# Loss of Y in leukocytes, dysregulation of autosomal immune genes and disease risks

Jan P. Dumanski,[1, 2, *] Jonatan Halvardson,[1] Hanna Davies,[1, #] Edyta Rychlicka-Buniowska,[3, #] Jonas Mattisson,[1, #] Behrooz Torabi Moghadam,[1, #] Noemi Nagy,[4] Kazimierz Węglarczyk,[5] Karolina Bukowska-Strakova,[5] Marcus Danielsson,[1] Paweł Olszewski,[3] Arkadiusz Piotrowski,[2] Aleksandra Ambicka,[6] Marcin Przewoźnik,[6] Łukasz Bełch,[7] Tomasz Grodzicki,[8] Piotr L. Chłosta,[7] Stefan Imreh,[9] Vilmantas Giedraitis,[10] Lars Lannfelt,[10] Lena Kilander,[10] Jessica Nordlund,[11] Adam Ameur,[1] Ulf Gyllensten,[1] Åsa Johansson,[1] Alicja Józkowicz,[12] Maciej Siedlar,[5] Alicja Klich-Rączka,[8] Janusz Jaszczyński,[13] Stefan Enroth,[1] Jarosław Baran,[5] Martin Ingelsson,[10] Janusz Ryś,[6] Lars A. Forsberg,[1, 14, *]

[1] Dept. of Immunology, Genetics and Pathology and Science for Life Laboratory, Uppsala University, Uppsala, Sweden

[2] Faculty of Pharmacy and 3P Medicine Laboratory, International Research Agendas Programme, Medical University of Gdańsk, Gdańsk, Poland

[3] International Research Agendas Programme, 3P Medicine Laboratory, Medical University of Gdańsk, Gdańsk, Poland

[4] Dept. of Cell and Molecular Biology, Karolinska Institutet, Solnavägen 9, Stockholm, Sweden

[5] Dept. of Clinical Immunology, Institute of Paediatrics, Jagiellonian University, Collegium Medicum, Wielicka 265, 30-663 Kraków, Poland

[6] Dept. of Tumour Pathology, Maria Skłodowska-Curie Memorial Cancer Centre and Institute of Oncology, Kraków, Branch, Kraków, Poland

[7] Dept. and Clinic of Urology, Jagiellonian University, Collegium Medicum, ul. Grzegórzecka 18, 31-531 Kraków, Poland

[8] Dept. and Clinic of Internal Medicine and Gerontology, Jagiellonian University, Collegium Medicum, ul. Sniadeckich 10, 31-531 Kraków, Poland

[9] Dept. Oncology-Pathology, Karolinska Institutet, 171 64 Solna, Sweden

[10] Dept. of Public Health and Caring Sciences / Geriatrics, Uppsala University, 751 85 Uppsala, Sweden

[11] Dept. of Medical Sciences, Uppsala University, 751 85 Uppsala, Sweden

[12] Dept. of Medical Biotechnology, Faculty of Biochemistry, Biophysics and Biotechnology, Jagiellonian University, Gronostajowa 7, 30-387 Kraków, Poland

[13] Dept. of Urology, Maria Sklodowska -Curie Memorial Cancer Centre and Institute of Oncology, Kraków Branch, Kraków, Poland

[14] Beijer Laboratory of Genome Research, Uppsala University, Uppsala, Sweden

*Correspondance: jan.dumanski@igp.uu.se and lars.forsberg@igp.uu.se

# these authors contributed equally

## Abstract

Mosaic loss of chromosome Y (LOY) in blood is linked to increased risk for morbidity and mortality in men. LOY is the most common acquired mutation and is associated with diseases such as cancer and Alzhemer's disease. We studied DNA, RNA and proteins in bulk, sorted- and single-cells *in vivo* and *in vitro*. We show that Alzheimer's disease and prostate cancer patients had more LOY in NK cells and CD4+ T-lymphocytes, respectively. Furthermore, gene expression was profoundly altered in cells with LOY in a pleiotropic fashion and autosomal genes important for normal immune cell functions showed LOY associated transcriptional effect. Proteomic analysis also indicated that LOY leaves a footprint in the plasma proteome. We provide the first mechanistic explanation for the associations between LOY in blood and risk for disease in other organs.

## Keywords

Post-zygotic mutation; somatic mutation; immune-surveillance; Alzheimer's disease; neurodegeneration; cancer; prostate cancer; single-cell; RNA sequencing; droplet digital PCR; ddPCR; cell sorting; fluorescent activated cell sorting; FACS; immune system; inflammation; sex chromosomes; sexual dimorphism; pleiotropy; mosaicism; loss of chromosome Y; LOY; aneuploidy; oligo-clonality; ocLOY; male-specific region of chromosome Y; MSY; pseudo-autosomal region; PAR; aberrant clonal expansion; ACE; LOY associated transcriptional effect; LATE; transcriptome; proteome; genome; genomics; *CXCL5* gene; *LY6E* gene; *CD99* gene; *LAG3* gene; *LY6E* gene; *ANXA1* gene; *CSF1R* gene; *APBA3* gene; *CSF2RA* gene; *JAK3* gene; *IL15* gene; *CD3D* gene; *CD3G* gene; *CADM1* gene; *PYCARD* gene; *LILRB2* gene; *LILRA1* gene; *KLRG1* gene; *HLA-DRA* gene; *CCL5* gene; *TCL1A* gene

## Introduction

Chromosome Y is required in mammals to override the development of the default female sex. The human X and Y chromosomes evolved from a pair of autosomes over the past 300 million years. During the early history of chromosome Y, a genetic decay with gene loss has occurred. However, the mammalian chromosome Y has been remarkably stable for the past 25 million years [1, 2]. The molecular studies of chromosome Y are challenging mainly due to its rich content of highly repetitive, low-level redundancy and palindromic sequences. Although the male DNA was included in the early genome sequencing projects, the Y chromosome was largely ignored, due to difficulties in sequence assembly. Nevertheless, chromosome Y contains genes important for a wide range of functions (**Fig. 1)** [2], suggesting that it is far from a gene desert. However, the notion of chromosome Y being a genetic wasteland has been, and perhaps still is, difficult to erase.

Since the earliest cytogenetic analyses, LOY has been known to occur in hematopoietic cells [3, 4], resulting in a complete lack of almost 2% of the haploid nuclear genome in a cell with this aneuploidy. However, the phenotypical consequences of LOY, on the level of a cell or an entire organism, have remained unclear. The prevailing consensus under several decades suggested that LOY should be considered phenotypically neutral and related to normal aging [5-7]. This further added to the misconception surrounding chromosome Y as only being important for male sex determination and other male-specific traits, but largely insignificant beyond that.

Recent epidemiological analyses have challenged the view that LOY in blood cells is phenotypically neutral. Studies have identified increased risks for men with LOY in connection with all-cause mortality [8, 9], Alzheimer’s disease [10], various forms of cancer [8, 11-13], autoimmune conditions [14, 15], age-related macular degeneration [16], cardiovascular disease [17], type 2 diabetes and obesity [9]. Furthermore, it has been known for centuries that males have a shorter life expectancy [18-20]. Hence, LOY as a male specific risk factor, showing reproducible associations with various common disorders, could help explaining this difference.

LOY in blood is the most frequent post-zygotic mutation, detectable in 20% of the UK Biobank male population [21], reaching 40% in 80 years old males [22], and 57% in 93 years old males [23]. Furthermore, LOY is not restricted to the hematopoietic system, since it has also been described in other non-cancerous tissues, although with much lower frequencies [17, 23]. Our understanding of why LOY occurs is also limited. Strong associations with age and smoking have been reported [8, 10, 24, 25] and these factors are related to a continuously increasing mutational load throughout the genome in all somatic cells, eventually also affecting genes that are responsible for correct segregation of chromosomes. Inherited genetic predisposition for LOY has also been described [21, 25, 26].

It is not known whether associations between LOY and increased risks for disease represent causal relationships, and if so, what the underlying mechanisms could be. The challenge that we address here is to move from epidemiological associations to mechanistic explanations on multiple levels of analysis such as DNA, mRNA and proteins. Specifically, we studied: (i) frequency of LOY in subsets of leukocytes from patients with prostate cancer (PC), Alzheimer's disease (AD) as well as controls; (ii) changes in transcriptomes of single cells and cellular subsets in bulk; and (iii) alterations of plasma protein levels in men with LOY.

## Different distribution of LOY in six types of sorted leukocytes among men with AD and PC

Studies of associations between LOY and various outcomes have until now focused on analyses of DNA from peripheral blood. Our investigation of a possible causality behind associations between LOY in blood and risk for diseases in other organs started with investigation of LOY in six major types of leukocytes sorted with a Fluorescent Activated Cell Sorter, in patients with AD, PC and healthy age-matched controls. Our working hypothesis was that there might be a difference in LOY frequency in various types of leukocytes between patients with AD and PC. The six subsets of sorted leukocytes were: CD19+ B lymphocytes; CD4+ T lymphocytes; CD8+ T-cytotoxic lymphocytes; Natural Killer (NK) cells; granulocytes; and monocytes. In total, we sorted leukocytes from 408 subjects; 228 patients with diagnosis of AD or PC and 180 healthy controls. We assessed the level of LOY from DNA in each subset using the most established method, i.e. SNP-arrays [8, 10, 24, 27]. The AD analysis included 215 subjects (121 patients and 94 age-matched controls) that were ≥70 years of age and free from cancer. The examination of PC included 263 subjects; 107 patients and 156 age-matched controls that were ≤80 years of age and free from AD (**Figs. 2** and **S1-S6**). The details of the studied subjects are shown in **Table S1**. Two general conclusions could be drawn from comparisons of the level of LOY in each cell type in patients and controls. One is the overall trend that the degree of LOY mosaicism was higher in patients in all studied cell types, as shown in 12 pairwise comparisons from patients with AD and PC versus controls (**Fig. 2**). This is well in line with previous studies using whole blood DNA. The overall level of LOY, however, varied substantially between cell types with highest frequencies in the myeloid lineage and considerably lower levels in lymphoid cells (**Figs. 2,** and **S1–S7**). The frequency of LOY in six cell types using four different cut-offs for LOY scoring is shown in **Fig. S7**. The second general conclusion is that LOY was not equally distributed between different subsets of leukocytes in patients with AD or PC vs controls. It has been discussed whether the associations between LOY in leukocytes with various diseases and its strong association with increasing age, could merely reflect a stochastic age-dependent process of accumulation of various mutations in all cells [28]. Our above results cells contradict this hypothesis.

We first used unadjusted logistic regression models with the continous mLRRY as response variable [8, 10], which showed that men diagnosed with AD had more LOY in NK cells compared with controls (*p*=0.0109, **Fig. 2A**) and that men diagnosed with PC had more LOY in CD4+ T lymphocytes (*p*=0.0312, **Fig. 2D**). This result was confirmed by logistic regression models fitting the effects from age and smoking which are known LOY-associated confounders (**Fig. 2B** and **E**). Remarkably, in NK cells, LOY was 4.4 times more common in AD patients than in controls, using a cut-off representing ≥40% of NK cells with aneuploidy (**Fig. S4A).** This could be compared to previous study of AD and LOY scored in whole blood DNA, showing an adjusted hazard ratio of 5.3 for AD-free follow up [10]. Furthermore, in PC patients we observed that CD4+ T-lymphocytes were more commonly affected with LOY. However, the level of LOY in this sorted subset was considerably lower than in NK cells from AD patients. This might reflect that CD4+ lymphocytes is a heterogeneous group of cells and a possibility that LOY in only a fraction of these cells is connected with an increased risk for disease in PC patients. Consequently, further studies should focus on specific subsets within CD4+ lymphocytes.

NK cells are an important part of the host defense and the immune system has been suggested as severely disturbed in AD [29, 30]. Furthermore, NK cells have been proposed as involved in the pathogenesis of AD in studies performed in humans and animal models [31-39]. However, the role of NK cells in AD has not been sufficiently explored, especially when the accumulation of post-zygotic mutations acquired during lifetime (such as LOY) is considered. The presented results support the possibility that LOY in certain types of leukocytes contributes to increased risk for specific diseases. Furthermore, the investigations of LOY in the context of associations with various diseases have until now typically used whole blood DNA. Overall, our results suggest that this strategy should be revised and we ought to consider using in future LOY studies DNA from the subsets of leukocytes that are most relevant for the disease under investigation. In conclusion, we provide the first indication that LOY in NK cells and CD4+ lymphocytes might contribute to the pathogenesis of AD and cancer, respectively, presumably by impairment of normal immune system functions.

Our results also suggested that cell clones with LOY often might be a result of oligo-clonal cell expansions in peripheral blood. In **Figure 2C** and **F**, we show data for oligo-clonality of LOY (ocLOY) in men with a diagnosis of AD or PC vs controls. OcLOY was defined as having LOY in ≥20% of cells in at least two of the six types of sorted leukocytes. We identified 52 subjects according to this definition and found that a majority of men with ocLOY had diagnosis of AD or PC (**Figs. S8-S10**). In the AD cohort, ocLOY was significantly higher in patients (OR=2.64, *p*=0.0127) with 24.0% of patients (median age=80, range=70-95) compared to 10.6% of controls (median age=75, range=70-95). A comparable difference was observed in the PC cohort in which 9.3% of the patients (median age=66, range=51-80) displayed compared to 5.1% of controls (median age=69, range=45-79). An overall analysis combining all AD and PC patients showed a significantly higher level of ocLOY compared to controls (OR=2.74, *p*=0.0026). Since the median age was slightly higher in patients than controls, we also investigated the frequency of ocLOY in a subset of subjects between 70 and 80 years of age. This analysis confirmed that OcLOY was more common (OR=2.87, *p*=0.0462) in patients (N=110, 18.2% with OcLOY, median age=76, range=70-80) than controls (N=70, 7.1% with OcLOY, median age=74, range=70-79). These results may suggest that the overall load of LOY in leukocytes may also contribute to disease development, in addition to involvement of specific cellular subsets with LOY in patients with AD and PC (i.e. NK cells and CD4+ T-lymphocytes, respectively).

## The frequency of cells with LOY can be estimated from mRNA transcriptome

Our next goal was to investigate mRNA expression in samples of sorted cells and single-cells as well as in cells *in vitro*. The first step was to estimate the level of LOY from mRNA datasets. We studied transcriptome using two independent technologies, namely single cell RNA sequencing (scRNAseq) of peripheral blood mononuclear cells (PBMCs) and sequencing of bulk RNA (RNAseq) from the sorted leukocytes and 13 lymphoblastoid cell lines (LCLs). The latter were derived from single cell clones and encompassed five LCLs with 100% LOY cells as well as eight LCLs containing essentially 100% of cells with chromosome Y (see below).

For scRNAseq results, each sequenced cell with expression of autosomal genes, but without any transcripts from genes located in the male specific region of chromosome Y (MSY) was considered as a LOY cell. The single cell analyses of PBMCs collected from 29 men (26 diagnosed with AD) generated a scRNAseq dataset encompassing 73,606 cells analyzed using established algorithms, as described in methods and elsewhere [21]. The distribution and level of LOY estimated from scRNAseq in four major cell types (i.e. NK cells, monocytes, B- and T lymphocytes) is shown in **Figure 3A** and **B.** The level of LOY in NK cells, monocytes, B- and T lymphocytes in the pooled scRNAseq dataset were 27% (range 7-87%), 23% (range 7-87%), 7% (range 2-40%) and 3% (range 1-6%), respectively. Noteworthy, we identified single cells with LOY in all subjects studied by scRNAseq. Concerning bulk RNAseq data, the level of LOY was estimated by comparing the stabilized mean number of mRNA transcripts from MSY genes with the corresponding values of autosomal genes. The resulting ratios were rescaled to fit between 0 and 100, to produce estimates of LOY. This analysis was performed for 134 samples (NK cells, monocytes and granulocytes) derived from cell sorting of 51 subjects (**Fig. 3C** and **D**).

In order to evaluate the mRNA-based LOY estimates, we performed pairwise comparisons for scRNAseq and RNAseq against the DNA-based LOY scoring, using the same set of samples (**Fig. 4**). We also compared the two mRNA-based methods and the Pearson's correlation coefficients for these pairwise analyses were 0.93, 0.92 and 0.91, respectively. This shows that the performance of

LOY estimations from scRNAseq and RNAseq is reliable. Concerning the 13 LCLs, LOY estimation was performed using two methods; array-based SNP genotyping and droplet digital PCR (ddPCR); the latter utilizing a common polymorphism within amelogenin gene located on chromosomes X and Y [40]. This ddPCR-assay is a recent methodological development allowing rapid scoring of LOY and has been described elsewhere [41]. The estimations of the level of LOY from both technologies were highly concordant with Pearson's correlation coefficient of 0.9961. Details of the 13 cell lines are shown in **Table S1**.

## LOY causes a reduced level of expression of genes from chromosome Y

The study of **L**OY **A**ssociated **T**ranscriptional **E**ffect (LATE) started with analyses of the expression of genes located on chromosome Y. According to Ensembl (v. 95), the Y chromosome contains 64 protein coding genes: 45 in MSY and 19 in the pseudo-autosomal regions (PARs). We detected in total expression of 20 protein coding genes from chromosome Y in leukocytes studied with RNAseq and scRNAseq (**Fig. 1**). The determination of normally expressed genes from chromosome Y was based on observations in samples and cells without LOY. Hence, in the bulk RNAseq dataset, genes with at least 10 sequencing reads in more than 75% of the samples without LOY were considered normally expressed. For the scRNAseq data, expression in at least 10% of single cells without LOY was required to be defined as a normally expressed gene. Of the 20 protein coding genes located on chromosome Y and showing expression using above criteria, seven are located in the MSY and 13 in the PARs.

After establishing which MSY genes were normally expressed in the studied cell subsets, we investigated the level of expression of these genes in relation to the level of LOY in each sample. The analyses of expression of MSY genes were performed with the dataset generated with bulk RNAseq from sorted cells. Similar analyses with the scRNAseq data could not be performed, since single cells with LOY are per definition lacking transcripts from MSY genes. The analyses of the bulk RNAseq dataset showed that the normalized abundance of transcripts from the MSY genes clearly decreased with increasing level of LOY (**Fig. 5A**). A corresponding analysis of the genes located in the PAR also showed a decrease in transcript abundance with increasing levels of LOY. However, this LOY associated decrease was not as distinct as for the MSY genes, which can be explained by sustained expression of the chromosome X-copy of PAR genes (**Fig. 5B**). The level of expression in relation to LOY for the normally expressed genes located on chromosome Y is shown in **Figure S11** and data is provided in **Table S2**. The effect on gene expression from a continuous LOY estimate was quantified using an established algorithm calculating the size and direction of the differential expression (DE) using regression analysis [42]. DE-values close to zero suggest no change and positive and negative DE-values indicate over- and under-expression, respectively. Among the PAR genes showing a lower expression as an effect of LOY is the *CD99* gene **(Fig. 6B** and **D, Table S2)**. Its product is a cell surface glycoprotein involved in processes such as leukocyte migration, cell adhesion and apoptosis; e.g. by functioning as a diapedesis-mediating receptor central for migration of monocytes through endothelial junctions [43, 44]. A lower expression of *CD99* in leukocytes with LOY might therefore mitigate extravasation and thus impair the recruitment and movement of leukocytes from circulation towards somatic tissues and sites of disease. In summary, samples with LOY showed a lower abundance of mRNA transcripts from genes located in MSY and PARs. This provides proof of principle and serve as an internal control that LATE can be confidently estimated by transcriptome analyses.

## Effects beyond the Y chromosome; autosomal LATE genes

In order to investigate whether LOY causes transcriptional effects of genes located on other chromosomes, we first determined the autosomal genes that are normally expressed in each studied cell type and subsequently measured their expression level in relation to LOY. The criteria used to identify normally expressed autosomal genes were the same as for the analyses of genes located on chromosome Y (see above). To determine differential expression of these genes as an effect of LOY we used two approaches: for the dataset generated with RNAseq we used the DESeq2 package in R [42] and for the scRNAseq dataset, negative binomial testing was applied as implemented in the R library Seurat [45]. Panel A in **Table 1** displays the number of normally expressed genes as well as the number of LATE genes (using two stringency levels) in the different cell types studied by the two mRNA-based technologies. In **Figure 6**, an additional criterion was applied to identify only the highly expressed LATE genes.

Two types of global LATE analyses were performed and each showed that the overall autosomal gene expression might be affected by LOY. The first was LATE study of *in vitro* cultured LCLs with and without LOY (**Fig. 5 C**). The LCLs were generated using Epstein–Barr virus transformation of lymphocytes collected from male donors with LOY in whole blood. DNA array as well as validation experiments using ddPCR were performed on viable LCLs and identified five monoclonal LCLs with 100% LOY cells as well as eight LCLs in which all cells carried chromosome Y. RNAseq of these 13 LCLs was performed and expression was investigated by principal components analysis. The first two principal components explained 45% and 19% of the total variance and a Kolmogorov–Smirnov test, comparing the variance in PC1 for LCLs with and without LOY, showed that the gene expression was affected in LCLs with LOY (D=1.0, $p$=0.0016). However, these results should be interpreted with caution since all LCLs with LOY in this analysis originated from the same donor (**Fig. 5C and Fig. S12**). It is nonetheless interesting that LCL1_2 clustered closer to the LCLs without LOY originating from other subjects, than to the other LCLs (with LOY) from this donor.

The second analysis of genome-wide LATE was performed using the dataset generated by scRNAseq from PBMCs collected *in vivo* (**Fig. 5D**). The hypothesis was that changes in expression as an effect of LOY would be larger in autosomal genes that are normally co-expressed with MSY genes, compared with genes that are not co-expressed with MSY genes. We mined the SEEK database for autosomal genes normally expressed in leukocytes and showing co-expression with six MSY genes expressed in all types of studied leukocytes (**Fig. 1**, **Figs. 5A** and **S11**) and identified 275 co-expressed genes and 12852 control genes, i.e. without evidence of co-expression. The results showed a significantly larger dysregulation of the co-expressed genes, compared with control genes (Wilcoxon rank sum test: $p$=0.0021). Similar analyses were also performed separately within monocytes, NK cells, T- and B lymphocytes, in which 142, 116, 95 and 122 co-expressed autosomal genes as well as 3677, 3230, 2752 and 3075 autosomal genes without co-expression were identified, respectively. This showed that the global LATE was strongest in monocytes and T lymphocytes and weaker in NK cells and not detectable in B lymphocytes (**Fig. 5D**).

Hence, both types of global LATE comparisons described above supported an overall effect from LOY on expression of autosomal genes and the next step was investigation of specific autosomal LATE genes. In total, we identified 4026 LATE genes showing varying degree of differential expression in sorted leukocytes and LCLs studied by RNAseq as well as in PBMCs using scRNAseq. The analyses of cells *in vivo* identified 3366 LATE genes (3110 by RNAseq and 337 by scRNAseq) and 904 LATE genes by RNAseq in the LCLs (**Table 1A**). Complete lists of all 4026 LATE genes identified in these datasets are provided in **Table S2**. This table shows the level of normal and differential expression, the unadjusted p-values and significance levels after correction for multiple testing, by cell type and technology. Furthermore, **Figs. S13-S16** display the observed differential expression of LATE genes in different cell types. A clear difference between bulk RNAseq and scRNAseq is that expression of about 3 times more genes could be detected with the former method. The main reasons for this difference might be because a gene is not expressed in a particular cell at the time of sequencing or a failure to be sequenced due to technological limitations. For example, scRNAseq has lower depth of sequencing and measures only transcripts tagged with polyA-tail. In contrast, in bulk RNA sequencing, a much larger number of cells are studied simultaneously, yielding transcript reads also from genes with lower normal expression and from genes with temporal variation in expression.

Comparing the number of LATE genes within cell types with the fraction of LATE genes shared between cell types provide interesting insights regarding possible pleiotropic effects of LOY (**Tables 1B** and **S3**). Pleiotropy can be defined as an effect from one feature at the genetic level to multiple characteristics at the phenotypic level [46]. Both in the bulk RNAseq and in the scRNAseq analyses, the fraction of LATE genes was substantially larger within specific cell types than the fraction of LATE genes shared between different subsets (Kruskal-Wallis Chi-squared= 6.5, $p$=0.0105). For instance, up to 15% of normally expressed genes showed LATE in NK cells, but only about 2% are shared between NK vs granulocytes and NK vs monocytes. Moreover, the fraction of shared LATE genes was larger among the *in vivo* studied cells compared with the LCLs (Kruskal-Wallis Chi-squared= 3.9, *p*=0.0495). These results suggest that LOY may preferentially cause expression differences of autosomal LATE genes in a cell type specific manner. This pleiotropic effect can to some extent be explained by the fact that different cell types normally express distinct sets of genes. Hence, future studies of LATE should be performed within specific cell types, rather than less informative studies on samples of mixed leukocytes. However, we also identified many LATE genes that were dysregulated in more than one cell type. For example, among the 3110 genes with LATE identified in the RNAseq data originating from the *in vivo* collection, 338 LATE genes were detected in two cells types and 31 were shared between all three types of studied leukocytes. Similarly, for the 337 LATE genes identified by scRNAseq, 25 genes showed LATE in more than one cell type and two of these LATE genes were shared between all four cell types studied (**Table S2**).

To investigate the biological functions of the identified LATE genes we first performed Gene Ontology (GO) analyses of the unique 3360 LATE genes discovered *in vivo* using a stringency level of *p*<0.05 (**Table 1**). We found these genes to be significantly enriched in 7 GO categories directly linked to immune system functions (**Table S4A**). For example, of the 3360 analyzed LATE genes we found 617 genes supporting the GO category *immune system process* and 423 supporting *immune response*. We also performed a corresponding GO analysis including only the 521 LATE genes showing differential expression using a more stringent threshold (i.e. FDR<0.1, **Table 1**) and this analysis validated the link between LATE genes and central immune system functions (**Table S4B**). Furthermore, for a subset of these LATE genes, we reviewed the literature for known functions and found many of them to have connections with the immune system and/or development of AD and/or cancer. A brief summary of relevant LATE genes is provided in **Table S5**, and among these genes are: *LAG3, LY6E, ANXA1, CSF1R, APBA3, CSF2RA, JAK3, IL15, CD3D, CD3G, CADM1, PYCARD, LILRB2, LILRA1, KLRG1, HLA-DRA* and *CCL5.* One of the LATE genes that showed the strongest downregulation was the autosomal *LAG3* (Lymphocyte-activation gene 3*)* in NK cells (**Fig. 6**). The LAG3 protein is a cell surface molecule that functions as an immune checkpoint receptor by binding its main ligand MHC class II with higher affinity than CD4. Cellular proliferation and immune cell activation is regulated by a balance between CD4/LAG3, in a similar fashion to the CTLA4 and PD1 immune checkpoints, where *LAG3* expression suppresses cell activity [47, 48]. The observed low expression of *LAG3* in LOY cells might disrupt the CD4/LAG3 balance, which is noteworthy in the context of immune-surveillance and the increased risk for cancer observed in men with LOY.

In the NK cells and monocytes, studied by both RNAseq and scRNAseq, we identified 206 and 60 genes, respectively, showing high expression and profound LATE in RNAseq (**Fig. 6**). Furthermore, 18 genes in NK cells and 9 genes in monocytes were supported by both technologies. Among these, a handful were detected as LATE genes in both cell types; such as the PAR gene *CD99*, the MSY genes *EIF1AY* and *RPS4Y1*, as well as the autosomal gene *LY6E*. The *LY6E* gene (lymphocyte antigen 6 family member E) is clearly upregulated in expression studies (**Fig. 6**, **Table S2**) and it has the potential to inhibit inflammatory cytokines and disrupt inflammatory cascades. A survey of more than 130 published clinical studies found that increased expression of *LY6E* is associated with poor survival outcome in multiple malignancies [49] and it has also been found to be important for drug resistance and tumor immune escape in breast cancer [50].

### LOY in blood is associated with changes in the plasma proteome

We hypothesized that LOY might leave a footprint in the plasma proteome in subjects with high frequency of leukocytes with this aneuploidy. The ultra-sensitive Proximity Extension Assay (PEA) [51] has previously [52] been used in the NSPHS-cohort [53] to characterize protein abundance levels in plasma. In brief, abundance levels of 424 plasma proteins have been quantified using five commercially available PEA-panels (Olink Proteomics AB; INF, NEU, ONC2, CVD2 and CVD3), measuring abundance of 92 proteins each. We estimated the LOY status in 480 males in NSPHS from 30X whole genome sequencing data, showing that 13 (2.7%), 36 (7.5%) and 159 (33%) of male NSPHS-subjects had LOY in more than 20%, 10%, and 1.5% of leukocytes, respectively (**Fig. 7A**). By comparing the levels of plasma proteins in men with LOY in more than 20% of leukocytes with men without aneuploidy, using a statistical model adjusted for confounders such as age, smoking, various life style factors and medical history [54], we found many plasma proteins with changed abundance between the groups (**Fig. 7B**). After strict Bonferroni correction, only one protein remained significant, i.e. chemokine CXCL5, which was three times more abundant in plasma from men with LOY. CXCL5 is a proangiogenic chemokine and a strong attractant for granulocytes and is secreted by various immune- and specific non-immune cells. In the context of cancer, CXCL5 has also been associated with late-stage disease and promotion of metastases [55-57]. Since the *CXCL5* gene is not among the LATE genes, caution is advised regarding cause and effect of high CXCL5 plasma levels in subjects with LOY. This might reflect an ongoing cancer-related process in these individuals. Among the other top-hit signals, there was an overrepresentation of proteins involved in regulation of the immune system and the inflammation panel (INF) generated highest number of candidates.

Overall, we did not observe a large overlap between LATE genes detected at the mRNA level within leukocytes and corresponding changes in abundance of their respective plasma proteins (details not shown). However, one of the proteins showing a tendency for higher abundance in men with LOY was TCL1A (**Fig. 7B**). Variants of this gene has been suggested in GWAS (as one of numerous genes) to be associated with increased risk for LOY [21, 25, 26]. The present and previous study (**Table S2**)[21] show that *TCL1A* is predominantly expressed in CD19+ B lymphocytes and overexpressed in cells from this subset affected by LOY (**Figs. S14D** and **S16D**), while the other types of leukocytes did not show its expression. However, it is remarkable that increased abundance of the TCL1A protein was detected in plasma from men with LOY. This is because the intracellular location of TCL1A is the cell nucleus [58] and that CD19+ B lymphocytes represent only a few percent of leukocytes in circulation. Results from scRNAseq and protein data, supports that LOY-linked upregulation of *TCL1A* oncogene in B-lymphocytes plays a role in driving clonal expansion in this subset. *TCL1A* (T-Cell Leukemia/Lymphoma 1A) is a gene normally expressed in early stage lymphocytes involved in regulation of multiple signaling pathways. Dysregulation of *TCL1A* has been associated with lymphomagenesis and cancer progression, e.g. in T cell prolymphocytic leukemia, via translocation that brings *TCL1A* under control of *TCR* [58]. The observation of *TCL1A* upregulation in B lymphocytes with LOY may suggest a novel LOY-dependent mode of activation for this oncogene, which deserves further studies. Our protein-related results provide a proof of concept that LOY can influence plasma protein levels and suggests that LOY may have an impact on systemic homeostasis. In conclusion, regardless of cause-and-effect relationships, CXCL5 and additional near significance signals showed in **Figure 7** should be studied further using a larger collection of men with high level of LOY in leukocytes.

### LOY has profound effects on transcriptome in pleiotropic fashion

LOY has already been shown as the most common post-zygotic mutation considering data from blood DNA alone [23, 28]. Furthermore, LOY also occurs in other tissues [17, 23], but these two reports only scratched the surface of the topic and studies of other tissues/cell types should follow in order to establish the frequency of LOY in adult and aging men across the soma. Moreover, as we show here, LOY occurs frequently in blood as oligo-clonal expansions and this could be best explained by independent mutations occurring in progenitors for different lineages of hematopoietic cells, resulting in more than one expanding LOY-clone circulating in blood. Considering the above reasoning, LOY should probably be viewed as the most common human mutation of all categories.

The combined number of LATE genes derived from analyses of various cells is surprisingly large, which suggests that a loss of ~2% of the human genome (via LOY) has a profound impact on cellular homeostasis. Furthermore, we show a limited overlap in the number of LATE genes between different types of hematopoietic cells. We therefore hypothesize that LOY has pleiotropic effects, depending on which cell type (i.e. a distinct normal transcriptomic program) that is disturbed by this large mutation. As already mentioned, epidemiological studies have associated LOY in blood to a number of different diseases and understanding of the underlying mechanisms of LOY in leukocytes on the development of pathologic processes in other tissues is the major future challenge. In this perspective, the LOY-pleiotropy might be a useful concept.

In total, we show 4026 LATE genes with varying degree of differential expression. The analyses of leukocytes via sorted- and single-cells *in vivo* identified 3360 LATE genes. We applied several steps of increasing the stringency for selection of genes with strong statistical support; i.e. robust differences in expression between cells with and without LOY. **Table S5** shows a short functional summary for 95 selected LATE genes that can be of interest with respect to continued research into immune-, cancer- and AD-related mechanisms. Many of these genes are involved in normal functions of the immune system (**Table S4**). In our initial papers suggesting the role of LOY in the development of cancer and AD [8, 10], we proposed that this aneuploidy might impair functions of the immune system, especially immune-surveillance that is responsible for elimination of abnormal cells and structures throughout the soma. In conclusion, the current work provides support for this hypothesis and indicates a role of LOY in distinct populations of leukocytes for the pathogenesis of cancer and Alzheimer’s disease.

## Acknowledgements

We acknowledge all patients and healthy donors of blood samples from Poland and Sweden. We thank Käthe Ström, Brygida Mazgaj, Bożena Lackowska and Niklas Dahl for help with collection of blood samples from patients and controls as well as Katarina Tegnér and Amanda Raine for assistance with scRNAseq library preparations. We acknowledge Darek Kędra for critical review of the manuscript. This work was supported by grants from the Hjärnfonden, Swedish Cancer Society, the Swedish Research Council, Alzheimerfonden, Konung Gustav V:s och Drottning Viktorias Frimurarestiftelse, the Science for Life Laboratory - Uppsala and the Foundation for Polish Science under the International Research Agendas Programme to J.P.D. as well as from the European Research Council ERC Starting Grant, the Swedish Research Council, and Kjell och Märta Beijers Stiftelse to L.A.F. Sequencing was performed by the SNP&SEQ and UGC Technology Platforms at Uppsala University. These facilities are part of the National Genomics Infrastructure (NGI) Sweden and Science for Life Laboratory. The SNP&SEQ Platform is also supported by the Swedish Research Council and the Knut and Alice Wallenberg Foundation.

## Author Contributions

Conceptualization, J.P.D, J.H., S.E. and L.A.F.; Methodology, J.P.D., J.H., E.R.-B, J.M., H.D., N.G., K.W., K.B.-S., M.D., J.N., J.J., A.K.-R., S.E., J.R., and L.A.F.; Software, J.H., J.M., B.T.-M., S.E. and L.A.F.; Investigation, all coauthors; Writing – Original Draft, J.P.D, J.H., E.R-B., H.D. and L.A.F.; Writing – Review & Editing, J.P.D., J.H., E.R-B., J.M., H.D., B.T.-M., K.B.-S., M.D., P.O., A.P., T.G., S.I., J.N., A.A., Å.J., A.J., S.E., M.I., J.R. and L.A.F.; Resources, all coauthors; Supervision, J.P.D., J.H., E.R-B., H.D., A.P., J.N., J.B., J.R., and L.A.F.; Funding Acquisition, J.P.D, U.G., Å.J., A.P., M.S., A.J., J.B., J.R. and L.A.F.; Project administration, J.P.D, and L.A.F.

## Declaration of Interests

J.P.D. and L.A.F. are cofounders and shareholders in Cray Innovation AB.

## References

1. Cortez D, Marin R, Toledo-Flores D, Froidevaux L, Liechti A, Waters PD, Grutzner F, Kaessmann H: Origins and functional evolution of Y chromosomes across mammals, Nature 2014, 508:488-493; (https://www.ncbi.nlm.nih.gov/pubmed/24759410)
2. Bellott DW, Hughes JF, Skaletsky H, Brown LG, Pyntikova T, Cho TJ, Koutseva N, Zaghlul S, Graves T, Rock S, Kremitzki C, Fulton RS, Dugan S, Ding Y, Morton D, Khan Z, Lewis L, Buhay C, Wang Q, Watt J, Holder M, Lee S, Nazareth L, Alfoldi J, Rozen S, Muzny DM,

Warren WC, Gibbs RA, Wilson RK, Page DC: Mammalian Y chromosomes retain widely expressed dosage-sensitive regulators, Nature 2014, 508:494-499; (https://www.ncbi.nlm.nih.gov/pubmed/24759411)

3. Jacobs PA, Brunton M, Court Brown WM, Doll R, Goldstein H: Change of human chromosome count distribution with age: evidence for a sex differences, Nature 1963, 197:1080-1081; (http://www.ncbi.nlm.nih.gov/pubmed/13964326)
4. Pierre RV, Hoagland HC: Age-associated aneuploidy: loss of Y chromosome from human bone marrow cells with aging, Cancer 1972, 30:889-894; (http://www.ncbi.nlm.nih.gov/pubmed/4116908)
5. Nowinski GP, Van Dyke DL, Tilley BC, Jacobsen G, Babu VR, Worsham MJ, Wilson GN, Weiss L: The frequency of aneuploidy in cultured lymphocytes is correlated with age and gender but not with reproductive history, Am J Hum Genet 1990, 46:1101-1111; (http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=2339703)
6. UKCCG: Loss of the Y chromosome from normal and neoplastic bone marrows. United Kingdom Cancer Cytogenetics Group (UKCCG), Genes, chromosomes cancer 1992, 5:83-88; (http://www.ncbi.nlm.nih.gov/pubmed/1384666)
7. Wiktor A, Rybicki BA, Piao ZS, Shurafa M, Barthel B, Maeda K, Van Dyke DL: Clinical significance of Y chromosome loss in hematologic disease, Genes, chromosomes & cancer 2000, 27:11-16; (http://www.ncbi.nlm.nih.gov/pubmed/10564581)
8. Forsberg LA, Rasi C, Malmqvist N, Davies H, Pasupulati S, Pakalapati G, Sandgren J, de Stahl TD, Zaghlool A, Giedraitis V, Lannfelt L, Score J, Cross NC, Absher D, Janson ET, Lindgren CM, Morris AP, Ingelsson E, Lind L, Dumanski JP: Mosaic loss of chromosome Y in peripheral blood is associated with shorter survival and higher risk of cancer, Nature Genetics 2014, 46:624-628; (http://www.ncbi.nlm.nih.gov/pubmed/24777449)
9. Loftfield E, Zhou W, Graubard BI, Yeager M, Chanock SJ, Freedman ND, Machiela MJ: Predictors of mosaic chromosome Y loss and associations with mortality in the UK Biobank, Sci Rep 2018, 8:12316; (https://www.ncbi.nlm.nih.gov/pubmed/30120341)
10. Dumanski JP, Lambert JC, Rasi C, Giedraitis V, Davies H, Grenier-Boley B, Lindgren CM, Campion D, Dufouil C, European Alzheimer's Disease Initiative I, Pasquier F, Amouyel P, Lannfelt L, Ingelsson M, Kilander L, Lind L, Forsberg LA: Mosaic Loss of Chromosome Y in Blood Is Associated with Alzheimer Disease, Am J Hum Genet 2016, 98:1208-1219; (http://www.ncbi.nlm.nih.gov/pubmed/27231129)
11. Ganster C, Kampfe D, Jung K, Braulke F, Shirneshan K, Machherndl-Spandl S, Suessner S, Bramlage CP, Legler TJ, Koziolek MJ, Haase D, Schanz J: New data shed light on Y-loss-related pathogenesis in myelodysplastic syndromes, Genes Chromosomes Cancer 2015, 54:717-724; (http://www.ncbi.nlm.nih.gov/pubmed/26394808)
12. Noveski P, Madjunkova S, Sukarova Stefanovska E, Matevska Geshkovska N, Kuzmanovska M, Dimovski A, Plaseska-Karanfilska D: Loss of Y Chromosome in Peripheral Blood of Colorectal and Prostate Cancer Patients, PLoS One 2016, 11:e0146264; (http://www.ncbi.nlm.nih.gov/pubmed/26745889)
13. Loftfield E, Zhou W, Yeager M, Chanock SJ, Freedman ND, Machiela MJ: Mosaic Y Loss Is Moderately Associated with Solid Tumor Risk, Cancer Res 2019, 79:461-466; (https://www.ncbi.nlm.nih.gov/pubmed/30510122)
14. Persani L, Bonomi M, Lleo A, Pasini S, Civardi F, Bianchi I, Campi I, Finelli P, Miozzo M, Castronovo C, Sirchia S, Gershwin ME, Invernizzi P: Increased loss of the Y chromosome in peripheral blood cells in male patients with autoimmune thyroiditis, Journal of autoimmunity 2012, 38:J193-196; (http://www.ncbi.nlm.nih.gov/pubmed/22196921)
15. Lleo A, Oertelt-Prigione S, Bianchi I, Caliari L, Finelli P, Miozzo M, Lazzari R, Floreani A, Donato F, Colombo M, Gershwin ME, Podda M, Invernizzi P: Y chromosome loss in male

patients with primary biliary cirrhosis, Journal of autoimmunity 2013, 41:87-91; (http://www.ncbi.nlm.nih.gov/pubmed/23375847)

16. Grassmann F, Kiel C, den Hollander A, Weeks D, Lotery A, Cipriani V, Weber B, International Age-related Macular Degeneration Genomics Consortium (IAMDGC): Y chromosome mosaicism is associated with age-related macular degeneration, European Journal of Human Genetics 2018, August 2018:(https://www.ncbi.nlm.nih.gov/pubmed/30158665)
17. Haitjema S, Kofink D, van Setten J, van der Laan S, Schoneveld A, Eales J, Tomaszewski M, de Jager S, Pasterkamp G, Asselbergs F, den Ruijter H: Loss of Y Chromosome in Blood Is Associated with Major Cardiovascular Events during Follow-up in Men after Carotid Endarterectomy, Circulation: Cardiovascular Genetics 2017, 10:e001544:(https://www.ncbi.nlm.nih.gov/pubmed/28768751)
18. Blatt Kalben B: Why Men Die Younger, North American Actuarial Journal 2000, 4:83-111; (http://dx.doi.org/10.1080/10920277.2000.10595939)
19. Austad SN: Why women live longer than men: sex differences in longevity, Gend Med 2006, 3:79-92; (https://www.ncbi.nlm.nih.gov/pubmed/16860268)
20. Cook MB, McGlynn KA, Devesa SS, Freedman ND, Anderson WF: Sex disparities in cancer mortality and survival, Cancer epidemiology, biomarkers & prevention 2011, 20:1629-1637; (http://www.ncbi.nlm.nih.gov/pubmed/21750167)
21. Thompson D, Genovese G, Halvardson J, Ulirsch J, Wright D, Terao C, Davidsson O, Day F, Sulem P, Jiang Y, Danielsson M, Davies H, Dennis J, Dunlop M, Easton D, Fisher V, Zink F, Houlston R, Ingelsson M, Kar S, Kerrison N, Kristjansson R, Li R, Loveday C, Mattisson J, McCarroll S, Murakami Y, Murray A, Olszewski P, Rychlicka-Buniowska E, Scott R, Thorsteinsdottir U, Tomlinson I, Torabi Moghadam B, Turnbull C, Wareham N, Gudbjartsson D, INTEGRAL-ILCCO, The Breast Cancer Association Consortium, CIMBA, The Endometrial Cancer Association Consortium, The Ovarian Cancer Association Consortium, The PRACTICAL Consortium, The Kidney Cancer GWAS Meta-Analysis Project, eQTLGen Consortium, BIOS Consortium, 23andMe Research Team, Kamatani Y, Finucane H, Hoffmann E, Jackson S, Stefansson K, Auton A, Ong K, Machiela M, Loh P-R, Dumanski J, Chanock S, Forsberg L, Perry J: Genetic predisposition to mosaic Y chromosome loss in blood is associated with genomic instability in other tissues and susceptibility to non-haematological cancers, submitted 2019, (http://dx.doi.org/10.1101/514026)
22. Zink F, Stacey SN, Norddahl GL, Frigge ML, Magnusson OT, Jonsdottir I, Thorgeirsson TE, Sigurdsson A, Gudjonsson SA, Gudmundsson J, Jonasson JG, Tryggvadottir L, Jonsson T, Helgason A, Gylfason A, Sulem P, Rafnar T, Thorsteinsdottir U, Gudbjartsson DF, Masson G, Kong A, Stefansson K: Clonal hematopoiesis, with and without candidate driver mutations, is common in the elderly, Blood 2017, 130:742-752; (https://www.ncbi.nlm.nih.gov/pubmed/28483762)
23. Forsberg L, Halvardson J, Rychlicka E, Danielsson M, Torabi Moghadam B, Mattisson J, Rasi C, Davies H, Lind L, Giedraitis V, Lannfelt L, Kilander L, Ingelsson M, Dumanski J: Mosaic loss of chromosome Y (LOY) in leukocytes matters, Nature Genetics 2018, (https://www.ncbi.nlm.nih.gov/pubmed/30374072)
24. Dumanski JP, Rasi C, Lonn M, Davies H, Ingelsson M, Giedraitis V, Lannfelt L, Magnusson PK, Lindgren CM, Morris AP, Cesarini D, Johannesson M, Tiensuu Janson E, Lind L, Pedersen NL, Ingelsson E, Forsberg LA: Smoking is associated with mosaic loss of chromosome Y, Science 2015, 347:81-83; (http://www.ncbi.nlm.nih.gov/pubmed/25477213)
25. Zhou W, Machiela MJ, Freedman ND, Rothman N, Malats N, Dagnall C, Caporaso N, Teras LT, Gaudet MM, Gapstur SM, Stevens VL, Jacobs KB, Sampson J, Albanes D, Weinstein S, Virtamo J, Berndt S, Hoover RN, Black A, Silverman D, Figueroa J, Garcia-Closas M, Real FX, Earl J, Marenne G, Rodriguez-Santiago B, Karagas M, Johnson A, Schwenn M, Wu X, Gu J, Ye Y, Hutchinson A, Tucker M, Perez-Jurado LA, Dean M, Yeager M, Chanock SJ: Mosaic

loss of chromosome Y is associated with common variation near TCL1A, Nat Genet 2016, 48:563-568; (http://www.ncbi.nlm.nih.gov/pubmed/27064253)

26. Wright DJ, Day FR, Kerrison ND, Zink F, Cardona A, Sulem P, Thompson DJ, Sigurjonsdottir S, Gudbjartsson DF, Helgason A, Chapman JR, Jackson SP, Langenberg C, Wareham NJ, Scott RA, Thorsteindottir U, Ong KK, Stefansson K, Perry JRB: Genetic variants associated with mosaic Y chromosome loss highlight cell cycle genes and overlap with cancer susceptibility, Nat Genet 2017, 49:674-679; (https://www.ncbi.nlm.nih.gov/pubmed/28346444)
27. Forsberg LA: Loss of chromosome Y (LOY) in blood cells is associated with increased risk for disease and mortality in aging men, Hum Genet 2017, 136:657-663; (https://www.ncbi.nlm.nih.gov/pubmed/28424864)
28. Forsberg LA, Gisselsson D, Dumanski JP: Mosaicism in health and disease - clones picking up speed, Nat Rev Genet 2017, 18:128-142; (https://www.ncbi.nlm.nih.gov/pubmed/27941868)
29. Speciale L, Calabrese E, Saresella M, Tinelli C, Mariani C, Sanvito L, Longhi R, Ferrante P: Lymphocyte subset patterns and cytokine production in Alzheimer's disease patients, Neurobiol Aging 2007, 28:1163-1169; (https://www.ncbi.nlm.nih.gov/pubmed/16814429)
30. Larbi A, Pawelec G, Witkowski JM, Schipper HM, Derhovanessian E, Goldeck D, Fulop T: Dramatic shifts in circulating CD4 but not CD8 T cell subsets in mild Alzheimer's disease, J Alzheimers Dis 2009, 17:91-103; (https://www.ncbi.nlm.nih.gov/pubmed/19494434)
31. Masera RG, Prolo P, Sartori ML, Staurenghi A, Griot G, Ravizza L, Dovio A, Chiappelli F, Angeli A: Mental deterioration correlates with response of natural killer (NK) cell activity to physiological modifiers in patients with short history of Alzheimer's disease, Psychoneuroendocrinology 2002, 27:447-461; (https://www.ncbi.nlm.nih.gov/pubmed/11911998)
32. Solerte SB, Cravello L, Ferrari E, Fioravanti M: Overproduction of IFN-gamma and TNF-alpha from natural killer (NK) cells is associated with abnormal NK reactivity and cognitive derangement in Alzheimer's disease, Ann N Y Acad Sci 2000, 917:331-340; (https://www.ncbi.nlm.nih.gov/pubmed/11268360)
33. Schindowski K, Peters J, Gorriz C, Schramm U, Weinandi T, Leutner S, Maurer K, Frolich L, Muller WE, Eckert A: Apoptosis of CD4+ T and natural killer cells in Alzheimer's disease, Pharmacopsychiatry 2006, 39:220-228; (https://www.ncbi.nlm.nih.gov/pubmed/17124644)
34. Jadidi-Niaragh F, Shegarfi H, Naddafi F, Mirshafiey A: The role of natural killer cells in Alzheimer's disease, Scand J Immunol 2012, 76:451-456; (https://www.ncbi.nlm.nih.gov/pubmed/22889057)
35. Araga S, Kagimoto H, Funamoto K, Takahashi K: Reduced natural killer cell activity in patients with dementia of the Alzheimer type, Acta Neurol Scand 1991, 84:259-263; (https://www.ncbi.nlm.nih.gov/pubmed/1950471)
36. Le Page A, Bourgade K, Lamoureux J, Frost E, Pawelec G, Larbi A, Witkowski JM, Dupuis G, Fulop T: NK Cells are Activated in Amnestic Mild Cognitive Impairment but not in Mild Alzheimer's Disease Patients, J Alzheimers Dis 2015, 46:93-107; (https://www.ncbi.nlm.nih.gov/pubmed/25720398)
37. Le Page A, Dupuis G, Frost EH, Larbi A, Pawelec G, Witkowski JM, Fulop T: Role of the peripheral innate immune system in the development of Alzheimer's disease, Exp Gerontol 2018, 107:59-66; (https://www.ncbi.nlm.nih.gov/pubmed/29275160)
38. Mrdjen D, Pavlovic A, Hartmann FJ, Schreiner B, Utz SG, Leung BP, Lelios I, Heppner FL, Kipnis J, Merkler D, Greter M, Becher B: High-Dimensional Single-Cell Mapping of Central Nervous System Immune Cells Reveals Distinct Myeloid Subsets in Health, Aging, and Disease, Immunity 2018, 48:380-395 e386; (https://www.ncbi.nlm.nih.gov/pubmed/29426702)
39. Korin B, Ben-Shaanan TL, Schiller M, Dubovik T, Azulay-Debby H, Boshnak NT, Koren T, Rolls A: High-dimensional, single-cell characterization of the brain's immune compartment, Nat Neurosci 2017, 20:1300-1309; (https://www.ncbi.nlm.nih.gov/pubmed/28758994)

40. Sullivan KM, Mannucci A, Kimpton CP, Gill P: A rapid and quantitative DNA sex test: fluorescence-based PCR analysis of X-Y homologous gene amelogenin, Biotechniques 1993, 15:636-638, 640-631; (https://www.ncbi.nlm.nih.gov/pubmed/8251166)
41. Danielsson M, Halvardson J, Davies H, Torabi Moghadam B, Mattisson J, Rychlicka-Buniowska E, Heintz J, Lannfelt L, Giedraitis V, Ingelsson M, Dumanski J, Forsberg L: Intra-individual changes in the frequency of mosaic loss of chromosome Y over time estimated with a new method, submitted 2019,
42. Love MI, Huber W, Anders S: Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2, Genome Biol 2014, 15:550; (https://www.ncbi.nlm.nih.gov/pubmed/25516281)
43. Schenkel AR, Mamdouh Z, Chen X, Liebman RM, Muller WA: CD99 plays a major role in the migration of monocytes through endothelial junctions, Nat Immunol 2002, 3:143-150; (https://www.ncbi.nlm.nih.gov/pubmed/11812991)
44. Vestweber D: How leukocytes cross the vascular endothelium, Nat Rev Immunol 2015, 15:692-704; (https://www.ncbi.nlm.nih.gov/pubmed/26471775)
45. Butler A, Hoffman P, Smibert P, Papalexi E, Satija R: Integrating single-cell transcriptomic data across different conditions, technologies, and species, Nat Biotechnol 2018, 36:411-420; (https://www.ncbi.nlm.nih.gov/pubmed/29608179)
46. Paaby AB, Rockman MV: The many faces of pleiotropy, Trends Genet 2013, 29:66-73; (https://www.ncbi.nlm.nih.gov/pubmed/23140989)
47. Huang CT, Workman CJ, Flies D, Pan X, Marson AL, Zhou G, Hipkiss EL, Ravi S, Kowalski J, Levitsky HI, Powell JD, Pardoll DM, Drake CG, Vignali DA: Role of LAG-3 in regulatory T cells, Immunity 2004, 21:503-513; (https://www.ncbi.nlm.nih.gov/pubmed/15485628)
48. Workman CJ, Cauley LS, Kim IJ, Blackman MA, Woodland DL, Vignali DA: Lymphocyte activation gene-3 (CD223) regulates the size of the expanding T cell population following antigen activation in vivo, J Immunol 2004, 172:5450-5455; (https://www.ncbi.nlm.nih.gov/pubmed/15100286)
49. Luo L, McGarvey P, Madhavan S, Kumar R, Gusev Y, Upadhyay G: Distinct lymphocyte antigens 6 (Ly6) family members Ly6D, Ly6E, Ly6K and Ly6H drive tumorigenesis and clinical outcome, Oncotarget 2016, 7:11165-11193; (https://www.ncbi.nlm.nih.gov/pubmed/26862846)
50. AlHossiny M, Luo L, Frazier WR, Steiner N, Gusev Y, Kallakury B, Glasgow E, Creswell K, Madhavan S, Kumar R, Upadhyay G: Ly6E/K Signaling to TGFbeta Promotes Breast Cancer Progression, Immune Escape, and Drug Resistance, Cancer Res 2016, 76:3376-3386; (https://www.ncbi.nlm.nih.gov/pubmed/27197181)
51. Assarsson E, Lundberg M: Development and validation of customized PEA biomarker panels with clinical utility. IN: Advancing precision medicine: Current and future proteogenomic strategies for biomarker discovery and development. Edited by Washington, DC, Science/AAAS, 2017, p. pp. 32-36; (http://www.sciencemag.org/collections/advancing-precision-medicine-current-and-future-proteogenomic-strategies-biomarker)
52. Enroth S, Maturi V, Berggrund M, Enroth SB, Moustakas A, Johansson A, Gyllensten U: Systemic and specific effects of antihypertensive and lipid-lowering medication on plasma protein biomarkers for cardiovascular diseases, Sci Rep 2018, 8:5531; (https://www.ncbi.nlm.nih.gov/pubmed/29615742)
53. Igl W, Johansson A, Gyllensten U: The Northern Swedish Population Health Study (NSPHS)--a paradigmatic study in a rural population combining community health and basic research, Rural Remote Health 2010, 10:1363; (https://www.ncbi.nlm.nih.gov/pubmed/20568910)
54. Enroth S, Bosdotter Enroth S, Johansson A, Gyllensten U: Effect of genetic and environmental factors on protein biomarkers for common non-communicable disease and use of personally

normalized plasma protein profiles (PNPPP), Biomarkers 2015, 20:355-364; (https://www.ncbi.nlm.nih.gov/pubmed/26551787)

55. Begley LA, Kasina S, Mehra R, Adsule S, Admon AJ, Lonigro RJ, Chinnaiyan AM, Macoska JA: CXCL5 promotes prostate cancer progression, Neoplasia 2008, 10:244-254; (https://www.ncbi.nlm.nih.gov/pubmed/18320069)
56. Soler-Cardona A, Forsthuber A, Lipp K, Ebersberger S, Heinz M, Schossleitner K, Buchberger E, Groger M, Petzelbauer P, Hoeller C, Wagner E, Loewe R: CXCL5 Facilitates Melanoma Cell-Neutrophil Interaction and Lymph Node Metastasis, J Invest Dermatol 2018, 138:1627-1635; (https://www.ncbi.nlm.nih.gov/pubmed/29474942)
57. Hu B, Fan H, Lv X, Chen S, Shao Z: Prognostic significance of CXCL5 expression in cancer patients: a meta-analysis, Cancer Cell Int 2018, 18:68; (https://www.ncbi.nlm.nih.gov/pubmed/29743818)
58. Paduano F, Gaudio E, Mensah AA, Pinton S, Bertoni F, Trapasso F: T-Cell Leukemia/Lymphoma 1 (TCL1): An Oncogene Regulating Multiple Signaling Pathways, Front Oncol 2018, 8:317; (https://www.ncbi.nlm.nih.gov/pubmed/30151355)

# MAIN TEXT FIGURES AND TABLES

## Figure 1

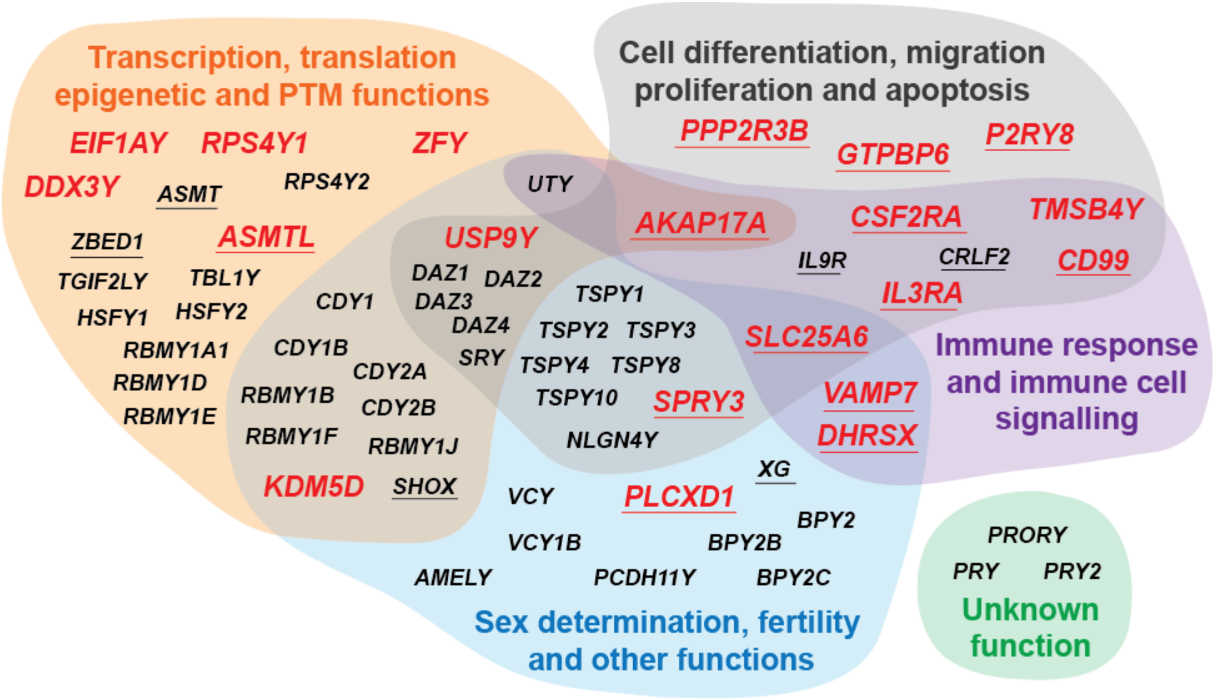
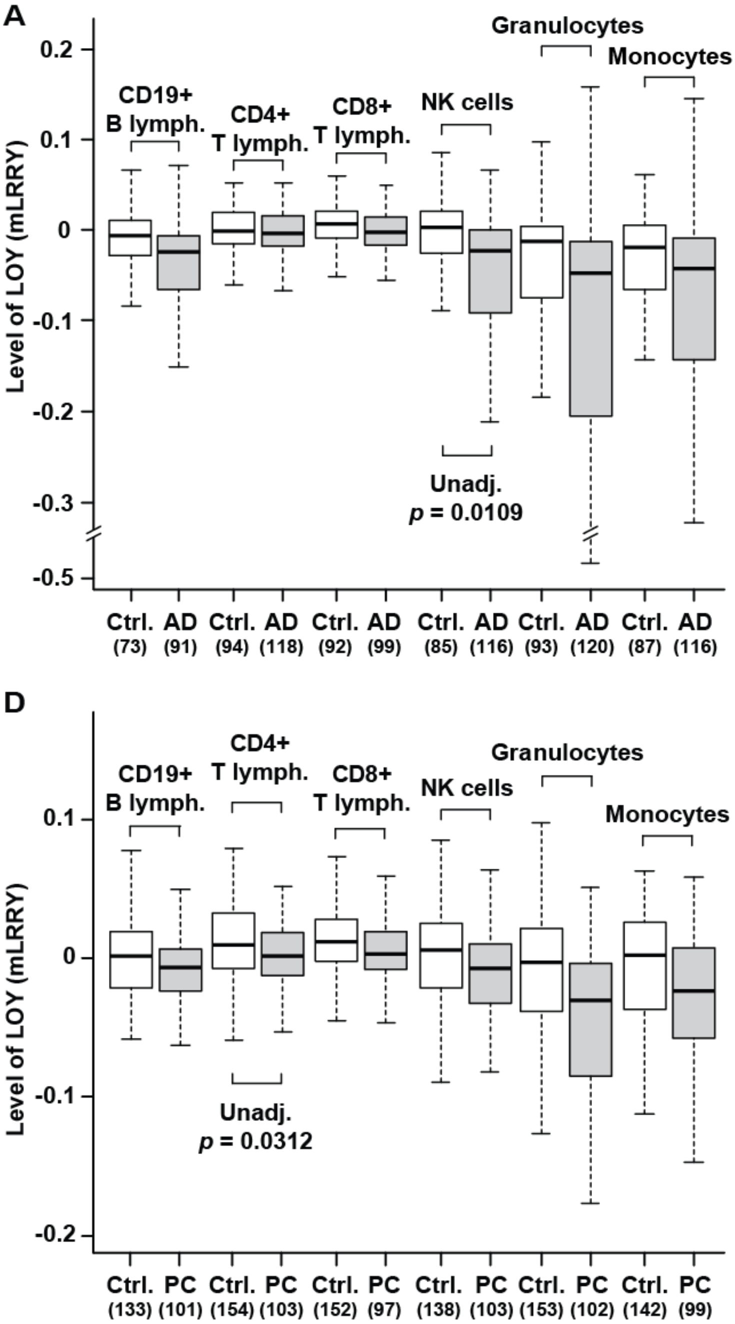


**Fig. 1.** Summary of normal functions of the 64 protein coding genes on chromosome Y, 19 located in the pseudo-autosomal regions (PAR, underlined) and 45 in the male specific region of chromosome Y (MSY). Red color indicates 20 genes expressed in leukocytes studied by us using RNAseq, 13 in PAR and 7 in MSY. Gene ontology (GO) analyses were performed for each chromosome Y gene to identify known biological functions and the identified GO terms were annotated into four functional categories. The categories were: "*Transcription, translation epigenetic and post-translational modifications (PTM) functions*" (orange area), "*Cell differentiation, migration proliferation and apoptosis*" (grey area), "*Sex determination, fertility and other functions*" (blue area) and "*Immune response and immune cell signaling*" (violet area). Functions for three genes is unknown (green area).

Figure 2



B

AD patients had more LOY only in NK cells vs. controls in LR models adjusted for age and smoking

| Cell type | Adj. *p* |
|---|---|
| CD19+ B lymphocytes | 0.6178 |
| CD4+ T lymphocytes | 0.7811 |
| CD8+ T lymphocytes | 0.4360 |
| NK cells | 0.0168 * |
| Granulocytes | 0.7307 |
| Monocytes | 0.8618 |

C

| Oligoclonality of LOY in men diagnosed with AD vs. controls | | AD | |
|---|---|---|---|
| | | Y | N |
| Oligoclonal (LOY in ≥2 cell types) | Y | 29 | 10 |
| | N | 92 | 84 |
| % oligoclonal subjects | | 24.0 | 10.6 |
| Fisher's exact test: OR = 2.64<br>95% C.I. = 1.16-6.45, *p* = 0.0127 | | | |

E

PC patients had more LOY only in T lymph. vs. controls in LR models adjusted for age and smoking

| Cell type | Adj. *p* |
|---|---|
| CD19+ B lymphocytes | 0.1927 |
| CD4+ T lymphocytes | 0.0331 * |
| CD8+ T lymphocytes | 0.0925 |
| NK cells | 0.1930 |
| Granulocytes | 0.2038 |
| Monocytes | 0.2832 |

F

| Oligoclonality of LOY in men diagnosed with PC vs. controls | | PC | |
|---|---|---|---|
| | | Y | N |
| Oligoclonal (LOY in ≥2 cell types) | Y | 10 | 8 |
| | N | 97 | 148 |
| % oligoclonal subjects | | 9.3 | 5.1 |
| Fisher's exact test: OR = 1.90<br>95% C.I. = 0.65-5.76, *p* = 0.2172 | | | |

**Fig. 2.** Comparison of the level of LOY in six populations of leukocytes in men diagnosed with Alzheimer's disease (AD) or prostate cancer (PC) vs controls. Leukocytes were sorted by FACS, followed by SNP-array genotyping of DNA from each cell fraction and calculation of the median Log R Ratio of the probes in the male specific part of chromosome Y (mLRRY). The numbers in parentheses under the X-axes in panels A and D denote the number of subjects studied for each cell type. Panels A and D show results from unadjusted analyses and panels B and E describe the results from adjusted logistic regression models. Panels C and F show results from investigation of LOY oligoclonality in men with AD and PC diagnoses vs controls. Abbreviations: Ctrl. = control, AD = Alzheimer's disease, PC = prostate cancer, LR = logistic regression, OR = odds ratio, lymph.= lymphocytes and NK = natural killer.

## Figure 3

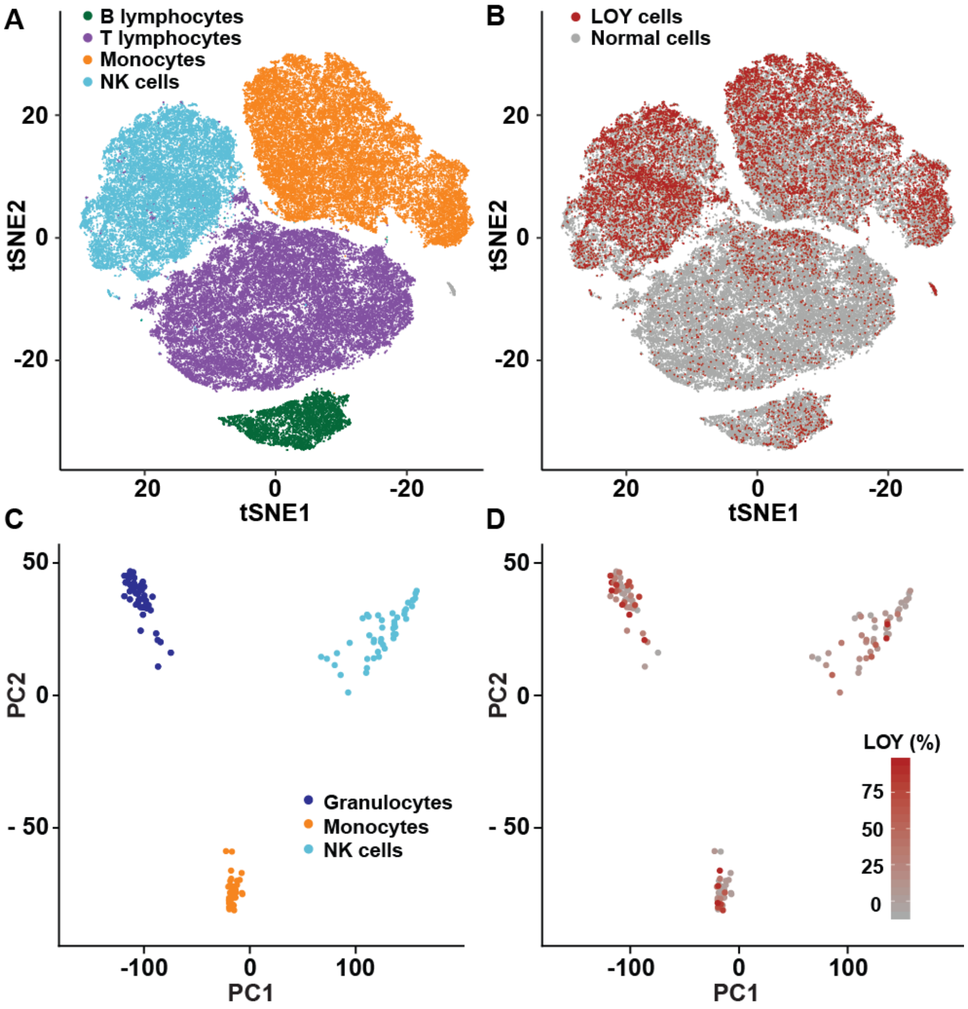


**Fig. 3.** Distribution and level of loss of chromosome Y (LOY) in leukocytes from aging men using two RNA-sequencing platforms. Panels A and B show results from single cell RNA sequencing of peripheral blood mononuclear cells (PBMCs) collected from 29 men, 26 diagnosed with Alzheimer's disease. The tSNE plot in panel A shows pooled data from 73,606 PBMCs, each dot representing a single cell, in four cell types that are distinguished by colors. Panel B shows the distribution of cells displaying LOY in different cell types by coloring LOY cells in red and normal cells in grey. Panels C and D show results from bulk RNA sequencing (RNAseq) performed in three cell types sorted by FACS from 51 individuals. Panel C displays a principal component analysis based on global gene expression where each dot represents one cell type in one individual. The distance between dots indicate similarity in gene expression. Panel D shows the level of LOY mosaicism in each sample by a gradient of red and grey where red indicates a high level of LOY.

## Figure 4

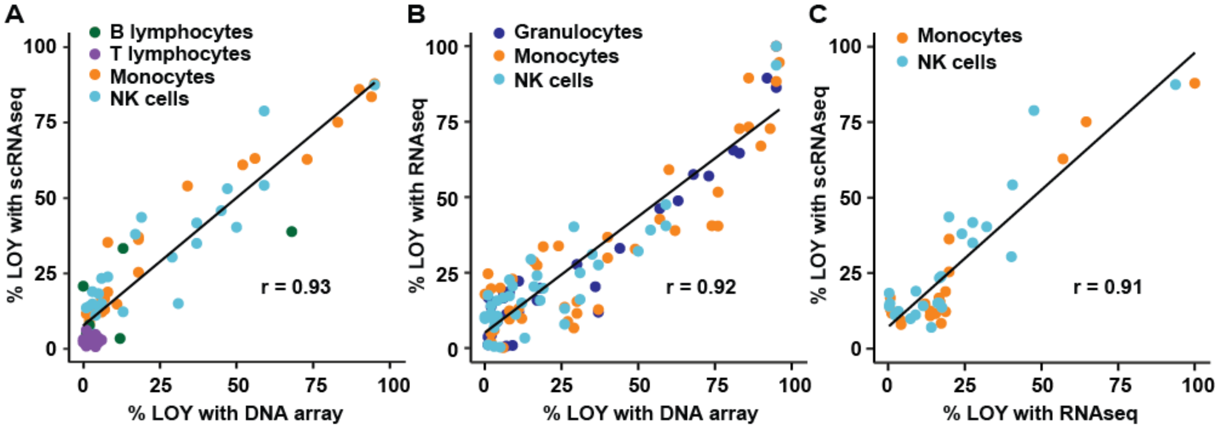


**Fig. 4.** Comparison of performance of two RNA-sequencing platforms for estimation of the level of LOY in leukocytes. From the same set of blood samples, LOY analysis was performed in pairwise comparisons between scRNAseq vs array-based SNP genotyping (panel A), RNAseq vs array genotyping (panel B) and scRNAseq vs RNAseq (panel C). For the single cell data, pseudo-bulk samples were created for these comparisons by pooling the single cell data from each cell type for each individual. Abbreviations: scRNAseq = single cell RNA sequencing, RNAseq = bulk RNA sequencing, r = Pearson's correlation coefficient.

## Figure 5

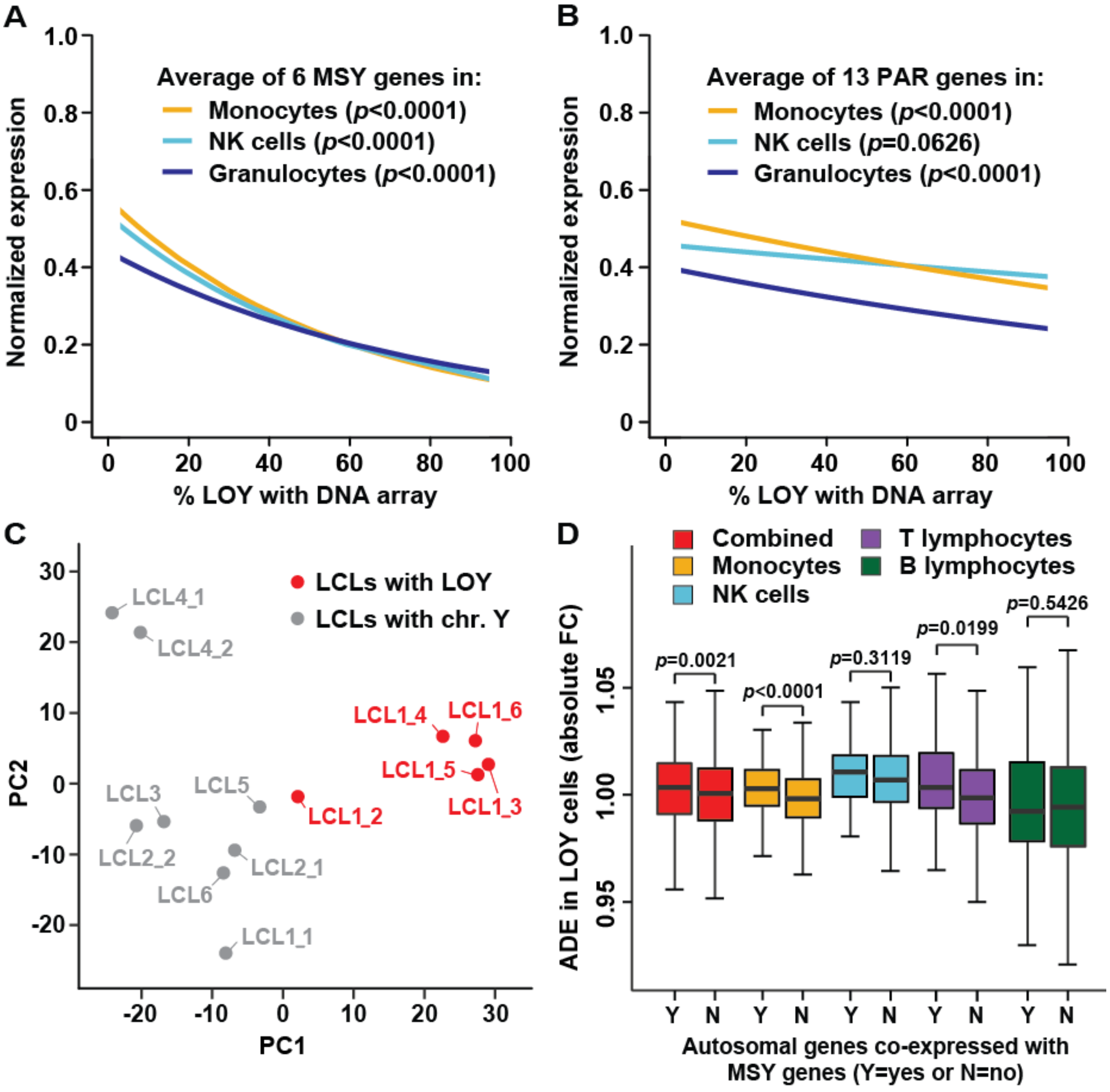


**Fig. 5.** LOY Associated Transcriptional Effect (LATE) observed *in vivo* and *in vitro*. Panels A and B show decreased level of expression of genes in the male-specific part of chromosome Y (MSY) and in the pseudo autosomal regions of chromosomes X and Y (PAR), respectively. Gene expression was estimated using RNAseq in three FACS sorted cell types *in vivo*. Panel C shows a principal component analysis plot of global gene expression in lymphoblastoid cell lines (LCLs). Red and grey dots represent LCLs with and without LOY, respectively, and donor identity are specified by numbers (i.e. LCL2_1 and LCL2_2 are from donor 2). Distance between dots indicates similarity in global gene expression and analysis of the variance in PC1 shows that LCLs with LOY have altered global autosomal expression compared to normal LCLs (Kolmogorov–Smirnov test: D=1.0, $p$=0.0016). Panel D shows results from co-expression analysis and autosomal differential expression (ADE) in single cells with LOY. Autosomal genes that are normally co-expressed with MSY genes displayed a higher level of differential expression in single cells with LOY compared with the control genes without normal co-expression with MSY genes (Wilcoxon rank sum test: $p$=0.0021).

Figure 6

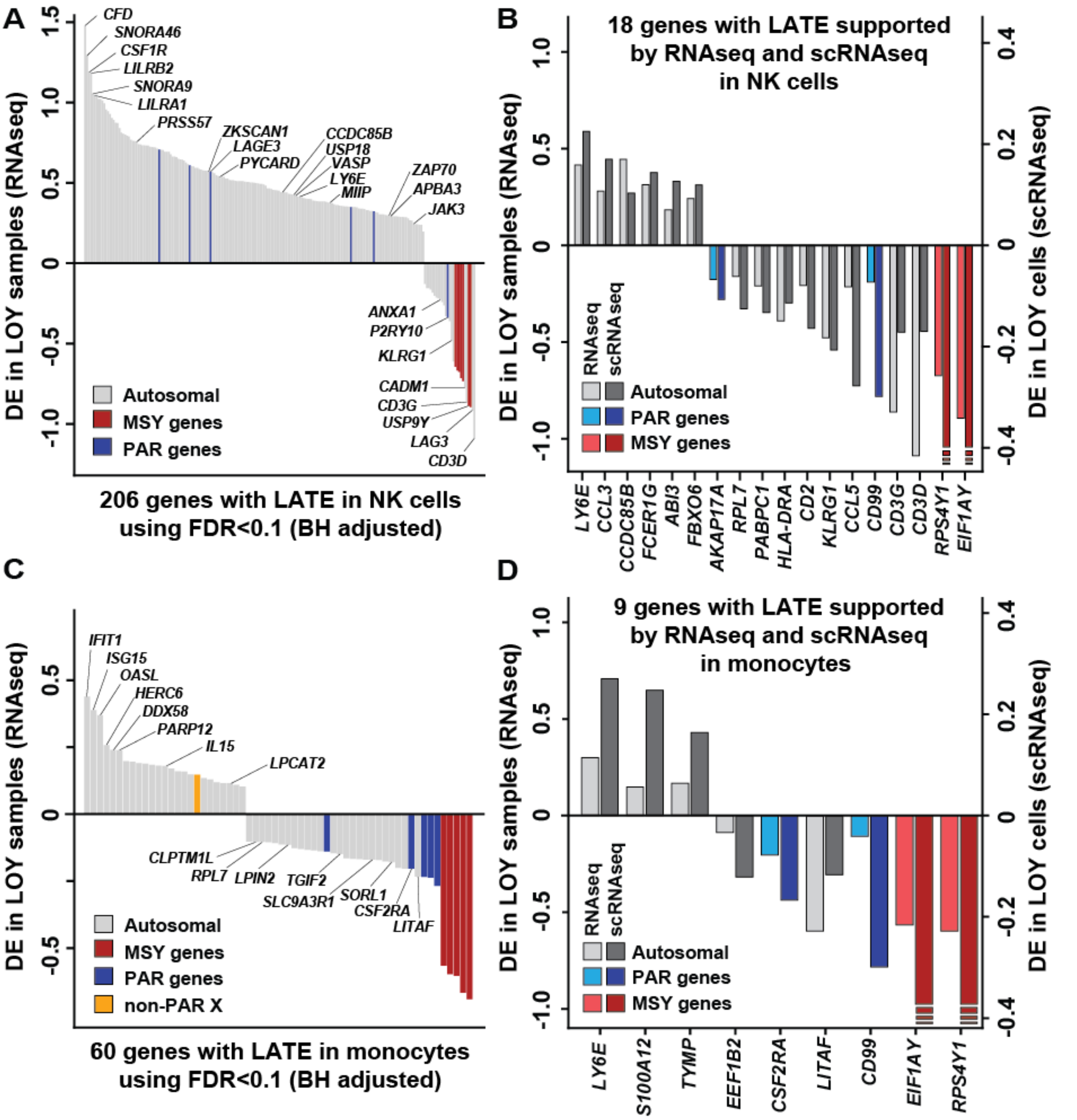


**Fig. 6.** Level of differential expression (DE) of genes showing substantial LATE in NK cells and monocytes *in vivo*. Panel A shows the level of DE in 206 genes with significant dysregulation after correction for multiple testing (FDR<0.1) and with at least an average of 100 reads per gene in RNAseq data from NK cells. Names are shown for LATE genes involved in immune system functions and/or cancer and/or development of Alzheimer's disease. Panel B shows DE in 18 genes detected with both RNAseq and scRNAseq in NK cells. Panels C and D show the corresponding results for monocytes, i.e. DE for 60 and 9 LATE genes detected in RNAseq and with both technologies, respectively.

## Figure 7

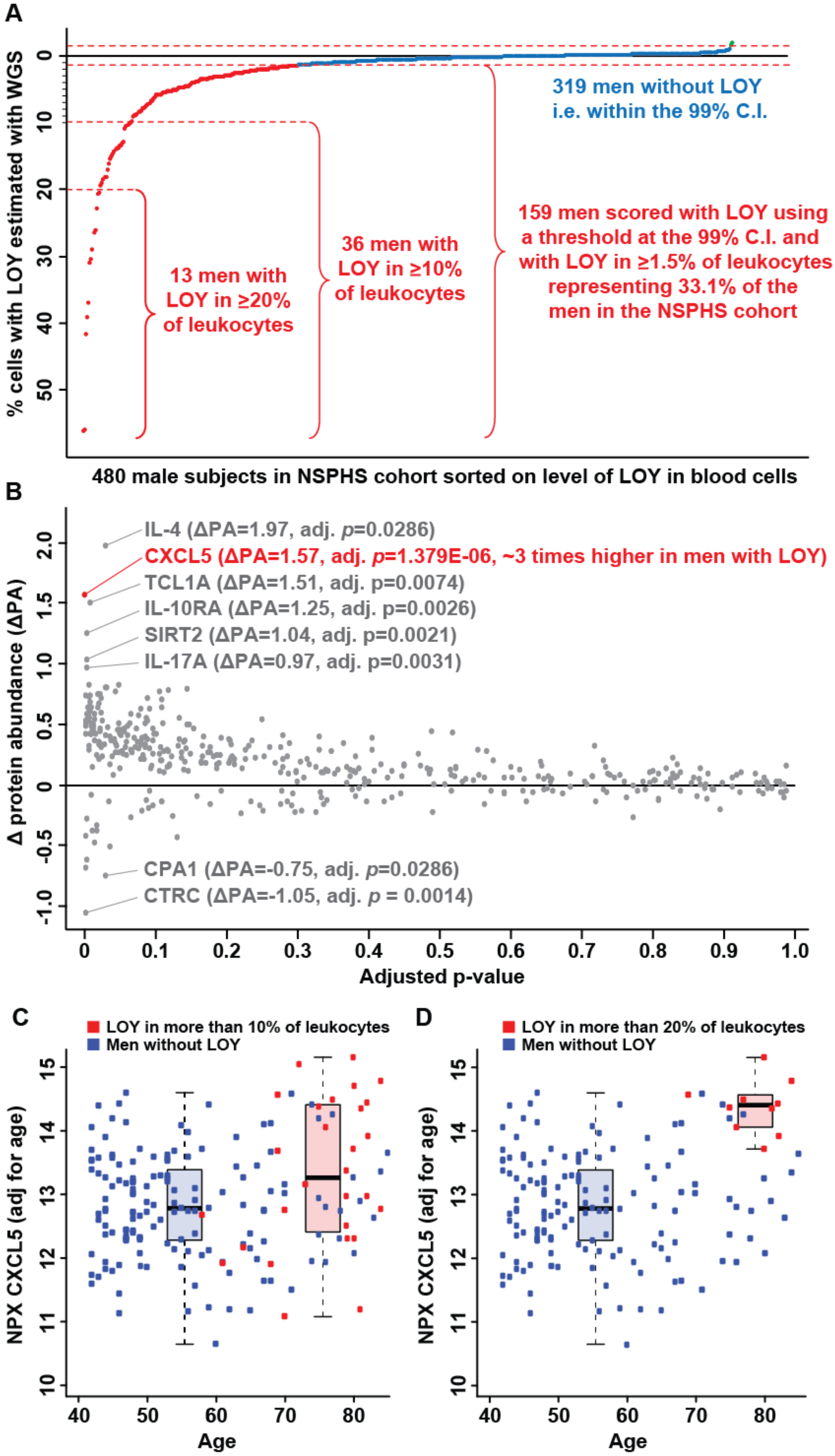

A
% cells with LOY estimated with WGS
0
10
20
30
40
50
319 men without LOY
i.e. within the 99% C.I.
13 men with
LOY in ≥20%
of leukocytes
36 men with
LOY in ≥10%
of leukocytes
159 men scored with LOY using
a threshold at the 99% C.I. and
with LOY in ≥1.5% of leukocytes
representing 33.1% of the
men in the NSPHS cohort
480 male subjects in NSPHS cohort sorted on level of LOY in blood cells
B
Δ protein abundance (ΔPA)
2.0
1.5
1.0
0.5
0
-0.5
-1.0
IL-4 (ΔPA=1.97, adj. p=0.0286)
CXCL5 (ΔPA=1.57, adj. p=1.379E-06, ~3 times higher in men with LOY)
TCL1A (ΔPA=1.51, adj. p=0.0074)
IL-10RA (ΔPA=1.25, adj. p=0.0026)
SIRT2 (ΔPA=1.04, adj. p=0.0021)
IL-17A (ΔPA=0.97, adj. p=0.0031)
CPA1 (ΔPA=-0.75, adj. p=0.0286)
CTRC (ΔPA=-1.05, adj. p = 0.0014)
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
Adjusted p-value
C
🟥 LOY in more than 10% of leukocytes
🟦 Men without LOY
NPX CXCL5 (adj for age)
10
11
12
13
14
15
40
50
60
70
80
Age
D
🟥 LOY in more than 20% of leukocytes
🟦 Men without LOY
NPX CXCL5 (adj for age)
10
11
12
13
14
15
40
50
60
70
80
Age

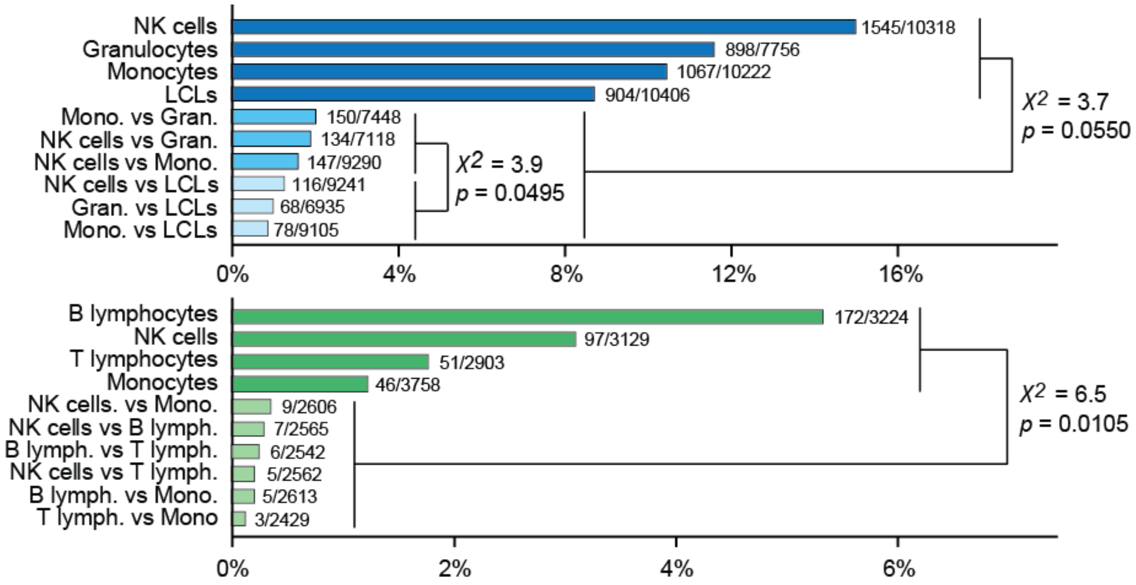
**Fig. 7.** Analyses of changes of plasma protein levels in men with LOY from the NSPHS cohort. Panel A shows result from analysis of LOY in blood DNA collected from 480 men studied with 30X whole genome sequencing. The level of LOY plotted on the Y-axis was estimated in each subject (represented by dots) by comparing the observed number of sequencing reads from the MSY in relation to the number of reads from the autosomes. The calculated ratio represents the percentage of cells with LOY, with a value close to zero in men carrying a Y chromosome in all studied cells. We used two thresholds for LOY scoring (i.e. ≥10% or ≥20% of cells with LOY) for identification of individuals with LOY for downstream analyses of protein abundance and comparisons with the 319 control subjects, defined using calculation of the 99% confidence interval (C.I.) of experimental variation. Panel B shows analysis of 424 plasma proteins in 13 men with LOY in ≥20% of leukocytes vs 319 controls. Protein levels were measured by the normalized protein expression (NPX). The difference in protein abundance (i.e. Δ PA) was estimated for each protein by subtraction of the observed mean NPX in the LOY group (adjusted for age) by the corresponding mean NPX in the controls. Abundance of each protein is visualized with grey dots and positive or negative values indicate higher or lower abundance in men with LOY vs controls, respectively. Panels C and D show individual measurements (NPX adjusted for age) of the CXCL5 for each of the subjects using the thresholds at ≥10% and ≥20% of leukocytes with LOY, respectively. The boxes on the X-axis are according to the mean age in each group.

**Table 1.** The number of expressed genes and the number of genes showing LOY Associated Transcriptional Effect (LATE) within and between different cell types detected using bulk RNAseq as well as scRNAseq technologies.

**A.** The number of expressed genes and the number of LATE genes **within** cell types detected using each and by both technologies.

| | **n. genes detected using RNAseq** | | **n. genes detected using scRNAseq** | | **n. LATE genes supported by RNAseq and scRNAseq (*p*)$^{\alpha}$** |
|---|---|---|---|---|---|
| **Cell type** | **n. genes in normal cells** | **n. LATE genes (*p*/adj. *p*)$^{\alpha}$** | **n. genes in normal cells** | **n. LATE genes (*p*/adj. *p*)$^{\alpha}$** | |
| **NK cells** | 10318 | 1545/312 | 3129 | 97/41 | 18 (β) |
| **Monocytes** | 10222 | 1067/98 | 3758 | 46/31 | 9 (γ) |
| **Granulocytes** | 7756 | 898/72 | - | - | - |
| **T lymphocytes** | - | - | 2903 | 51/16 | - |
| **B lymphocytes** | - | - | 3224 | 172/4 | - |
| **LCLs** | 10406 | 904/809 | - | - | - |

**B.** The fraction of LATE genes **within** cell types (n. LATE / n. expressed) and **between** different cell types (n. shared LATE / n. shared expressed) identified using RNAseq (blue) and scRNAseq (green).



**α)** *p* denotes number genes identified at α-level 0.05 and adj. *p* after correction for multiple testing (FDR<0.1)

**β)** LATE genes in NK cells supported by RNAseq and scRNAseq: *ABI3, AKAP17A, CCDC85B, CCL3, CCL5, CD2, CD3D, CD3G, CD99, EIF1AY, FBXO6, FCER1G, HLA-DRA, KLRG1, LY6E, PABPC1, RPL7* and *RPS4Y1* (see **Fig. 6B**)

**γ)** LATE genes in monocytes supported by RNAseq and scRNAseq: *CD99, CSF2RA, EEF1B2, EIF1AY, LITAF, LY6E, RPS4Y1, S100A12* and *TYMP* (see **Fig. 6D**)

Abbreviations: n. = number, chr. = chromosome, sc = single cell, lymph. = lymphocyte, NK = natural killer, Gran. = Granulocytes, Mono. = Monocytes, LCLs = lymphoblastoid cell lines, $X^2$= Kruskal-Wallis Chi-squared

# METHODOLOGY

## METHOD DETAILS – WET LAB

### *1. Sample collection*

Blood samples for DNA and mRNA analyses were collected from 408 male subjects in Uppsala, Sweden and in Kraków, Poland. In Uppsala, samples from the cohorts Uppsala Longitudinal Study of Adult Men (ULSAM), Alzheimer's disease cohort (UAD) and controls (M) were collected during January 2015 to May 2018, at the Geriatric/Memory Clinic, Uppsala Academic Hospital. In Kraków, samples from prostate cancer patients (KP) were collected from March 2015 to May 2018 at the Centre of Oncology, Maria Skłodowska-Curie Memorial Institute, Kraków Branch and the Department and Clinic of Urology of the Jagiellonian University Collegium Medicum in Kraków. Samples from patients with Alzheimer's disease (KAD) were collected from January 2017 to May 2018 at the Clinic of Internal Diseases and Gerontology of the Jagiellonian University, Collegium Medicum in Kraków. Control samples (M) were collected from December 2015 to May 2018 from the general population of Kraków and Uppsala. Blood samples for DNA and protein analyses were collected from 480 male participants of The Northern Swedish Population Health Study (NSPHS) during 2006 and 2009 (Assarsson and Lundberg, 2017). The criteria for recruitment of prostate cancer patients was advanced stage prostate cancer; i.e. Gleason grade 7 or higher. The prostate cancer patients were recruited before treatment, or during the first stage of treatment. As for Alzheimer's disease, patients with ongoing clinically and radiologically confirmed diagnosis were recruited. The clinical stage for recruited Alzheimer's patients was intermediate or severely advanced disease.Written informed consent was obtained from all individuals included in the study. The research was approved by the local research ethics committees of Uppsala Region in Sweden (Regionala Etikprövningsnämnden, Dnr 2013/350, Dnr 2015/092, Dnr 2015/458, Dnr 2015/458/2 with update from 2018) and by the Bioethical Committee of the Regional Medical Chamber in Kraków, Poland (NR: 6/KBL/OIL/2014).

### *2. Preparation and sorting of blood cells with FACS*

We implemented two strategies for preparation of blood cells for sorting using Fluorescence Activated Cell Sorting (FACS), as described in sections 2.1 and 2.2 below.

### *2.1 Isolation and labeling of peripheral blood mononuclear cells (PBMCs) for sorting of T- and B lymphocytes as well as NK cells.*

16 ml of whole blood were collected into two BD Vacutainer® CPT™ Mononuclear Cell Preparation Tubes (BD). Blood samples were centrifuged within 2 hours of blood collection, following manufacturer's instructions. Isolated peripheral blood mononuclear cells (PBMCs) were stained with 20 µl of BD Multitest™ 6-color TBNK reagent (BD) and incubated for 20 min. at 4ºC. After incubation, PBMCs were washed with PBS and cell pellets were resuspended in 1 ml of PBS containing 3 mM EDTA.PBMCs were filtered through 40 µm cell strainer to obtain real single cell suspensions by removing cell aggregates (clumps).

### *2.2. Isolation and labeling of white blood cells (WBCs) for sorting of granulocytes, monocytes and B lymphocytes*

16 ml of whole blood were collected into two BD Vacutainer® K2 EDTA tubes (BD). Red blood cells were lysed using 1× BD Pharm Lyse™ lysing solution (BD) added to blood samples up to 50 ml. Samples were incubated at room temperature for 10 min. The lysing step was repeated and then cells were then washed with PBS. Isolated white blood cells (WBCs) were stained with following antibodies: 5 µl of PE-labeled CD14, clone MφP9 (BD) and 20 µl of APC-labeled CD19, clone HIB19 (BD) and incubated for 20 min. at 4ºC. After incubation, WBCs were washed with PBS and cell pellets were resuspened in 2 ml of PBS containing 3 mM EDTA. WBCs were filtered through 40 µm cell strainer to obtain real single cell suspensions by removing cell aggregates (clumps).

### *2.3 Sorting of target cell populations with FACS*

The target populations of cells were isolated using FACS Aria III (Becton Dickinson) at Uppsala University, FACS Aria II (Becton Dickinson) at Jagiellonian University Collegium Medicum or MoFlo (Beckman Coulter) at Jagiellonian University. Data was acquired and analyzed using BD FACSDiva™ Software (Becton Dickinson) with FACS Aria III and FACS Aria II or Summit™ Software System (Cytomation) with MoFlo. Compensation was determined using cells stained with single antibody from BD: FITC–labeled CD3, clone UCHT1; PE-labeled CD14, clone MφP9; PerCP-Cy™5.5–labeled CD45, clone HI30; PE-Cy™7–labeled CD4, clone SK3; APC-labeled CD19, clone HIB19; APC-Cy7–labeled CD8, clone SK1.

Live cells were sorted based on their FSC and SSC. CD4+ T cells were defined as CD45+CD3+CD8-CD4+; CD8+ T cells were defined as CD45+CD3+CD4-CD8+; B cells were defined as CD45+CD3-CD19+; NK cells were defined as CD45+CD3-CD4-CD16+CD56+; monocytes were defined based on their size and as CD14+; granulocytes were defined based on their size and granularity and additional B cells were defined based on their size and as CD19+. Cells were sorted to achieve the purity of above 96%. At least 200 000 cells of each type were sorted. After sorting, cells were centrifuged for 5 min. at 400 RCF and 1 ml of RNAProtect (Qiagen) was added to cell pellets. Resuspended cells were centrifuged for 5 min. at 400 RCF, the supernatant was removed and the cell pellets were frozen immediately on dry ice for further processing.

### *3. Establishment of lymphoblastoid cell lines (LCLs)*

Peripheral blood mononuclear cells (PBMC) were isolated from blood by Ficoll-Paque gradient separation. PBMC were infected with EBV by incubating them with supernatant1 of the virus producing B95-8 line for 90 min at 37 ºC. Thereafter the cells were washed and re-suspended in complete RPMI-1640 medium (supplemented with 10% heat inactivated FCS, penicillin and streptomycin). 1 µg/ml Cyclosporine A was added to the cultures in order to inhibit T cell expansion and function. After 3 weeks of culturing, transformed cells were plated into microtitre plates with U-shaped bottom at a 1 cell/well density. In each well, gamma-irradiated human fibroblasts (5000 rad) served as feeder cells. Growing clones were expanded and further analyzed. Clones showing either 100% LOY or 100% normal cells were selected for further study. The percentage of LOY was determined using both genotyping and ddPCR technologies (see below).

### *4. LOY analyses using DNA from sorted cells and LCLs*

#### *4.1 DNA extraction*

DNA was extracted from cell pellets of sorted cells using an in-house protocol. The cell pellet was resuspended in lysis buffer containing 10 mM EDTA, 10 mM Tris-HCL (pH 7.9), 50 mM NaCl, 1% N-Lauroylsarcosine sodium salt (Sigma) with 10 mg/ml proteinase K (Sigma) and incubated for 2 hours in 50ºC. The DNA was then precipitated using Sodium Acetate (pH 5.4) and 96% Ethanol, washed with 80% Ethanol and resuspended in MQ water. DNA was extracted from whole blood using QIAmp DNA Blood Midi kit (Qiagen) according to manufacturer's protocol. The concentration of DNA was measured using Quant-iT™ PicoGreen® dsDNA

Assay Kit (Thermo Fisher Scientific) according to manufacturer’s instructions, and analyzed using plate reader Infinite M200 (Tecan Diagnostics). DNA was stored in -20ºC freezer.

### *4.2 Genotyping using SNP-arrays*

All genotyping experiments were performed following the manufacturer’s instructions at the Science for Life technology platform SNP&SEQ at Uppsala University, Sweden. DNA extracted from the 2661 FACS sorted populations (collected from 408 subjects) was genotyped using three different versions of Illumina SNP-arrays (242 samples on the InfiniumCoreExome-24v1-1, 60 samples on the InfiniumOmniExpressExome-8v1-3 and 2359 samples on the InfiniumQCArray-24v1). The DNA extracted from the 13 LCLs was genotyped using the InfiniumQCArray-24v1.

### *4.3 Experiments using digital droplet PCR (ddPCR)*

Additional quantification of the level of LOY for validation was performed using previously described procedure [1]. Briefly, extracted DNA from the established LCL cell lines was digested for 15 min with HindIII enzyme (Thermo Fischer) in 37 °C. After this, 50 ng of digested and diluted DNA was added together with PCR primers and probes targeting a known difference between the AMELX and AMELY gene assay C_990000001_10 (Thermo Fisher ) to ddPCR supermix for probes no dUTP (Bio-Rad). Droplets were then generated using an automated droplet generator (Bio-Rad). Following this, the digested DNA was amplified using PCR. Droplets fluorescence intensity in two channels FAM and VIC was measured using the Bio-Rad’s QX200 Droplet Reader.

## ***5. Experiments using bulk RNA from sorted cells and LCLs***

### *5.1 RNA extraction*

RNA was extracted from cell pellets using RiboPure™ RNA Purification Kit (Thermo Fisher Scientific) according to the manufacturer’s instructions. A possible DNA contamination was removed using TURBO DNA-free™ Kit (Thermo Fisher Scientific) following the manufacturer’s protocol. The RNA was then concentrated using GeneJET RNA Cleanup and Concentration Micro Kit (Thermo Fisher Scientific) according to manufacturer’s instructions. RNA quality was measured with RNA Pico Chip (Agilent Technologies) following manufacturer’s protocol, and analyzed using Agilent 2100 Bioanalyzer (Agilent Technologies). RNA was kept on ice during handling and was stored in -80 ºC freezer.

### *5.2 Bulk RNA sequencing (RNAseq)*

Bulk RNA sequencing was performed by the Science for Life technology platform at Uppsala Genome Centre (Uppsala University, Sweden). The applied method quantifies the amount of mRNA of different genes (about 20,000) by amplicon-specific primer pairs (usually targeting exon 1 and exon 2 of the spliced mRNA) in PCR amplification followed by next-generation sequencing. For each sample, an average of 10 ng of RNA (1 - 100 ng) from the 150 FACS sorted samples was used. Library preparation was performed using the Ion Ampliseq Human Gene Expression kit (Thermo Fisher Man0010742). Briefly, cDNA libraries were first constructed followed by amplification using Human Gene Expression Core panel (Thermo Fisher). For samples with a low amount of starting RNA the number of amplification cycles was increased (to 11 - 16 cycles, depending on RNA abundance in the specimen). Then, IonCode (Thermo Fisher) barcodes were ligated to the amplified mRNA. Amount of RNA in each library was estimated using a Fragment Analyzer instrument (Agilent). Libraries were loaded onto Ion 550 chips following recommendations from the manufacturer (MAN0017275) and sequenced using an Ion S5 XL instrument. In total, 51 monocyte and granulocyte samples and 48 NK cell samples were sequenced. Out of these samples, 2 granulocyte and 2 NK cell samples failed to produce sequencing results.

## ***6. Experiments using RNA from single cells***

### *6.1 Sample preparation*

Whole blood was collected from Swedish men using BD Vacutainer CPT tubes (BD Biosciences) following the manufacturer's instructions and stored on ice. PBMCs were isolated using density gradient centrifugation and resuspended in 0.04% BSA 1X PBS solution. Concentrations were measured using an EVE cell counter (NanoEnTek, Seoul) and cells were diluted to $10^6$ cells/ml. The cells were then delivered to the Science for Life SNP&SEQ Technology platform at Uppsala University, Sweden.

### *6.2 Single cell RNA sequencing (scRNAseq)*

The collected PBMCs were loaded on a 10X Genomics Chromium Single Cell 3' Chip v2 for sequence library generation using protocol CG00052 (10X Genomics). The applied method quantifies the amount of mRNA of different genes in single-cells using a 3'-end protocol. It allows for gene expression profiling of up to 10,000 individual cells per sample by construction of Gel Bead Emulsions (GEMs). Each GEM contains a single cell together with reagents such as (i) an Illumina R1 sequence (read 1 sequencing primer), (ii) a 16 nt 10x Barcode, (iii) a 10 nt Unique Molecular Identifier (UMI), and (iv) a poly-dT primer sequence. Incubation of the GEMs produces barcoded, full-length cDNA from poly-adenylated mRNA. GEMs are thereafter broken and the pooled fractions are recovered and sequenced. The above steps (including sampling and sample preparation) were performed within 5 hours. The single cell libraries were sequenced on Illumina HiSeq2500 and NovaSeq instruments, according to manufacturer's instructions. All single cell library preparation and sequencing were performed at the Science for Life technology platform SNP&SEQ, Uppsala University, Sweden.

### *7. Measurements of plasma protein abundance in men with LOY*

To investigate if LOY leaves a footprint in the plasma proteome previously generated data from NSPHS cohort was analyzed. The fraction of cells with LOY was estimated from sequencing data and protein abundance in the same subjects was measured using Proximity Extension Assays (PEA).

#### *7.1 Whole genome sequencing (WGS)*

From all 480 men in the NSPHS cohort, 30X whole genome sequencing (WGS) was available for LOY analysis and generated as follows. Sample DNA was fragmented using a Covaris E220 (Covaris Inc., Woburn, MA, USA) to an insert size of 320 bp. Sequencing libraries were prepared using 1.1/1 $\mu$g DNA using the TruSeq DNA PCR free sample preparation kit (illumina Inc. guide 15036187). Libraries were sequenced using an Illumina HiSeq X instrument (HiSeq Control Software 3.3.39/RTA 2.7.1, v2.5 sequencing chemistry). The sequencing was performed at the Science for Life SNP&SEQ technology platform, Uppsala University, Sweden.

#### *7.2 Proximity Extension Assays (PEA)*

From the same set of 480 NSPHS men, plasma protein abundance of 441 proteins was available. This data was generated using five different PEA-panels commercially available from Olink Proteomics, Uppsala, Sweden (www.olink.com). Specifically, panels CVD II, CVD III, INF I, ONC 2 and NEU I were analyzed by Olink company. The sample preparation and PEA workflow has been described previously (Enroth et al., 2018).

## QUANTIFICATION AND STATISTICAL ANALYSIS

### *8. Quantification of LOY from DNA*

*8.1 LOY analysis using SNP-array data*

All included experiments passed strict quality control at the genotyping facility and in addition, we calculated for every experiment the sdLRR1, i.e. the standard deviation of Log R Ratio of probes on chromosome 1. Only samples with sdLRR1 <0.28 were considered to have good quality and included in downstream analyses, as recommended by Illumina (Technical note from Illumina, https://www.illumina.com/content/dam/illuminamarketing/documents/products/appnotes/appnote_cnv_loh.pdf ). For each sample, we calculated the median of the log R ratio values of the SNP-array probes located in the male specific region of chromosome Y (MSY), i.e. the continuous mLRRY, as described previously [2]. To correct for genotyping batch effects, we calculated the local regression median (LRM) in each batch and the mLRRY value for each subject was thereafter corrected by the batch-specific LRM. The percentage of normal cells without LOY in each sample was estimated using a formula described previously, i.e. $100*(2^{2*\mathrm{mLRRY}})$ [1].

*8.2 LOY analysis using ddPCR data*

The data generated by the Bio-Rad's QX200 Droplet Digital PCR System was analyzed in Bio-Rad's software QuantaSoft (version 1.7.4.0917) as described elsewhere [1]. Briefly, the amount of DNA of *AMELY* was divided by the amount of DNA for *AMELX* in each sample. The ratio represents an unbiased estimate of the level of LOY in the examined samples.

*8.3 LOY analysis from WGS data*

We used for scoring of LOY from whole genome sequencing data, a pipeline similar to previously described [3]. Briefly, sequencing raw reads were mapped and aligned using BWA-MEM 0.7.12 and v. GRCh37 of the human reference, followed by sorting and indexing by Samtools 0.1.19. Alignments from different lanes and flow cells were merged using Picard (1.120). In order to estimate the amount of LOY in each male sample the software ControlFREEC (version *FREEC-11.5*) [4] was used. ControlFREEC calculates read depth ratios for genomic windows from WGS-data while taking mappability for each region into account. As the ratio produced by ControlFREEC will equal to 1 for any region with two chromosomes and without any traces of CNVs, a Y chromosome region with the same conditions for a male subject without LOY will be 0.5. The Y-chromosome ratio was therefore multiplied by 2 for each male in order to get a Y chromosome ratio between 0 and 1. For the ControlFREEC settings, it was given the same references file that was used to create the bam-files as well as the 2 mismatches, hg19 and read length of 100 base pairs mappability file linked in programs manual (ControlFREEC v11.5). The settings were kept as default, except for the window-size set to 50.000 base pairs.

## *9. Quantification of LOY from RNA*

### *9.1 Analysis pipeline of bulk RNAseq data from sorted cells*

For each of the FACS sorted samples analyzed using the Ion Ampliseq Human Gene Expression kit, raw read-counts were read into the R (v3.5.3) software, and one matrix of read-counts was created per cell type. Outliers were searched for within each cell type using principal component plots. Only among the monocyte samples, eight outliers were identified and removed. To remove low quality expression data, the count matrices were filtered by creating two groups of samples (i.e. more and less than 50% LOY). In each group, only genes with at least 10 reads in one third of the samples were used.

In order to identify the normally expressed genes in the sorted cell fractions, we focused on samples without LOY. In these, a minimum of 5 reads in at least one third of the samples was required to be considered a normally expressed gene. In order to estimate the level of LOY in each sample in the RNAseq data, a variance-stabilizing transformation was first applied (DESeq2, varianceStabilizingTransformation function) to the count matrix in order to make the expression more coherent. A mean expression was then calculated for all genes on the autosomal chromosomes, and compared to the mean expression of the MSY-genes, generating a fraction of expression for the Y-chromosome. The calculated fractions were rescaled to fit between 0 and 100. This was done per cell-type, to remove cell type specific effects on expression levels.

We estimated the LOY associated transcriptional effects (LATE) as follows. The R library DESeq2 (v1.22.2) was used to identify genes with differential expression in the leukocytes derived from blood for each patient. As LOY is a continuous trait, the percentage of LOY estimated from SNP-arrays for each sample was used for statistical testing. To decrease the impact of read-count outliers the fraction of LOY was binned using 20% intervals between <20% - 80%< and these bins were used as a continuous input variable for DESeq2 [5]. To avoid batch effects batches were introduced as a covariate in the analysis. Independent filtering of genes in DESeq2 was done using 0.1 as a value for the alpha parameter, which was also used as a cutoff for the DESeq2 adjusted p-values.

To investigate the relation between LOY and expression level of MSY genes, we applied a non-linear least square model, using nls() function in R, on average expression of all selected MSY genes (6 in panel A and 13 in panel B) and percentage of LOY in samples of each cell type. For each cell type, the expression level of each sample has been divided by the maximum expression level among all samples in order to keep all values in the same scale of 0 to 1.

### *9.2 Analysis pipeline of bulk RNAseq data from LCLs*

We used R (version 3.5.2) and the edgeR (version 3.24.3) library to compare the expression between LOY and non-LOY group of samples in two steps. As all the clones with LOY were derived from the same individual (this individual also gave rise to one non-LOY clone) we designed the analysis to remove inter and intra individual differential expression effects.
As step one, in order to identify the intra-individual differences, we targeted the differential expression genes among all non-LOY clones by comparing each non-LOY individual's clones with all other non-LOY clones. The union of all differentially expressed genes resulted out of each comparison was gathered into List1 (**Figure S12**). This gave us a set of genes that are differentially expressed between subjects but not due to LOY status. As step two, we targeted the differentially expressed genes regarding LOY status in two steps: first by comparing the non-LOY versus LOY clones derived from the same individual; second by comparing all non-LOY versus LOY clones and the overlap of the results of steps one and two was gathered into List2. The final list of LATE genes was generated by comparing List2 and List1 and extracting the differentially expressed genes that exist in List2 but not in List1, to remove the intra-individual differentially expressed genes from the LATE genes.

### *9.3 Analysis pipeline of single cell scRNAseq data*

Sequence reads were mapped to the hg19 version of the human genome using CellRanger v2 (10X Genomics) using recommended settings. The produced raw reads matrices were analyzed using R (v3.4) and the R library Seurat (v2.2) [6]. For each sample the read count matrix was used to create a Seurat object in R. After this all data was log normalized and variable genes were identified using standard settings from the Seurat manual. Data was then scaled and the principal components were calculated. Then, the number of informative principle components for each sample was identified using an elbow plot and cells were then clustered using the findclusters Seurat function and informative principle components.

A tSNE plot was produced with the resulting clusters of the nearest neighbor clustering indicated by the color of each cell in the tSNE plot, to confirm consistency between the two methods. An in house R script was then used to calculate the level of expression of cell marker genes in each cluster and cell types was assigned using the strongest expression signal. These results were then manually inspected using heatmaps in order to confirm the automated cell-type assignment.

For each sequenced cell, the number of reads mapping to genes in the MSY was counted. Cells having a LOY transcriptional profile were identified as cells having no expression from genes in the MSY. To calculate percentage of LOY in each sample and cell-type the fraction of cells having LOY was calculated for each specific cell type. Following this step, all individuals were merged together creating one object for each studied cell type (monocytes, NK cells, T- and B lymphocytes). The procedure outlined above was repeated for the merged data-sets and the cells were filtered only keeping cells with at least 800 expressed genes but no more than 2000 expressed genes. This was done both to remove experiments where two cells were sequenced together, and to remove cells with a generally low expression of genes, where the LOY estimate may be erroneous. Cells with more than 5% mitochondrial RNA were also excluded, in order to remove possibly apoptotic and damaged cells from the analysis.

Normally expressed genes in the scRNAseq data were defined by the expression in at least 10% of single cells without LOY. After establishing the normally expressed genes in each cell type in the scRNAseq data, we further filtered the cells to only include only cells having more than 2500 UMIs. This was done in order to remove possible false positive LOY cells, and only keep cells where a robust and high number of UMIs was found. We then identified LATE genes using the Seurat FindMarkers function. We used this function to test differences in gene expression in cells with and without LOY using the built-in negative binomial test.

### *9.4 Co-expression analysis*

We investigated if changes in expression as an effect of LOY would be larger in autosomal genes that are normally co-expressed with MSY genes, compared with genes that are not co-expressed with MSY genes. Six MSY genes that are normally expressed in lymphocytes were used in this analysis (i.e. *RPS4Y1, ZFY, USP9Y, DDX3Y, KDM5D* and *EIF1AY*). Genes normally co-expressed with these MSY-genes were identified using the SEEK database [http://seek.princeton.edu/]. After this we calculated the fold change of the expression between LOY and normal cells for each cell-type. We then selected the top 300 co-expressed genes retrieved from the SEEK database and for each of these genes we identified if they were expressed in our scRNAseq data or not. Following this the fold changes for genes found to be co-expressed and genes that were not co-expressed was used to produce plots and p-values (Wilcoxon rank sum test).

### *10. Gene ontology (GO) functional annotation*

For annotation of known functions of chromosome Y genes for **Fig. 1**, a list of all genes located on the Y chromosome together with associated GO categories was downloaded from Ensembl (v. 95) using the R library bioMart (v2.38). The list was filtered for protein coding genes and annotated as either positioned in the male specific part of the Y chromosome (MSY) or in the pseudo autosomal regions (PAR). The GO categories for each gene were inspected manually and genes were assigned into four different categories: "*Transcription, translation epigenetic and PTM functions*", "*Cell differentiation, migration proliferation and apoptosis*", "*Sex determination, fertility and other functions*", "*Immune response and immune cell signaling*".

We also performed GO analyses of the LATE genes identified using a stringency level of $p$<0.05 (n=3360) as well as for LATE genes showing differential expression using the more stringent significance level of FDR<0.1 (n=521) in the samples studied *in vivo* using scRNAseq and RNAseq (**Table S2**). The analyses were performed using GO Ontology database (http://geneontology.org/) on the 2019-05-24 using the "*GO biological process complete*" option, the Fisher's exact test, and the FDR correction method. As input we used the 521 and 3360 LATE genes showing differential expression as an effect of LOY and as reference background, we used all the expressed genes identified *in vivo* in the scRNAseq and RNAseq experiments (n=11859).

### *11. Analysis of protein abundance*

Analysis of the protein data was performed using an adapted Personally Normalized Plasma Protein Profiles (PNPPP) method [7]. Briefly, variance in protein abundance levels not related to LOY was accounted for by first examining a set of 159 covariates (anthropometrics, lifestyle variables, medication and genetic markers) using ANOVA in individuals not affected by LOY. The genetic marker was defined as the dosages for the top-hit for each protein as reported in Enroth et al 2018 [8]. Age was always included as a covariate due to the high correlation between LOY and age. The models generated using the individuals not affected by LOY were then used to adjust the protein abundance levels in all individuals, including those affected by LOY. In a second step, adjusted protein levels were compared between the two groups and a two-sided Spearman's ranked test was used to calculate p-values. A Bonferroni adjusted q-value of 0.05 was used to determine statistically significant differences. Out of the 441 proteins 16 had <20% of samples above the limit of detection and were excluded from the final analysis.

### *12. Statistical analyses*

Statistical analyses such as logistic regression, linear regression, Fisher extact test, Kolmogorov–Smirnov test and Wilcoxon rank sum test was performed using R version 3.3.1 (http://www.rproject.org/), using two-sided tests and α-levels at 0.05 and default parameters, unless otherwise specified in the text. When appropriate, correction of p-values for multiple testing was applied, such as Bonferroni correction in the protein analyses (**Figure 7**). In order to define the LATE genes, we applied three stringency levels. The first was based on standard p-value at α-level 0.05 and the second was a more strict p-value correcting for multiple testing (FDR<0.1) using Benjamini and Hochberg (BH) correction (**Table 1**). Furthermore, an additional criterion was applied to identify only the highly expressed LATE genes, i.e. using threshold of at least an average of 100 reads per gene in RNAseq data (**Figure 6**).

## REFERENCES FOR METHODOLOGY

1. Danielsson M, Halvardson J, Davies H, Torabi Moghadam B, Mattisson J, Rychlicka-Buniowska E, Heintz J, Lannfelt L, Giedraitis V, Ingelsson M, Dumanski J, Forsberg L: Intra-individual changes in the frequency of mosaic loss of chromosome Y over time estimated with a new method, submitted 2019, (doi: https://doi.org/10.1101/631713)
2. Forsberg LA, Rasi C, Malmqvist N, Davies H, Pasupulati S, Pakalapati G, Sandgren J, de Stahl TD, Zaghlool A, Giedraitis V, Lannfelt L, Score J, Cross NC, Absher D, Janson ET, Lindgren CM, Morris AP, Ingelsson E, Lind L, Dumanski JP: Mosaic loss of chromosome Y in peripheral blood is associated with shorter survival and higher risk of cancer, Nature Genetics 2014, 46:624-628; (http://www.ncbi.nlm.nih.gov/pubmed/24777449)

3. Dumanski JP, Lambert JC, Rasi C, Giedraitis V, Davies H, Grenier-Boley B, Lindgren CM, Campion D, Dufouil C, European Alzheimer's Disease Initiative I, Pasquier F, Amouyel P, Lannfelt L, Ingelsson M, Kilander L, Lind L, Forsberg LA: Mosaic Loss of Chromosome Y in Blood Is Associated with Alzheimer Disease, Am J Hum Genet 2016, 98:1208-1219; (http://www.ncbi.nlm.nih.gov/pubmed/27231129)
4. Boeva V, Zinovyev A, Bleakley K, Vert JP, Janoueix-Lerosey I, Delattre O, Barillot E: Control-free calling of copy number alterations in deep-sequencing data using GC-content normalization, Bioinformatics 2011, 27:268-269; (http://www.ncbi.nlm.nih.gov/pubmed/21081509)
5. Love MI, Huber W, Anders S: Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2, Genome Biol 2014, 15:550; (https://www.ncbi.nlm.nih.gov/pubmed/25516281)
6. Butler A, Hoffman P, Smibert P, Papalexi E, Satija R: Integrating single-cell transcriptomic data across different conditions, technologies, and species, Nat Biotechnol 2018, 36:411-420; (https://www.ncbi.nlm.nih.gov/pubmed/29608179)
7. Enroth S, Bosdotter Enroth S, Johansson A, Gyllensten U: Effect of genetic and environmental factors on protein biomarkers for common non-communicable disease and use of personally normalized plasma protein profiles (PNPPP), Biomarkers 2015, 20:355-364; (https://www.ncbi.nlm.nih.gov/pubmed/26551787)
8. Enroth S, Maturi V, Berggrund M, Enroth SB, Moustakas A, Johansson A, Gyllensten U: Systemic and specific effects of antihypertensive and lipid-lowering medication on plasma protein biomarkers for cardiovascular diseases, Sci Rep 2018, 8:5531; (https://www.ncbi.nlm.nih.gov/pubmed/29615742)