess cell counter from Thermo Fisher and transferred to 6-well plates, where each well was seeded with 6×10^5 cells. After 24h, the cells were transfected using the Lipofectamine 2000 Transfection Reagent from Thermo Fisher with 2.5µg DNA and 12µl Lipofectamine. The cells were either transfected with no DNA present (denoted as empty) or with the plasmids

previously described. Transfected cells were allowed to grow for 48 hours before being harvested.

Western immunoblotting

The transfected cells grown for 48 hours were removed from the growing surface by flushing, centrifuged at 300g for 7 minutes, then resuspended in Dulbecco's Phosphate Buffered Saline (PBS) purchased from Merck (Cat#: D8537). The cells were then centrifuged a second time before the cell pellet was resuspended in cOmplete lysis buffer from Merck (Cat#: 04719956001) and lysed for 30 minutes on ice. After lysing, the cellular debris was spun down using a microcentrifuge at 21130g for 10 minutes, before lysates were pipetted off. 13µl of each sample lysate was mixed with 5µl 4X NuPage LDS sample buffer from Merck (Cat#: NP0007), and 2µl 10X NuPage Sample Reducing Agent (Cat#: NP0009). The protein lysates were then heated at 70°C for 10 minutes, before being applied on a NuPage 10% Bis-Tris gel (Cat#: NP0301BOX). In addition to the various samples, the SeeBlue Plus2 Pre-Stained Protein Standard (Cat#: LC5925) was applied on the gel. The Lithium dodecyl sulphate polyacrylamide gel electrophoresis (LDS-PAGE) was performed using 20X NuPage MOPS running buffer (Cat#: NP0001) diluted in Milli-Q water to 1x for the main chamber, and 1x MOPS with added 2.5% (v/v) 1X NuPage antioxidant (Cat#: NP0005) for the inner chamber. The electrophoresis was accomplished using 180V for 1 hour and 10 minutes. Afterwards, the proteins in the gel were transferred to a polyvinylidene difluoride (PVDF) membrane using the iBlot dry blotting system from Thermo Fisher (Cat#: IB401002) by using program 3 for 7 minutes. The membrane was cut into pieces according to the antibodies used, then washed three times in Tris Buffered Saline (TBS) with 0.1% (v/v) Tween20 (TBS-T) for 5 minutes each. After the wash, the membranes were blocked for 1 hour on a shaker in 5% (w/v) Blotting-Grade Blocker milk powder from Bio-Rad (Cat#: 1706404) in TBS-T. After blocking, the membranes were again washed using the previous method, before being incubated with primary antibody in TBS-T with 5% (w/v) Bovine Serum Albumin (BSA) overnight on a shaker in a cold room. The primary antibodies used were mouse α-GAPDH (Cat#: MAB374, Merck) in a 1:500 dilution, goat α-GAPDH (Cat#: SC-48167, Santa Cruz) in a 1:500 dilution, mouse α-DDK/FLAG (Cat#: TA50011/OTI4C5, OriGene) in a 1:2000 dilution, mouse α-Myc (Cat#: R950-25, Thermo Fisher) in a 1:1000 dilution, goat α-AIRE (Cat#: PAB7040, Abnova) in a 1:1000 dilution, mouse α-KRT14 (Cat#: sc-53253 AF647, Santa Cruz)

in a 1:500 dilution, and mouse α-S100A8 (Cat#: AM31838FC-N, OriGene) in a 1:500 dilution. After probing with primary antibodies, the membranes were rewashed, before being incubated with the secondary antibodies in TBS-T with 5% (w/v) BSA for 1 hour on a plate shaker. A goat α-Mouse antibody (Cat#: 626520, Thermo Fisher) conjugated with Horse Radish Peroxidase (HRP) in a 1:2000 dilution was used for the membranes probed with mouse antibodies, while a rabbit α-Goat (Cat#: 611620, Thermo Fisher) antibody conjugated with HRP in a 1:2000 dilution was used against the membranes probed with goat antibodies. The membranes were rewashed using the previous method, then soaked in Pierce ECL Western Blotting Substrate (Cat#: 32106) and imaged using a Bio-Rad Chemidoc.

Flow cytometry

Transfected cells grown for 48 hours were harvested by flushing, centrifuged at 300g for 5 minutes, and resuspended in PBS (Cat#: D8537). The cells were fixed and permeabilised using a BioLegend True-Nuclear Transcription Factor Staining buffer set (Cat#: 424401) according to the True-Nuclear Transcription Factor Staining Protocol for 5ml tubes. Cells were stained using antibodies consisting of rat α-FLAG/DDK conjugated to APC from BioLegend (Cat#: 637307) in a 1:40 dilution, in a Cell staining buffer, consisting of 5% (v/v) FBS in PBS. The cells were subsequently analysed using a BD Biosciences Accuri C6 flow cytometer. The 640nm red laser was used to excite the APC conjugated antibodies, while the signal from these was detected using a 675nm/25 bandpass filter and the FL4 detector. Results were analysed using the FlowJo v10.2 software.

RNA isolation

Transfected cells grown for 48h were harvested by flushing, then transferred to 15ml tubes where they were centrifuged at 300g for 7 minutes before being resuspended in cold PBS. RNA was isolated using Qiagen RNeasy Mini Kit (Cat#: 74106) according to the RNeasy Mini Handbook. A Qiagen QIAshredder spin column (Cat#: 79656) was used to homogenise the samples, and a Qiagen RNase free DNase solution (Cat#: 79254) was used in order to digest any genomic DNA. Afterwards, the RNA samples were stored at -80°C. A NanoDrop microvolume spectrophotometer from Thermo Fisher was used to check the RNA concentration. In order to be sure that high-quality RNA was used, the samples were also analysed using an Agilent Bioanalyzer 2100 with an Agilent RNA 6000 Nano kit (Cat#: 5067-1511).

Real-time quantitative PCR

Previously isolated RNA was analysed using qPCR in a two-step process. The first step consisted of using a Superscript VI VILO cDNA Synthesis Kit with added ezDNase (Cat#: 11766050) to turn mRNA into cDNA according to the Reverse transcription protocol for SuperScript IV VILO Master Mix with ezDNase enzyme. In order to maximise genomic DNA digestion of the ezDNase process, 5 minutes incubation time was used. To analyse the cDNA a variety of TaqMan probes and a TaqMan Universal PCR Master Mix (Cat#: 4304437) were used according to the TaqMan Universal PCR Master Mix User Guide and subsequently analysed using an Applied Biosystems 7900HT Fast Real-Time PCR System and the SDS 2.3 software. Three technical controls for each of three biological replicates were used, in addition to no template control (NTC) and no reverse transcriptase control (-RT). TaqMan probes were purchased from Thermo Fisher and consisted of probes for the genes *GAPDH* (Cat#: Hs99999905_m1), *AIRE* (Cat#: Hs00230829_m1), *CCNH* (Cat#: Hs00236923_m1), *IGFL1* (Cat#: Hs01651089_g1), *KRT14* (Cat#: Hs00265033_m1), *S100A8* (Cat#: Hs00374264_g1), *IL10RA* (Cat#: Hs00155485_m1), *SLC3A2* (Cat#: Hs00374243_m1), *ERMAP* (Cat#: Hs00367924_m1), *ARMC5* (Cat#: Hs01000278_m1), *SLC7A11* (Cat#: Hs00921938_m1), *INHBE* (Cat#: Hs00368884_g1), and *CHAC1* (Cat#: Hs00225520_m1). The qPCR results were processed using the ΔΔCT method in Microsoft Excel 2016, and normalised against the *GAPDH* housekeeping gene, then fold change was calculated between the transfected cell populations and empty vector.

RNAseq

RNA previously isolated was delivered to the Genomics Core Facility at the Department of Clinical Science of the University of Bergen, and RNAseq was performed by them. Two different library prep kits were used, the Illumina TruSeq Stranded Total RNA Library Prep GOLD kit (Cat#: 20020599) and the Illumina TruSeq Stranded mRNA Library Prep kit (Cat#: 20020595), together with the Illumina TruSeq RNA CD Index Plate index adapters (Cat#: 20019792). Sequencing was performed with three biological replicates using an Illumina HiSeq 4000 sequencer, with a read depth of approximately 100 million reads per sample. After sequencing, the resulting fastq files were analysed using the FastQC software to check the quality of the data.[74] In order to clean the dataset and remove the highly transcribed (up to 20%) *AIRE* reads, the aligner bowtie2 was used with the *AIRE* cDNA sequence from the

plasmid, and the resulting non-aligned reads were retained in new fastq files.[75] In bowtie2 the options --local --phred33 and --un-conc-gz was used. To align the reads from the fastq files to the GRCh38.p10 human reference transcriptome the pseudoaligner Kallisto was used with the default options.[76] After alignment, the transcript alignment was imported and summarised into gene alignment using the tximport R-package, and then subsequently annotated with gene names using the R-package EnsDb.Hsapiens.v86.[77, 78] The resulting dataset was analysed for differential expression using the DESeq2 R-package with default options.[79] In order to visualise the dataset in volcano plots and histograms, the R-package ggplot2 was used.[80] To investigate the gene ontology of the differentially expressed genes, the Gene Ontology enRIchment anaLysis and visuaLizAtion tool (GOrilla) was used to generate enriched gene ontology data, while REVIGO was used for visualisation.[81, 82]

4. Results

Confirming the expression of AIRE and AIRE mutants

In order to identify AIRE activity reporter genes usable in a deep mutational scanning assay and to study the transcriptome of AIRE expressing cells, human HEK293FT embryonic kidney cells were transfected with plasmids containing *AIRE* and *AIRE* mutants. The cells were either left untransfected (Empty), transfected with pCMV6 plasmids containing the FLAG and MYC tagged *AIRE* wildtype (Wt) or transfected with the same plasmids with the recessive nonsense mutation R257X or the dominant missense mutation C311Y. Two methods were performed to confirm the successful transfection and AIRE expression, western blot and flow cytometry. Western blot confirms the expression of the Wt with a distinct band in the 64 kDa region using both anti-FLAG and anti-AIRE antibodies (**Fig. 8**).

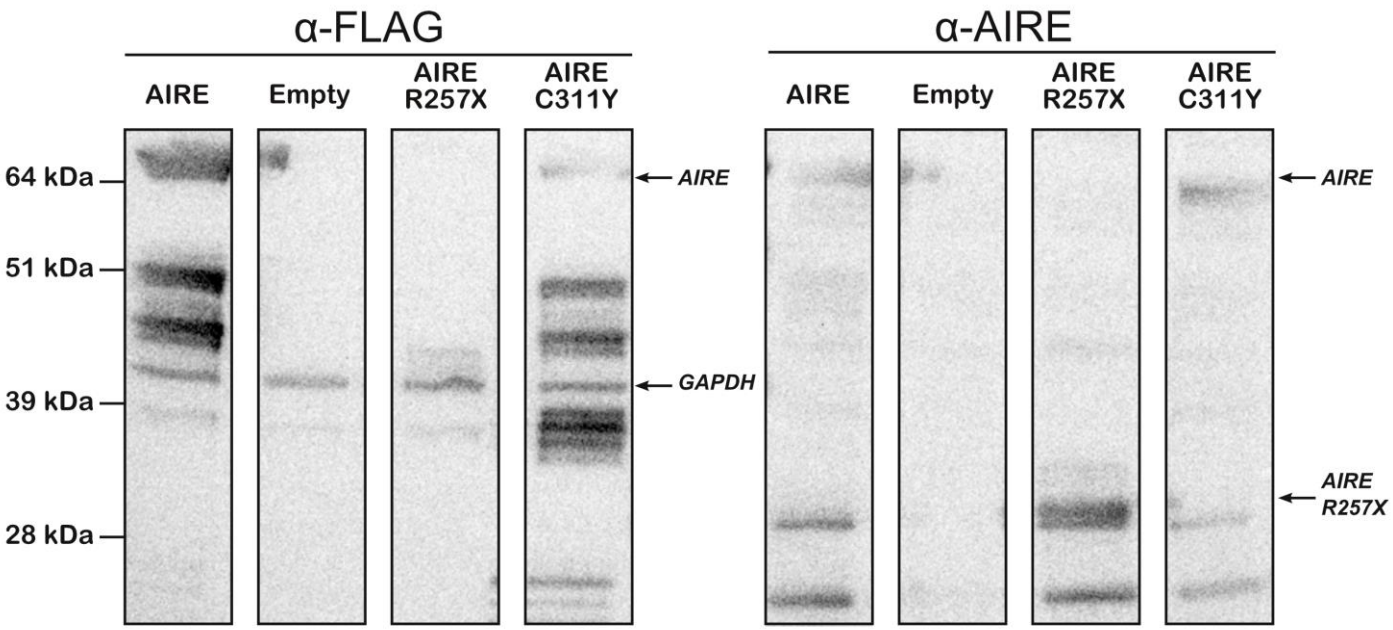


Figure 8 Confirmation of wildtype and mutant AIRE expression using Western blot. Lysates isolated from untransfected HEK293FT cells (Empty) or cells transfected with *AIRE* wildtype or the mutants R257X and C311Y. Both anti-FLAG and anti-AIRE antibodies were used in order to confirm AIRE bands, and to detect the truncated R257X mutant which lacks the FLAG tag. Skewed blot is responsible for any marking at the top of the Empty cell lysate. Detection of GAPDH in the anti-flag part of the blot only, was used as a loading control. The image is representative of two independent experiments.

Although slightly larger than the consensus ~57 kDa, the *AIRE* construct also contains the C-terminal recognition tags FLAG and Myc, increasing its size. In addition to this band, some bands appear in the AIRE Wt fraction using the C-terminal FLAG antibody, and two bands appear when using the N-terminal AIRE antibody, suggesting some degradation of AIRE. The population transfected with the AIRE R257X mutant shows no bands using the FLAG

antibody, yet gets a band at approximately 30 kDa unique to this population with the AIRE antibody. This size is consistent with the R257X mutation as this mutation is a nonsense mutation that leads to a truncated AIRE protein, with a theoretical molecular weight of 27.7 kDa (estimated using ProtParam[83]). The premature stop codon is also responsible for the lack of a FLAG sequence in this *AIRE* mutant. The cell population transfected with the C311Y mutant shows a weak band at the 65 kDa position, with the majority of AIRE in various bands between ~35-50 kDa, concordant with considerable degradation of this mutant.

Flow cytometric analysis of cells transfected with *AIRE* Wt or the positive control *21-hydroxylase* (21OH) in a similar vector shows an increase of 40.8% and 55.5% FLAG-positive cells respectively, compared to untransfected cells (**Fig. 9**) when using anti-FLAG primary antibodies conjugated to the fluorochrome APC. Initial gating for cells can be seen in **a**, while **b** shows the gating for FLAG-positive cells, and **c** shows a histogram overview of the same FLAG gate. Transfection efficiency of AIRE Wt has subsequently increased to ~65% while analysis of the C311Y mutant shows a comparable increase in 66% of FLAG-positive cells compared to the untransfected population (**Fig. 10**).

Confirming AIRE activity using qPCR of known AIRE-responsive genes

Having confirmed AIRE expression with both western blot and flow cytometry, AIRE activity was investigated using qPCR. To find AIRE reporter gene candidates using RNAseq, AIRE does not only need to be present but actively inducing target gene expression in cell populations. From the literature, we selected AIRE-induced genes *Insulin Growth Factor-Like Family Member 1* (*IGFL1*), *Keratin 14* (*KRT14*) and *S100 Calcium Binding Protein A8* (*S100A8*) as activity probes, while *Cyclin H* (*CCNH*) was selected as a negative control not affected by AIRE. The housekeeping gene *Glyceraldehyde-3-Phosphate Dehydrogenase* (*GAPDH*) was used to normalise expression among populations. Substantial *AIRE* overexpression was found, with a fold change increase of around 250 thousand times in the *AIRE* Wt and the C311Y mutant transfected cell populations compared to untransfected cells, and around 550 thousand times increased expression in the R257X mutant (**Fig. 11**). The AIRE un-regulated control gene *CCNH* shows a fold change hovering around 1 for all populations, though slightly higher in the two mutant populations.

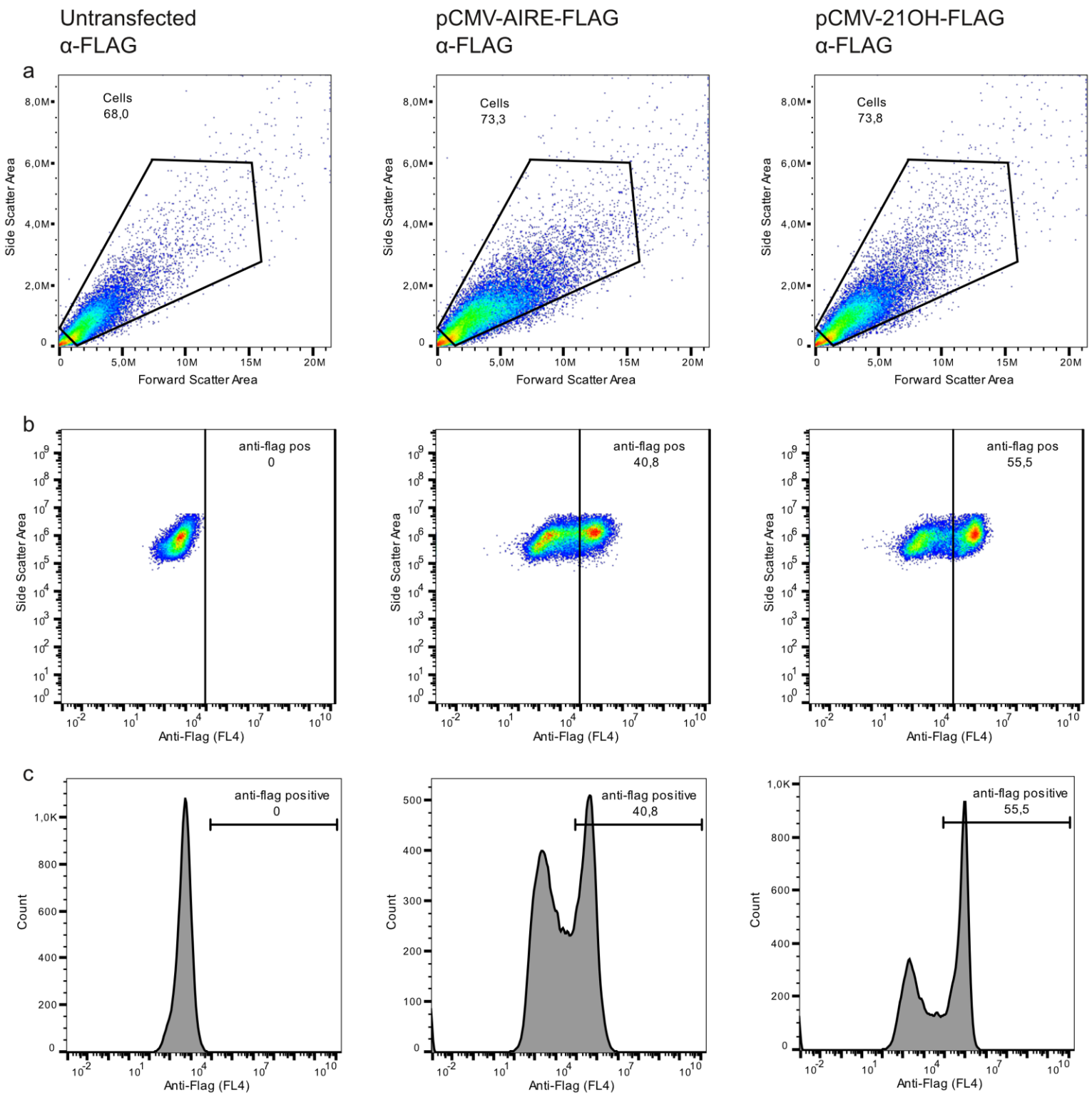


Figure 9 Confirmation of AIRE expression using flow cytometry analysis. Untransfected cells, or cells transfected with *AIRE* Wt or the positive control *21-hydroxylase* (*21OH*) were permeabilised and stained with anti-FLAG antibodies. Initial gating strategy shown in **a**, while **b** shows gating for the presence of the FLAG tag (APC/FL4), either on AIRE or 21OH. Histogram of gating for FLAG positive cells shown in **c**. Transfection efficiency estimated as ~41% for *AIRE* and ~56% for *21OH*.

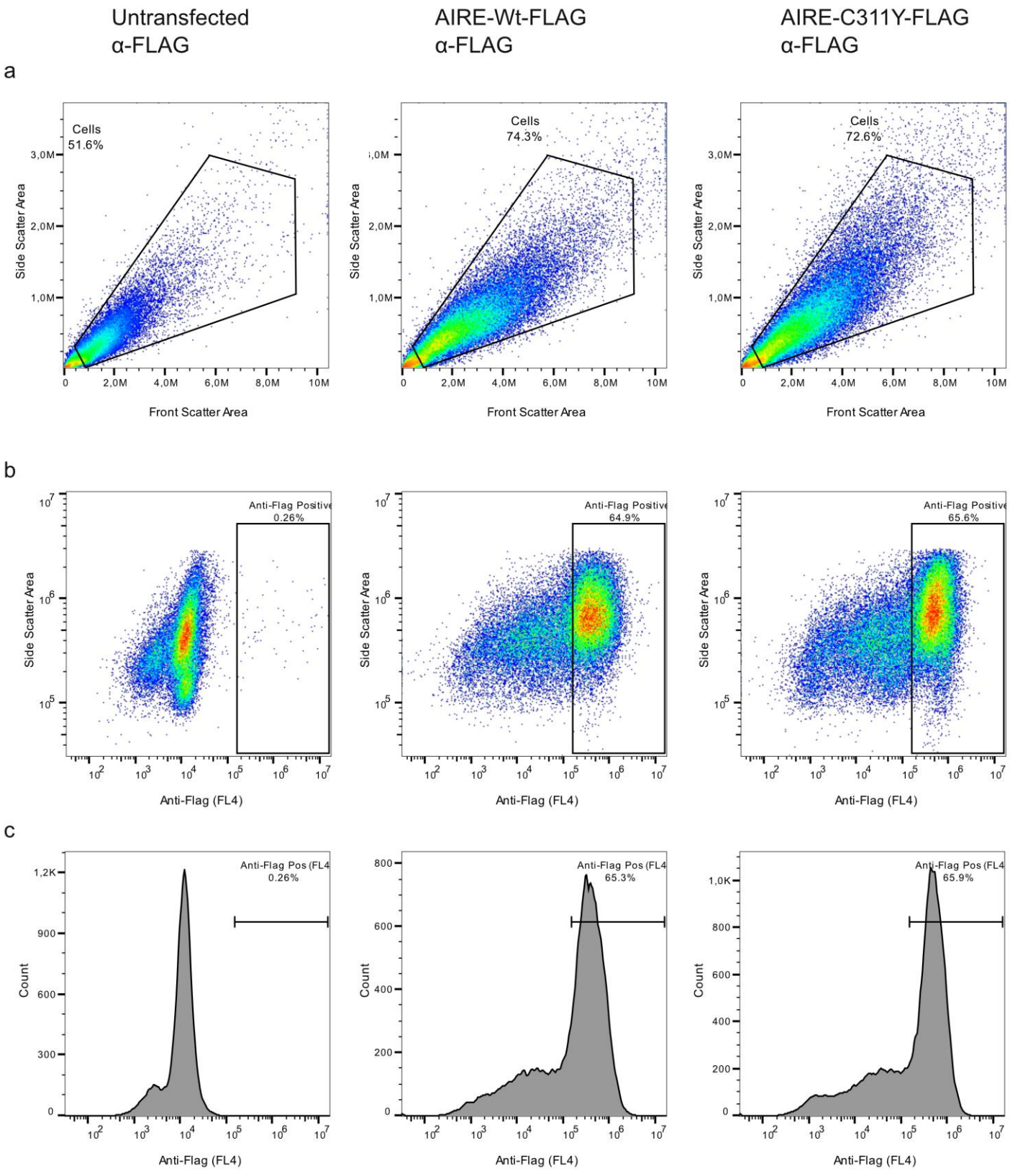


Figure 10 Confirming expression of the *AIRE* mutant C311Y using flow cytometry. Untransfected HEK293FT cells or cells transfected with *AIRE* Wt or the C311Y mutant permeabilised and stained using anti-FLAG antibodies. Initial gating strategy shown in **a**, while gating for the presence of FLAG (APC/FL4) shown in **b**, histogram of gating for the FLAG positive population shown in **c**. Transfection efficiency estimated as ~65% for the wildtype and ~66% for the C311Y mutant. Flow cytometry analysis of the R257X mutant using anti-FLAG antibodies was not possible because of its truncated nature.

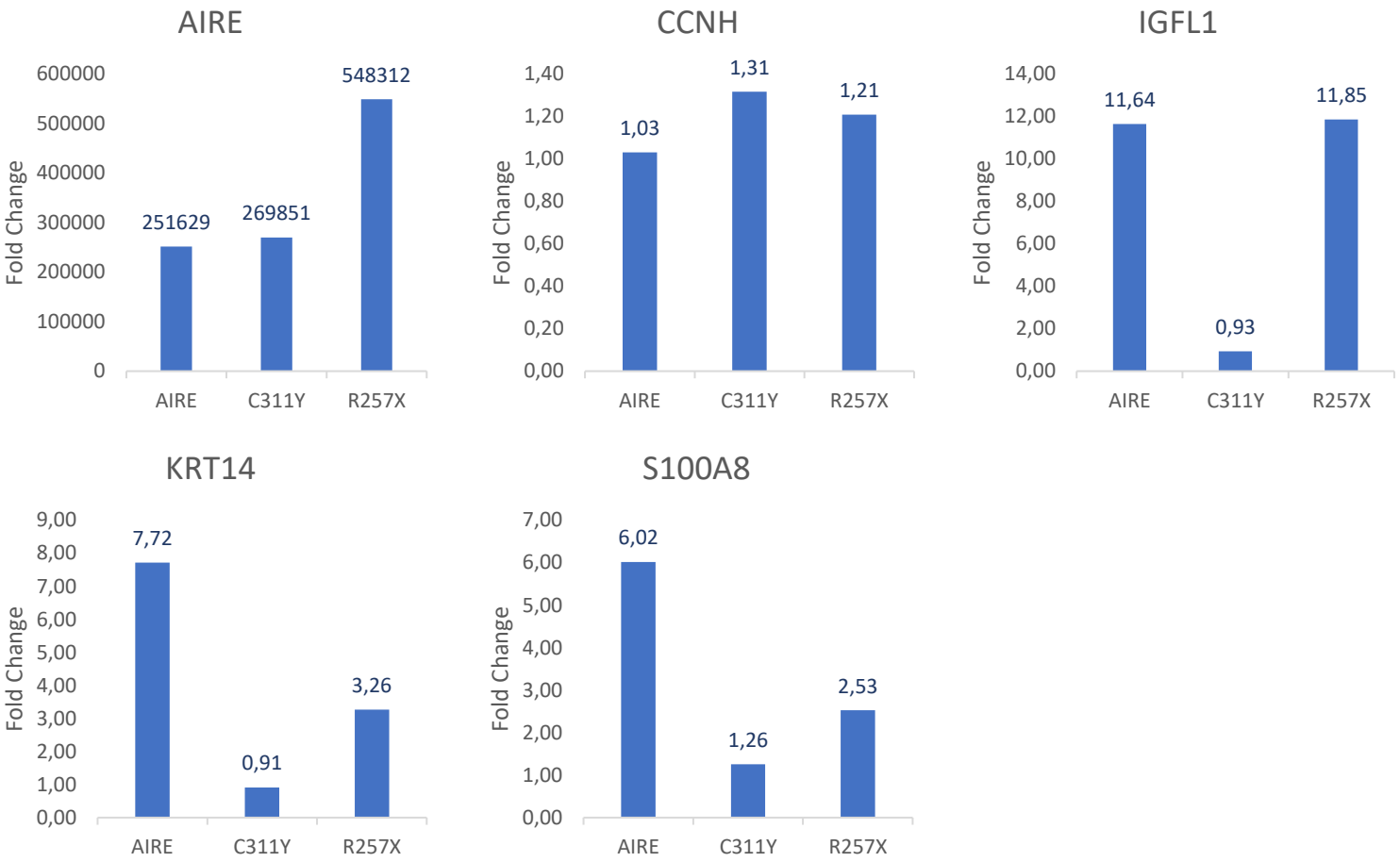


Figure 11 Confirming AIRE activity. RNA from untransfected cells and cells transfected with *AIRE* wildtype or the *AIRE* mutants C311Y and R257X were used in qPCR analyses. TaqMan probes were used for *AIRE*, the negative control *CCNH*, and the purported AIRE regulated genes *IGFL1*, *KRT14* and *S100A8*. Results were normalised against the housekeeping gene *GAPDH*. The ΔΔCT method was used to compare the *AIRE* Wt and mutants with the untransfected control and the difference of expression was calculated as Fold Change. The experiment shown is the result of 3 technical and 3 biological replicates.

Of the AIRE influenced genes; *IGFL1* shows a fold change increase of 11.6 times in the *AIRE* Wt compared to the untransfected cell population, and although this decreases to 0.9 in the C311Y mutant, the R257X mutant shows similar expression to Wt with a fold change of 11.9. The *KRT14* gene shows an increase of 7.7 in the Wt compared to the untransfected cells, a reduction to 0.9 in the C311Y mutant, and slightly less reduction in the R257X mutant at 3.3. The *S100A8* gene increases to 6.0 in the Wt compared to the untransfected cells, with a decrease to 1.3 in the C311Y mutant and 2.5 in the R257X mutant. The increase in gene expression of all three AIRE reactive genes in the Wt suggests that AIRE is functional. In addition, the reduced expression of *KRT14* and *S100A8* in the *AIRE* mutants make them potential reporter genes.

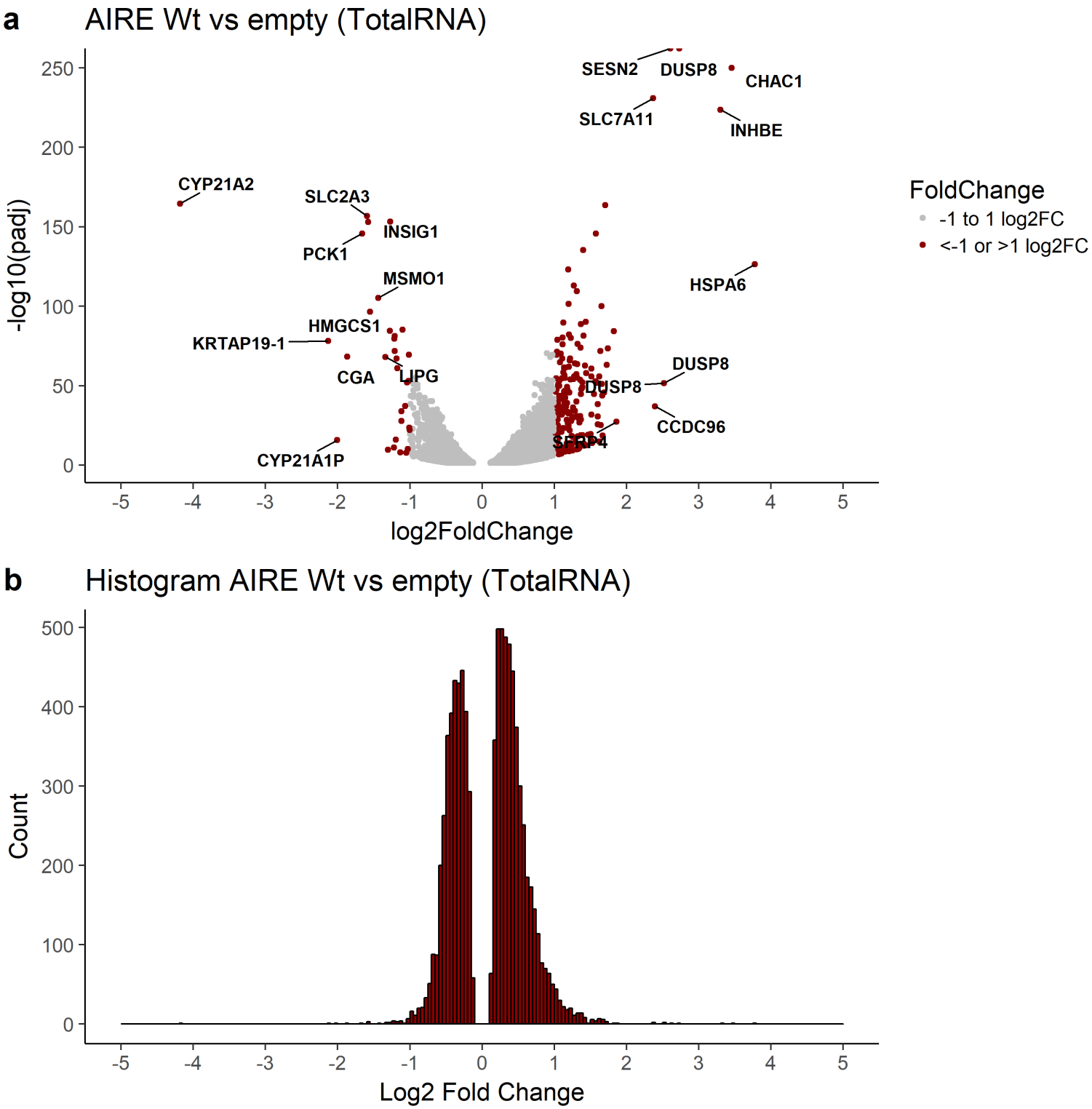
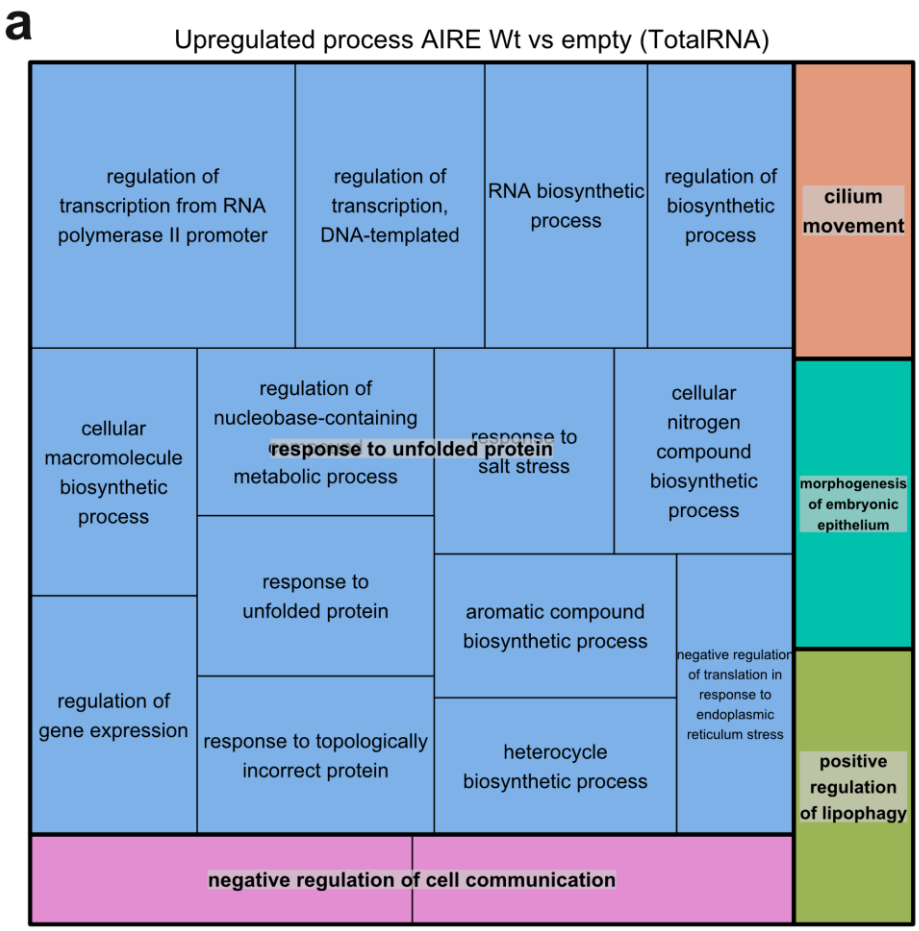


Figure 12 RNAseq analysis comparing the transcriptome of *AIRE* transfected HEK293FT cells with untransfected cells. Genes with lower than 0.05 FDR adjusted p-values (padj) were selected for visualisation. Volcano plot in **a** of the differential expression between the populations in log2 Fold Change on the X-axis and significance in -log10 FDR adjusted p-values on the y-axis. Genes with a log2 fold change above 1 or below -1 are marked in red. Names are given for the top and bottom 10 genes in log2 foldchange. Histogram of the differential expression in **b**. RNAseq was performed with three biological replicates, and library preparation was performed using a TotalRNA with ribosomal RNA reduction kit.



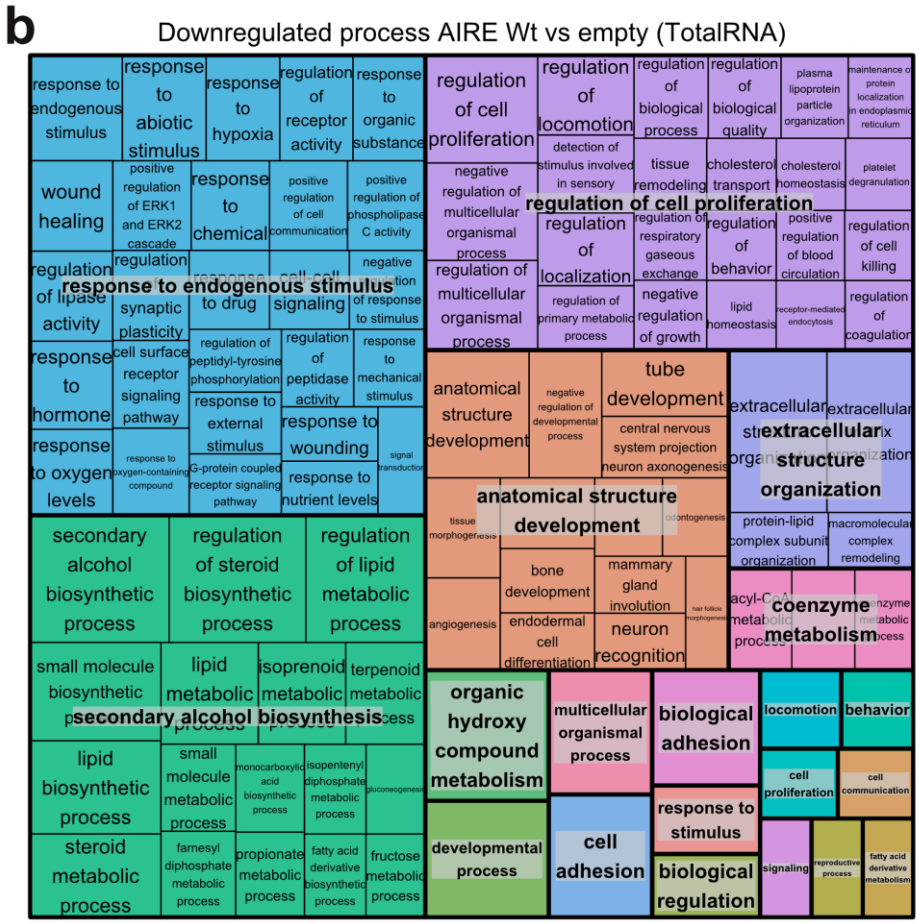


Figure 13 Gene ontology overview of differentially expressed genes between *AIRE* Wt and untransfected cells. Significant genes (<0.05 padj) were separated into ranked lists where one was upregulated and one downregulated. These lists were used as input for GOrilla with a significance cut-off at 10^{-3} and the resulting gene ontology process list was used as input for Revigo, and visualised with the Treemap R-package. The area of each category corresponds with its significance in -log10 p-value. The lists contained 4852 and 3632 upregulated (**a**) and downregulated genes (**b**) respectively.

Table 1 *AIRE* reporter gene candidate selection strategy. Genes from the RNAseq were selected according to high expression in FPKM and higher expression in AIRE transfected cells compared to untransfected in fold change. Genes were subsequently selected for their significance in FDR. Additional criteria were a preference for membrane bound proteins for easier flow cytometry detection, a link with APS-1 symptoms, and detection in previous studies using microarray.[43]

Gene	FPKM	Fold Change	FDR	Membrane bound	APS-1 related	Abramson 2010 (microarray)
INHBE	10.87	9.54	0.0054	No	**Yes (Ovarian Failure)**	**Yes (5.6)**
IL10RA	8.86	2.80	0.0140	**Yes**	No	**Yes (0.9)**
ERMAP	9.65	1.62	0.0347	**Yes**	No	No
SLC3A2	14.09	1.50	0.0151	**Yes**	No	No
SLC7A11	12.14	4.78	0.0021	**Yes**	No	**Yes (5.4)**
CHAC1	11.01	9.97	0.0049	No	No	**Yes (2.8)**

Identifying new AIRE activity reporter genes using RNA sequencing

Having confirmed that AIRE is present and active in the transfected cell populations, transcriptome analysis using RNAseq was performed to identify additional possible reporter genes for the deep mutational scanning assay. RNA was isolated from *AIRE* transfected and untransfected cells, confirmed as high quality (RIN >9) using a bioanalyzer 2000 and prepped for sequencing using a TotalRNA with ribosomal RNA depletion library preparation kit. Sequencing was performed using an Illumina NextSeq 4000 sequencer, and data were quality checked using FastQC. Alignment to the human reference genome (GRCh38) was performed using the pseudoaligner Kallisto, and differential expression analysis was accomplished using DEseq2. The resulting volcano plot of 8484 genes with less than 0.05 FDR adjusted p-values comparing the *AIRE* Wt against the untransfected transcriptome (Empty) can be seen in **figure 12**. Both upregulated and downregulated genes are apparent, with 4852 upregulated genes and 3632 downregulated genes, although the majority of highly differentially expressed genes were upregulated. Gene ontology enrichment of the significantly differentially expressed genes shows upregulation of genes related to the unfolded protein response (UPR), and a downregulation of a variety of pathways (**Fig. 13**). In order to select some candidate reporter genes from amongst the subset of significantly upregulated genes, a selection strategy was needed. The principal criteria chosen were genes with high and significant differential expression, high absolute expression, and encoding a membrane-bound protein to facilitate easier FACS sorting. Additionally, some factors strengthened a candidate genes case, such as being known as related to APS-1 manifestations or appearing in previous microarray studies of AIRE transcriptomes[43]. The resulting genes can be seen in **Table 1**.

qPCR analysis for validation and evaluation of candidate reporter genes from transcriptome comparisons

The previously identified candidate reporter genes with high expression in AIRE transfected cell populations and low expression in untransfected cells were investigated using qPCR with RNA from untransfected cells and cells transfected with *AIRE* Wt and the mutants R257X and C311Y (**Fig. 14**). *Interleukin 10 Receptor Subunit Alpha* (*IL10RA*) exhibits slightly increased expression in the *AIRE* Wt and C311Y mutant populations, while R257X shows a substantial increase in fold change difference from untransfected cells. Similarly, *Solute Carrier Family 3 Member 2* (*SLC3A2*) shows a gradual increase in expression from the *AIRE* Wt to the C311Y and finally the R257X mutant with the most substantial difference in expression. The same pattern can be seen in *Erythroblast Membrane Associated Protein* (*ERMAP*) and *Armadillo Repeat-Containing Protein 5* (*ARMC5*) with R257X having the highest expression but with *AIRE* Wt having little if any difference from untransfected cells. *Solute Carrier Family 7 Member 11* (*SLC7A11*) does have increased expression in *AIRE* Wt compared to empty, and reduced expression in C311Y compared to Wt, yet R257X is still the highest expressing population. *Inhibin Beta E Subunit* (*INHBE*) and *ChaC Glutathione Specific Gamma-Glutamylcyclotransferase 1* (*CHAC1*) show little or no difference between the three cell populations even if they all have substantially increased expression compared to the untransfected cells. Comparing the results of the RNAseq with the qPCR, the values correlate quite well, although the RNAseq values are generally lower (**Fig. 15**). As none of these candidate reporter genes exhibits the preferred properties, namely increased expression in cells transfected with *AIRE* Wt and reduced expression compared to Wt in cells transfected with the *AIRE* mutants, none of them appears suitable as reporter genes in a deep mutational scanning assay.

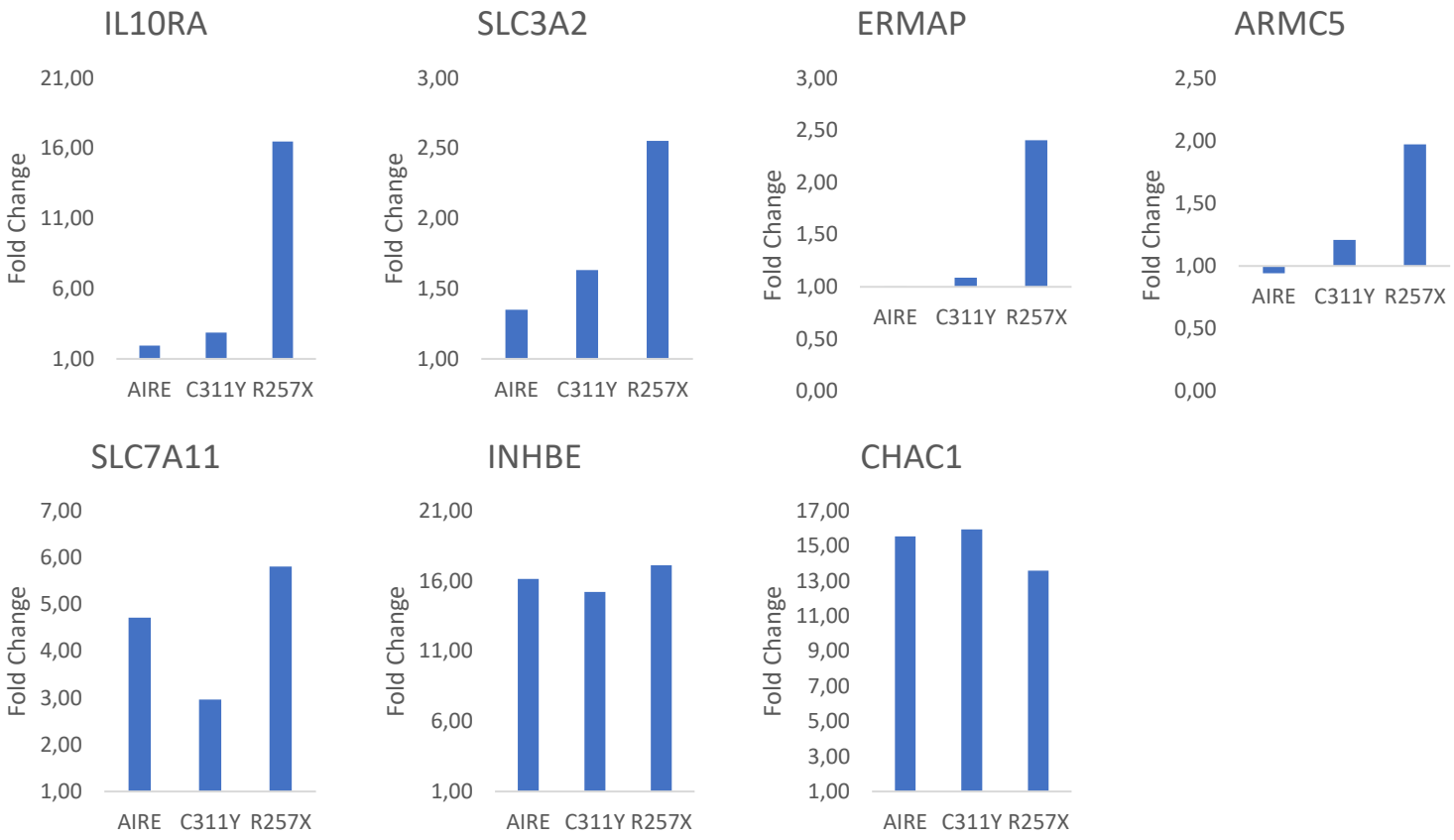


Figure 14 Evaluating reporter genes using qPCR analyses. RNA isolated from untransfected cells, cells transfected with *AIRE* wildtype, or the *AIRE* mutants C311Y and R257X was used to evaluate the suitability of the previously selected reporter gene candidates. TaqMan probes were utilised for the previously identified candidate reporter genes *IL10RA*, *SLC3A2*, *ERMAP*, *ARMC5*, *SLC7A11*, *INHBE*, and *CHAC1*. All gene expression was normalised using the housekeeping gene *GAPDH*. The ΔΔCT method was utilised to calculate the difference of expression in Fold Change between the transfected cells and the untransfected control. Three technical and three biological replicates for each cell population were used.

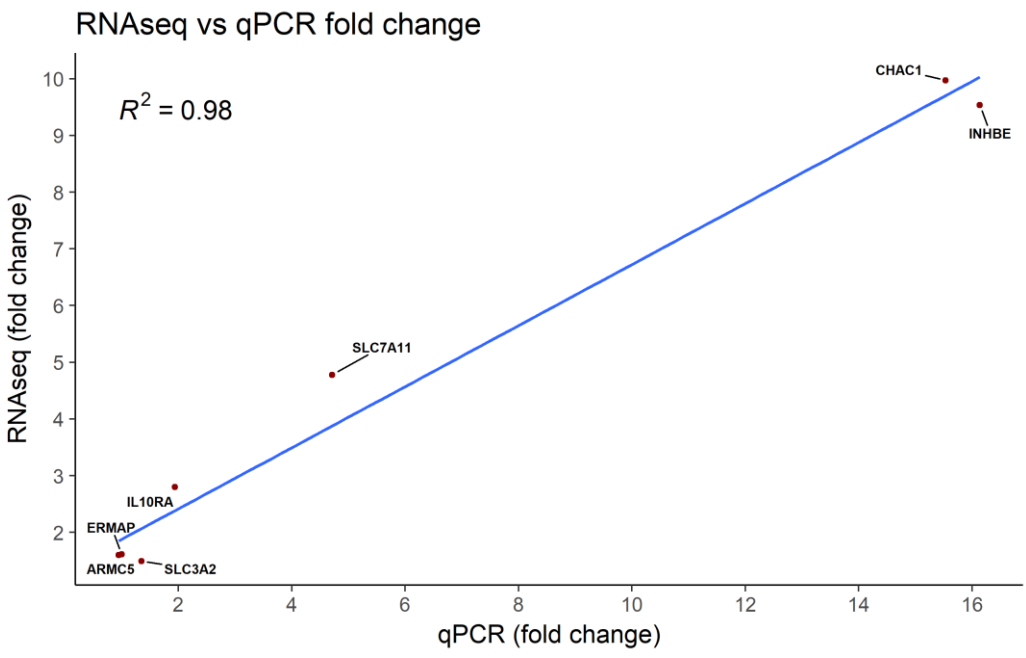


Figure 15 Comparison of RNAseq and qPCR fold change. Comparison of fold change between RNAseq and qPCR. Data from the genes *CHAC1, ERMAP, INHBE, SLC3A2, SLC7A11, ARMC5, IL10RA*. RNAseq data taken from *AIRE* Wt vs empty TotalRNA (**Fig. 12**) and qPCR data taken from **figure 14**.

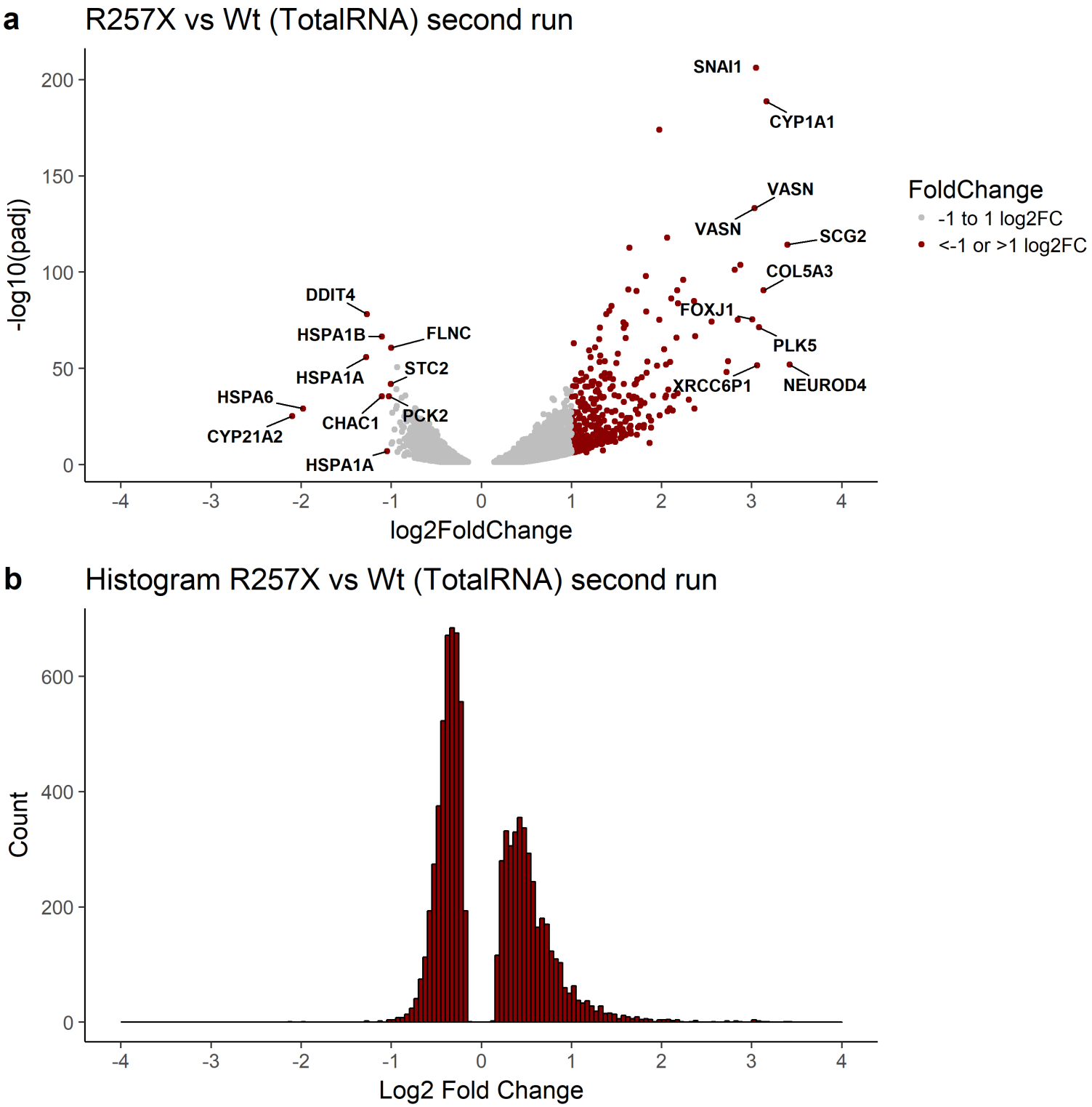


Figure 16 Differential gene expression analysis from RNAseq of R257X and *AIRE* Wt transfected cells. Genes with a significance threshold of less than 0.05 FDR adjusted p-values (padj) visualised in a volcano plot (**a**) and histogram (**b**). The volcano plot ranks the differential expression in log2 Fold Change on the x-axis and the significance in -log10(padj) on the y- axis. Genes with higher than 1 and lower than -1 log2 Fold Change are marked in red. The top 10 and bottom 10 log2 Fold Change genes are marked with names. RNAseq was performed with three biological replicates, and library preparation performed using a TotalRNA with ribosomal RNA reduction kit.

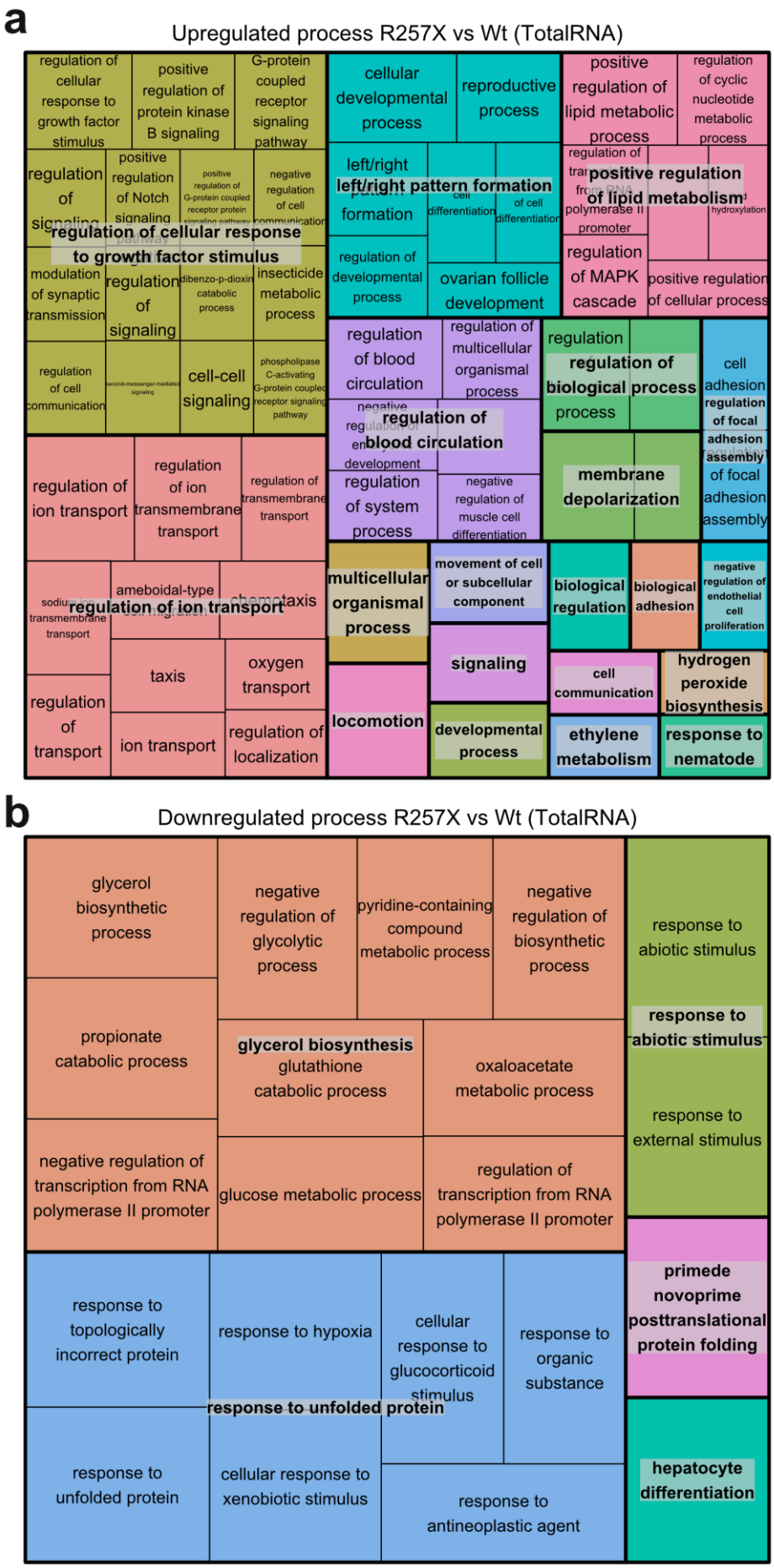


Figure 17 Gene ontology overview of differentially expressed genes between R257X and *AIRE* Wt transfected cells. Significant genes (<0.05 padj) were separated into ranked lists where one was upregulated and one downregulated. These lists were used as input for GOrilla with a significance cut-off at 10^{-3} and the resulting gene ontology list was used as input for Revigo, and visualised with the Treemap R-package. The area of each category corresponds with its significance in -log10 p-value. The lists contained 3947 and 4442 upregulated (**a**) and downregulated (**b**) genes respectively.

Finding reporter genes using RNAseq of AIRE mutants

To narrow down the population of genes suitable for use as AIRE activity reporter genes, we performed a new series of RNAseq experiments. To find a subset of genes that were upregulated in *AIRE* Wt transfected cells compared to untransfected cells, but downregulated in *AIRE* mutants compared to the Wt, the *AIRE* mutants R257X and C311Y were analysed and compared with Wt. The subset of genes with a significance of less than 0.05 FDR adjusted p-value (padj) was visualised using volcano plots. The 8389 significant differentially expressed genes when comparing the R257X mutant and *AIRE* Wt can be seen in **figure 16** where 3947 genes show upregulation and 4422 genes show downregulation. Despite the majority of genes showing downregulation, the majority of highly differentially expressed genes with a log2 fold change of higher than 1 or lower than -1 is situated within the upregulated population. **Figure 17** visualises the enriched gene ontology terms of the biological processes involving these genes. A variety of processes are upregulated, but with the most significant processes in the downregulated population being glycerol biosynthesis and UPR. Comparing the transcriptome of *AIRE* mutant C311Y with the Wt found 2540 differentially expressed genes with a significance of less than 0.05 FDR adjusted p-value (padj) with 759 upregulated genes and 1782 downregulated genes. Overall, there was little difference between the cell populations, with only one gene with a log2 fold change below -1 (**Fig. 18**).

In an attempt to increase the likelihood of finding differentially expressed genes with a low absolute expression and comparing the performance between different RNAseq methods, we next set up an equivalent experiment using an mRNA library preparation kit with polyA tail capture. The RNA isolated from transfected and untransfected cells were then analysed using this preparation method. The comparison of differentially expressed genes with a significance of less than 0.05 FDR adjusted p-values (padj) between R257X and Wt can be seen in **figure 19**, with 10073 significant genes of which 4657 were upregulated, and 5417 were downregulated. Similar to the first RNAseq of these populations (**Fig. 16**), the majority of highly differentially expressed genes with a log2 fold change of higher than 1 and lower than -1 are seen in the upregulated subset.

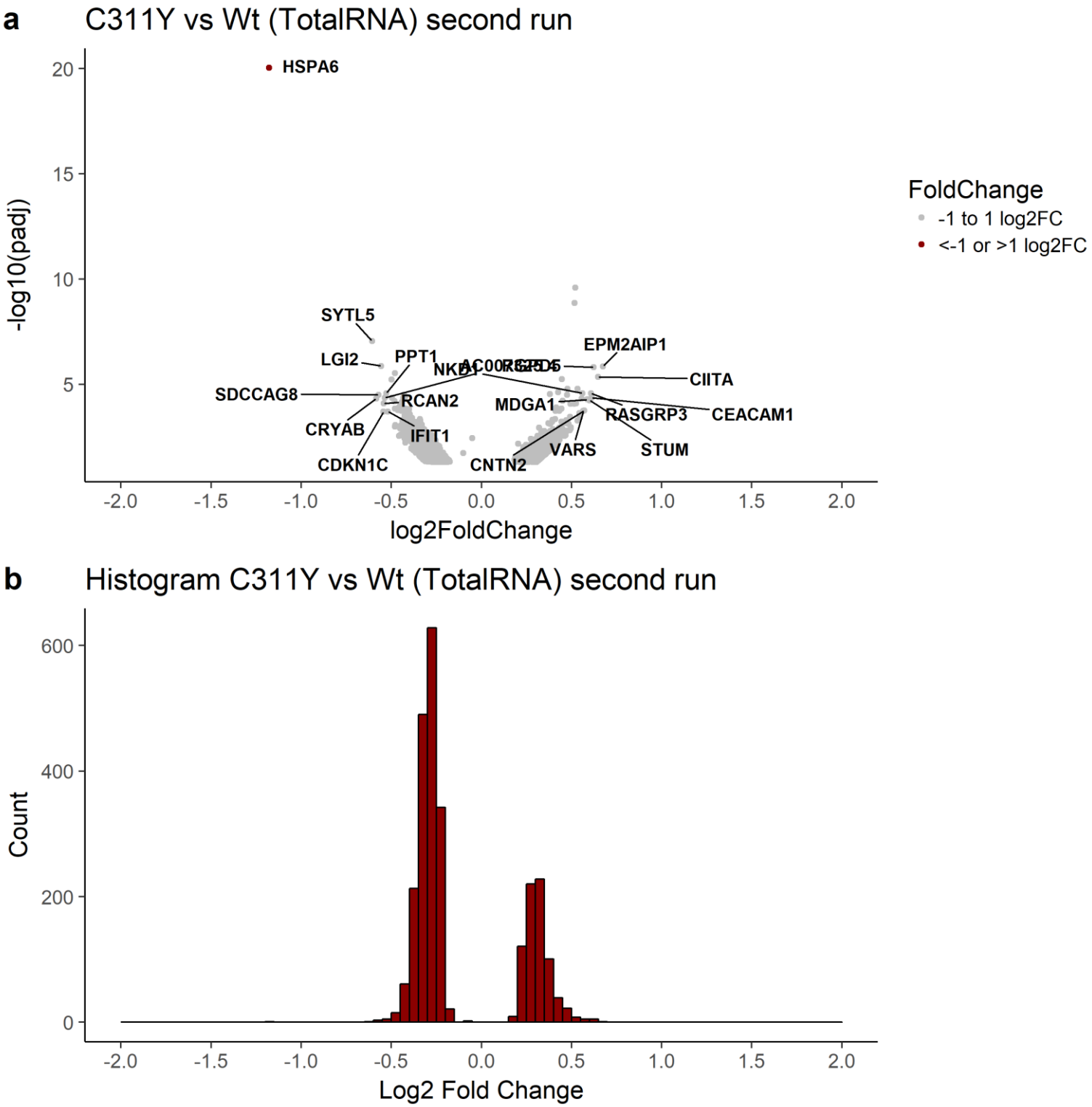


Figure 18 Differential gene expression analysis from RNAseq of C311Y and *AIRE* Wt transfected cells. Genes with a significance threshold of less than 0.05 FDR adjusted p-values (padj) visualised in a volcano plot (**a**) and histogram (**b**). The volcano plot ranks the differential expression in log2 Fold Change on the x-axis and the significance in -log10(padj) on the y-axis. Genes with higher than 1 and lower than -1 log2 Fold Change are marked in red. The top 10 and bottom 10 log2 Fold Change genes are marked with names. RNAseq was performed with three biological replicates, and library preparation performed using a TotalRNA with ribosomal RNA reduction kit.

Figure 20 shows the significantly differentially expressed genes between C311Y and Wt; a total of 2299 significant genes were found, of which 1012 genes were found to be upregulated, and 1287 were downregulated. However, no genes had a higher log2 fold change than 1 or lower than -1 in this comparison. The RNAseq approach used in this experiment is unable to detect an identifiable highly differentially expressed downregulated population in the *AIRE* mutants. Thus, the focus changed to evaluate and compare the RNAseq setup.

Comparing RNAseq library preparation methods

In order to compare the two library prep methods, transcriptomes of *AIRE* Wt were compared with untransfected (empty) cells and significantly differentially expressed genes were visualised in **figure 21**. A total of 4443 genes were found to be differentially expressed, where 2108 genes were found to be upregulated, and 2335 were downregulated. Similar to the comparison using TotalRNA library preparation (**Fig. 12**), the mRNA method shows more highly differentially expressed genes in the upregulated portion, though the number of genes with a log2 fold change above 1 or below -1 is less than the first RNAseq. A comparison of the log2 fold change distribution of the mRNA compared to the TotalRNA RNAseq, plotted in **figure 22a** shows that the mRNA population trends closer to 0 compared to the TotalRNA population, with generally lower differential expression in log2 fold change. Comparing the absolute expression in transcripts per million (TPM) both the Wt (**Fig. 22b**) and untransfected (**Fig. 22c**) populations show mostly similar expression, although a variety of genes show higher expression using the mRNA library kit.

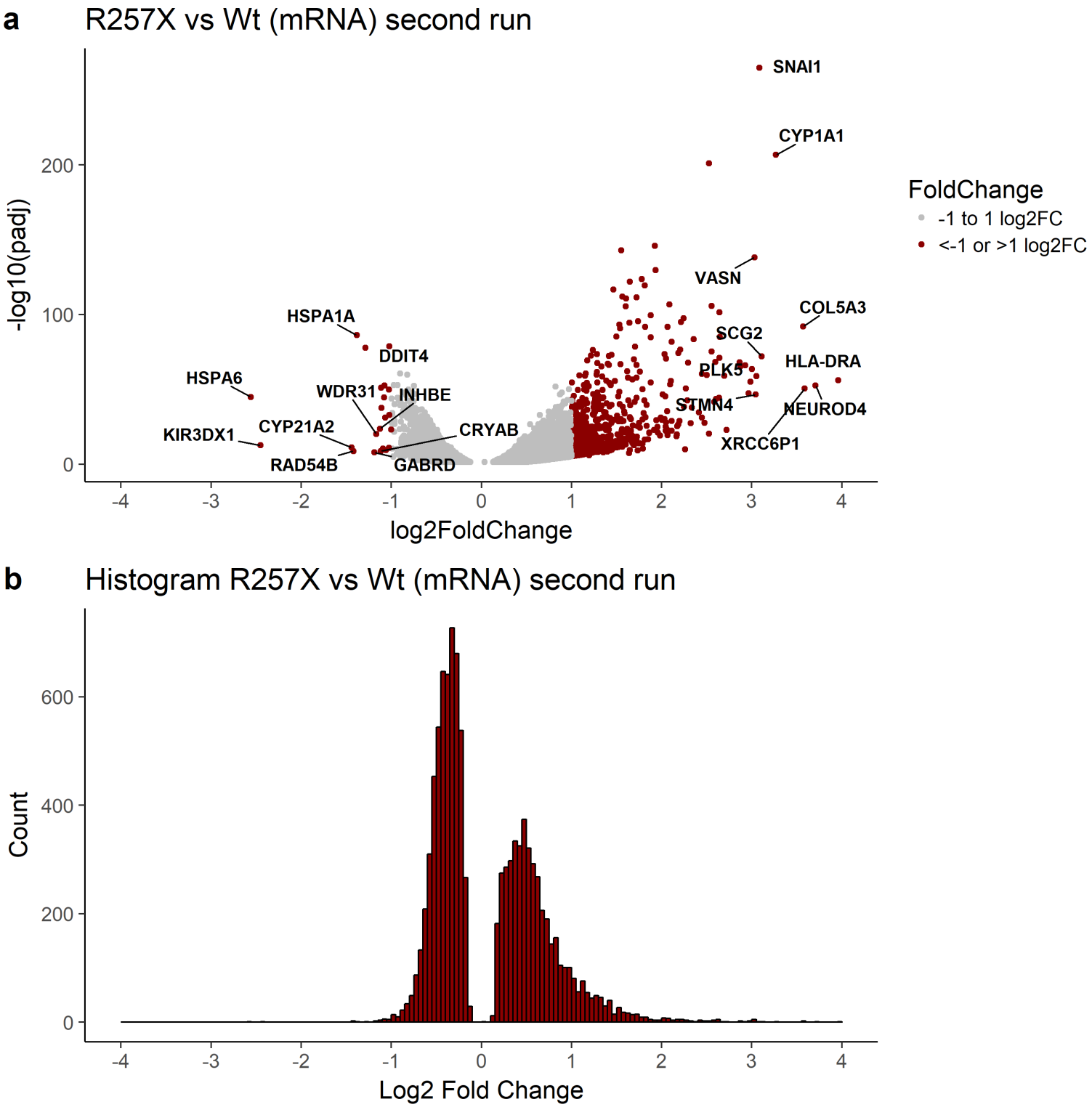


Figure 19 Differential gene expression analysis from RNAseq of R257X and *AIRE* Wt transfected cells. Genes with a significance threshold of less than 0.05 FDR adjusted p-values (padj) visualised in a volcano plot (**a**) and histogram (**b**). The volcano plot ranks the differential expression in log2 Fold Change on the x-axis and the significance in -log10(padj) on the y-axis. Genes with higher than 1 and lower than -1 log2 Fold Change are marked in red. The top 10 and bottom 10 log2 Fold Change genes are marked with names. RNAseq was performed with three biological replicates, and library preparation performed using a mRNA with polyA tail capture kit.

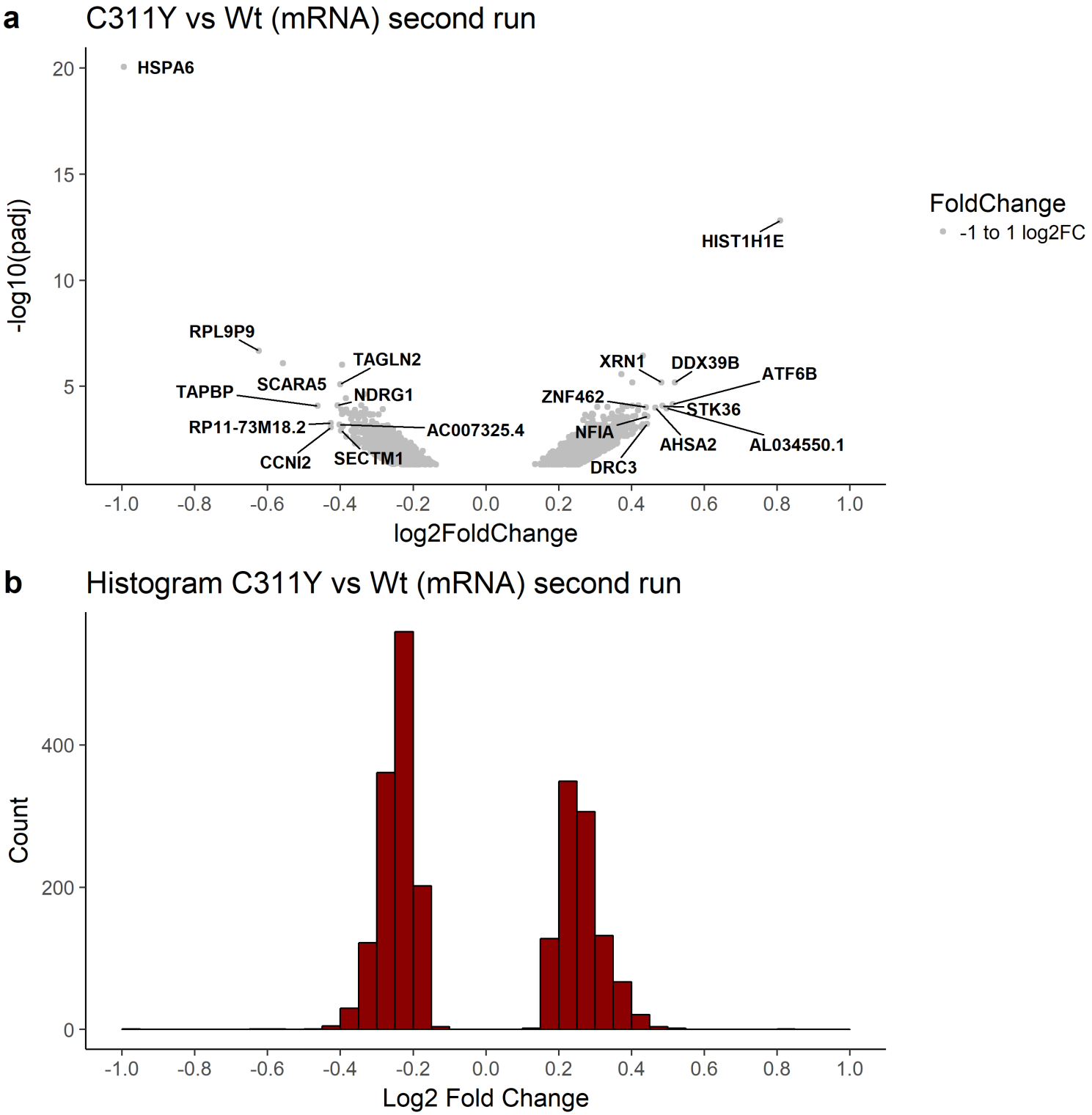


Figure 20 Differential gene expression analysis from RNAseq of C311Y and *AIRE* Wt transfected cells. Genes with a significance threshold of less than 0.05 FDR adjusted p-values (padj) visualised in a volcano plot (**a**) and histogram (**b**). The volcano plot ranks the differential expression in log2 Fold Change on the x-axis and the significance in -log10(padj) on the y-axis. The top 10 and bottom 10 log2 Fold Change genes are marked with names. RNAseq was performed with three biological replicates, and library preparation performed using a mRNA with polyA tail capture kit.

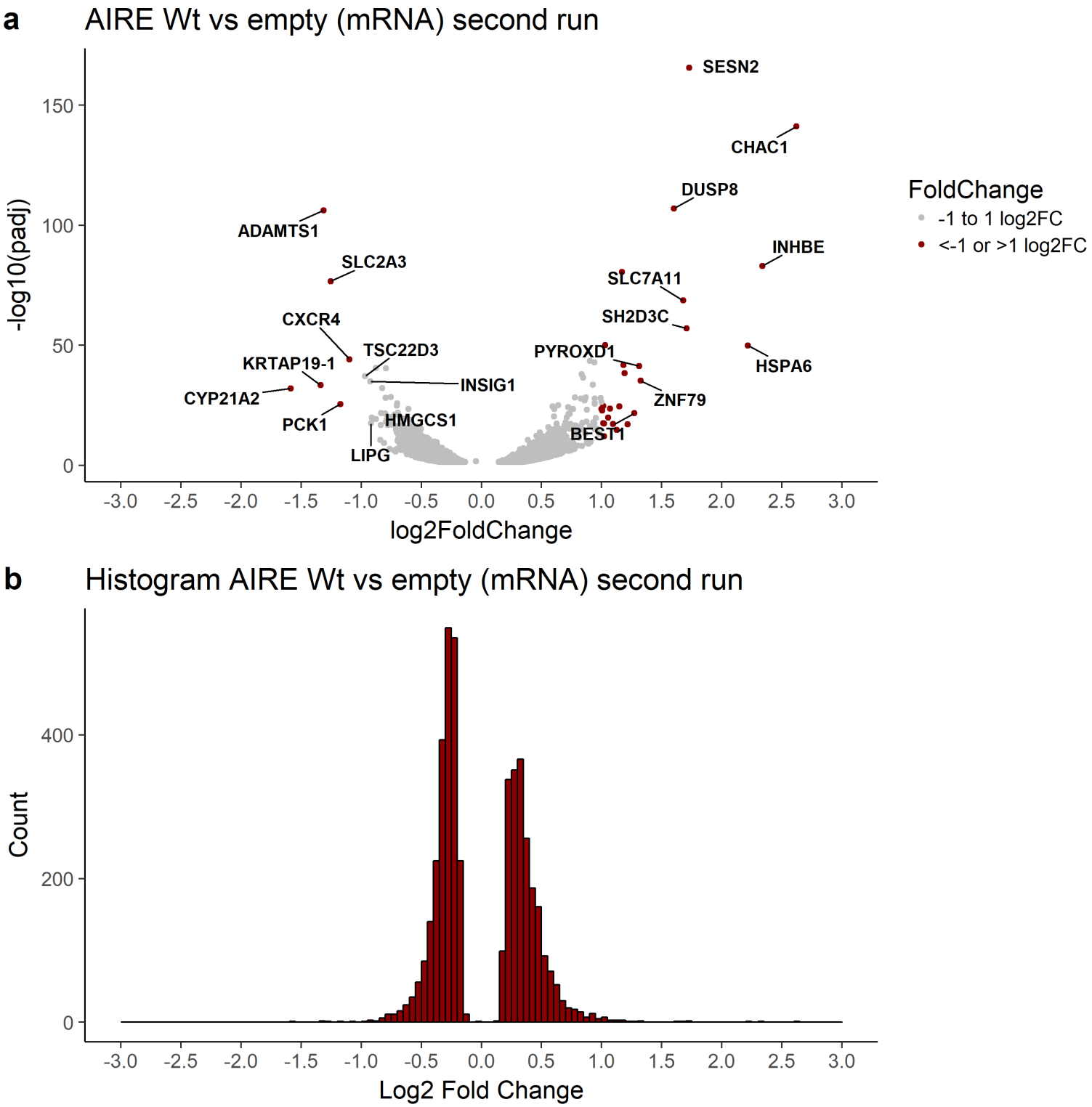
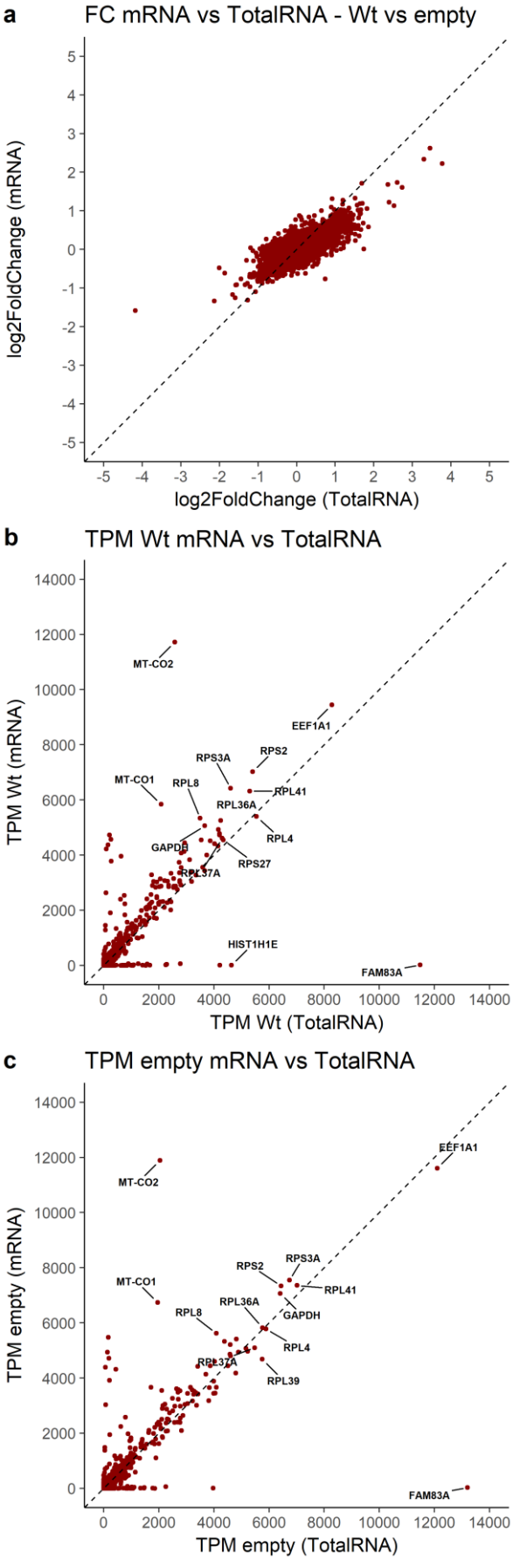


Figure 21 Differential gene expression analysis from RNAseq of *AIRE* Wt and untransfected (empty) cells. Genes with a significance threshold of less than 0.05 FDR adjusted p-values (padj) visualised in a volcano plot (**a**) and histogram (**b**). The volcano plot ranks the differential expression in log2 Fold Change on the x-axis and the significance in -log10(padj) on the y-axis. Genes with higher than 1 and lower than -1 log2 Fold Change marked in red. Top 10 and bottom 10 log2 Fold Change genes marked with names. RNAseq performed with three biological controls, and library preparation performed using a mRNA with polyA tail capture kit.



Confirming and evaluating RNAseq

In order to confirm the previous RNAseq experiments and to evaluate how representative they were, a new independent set of experiments were performed. Plasmids were sequenced to confirm their gene inserts, plasmids were reamplified, new cells were transfected, and RNA was subsequently isolated. RNAseq was performed again using the mRNA library preparation kit with RNA isolated from *AIRE* Wt, R257X and C311Y transfected cells.

Differential gene expression between the R257X mutant and *AIRE* Wt transfected cells, found 8329 genes with a significance of less than 0.05 FDR adjusted p-values. Of these significantly differentially expressed genes, 3658 genes were found to be upregulated, and 4673 genes were downregulated. These genes are visualised in **figure 23**. The majority of highly differentially expressed genes with a log2 fold change above 1 or below -1 were found to be upregulated, similarly to previous differential expression analysis of this comparison. Comparing the transcriptomes of C311Y and Wt found just 49 differentially expressed genes, where 5 of these were upregulated, and 44 were downregulated (**Fig. 24**). None of these genes was highly differentially expressed.

Figure 22 Comparison of RNAseq runs with different preparation method. RNAseqs with mRNA and TotalRNA library kit setups compared. The log2 fold change of first and second RNAseq of *AIRE* Wt vs untransfected (empty) is compared in **a**. The transcripts per million (TPM) of each gene in the *AIRE* Wt population is compared in **b**, and the TPM of the genes in the untransfected population is compared in **c**.

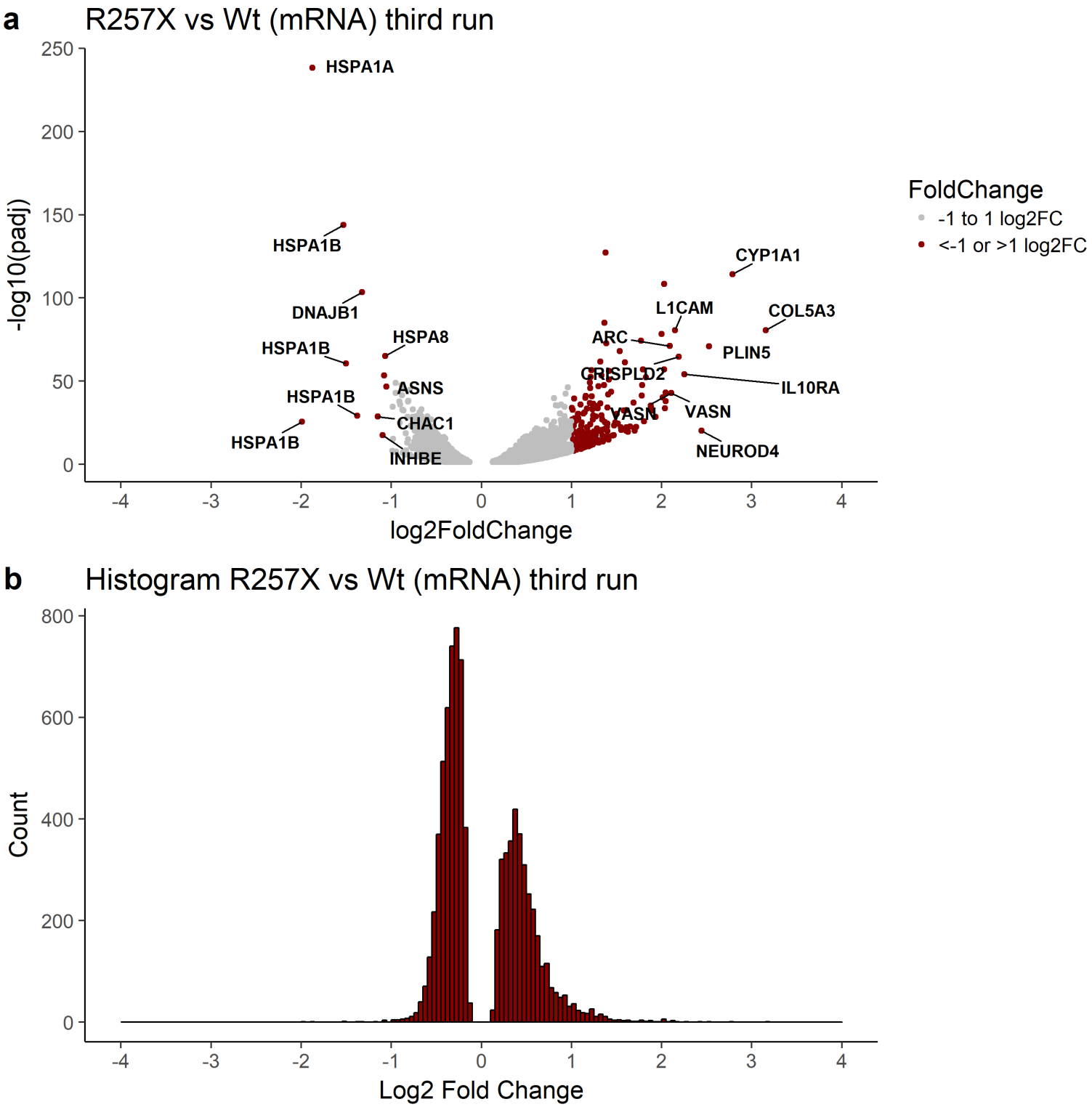


Figure 23 Differential gene expression analysis from RNAseq of R257X and *AIRE* Wt transfected cells. Genes with a significance threshold of less than 0.05 FDR adjusted p-values (padj) visualised in a volcano plot (**a**) and histogram (**b**). The volcano plot ranks the differential expression in log2 Fold Change on the x-axis and the significance in -log10(padj) on the y-axis. Genes with higher than 1 and lower than -1 log2 Fold Change are marked in red. The top 10 and bottom 10 log2 Fold Change genes are marked with names. RNAseq was performed with three biological replicates, and library preparation performed using a mRNA with polyA tail capture kit.

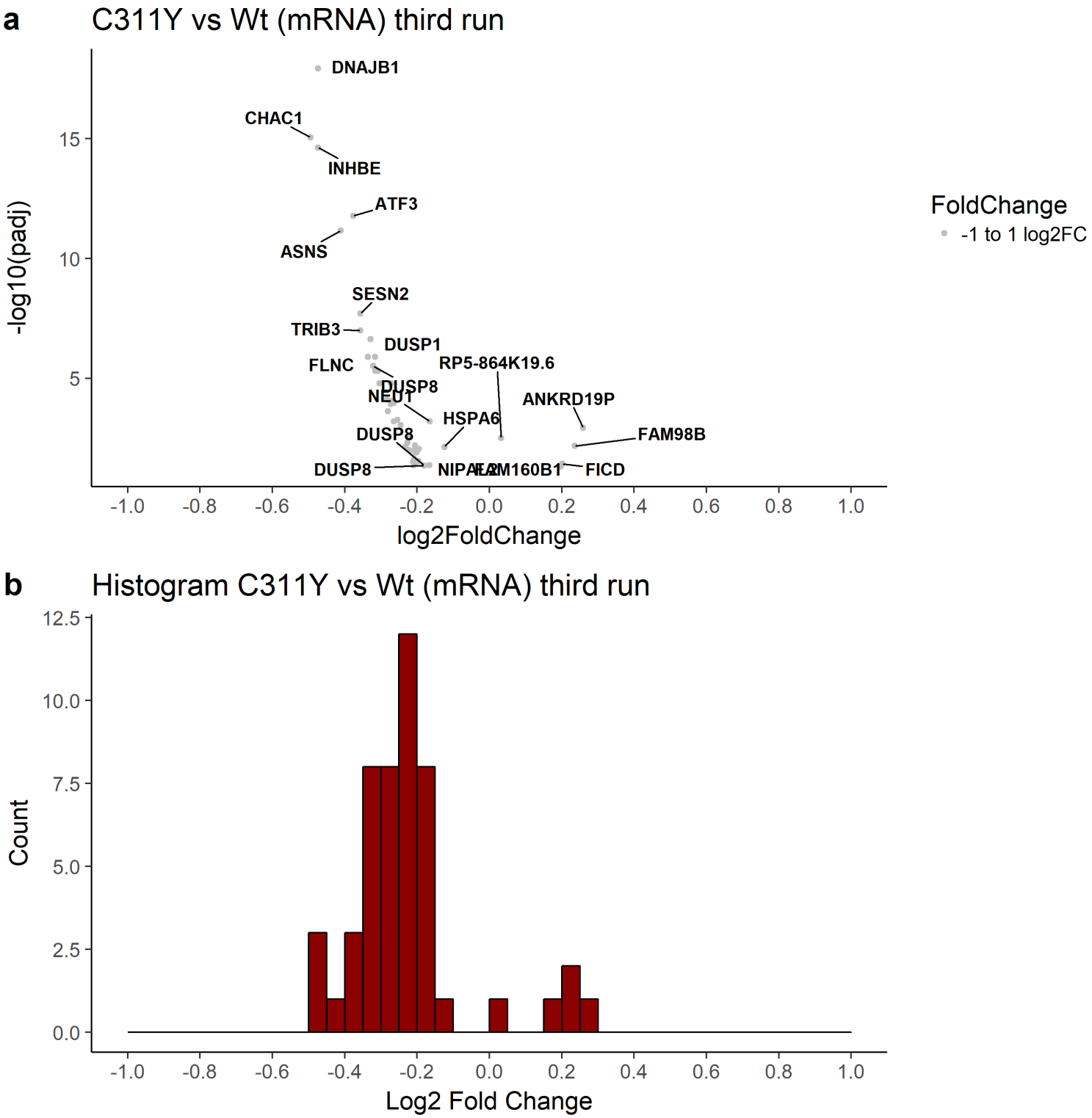


Figure 24 Differential gene expression analysis from RNAseq of C311Y and *AIRE* Wt transfected cells. Genes with a significance threshold of less than 0.05 FDR adjusted p-values (padj) visualised in a volcano plot (**a**) and histogram (**b**). The volcano plot ranks the differential expression in log2 Fold Change on the x-axis and the significance in -log10(padj) on the y-axis. The top 10 and bottom 10 log2 Fold Change genes are marked with names. RNAseq was performed with three biological replicates, and library preparation performed using a mRNA with polyA tail capture kit.

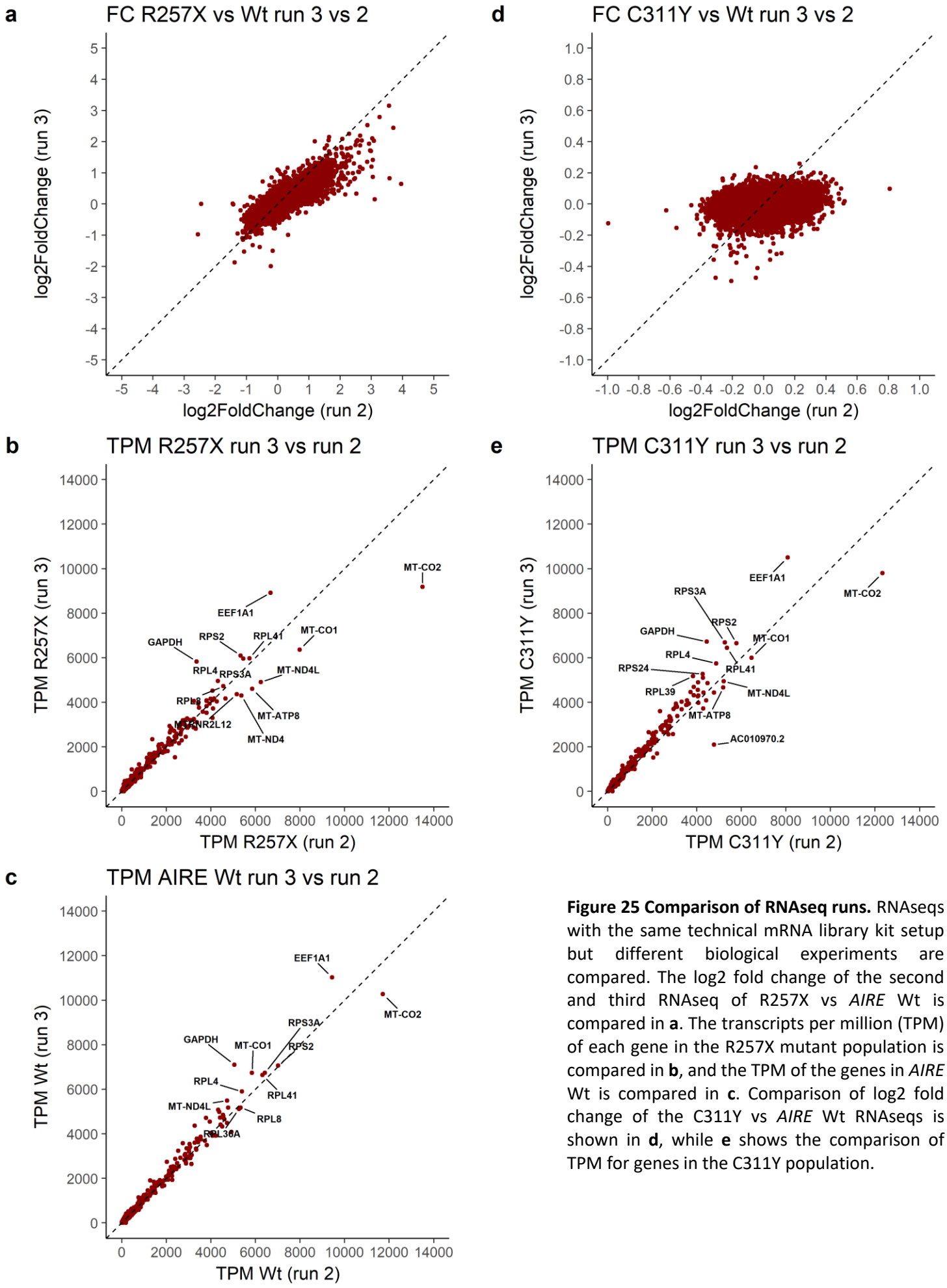


Figure 25 Comparison of RNAseq runs. RNAseqs with the same technical mRNA library kit setup but different biological experiments are compared. The log2 fold change of the second and third RNAseq of R257X vs *AIRE* Wt is compared in **a**. The transcripts per million (TPM) of each gene in the R257X mutant population is compared in **b**, and the TPM of the genes in *AIRE* Wt is compared in **c**. Comparison of log2 fold change of the C311Y vs *AIRE* Wt RNAseqs is shown in **d**, while **e** shows the comparison of TPM for genes in the C311Y population.

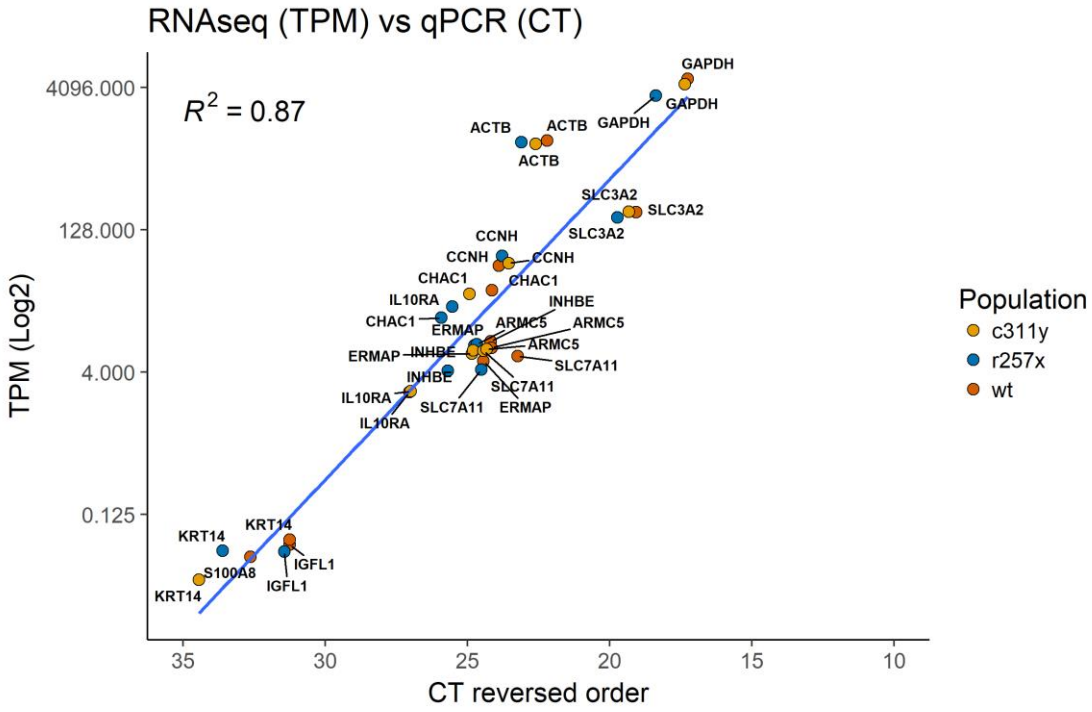


Figure 26 Comparison of RNAseq and qPCR. Comparison of transcript per million (TPM) from RNAseq with CT values from qPCR analysis of the same biological experiments. Data from the genes *IGFL1, KRT14, S100A8, CHAC1, ERMAP, INHBE, SLC3A2, SLC7A11, ARMC5, IL10RA, GAPDH, ACTB, CCNH*. The second run of RNAseqs with R257X and C311Y vs Wt using TotalRNA library kit used for TPM and count values.

Comparing the RNAseq of the same technical setups but with different biological experiments, the log2 fold change values of significant genes in the R257X compared to Wt trend slightly more towards 0 in the last RNAseq, although this deviation is very modest (**Fig. 25a**). The TPMs of the genes in the R257X transfected cell population are remarkably stable across experiments as seen in **figure 25b**. This is also true for the Wt transfected cells, although with a slight increase of TPM in genes for the last RNAseq (**Fig. 25c**). Comparing the differentially expressed genes between two analyses of the C311Y and Wt populations, the majority of genes are not differentially expressed in the last RNAseq (**Fig. 25d**). However, when looking at the absolute expression in TPM for genes in the C311Y population, the RNAseqs correlates very well, with a slightly higher expression in the last RNAseq (**Fig. 25e**). In order to compare how well the RNAseq analyses correlates with the previously performed qPCR, absolute expression values from RNAseq were compared to CT values from qPCR of the same RNA isolates (**Fig. 26**). The TPM from RNAseq correlates well with CT values, with an R^2 of 0.87. Importantly the previously reported gene candidates *KRT14* and *S100A8* are not significantly identified using RNAseq, with a TPM of 0.07 and 0.04 respectively.

5. Discussion

AIRE, the autoimmune regulator, is a potent transcriptional regulator able to induce expression of thousands of repressed genes. This makes AIRE crucial for the process of negative selection, in which T-cells get controlled for their ability to recognise the body's own proteins. Mutations in *AIRE* are responsible for the rare autoimmune disorder APS-1. Such mutations have been found within all of AIREs functional domains, yet the patient phenotypes differ in their severity depending on the underlying mutation. Most AIRE mutations are inherited recessively, but some dominant mutations have been identified.

AIRE expression is tightly controlled and occurs primarily in mTECs in the medulla of the thymus. The AIRE protein consists of a number of functional domains; a CARD protein dimerisation domain, a SAND general DNA interaction domain, and two PHD zinc fingers for protein interaction. AIRE functions by recognising and inducing transcription of repressed genes. This recognition is performed by PHD zinc fingers interacting with unmethylated histone tails, or by recognition of the repressive ATF7ip-MBD1 complex. AIRE then recruits a number of proteins into an efficient transcriptional machinery. Because AIRE does not target specific genes or promoters like other transcriptional regulators but instead targets any repressed gene in a cell, AIRE induces different genes depending on which cell type it is active in. In addition, AIRE activity is stochastic, in that different cells will have different genes induced by AIRE. However, all AIRE regulated genes will be induced on a population level.

APS-1 is characterised by the three main manifestations of Addison's disease, hypoparathyroidism, and chronic mucocutaneous candidosis. However, APS-1 patients may not have all of these symptoms and may exhibit a variety of other symptoms unique to that patient. This means that some APS-1 patients may be undiagnosed, or diagnosed with other autoimmune diseases. With the increasing number of large-scale sequencing and GWAS studies of patients, novel AIRE mutations, especially mutations with a moderate phenotypic effect, may be found and linked to other diseases.

While AIRE has been a focus of much research since its identification in 1997, many questions still remain regarding AIRE function and interaction. The characterisation of AIRE mutations may improve our understanding of AIRE function, allow for better diagnosis of a number of

patients with an atypical APS-1 phenotype, and inform large-scale GWAS and sequencing efforts.

The current thesis is part of a larger project where the long-term aim is to develop a multiplexed screening assay for AIRE function, by the use of a deep mutational scanning approach. This would allow characterisation of the functional effect of all theoretically possible mutations in AIRE. As a means to identify possible reporter genes for AIRE activity that could be used as a part of this multiplexed assay, we transfected HEK293FT human embryonic kidney cells with *AIRE* and *AIRE* mutants and analysed the transfected populations using RNAseq and qPCR.

Establishing an experimental system for functional investigation of AIRE variants

In order to investigate AIRE in a high throughput assay, a highly expressing and fast-growing cell system is needed. To that effect, HEK293FT cells were transfected with pCMV6 plasmids containing FLAG and Myc tagged AIRE, and the AIRE mutants R257X and C311Y.

The western blot confirmed AIRE and AIRE mutant expression in this cell-system (**Fig. 8**), with AIRE exhibiting a band approximately 64 kDa, slightly higher than the estimated 61 kDa of AIRE with FLAG and Myc tags (ProtParam[83]). The R257X mutant was detected using N-terminal recognising anti-AIRE antibodies, exhibiting a major band at approximately 30 kDa, consistent with estimates (ProtParam[83]) and previous studies.[36] This band was unique for the R257X mutant. As expected, the R257X mutant was not detected using anti-FLAG antibodies since its premature stop codon prevents the expression of the C-terminal FLAG tag. To our knowledge, the C311Y mutant has not been previously investigated using western blotting but was identified using both anti-FLAG and anti-AIRE antibodies with a weak band at the consensus size of 64 kDa, yet with substantial degradation.

Detection of FLAG-positive cells using flow cytometry confirmed the presence of AIRE and allowed for a quantification of the percentage of cells successfully transfected. Initial transfection showed a moderate transfection efficiency of approximately 41%, comparable to but less than the positive control 21OH at approximately 56% transfection efficiency (**Fig. 9b and 9c**). However, the transfection efficiency has thereafter increased to approximately 65% for *AIRE* Wt with a conservative gating strategy, and comparably 66% for the C311Y *AIRE*

mutant (**Fig. 10b and 10c**). The reason for this increase is uncertain but may reflect increased proficiency with transfection and flow-cytometry in general.

Characterising AIRE protein degradation as a result of overexpression and mutant instability

Our setup of using two different antibodies with binding affinity on each extremity of the AIRE protein in the western blot allows for estimation of the location of the exhibited protein degradation. AIRE Wt exhibited some degree of degradation with three products located approximately between 40 and 50 kDa using the C-terminal antibodies, and a band at 30 kDa and about 20 kDa detectable using the N-terminal antibodies (**Fig. 8**). These band sizes suggest an N-terminal degradation located between the PHD1 and CARD domain. The R257X mutant is only detectable using the N-terminal antibody, yet in addition to the dominant band slightly above 30 kDa, it does exhibit the same 30 and 20 kDa degraded bands as detected in the Wt blot. The C311Y blot exhibited considerable degradation, with a weak consensus band in both blots, but with many degraded bands in the C-terminal blot with a size between 35 and 50 kDa, and a few bands in the sub-25 kDa size. In addition, C311Y exhibited the two 30 and 20 kDa bands in the N-terminal blot, consistent with both other AIRE variants. Again, this suggests degradation from the N-terminal side between the CARD domain and the first PHD finger, in addition to some degradation responsible for a construct only containing one or two PHD domains and the FLAG sequence. It is possible that the two sub-30 kDa bands detected using N-terminal anti-AIRE antibodies is due to unspecific binding, yet these bands do not seem to appear using the negative control of the untransfected cells.

The substantial degradation of the C311Y mutant suggests that this is a very unstable protein. Causative for this instability may be the removal of the stabilising zinc-binding motif of the primary PHD zinc finger previously reported in this mutant.[57] As such, this instability may at least partially point to how this mutant is dominant, by producing a full-length AIRE protein without the necessary fold for the action of PHD1 and PHD2 in the AIRE multimer.

Confirming functional AIRE by AIRE gene induction

After successfully establishing the cell system as AIRE and AIRE mutant expressing, the cell system was evaluated for AIRE gene induction, to confirm a functional AIRE protein.

qPCR analysis of *AIRE* and *AIRE* mutants showed an extremely high fold change of AIRE in transfected cells (**Fig. 11**), and RNAseq data confirms that AIRE is responsible for up to 20% of all reads in the experiment (data not shown). In fact, AIRE Wt and the C311Y mutant expression were at 251 and 269 thousand times that of untransfected cells respectively, while the AIRE mutant R257X had a two-fold higher expression than Wt. Gene ontology enrichment of these populations showed that the UPR was upregulated in the *AIRE* Wt transfected populations compared to untransfected cells (**Fig. 13a).** It could be speculated that this may be a response to the extreme transcription of AIRE in these cells saturating the endoplasmic reticulum (ER) and is consistent with the substantial degradation found for the Wt and C311Y variants using western blot. However, when analysing the gene ontology enrichment of the comparison between the R257X mutant and the AIRE Wt, the UPR is amongst the downregulated population of genes (**Fig. 17b**). Thus, it seems that the UPR effect is restricted to the Wt and possibly the C311Y mutant which has overall similar expression. The reason for this discrepancy is difficult to disentangle based on the current experiments. The R257X mutant encodes a truncated protein, with an estimated and observed size at approximately 30 kDa (**Fig. 8**) and thus may be more efficiently translated, have fewer problems in folding, and may activate the UPR to a lesser degree than the AIRE Wt. However, it may also be the opposite, that the triggering of the UPR to a lesser extent, allows for increased transcription of the R257X mutant. Alternatively, it may just reflect the fact that a shorter DNA sequence allows for faster transcription, and the higher fold change of R257X may not reflect an actual increase in protein synthesis.

The qPCR analysis of the known AIRE-dependant genes *IGFL1, KRT14* and *S100A8* showed an increased transcription in AIRE Wt populations compared to untransfected cell populations (**Fig. 11**). This result is consistent with the literature, where these genes are described as highly AIRE-dependant in Wt transfected HEK293T cells.[43] Furthermore, the degradation of AIRE Wt, as evident from the western blots, does not seem to influence the functional activity of AIRE. When comparing the transcription of these genes in the C311Y mutant, all three genes had lower expression. However, while the R257X mutant showed decreased transcription in *KRT14* and *S100A8,* the transcription of *IGFL1* is not reduced. The results of *KRT14* and *S100A8* are consistent with previous studies involving *AIRE* and *AIRE* mutant transfected 4D6 cells, while *IGFL1* expression was previously reported to be reduced with

R257X.[57] The reason for this discrepancy is unclear, but explanations may involve cell-specific differences: Because of AIRE's targeting of epigenetically or actively repressed genes any difference in which genes are repressed between differentiated cells will lead to a different AIRE induced transcriptome.[28, 35] Thus, any reporter gene system will by necessity be cell specific, and while being able to accurately predict AIRE mutational effect, will not be as representative for actual transcriptome changes in patient cells.

These results show that *KRT14* and *S100A8* exhibit the desired properties of reduced expression for the *AIRE* mutants compared to Wt and may be taken forward as potential AIRE reporter genes in a functional screening assay using HEK293FT cells.

However, both of these candidates exhibited only the first of our preferred reporter gene traits:

- High and significant differential expression between wildtype and mutants
- High expression in absolute terms
- Encoding cell surface proteins for easier FACS sorting.

Thus, in order to identify unknown AIRE regulated genes suiting these criteria, RNA sequencing was utilised.

Transcriptome sequencing of wildtype AIRE to identify candidate reporter genes

Having confirmed that AIRE is highly transcribed, translated, and able to induce the expression of known AIRE target genes, we next set out to identify novel reporter genes using transcriptome sequencing.

The first RNAseq performed, comparing the *AIRE* Wt transfected population with untransfected HEK293FT cells, showed a large number of significantly upregulated genes (**Fig. 12**), albeit most of them with a relatively modest log2 fold change. A similar experiment using stably transfected HEK293 cells with doxycycline-induced *AIRE* Wt expression reported 691 upregulated genes, and one downregulated gene with a threshold of 1 and -1 log2 fold change respectively.[84] This is comparable to our result, which found 219 upregulated genes and 32 downregulated genes using the same thresholds. A large number of upregulated

genes are consistent with previous studies using mTEC cells from Wt and knockout mice, yet our result does have a more substantial number of downregulated genes with lower differential expression.[47] However, this may be explained by the aforementioned cell-specific differences in AIRE transcriptomes.

Evaluation of candidate reporter genes identified from RNAseq analysis using qPCR

We selected putative reporter genes from the list of upregulated genes in the previous RNAseq, according to the selection strategy outlined in **table 1**. These genes were subsequently investigated using qPCR with gene-specific TaqMan probes. However, the result of this investigation did not yield any viable reporter genes fulfilling our preferred properties (**Fig. 14**). While the results of the qPCR did correlate well with the RNAseq result in the comparison between *AIRE* Wt and untransfected cells (**Fig. 15**), they did not show decreased expression in the *AIRE* mutants compared to the Wt. Hence, these genes may be upregulated not because of the transcriptional activity of AIRE, but rather as a response to the transfection. In retrospect, it might have served as a better control to use a plasmid without an AIRE insert in the comparison with *AIRE* Wt.

Another notable finding regarding the qPCR result was the increased expression of some of these candidate reporter genes in the R257X mutant compared to the Wt. This might further strengthen the suspicion that these genes are upregulated as a result of the transfection, or as a more unspecific result of the very high level of *AIRE* expression found for this variant (**Fig. 11**).

A reassuring finding of these analyses came from the comparison of all qPCR results with their comparative RNAseq results of the same biological RNA sample. We found a very robust correlation between the RNAseq and qPCR results with an R^2 of 0.87 indicating that the results from RNAseq are very robust (**Fig. 26**).

Change of RNAseq strategy to isolate transcriptome differences between AIRE wildtype and mutants

As a consequence of our inability to identify reporter genes merely using the *AIRE* Wt and untransfected cell comparison, the experimental strategy was changed to attempt to identify reporter genes using a comparison between *AIRE* mutants and the Wt. This would allow for

identification of genes that are the result of AIRE expression, yet are impaired when AIRE function is hampered. The Wt and the known mutants R257X and C311Y were investigated using an identical RNAseq approach as in the initial experiment. Differential gene expression analysis of the R257X mutant showed a majority of upregulated genes compared to the Wt (**Fig. 16**), while the C311Y mutant exhibited a remarkably similar transcriptome to the Wt (**Fig. 18**). To the best of our knowledge, neither of these *AIRE* mutants has previously been investigated using transcriptome analysis, either with microarray or RNAseq analysis and so these unexpected results are difficult to judge with respect to previous literature.

As a means to increase detection of genes with low expression in these populations, a change in methodology was made from the TotalRNA library preparation, which involves a step that removes the highly expressed ribosomal RNA, to a mRNA library preparation method using mRNA isolation with magnetic oligo dT beads. This approach ensures that any read in the sequencing is of the exome. However, neither method removes very highly expressed gene transcripts, such as the transfected *AIRE*, which in our case would be responsible for up to 20% of reads. RNAseq using the mRNA library prep protocol was performed on the previously investigated RNA samples. The results were overall very similar to using the TotalRNA library preparation in the R257X (**Fig. 19**) and C311Y (**Fig. 20**) mutants compared to the Wt.

In an attempt to further elucidate how stable the overall transcriptome results are between using two different library preparation methods, the RNAseq comparison between *AIRE* Wt and untransfected cells were redone with a new biological experiment using the mRNA preparation method. (**Fig. 21**). This comparison yielded a very similar result, albeit with a general reduction of the highly differentially expressed genes compared to the previous experiment (**Fig. 12**). However, when comparing the log2 fold change and the TPM between the two experiments, the results were (apart from a few outliers) very similar, though with the log2 fold change exhibiting a general trend towards a slightly less absolute difference in the mRNA population compared to the initial TotalRNA sequencing (**Fig. 22**).

Evaluating the replicability of RNAseq

As a means to validate our results of the comparisons between the mutants and the Wt, new biologically independent experiments were performed and sequenced with an identical technical setup. Both the R257X (**Fig. 23**) and C311Y (**Fig. 24**) showed similar patterns as

previous experiments, although with a reduction in the number of highly differentially expressed genes. This was especially apparent in the C311Y population where only 49 genes were found to be significantly differentially expressed compared to AIRE Wt. However, when comparing the TPM and log2 fold change values across experiments, the TPM values correlate well between both R257X and C311Y experiments (**Fig. 25**). Similarly, the log2 fold change of the two R257X RNAseqs correlates well, but with a slight tilt towards increased highly differential expression in the first experiment. The log2 fold change comparison of the C311Y shows that there are overall low levels of significantly differentially expressed genes in both biological replicates, and in particular for the last RNAseq.

Methodological optimisation

We were not able to identify possible reporter genes using the current approach, and as such, it may require some methodological changes to our protocol.

Our comparisons between the TotalRNA and mRNA library preparation methods show that the results are similar between the two. Although, the initial experiments for the R257X mutant compared with Wt suggested that mRNA (**Fig. 19**) preparation might yield an increase in the number of significantly differentially expressed genes compared to TotalRNA (**Fig. 16**), the subsequent experiment did not support that mRNA outperformed TotalRNA (**Fig. 23**).

One confounder in our attempt to increase our sensitivity to lowly expressed genes is the extreme overexpression of AIRE in our samples. AIRE is responsible for up to 20% of all reads in the RNAseqs of transfected cells, or in absolute terms up to 20 million reads in a total of 100 million reads. Thus, decreasing the amount of noncoding RNA may not have made much difference, as the reads would have a higher chance of populating to AIRE rather than any other transcript. In an attempt to properly evaluate the TPM estimates for genes between RNAseqs the inserted sequence from the plasmid was used to bioinformatically remove approximately 90% of AIRE reads in the samples before differential expression analysis. However, while this bioinformatical method works to reduce confounding reads in the final TPM estimation, it does not restore the lost sensitivity of those 20 million reads. Consequently, if using the same plasmid setup with extreme AIRE expression, a method to remove AIRE transcripts before RNAseq should be developed. One such method might be the use of AIRE or FLAG sequence-specific magnetic beads. Another alternative might be to

change the plasmid or promoter as a means to reduce the constitutive overexpression. One such alternative is the use of stably transduced cells where AIRE expression is under control of an inducible promoter, such as a doxycycline controlled promoter. This method has previously been used with success with AIRE.[84] However, while this might help in finding relevant candidate reporter genes, a drawback with this approach is that such stable transduction will likely not be possible to accomplish subsequently in the high throughput deep mutational scanning assay.

No transcriptome analysis has previously been done with the two mutants R257X and C311Y to the best of our knowledge. Thus, it is uncertain how representative these transcriptomes are, and how well they correspond to the physiological result of these mutations in vivo. The R257X mutant results in a truncated AIRE protein, which contains part of a general DNA interaction domain and the protein interaction domain CARD, thus it may still have some residual activity, consistent with our finding of a number of upregulated genes by this mutant. The R257X mutant may be degraded before translation in patient cells by the process of NMD.[85] However; this process is usually only triggered when the premature stop codon is located at least 50-55 nucleotides upstream of an exon-exon junction, which is not the case in the R257X mutant.[86] Notably, there are some exceptions to this rule, such as premature stop codons in the gene coding for the beta subunit of the T-cell receptor where a location 8-10 nucleotides upstream is sufficient for NMD.[87] The C311Y mutant may similarly not represent the physiological state in patients. As evident from the western blot experiments this mutant is substantially degraded at the protein stage, possibly by the UPR, and any remaining full-length AIRE may be the result of the extreme overexpression.

Nonetheless, these transcriptomes may be representative, yet any AIRE activity reporter genes might be so low in absolute transcripts that they are hard to identify using our RNAseq setup. This is further suggested by the fact that we were able to identify AIRE induced genes using qPCR that remained below the detectable limit of our RNAseq. Because AIRE activity has been reported to be largely stochastic, any processing of the dataset that is designed to remove noise may also remove any AIRE induced genes. DESeq2, used in our setup for differential expression analysis, makes certain assumptions about the dataset, and these assumptions may confound discovery of the unique transcripts of AIRE activity. Thus, it would have been useful to test out various analysis methods, such as baySeq differential expression

analysis, which does not make any assumptions about the dataset.[88] Another possibility is the use of the chemical etoposide, in an attempt to boost AIRE activity. Etoposide prevents religation of the DNA strands following TOP2 cutting, leading to DSB enrichment in the genome. The relaxation of the chromatin may allow AIRE to work more efficiently, or the DSB may lead to increased DNA-PK activity and has been used in previous studies to boost the number of differentially expressed genes.[42, 43, 84] However, etoposide does induce a large number of genes even without AIRE present, and so differentiating between AIRE and etoposide-induced genes may be difficult. In addition, our approach of using a low number of biological replicates and a large number of reads per sample might not have been optimal. It may not have increased the likelihood of discovering genes with low expression, but rather resulted in resequencing of the highly expressed transcripts. Instead, splitting the total amount of reads available amongst a greater number of biological replicates may decrease the noise floor and allow for identification of genes with lower differential expression. Indeed, such an approach has previously been evaluated and recommended.[89]

Concluding remarks

While we were able to establish an experimental system for investigation of AIRE activity, as confirmed by western blot, flow cytometry and qPCR, we were unable to identify any new AIRE inducible genes suitable as reporter genes using RNAseq.

However, we did confirm that the previously known AIRE inducible genes *KRT14* and *S100A8* have lower expression in two *AIRE* mutants compared to HEK293FT cells transfected with *AIRE* Wt (**Fig. 11**). This difference of expression in the mutants makes them possible reporter genes in this cell line. On the other hand, looking at the expression of these genes using RNAseq, the absolute expression in transcripts per million (TPM) was very low. This makes them challenging to use in a differential expression analysis. Pilot attempts to perform western blot analysis of transfected cells using KRT14 and S100A8 antibodies have thus far not been successful, as there was no visible band for KRT14, and only a very weakly expressed band for S100A8, in addition to mostly unspecific binding (data not shown).

In addition to identifying AIRE activity reporter genes, the process of RNAseq analysis as a tool in such an endeavour was investigated. Comparisons of the same population with the same technical RNAseq setup but in different biological experiments show a good correlation

between biological experiments when it comes to absolute expression in TPM (**Fig. 22b, 22c, 25b, 25c, 25e**). The absolute number of significantly differentially expressed genes and their log2 fold changes varies somewhat between experiments. However, the overall picture is very similar (**Fig 22a, 25a, 25d**). Comparing the different methods of RNAseq and qPCR, we found a highly linear correlation between the log2 of TPM in RNAseq and the reversed CT value from qPCR (**Fig. 26**) and the fold changes in both (**Fig. 15**). The sensitivity of a targeted qPCR experiment will, however, exceed the sensitivity of the more explorative transcriptome sequencing. Thus, AIRE reporter genes found using qPCR may not have enough reads for differential expression in RNAseq.

Future perspectives

With an inability to discover suitable reporter genes for AIRE activity, our approach to deep mutational scanning may need to be revised. The proposed stochastic nature of AIRE activity may limit the applicability of reporter genes in general, and the unspecific binding of AIRE to promoters of its target genes makes the development of a reporter system activated by AIRE challenging. It may be possible to increase the DNA binding specificity of AIRE by changing the SAND domain sequence. This AIRE variant may then be used as a part of a luciferase reporter system. However, this would be counterproductive as any mutations may then have a completely different effect than in the Wt sequence. Another approach is to use a proximity ligation assay to detect when AIRE and its partners are in proximity.[90] However, this indirect method does not necessarily detect activity, because AIRE partners may be recruited without AIRE actively inducing gene transcription depending on the mutation.

AIRE is a fascinating protein, with an extraordinary effect on the cells in which it is expressed. Albeit challenging, efforts to develop deep mutational scanning approaches for AIRE have the potential to improve APS-1 diagnostics, increase our understanding of AIREs biological role, and inform our knowledge of DNA repression, expression, and the nature of autoimmunity.

References

1. Aaltonen, J. *et al.* High-resolution physical and transcriptional mapping of the autoimmune polyendocrinopathy-candidiasis-ectodermal dystrophy locus on chromosome 21q22.3 by FISH. *Genome Res* **7**, 820-829 (1997).

2. Finnish-German, A.C. An autoimmune disease, APECED, caused by mutations in a novel gene featuring two PHD-type zinc-finger domains. *Nat Genet* **17**, 399-403 (1997).

3. Nagamine, K. *et al.* Positional cloning of the APECED gene. *Nat Genet* **17**, 393-398 (1997).

4. Meredith, M., Zemmour, D., Mathis, D. & Benoist, C. Aire controls gene expression in the thymic epithelium with ordered stochasticity. *Nat Immunol* **16**, 942-949 (2015).

5. Kyewski, B. & Klein, L. A central role for central tolerance. *Annu Rev Immunol* **24**, 571-606 (2006).

6. Heino, M. *et al.* Autoimmune regulator is expressed in the cells regulating immune tolerance in thymus medulla. *Biochem Biophys Res Commun* **257**, 821-825 (1999).

7. Gray, D., Abramson, J., Benoist, C. & Mathis, D. Proliferative arrest and rapid turnover of thymic epithelial cells expressing Aire. *J Exp Med* **204**, 2521-2528 (2007).

8. Rossi, S.W. *et al.* RANK signals from CD4(+)3(-) inducer cells regulate development of Aire-expressing epithelial cells in the thymic medulla. *J Exp Med* **204**, 1267-1272 (2007).

9. Bjorses, P. *et al.* Localization of the APECED protein in distinct nuclear structures. *Hum Mol Genet* **8**, 259-266 (1999).

10. Rinderle, C., Christensen, H.M., Schweiger, S., Lehrach, H. & Yaspo, M.L. AIRE encodes a nuclear protein co-localizing with cytoskeletal filaments:

altered sub-cellular distribution of mutants lacking the PHD zinc fingers. *Hum Mol Genet* **8**, 277-290 (1999).

11. Bjorses, P. *et al.* Mutations in the AIRE gene: effects on subcellular location and transactivation function of the autoimmune polyendocrinopathy-candidiasis-ectodermal dystrophy protein. *Am J Hum Genet* **66**, 378-392 (2000).

12. Kont, V. *et al.* DNA methylation signatures of the AIRE promoter in thymic epithelial cells, thymomas and normal tissues. *Mol Immunol* **49**, 518-526 (2011).

13. Haljasorg, U. *et al.* A highly conserved NF-kappaB-responsive enhancer is critical for thymic expression of Aire in mice. *Eur J Immunol* **45**, 3246-3256 (2015).

14. Herzig, Y. *et al.* Transcriptional programs that control expression of the autoimmune regulator gene Aire. *Nat Immunol* **18**, 161-172 (2017).

15. Yanagihara, T. *et al.* Intronic regulation of Aire expression by Jmjd6 for self-tolerance induction in the thymus. *Nat Commun* **6**, 8820 (2015).

16. Kelley, L.A., Mezulis, S., Yates, C.M., Wass, M.N. & Sternberg, M.J.E. The Phyre2 web portal for protein modeling, prediction and analysis. *Nature Protocols* **10**, 845 (2015).

17. Schrodinger, LLC. The PyMOL Molecular Graphics System, Version 2.0. 2015.

18. Pitkanen, J. *et al.* The autoimmune regulator protein has transcriptional transactivating properties and interacts with the common coactivator CREB-binding protein. *J Biol Chem* **275**, 16802-16809 (2000).

19. Bottomley, M.J. *et al.* The SAND domain structure defines a novel DNA-binding fold in transcriptional regulation. *Nat Struct Biol* **8**, 626-633 (2001).

20. Ferguson, B.J. *et al.* AIRE's CARD revealed, a new structure for central tolerance provokes transcriptional plasticity. *J Biol Chem* **283**, 1723-1731 (2008).

21. Ramsey, C., Bukrinsky, A. & Peltonen, L. Systematic mutagenesis of the functional domains of AIRE reveals their role in intracellular targeting. *Hum Mol Genet* **11**, 3299-3308 (2002).

22. Yoshida, H. *et al.* Brd4 bridges the transcriptional regulators, Aire and P-TEFb, to promote elongation of peripheral-tissue antigen transcripts in thymic stromal cells. *Proc Natl Acad Sci U S A* **112**, E4448-4457 (2015).

23. Surdo, P.L., Bottomley, M.J., Sattler, M. & Scheffzek, K. Crystal structure and nuclear magnetic resonance analyses of the SAND domain from glucocorticoid modulatory element binding protein-1 reveals deoxyribonucleic acid and zinc binding regions. *Mol Endocrinol* **17**, 1283-1295 (2003).

24. Yang, S., Bansal, K., Lopes, J., Benoist, C. & Mathis, D. Aire's plant homeodomain(PHD)-2 is critical for induction of immunological tolerance. *Proc Natl Acad Sci U S A* **110**, 1833-1838 (2013).

25. Koh, A.S. *et al.* Aire employs a histone-binding module to mediate immunological tolerance, linking chromatin regulation with organ-specific autoimmunity. *Proc Natl Acad Sci U S A* **105**, 15878-15883 (2008).

26. Org, T. *et al.* The autoimmune regulator PHD finger binds to non-methylated histone H3K4 to activate gene expression. *EMBO Rep* **9**, 370-376 (2008).

27. Chakravarty, S., Zeng, L. & Zhou, M.M. Structure and site-specific recognition of histone H3 by the PHD finger of human autoimmune regulator. *Structure* **17**, 670-679 (2009).

28. Heintzman, N.D. *et al.* Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nat Genet* **39**, 311-318 (2007).

29. Gaetani, M. *et al.* AIRE-PHD fingers are structural hubs to maintain the integrity of chromatin-associated interactome. *Nucleic Acids Res* **40**, 11756-11768 (2012).

30. Conti, E. & Kuriyan, J. Crystallographic analysis of the specific yet versatile recognition of distinct nuclear localization signals by karyopherin alpha. *Structure* **8**, 329-338 (2000).

31. Ilmarinen, T. *et al.* The monopartite nuclear localization signal of autoimmune regulator mediates its nuclear import and interaction with multiple importin alpha molecules. *FEBS J* **273**, 315-324 (2006).

32. Kumar, P.G. *et al.* The autoimmune regulator (AIRE) is a DNA-binding protein. *J Biol Chem* **276**, 41357-41364 (2001).

33. Saare, M., Rebane, A., Rajashekar, B., Vilo, J. & Peterson, P. Autoimmune regulator is acetylated by transcription coactivator CBP/p300. *Exp Cell Res* **318**, 1767-1778 (2012).

34. Chuprin, A. *et al.* The deacetylase Sirt1 is an essential regulator of Aire-mediated induction of central immunological tolerance. *Nat Immunol* **16**, 737-745 (2015).

35. Waterfield, M. *et al.* The transcriptional regulator Aire coopts the repressive ATF7ip-MBD1 complex for the induction of immunotolerance. *Nat Immunol* **15**, 258-265 (2014).

36. Halonen, M. *et al.* APECED-causing mutations in AIRE reveal the functional domains of the protein. *Hum Mutat* **23**, 245-257 (2004).

37. Bansal, K., Yoshida, H., Benoist, C. & Mathis, D. The transcriptional regulator Aire binds to and activates super-enhancers. *Nat Immunol* **18**, 263-273 (2017).

38. Jang, M.K. *et al.* The bromodomain protein Brd4 is a positive regulatory component of P-TEFb and stimulates RNA polymerase II-dependent transcription. *Mol Cell* **19**, 523-534 (2005).

39. Giraud, M. *et al.* An RNAi screen for Aire cofactors reveals a role for Hnrnpl in polymerase release and Aire-activated ectopic transcription. *Proc Natl Acad Sci U S A* **111**, 1491-1496 (2014).

40. Kanno, T. *et al.* BRD4 assists elongation of both coding and enhancer RNAs by interacting with acetylated histones. *Nat Struct Mol Biol* **21**, 1047-1057 (2014).

41. Giraud, M. *et al.* Aire unleashes stalled RNA polymerase to induce ectopic gene expression in thymic epithelial cells. *Proc Natl Acad Sci U S A* **109**, 535-540 (2012).

42. Zumer, K., Low, A.K., Jiang, H., Saksela, K. & Peterlin, B.M. Unmodified histone H3K4 and DNA-dependent protein kinase recruit autoimmune regulator to target genes. *Mol Cell Biol* **32**, 1354-1362 (2012).

43. Abramson, J., Giraud, M., Benoist, C. & Mathis, D. Aire's partners in the molecular control of immunological tolerance. *Cell* **140**, 123-135 (2010).

44. Derbinski, J., Schulte, A., Kyewski, B. & Klein, L. Promiscuous gene expression in medullary thymic epithelial cells mirrors the peripheral self. *Nat Immunol* **2**, 1032-1039 (2001).

45. Anderson, M.S. *et al.* Projection of an immunological self shadow within the thymus by the aire protein. *Science* **298**, 1395-1401 (2002).

46. Anderson, M.S. *et al.* The cellular mechanism of Aire control of T cell tolerance. *Immunity* **23**, 227-239 (2005).

47. Sansom, S.N. *et al.* Population and single-cell genomics reveal the Aire dependency, relief from Polycomb silencing, and distribution of self-antigen expression in thymic epithelia. *Genome Res* **24**, 1918-1931 (2014).

48. Org, T. *et al.* AIRE activated tissue specific genes have histone modifications associated with inactive chromatin. *Hum Mol Genet* **18**, 4699-4710 (2009).

49. Guerau-de-Arellano, M., Mathis, D. & Benoist, C. Transcriptional impact of Aire varies with cell type. *Proc Natl Acad Sci U S A* **105**, 14011-14016 (2008).

50. Villasenor, J., Besse, W., Benoist, C. & Mathis, D. Ectopic expression of peripheral-tissue antigens in the thymic epithelium: probabilistic, monoallelic, misinitiated. *Proc Natl Acad Sci U S A* **105**, 15854-15859 (2008).

51. Brennecke, P. *et al.* Single-cell transcriptome analysis reveals coordinated ectopic gene-expression patterns in medullary thymic epithelial cells. *Nat Immunol* **16**, 933-941 (2015).

52. Owen, J.A., Punt, J., Stranford, S.A., Jones, P.P. & Kuby, J. *Kuby immunology*, 7th edn. W.H. Freeman: New York, 2013.

53. Klein, L., Hinterberger, M., Wirnsberger, G. & Kyewski, B. Antigen presentation in the thymus for positive selection and central tolerance induction. *Nat Rev Immunol* **9**, 833-844 (2009).

54. Aschenbrenner, K. *et al.* Selection of Foxp3+ regulatory T cells specific for self antigen expressed and presented by Aire+ medullary thymic epithelial cells. *Nat Immunol* **8**, 351-358 (2007).

55. Ahonen, P. Autoimmune polyendocrinopathy--candidosis--ectodermal dystrophy (APECED): autosomal recessive inheritance. *Clin Genet* **27**, 535-542 (1985).

56. Cetani, F. *et al.* A novel mutation of the autoimmune regulator gene in an Italian kindred with autoimmune polyendocrinopathy-candidiasis-ectodermal dystrophy, acting in a dominant fashion and strongly cosegregating with hypothyroid autoimmune thyroiditis. *J Clin Endocrinol Metab* **86**, 4747-4752 (2001).

57. Oftedal, B.E. *et al.* Dominant Mutations in the Autoimmune Regulator AIRE Are Associated with Common Organ-Specific Autoimmune Diseases. *Immunity* **42**, 1185-1196 (2015).

58. Wolff, A.S. *et al.* Autoimmune polyendocrine syndrome type 1 in Norway: phenotypic variation, autoantibodies, and novel mutations in the autoimmune regulator gene. *J Clin Endocrinol Metab* **92**, 595-603 (2007).

59. Husebye, E.S., Perheentupa, J., Rautemaa, R. & Kampe, O. Clinical manifestations and management of patients with autoimmune polyendocrine syndrome type I. *J Intern Med* **265**, 514-529 (2009).

60. Neufeld, M., Maclaren, N.K. & Blizzard, R.M. Two types of autoimmune Addison's disease associated with different polyglandular autoimmune (PGA) syndromes. *Medicine (Baltimore)* **60**, 355-362 (1981).

61. Bruserud, O. *et al.* A Longitudinal Follow-up of Autoimmune Polyendocrine Syndrome Type 1. *J Clin Endocrinol Metab* **101**, 2975-2983 (2016).

62. Husebye, E.S., Anderson, M.S. & Kampe, O. Autoimmune Polyendocrine Syndromes. *N Engl J Med* **378**, 1132-1141 (2018).

63. Wang, C.Y. *et al.* Characterization of mutations in patients with autoimmune polyglandular syndrome type 1 (APS1). *Hum Genet* **103**, 681-685 (1998).

64. Bruserud, O., Oftedal, B.E., Wolff, A.B. & Husebye, E.S. AIRE-mutations and autoimmune disease. *Curr Opin Immunol* **43**, 8-15 (2016).

65. Araya, C.L. & Fowler, D.M. Deep mutational scanning: assessing protein function on a massive scale. *Trends Biotechnol* **29**, 435-442 (2011).

66. Fowler, D.M. & Fields, S. Deep mutational scanning: a new style of protein science. *Nat Methods* **11**, 801-807 (2014).

67. Melnikov, A., Rogov, P., Wang, L., Gnirke, A. & Mikkelsen, T.S. Comprehensive mutational scanning of a kinase in vivo reveals substrate-dependent fitness landscapes. *Nucleic Acids Res* **42**, e112 (2014).

68. Bonner, W.A., Hulett, H.R., Sweet, R.G. & Herzenberg, L.A. Fluorescence activated cell sorting. *Rev Sci Instrum* **43**, 404-409 (1972).

69. Tewhey, R. *et al.* Direct Identification of Hundreds of Expression-Modulating Variants using a Multiplexed Reporter Assay. *Cell* **165**, 1519-1529 (2016).

70. Heid, C.A., Stevens, J., Livak, K.J. & Williams, P.M. Real time quantitative PCR. *Genome Res* **6**, 986-994 (1996).

71. Wilhelm, B.T. *et al.* Dynamic repertoire of a eukaryotic transcriptome surveyed at single-nucleotide resolution. *Nature* **453**, 1239-1243 (2008).

72. Nagalakshmi, U. *et al.* The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science* **320**, 1344-1349 (2008).

73. Van Dilla, M.A., Trujillo, T.T., Mullaney, P.F. & Coulter, J.R. Cell microfluorometry: a method for rapid fluorescence measurement. *Science* **163**, 1213-1214 (1969).

74. Andrews, S. FastQC A Quality Control tool for High Throughput Sequence Data. 0.11.7 ed. http://www.bioinformatics.babraham.ac.uk/projects/fastqc/; 2010.

75. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S.L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* **10**, R25 (2009).

76. Bray, N.L., Pimentel, H., Melsted, P. & Pachter, L. Near-optimal probabilistic RNA-seq quantification. *Nat Biotechnol* **34**, 525-527 (2016).

77. Soneson, C., Love, M.I. & Robinson, M.D. Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences. *F1000Res* **4**, 1521 (2015).

78. Rainer, J. EnsDb.Hsapiens.v86: Ensembl based annotation package. R package version 2.99.0 ed. https://bioconductor.org/packages/release/data/annotation/html/EnsDb.Hsapiens.v86.html; 2017.

79. Love, M.I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* **15**, 550 (2014).

80. Wickham, H. Ggplot2 : elegant graphics for data analysis. (2009).

81. Eden, E., Navon, R., Steinfeld, I., Lipson, D. & Yakhini, Z. GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists. *BMC Bioinformatics* **10**, 48 (2009).

82. Supek, F., Bosnjak, M., Skunca, N. & Smuc, T. REVIGO summarizes and visualizes long lists of gene ontology terms. *PLoS One* **6**, e21800 (2011).

83. Gasteiger, E. *et al.* Protein Identification and Analysis Tools on the ExPASy Server. In: Walker, J.M. (ed). *The Proteomics Protocols Handbook*. Humana Press: Totowa, NJ, 2005, pp 571-607.

84. Guha, M. *et al.* DNA breaks and chromatin structural changes enhance the transcription of autoimmune regulator target genes. *J Biol Chem* **292**, 6542-6554 (2017).

85. Maquat, L.E. When cells stop making sense: effects of nonsense codons on RNA metabolism in vertebrate cells. *RNA* **1**, 453-465 (1995).

86. Nagy, E. & Maquat, L.E. A rule for termination-codon position within intron-containing genes: when nonsense affects RNA abundance. *Trends in Biochemical Sciences* **23**, 198-199 (1998).

87. S., C.M., S., L. & F., W.M. A splicing-dependent regulatory mechanism that detects translation signals. *The EMBO Journal* **15**, 5965-5975 (1996).

88. Hardcastle, T.J. & Kelly, K.A. baySeq: empirical Bayesian methods for identifying differential expression in sequence count data. *BMC Bioinformatics* **11**, 422 (2010).

89. Schurch, N.J. *et al.* How many biological replicates are needed in an RNA-seq experiment and which differential expression tool should you use? *RNA* **22**, 839-851 (2016).

90. Soderberg, O. *et al.* Direct observation of individual endogenous protein complexes in situ by proximity ligation. *Nat Methods* **3**, 995-1000 (2006).