

Genetic variants associated with cell-type-specific intra-individual gene expression variability reveal new mechanisms of genome regulation

Angli Xue^{1,2}, Seyhan Yazar¹, José Alquicira-Hernández^{1,3,4}, Anna S E Cuomo^{1,5}, Anne Senabouth¹, Gracie Gordon⁶, Pooja Kathail⁶, Chun Jimmie Ye^{6,7}, Alex W. Hewitt⁸, Joseph E. Powell^{1,9,*}

Affiliations:

¹ Garvan-Weizmann Centre for Cellular Genomics, Garvan Institute of Medical Research, Sydney, NSW, 2010, Australia

² School of Biomedical Sciences, University of New South Wales, Sydney, NSW, 2052, Australia

³ Institute for Molecular Bioscience, The University of Queensland, Brisbane, QLD, Australia

⁴ Division of Genetics, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, MA, USA

⁵ Centre for Population Genomics, Garvan Institute of Medical Research, Sydney, NSW, 2010, Australia

⁶ University of California, San Francisco, San Francisco, CA, USA

⁷ Parker Institute for Cancer Immunotherapy, San Francisco, CA, USA

⁸ Menzies Institute for Medical Research, University of Tasmania, Hobart, TAS, Australia

⁹ UNSW Cellular Genomics Futures Institute, University of New South Wales, Sydney, NSW, 2052, Australia

*Correspondence to: j.powell@garvan.org.au

Contents

Supplementary Notes 1-5

Supplementary Figures 1-20

Supplementary Note 1

Dissecting the mean-variance relationship in the single-cell gene expression data

We tried to quantify the relationship by visualize the Spearman's correlation coefficient (ρ_s) between within-individual mean and variance for all the genes. The mean and variance were calculated from the SCTransformed counts per individual, per gene, for each cell type. We observed that the median ρ_s ranged from 0.938 ~ 0.984 across 14 cell types (**Figure 2A**). Notably, there exist an obvious drop-off in the mean-correlation curve at the mean ranging from 0.01 to 0.1 (**Supplementary Figure 2**). Based on this observation, we predict that many of the identified veQTLs might also be significant in the eQTL association test. Yazar et al. (2021) performed eQTL analysis and identified 25,606 eQTLs with mean effects on gene expression within individuals in a cell-type-specific manner in the OneK1K cohort. Since this number is based on Spearman's rank correlation analysis, for fair comparison, we reran the analysis using the Matrix eQTL and compared the results with veQTL signals (**Methods**).

Supplementary Note 2

Comparison of different dispersion indices

There are multiple ways to calculate the dispersion of the within-individual gene expression across cells. The most common non-parametric indices are Fano factor, coefficient of variation (CV), and CV^2 .

$$Fano\ factor = \frac{\sigma^2}{\mu}$$

$$CV = \frac{\sigma}{\mu}$$

$$CV^2 = \frac{\sigma^2}{\mu^2}$$

From the equation, we know that these indices still preserve a negative relationship with the mean. Given a negative binomial (NB) distribution where $\sigma^2 = \mu + \mu^2 \times \theta$, we can further derive these indices as:

$$Fano\ factor = 1 + \mu \times \theta$$

$$CV = \sqrt{\frac{1}{\mu} + \theta}$$

$$CV^2 = \frac{1}{\mu} + \theta$$

From the derived equation, Fano factor is linearly correlated with the mean given fixed dispersion. The CV and CV^2 would become extremely large when the mean is small, which is very common in single-cell data. Based on these properties, none of these indices are expected to be good indicators for dispersion.

Another index to represent dispersion is the residuals from the linear regression of variance on the mean for each individual. Several studies have adopted this strategy before. However, this strategy has three main limitations. First, the strategy assumes a linear relationship between the mean and variance, but, given the properties of the NB distribution, this assumption is clearly violated. Second, if there is a true biological relationship between the mean and dispersion, this residual strategy will artificially remove it. Third, when the mean is very small, use an NB distribution; the variance will be very close to the mean (in the meantime, many intra-individual means will be zero under this scenario); thus, the linear regression will suffer from collinearity, and the residuals will be heavily skewed. A more detailed evaluation of the count distribution model was described below.

Supplementary Note 3

Evaluation of the fitness of different models for the count distribution

To evaluate the suitable distribution to model count data for intra-individual expression, we used several methods to assess the underlying count model across all genes and cell types. First, we estimated the intra-individual mean and variance for each individual across all genes. We found that, except for genes with zero expression, most intra-individual variance estimates are larger or equal to the intra-individual mean estimates.

Second, we ran a likelihood ratio test (LRT) to evaluate the goodness-of-fit between the Poisson and NB models for the intra-individual count distribution for all the genes, and tested them in every individual and each cell type. On average, 21.1% of genes showed nominal evidence (P -value < 0.05) favouring the NB model rather than the Poisson model (**Supplementary Figure 14**). This means we will need to at least use NB distribution to model the dispersion of the gene expression across cells. We also ran an LRT between NB and ZINB, which showed that only a very small proportion of genes (~1.46%) showed evidence in favour of ZINB distribution (**Supplementary Figure 15**). Based on these analyses, we chose to use NB model as the default model and estimate the dispersion parameter (θ) as the indicator for intra-individual variability.

Supplementary Note 4

The simulation for dispersion estimation using Cox-Reid adjusted MLE

To assess the performance of Cox-Reid adjusted MLE for dispersion parameter, we conducted a simulation with different numbers of cells (M), mean (μ), and dispersion (θ). The simulation parameters are as follows:

$M \in (10, 25, 50, 100, 500, 1000, 5000)$

$\mu \in (0.001, 0.01, 0.05, 0.1, 0.2, 0.3, 0.4, 0.5, 1, 1.5, 2, 2.5, 3, 5, 10, 30)$

$\theta \in (0, 0.1, 0.5, 1)$

We simulated an expression matrix with 1,000 genes x m cells and each gene follows the same Negative Binomial distribution. Each row is drawn from $X \sim NB(\mu, \theta)$, and the R code is `rnbinom(n = n, size = 1/theta, mu = mu)`. We then used `glmGamPoi::glm_gp()` to estimate the dispersion parameter simultaneously across all rows and obtained the mean of the dispersion estimates over 1,000 rows. The whole process is replicated for 100 times. For each true dispersion parameter, we visualized the simulation results by a boxplot of different means against the dispersion estimates across 100 replicates.

Supplementary Note 5

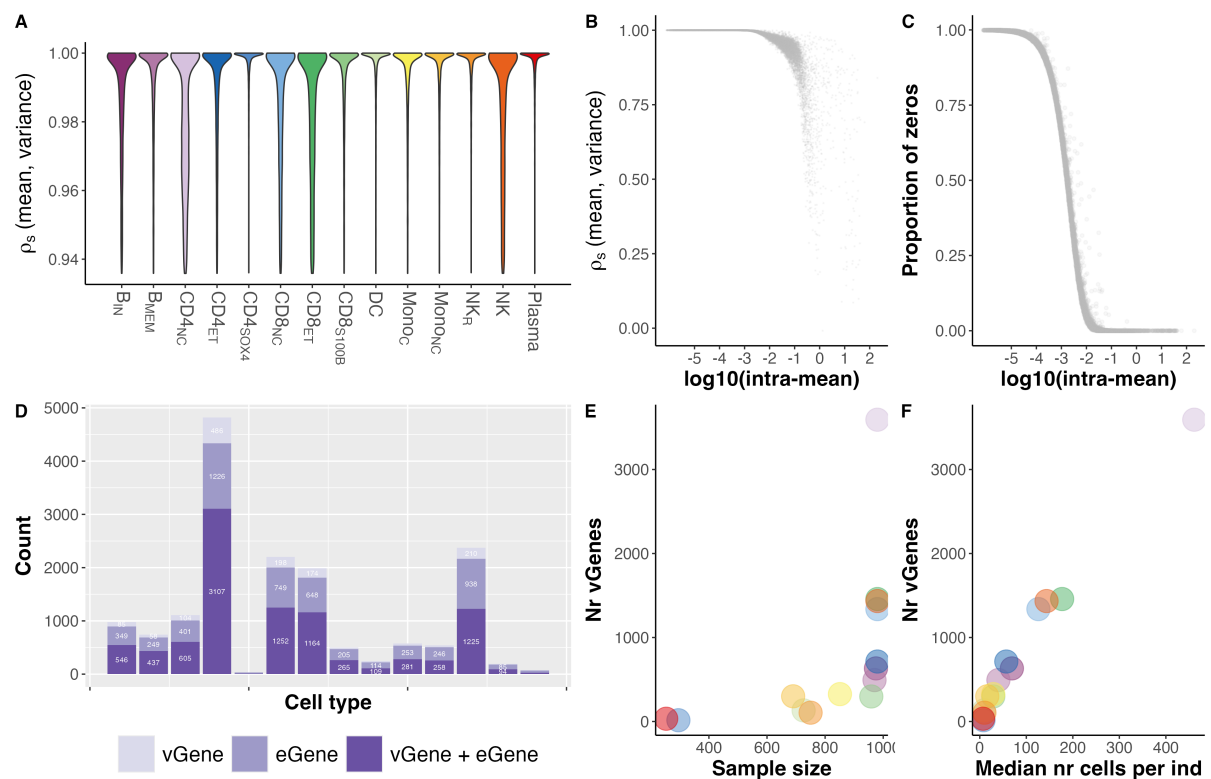
The special case of estimating dispersion when the mean is small

We have shown the moment estimators for mean and variance for NB distribution. It is easy to get moment estimate for dispersion $\theta = \frac{\sigma^2 - \mu}{\mu^2}$. The moment estimator for dispersion will become biased when mean is very small and/or mean and variance are very close. As we know that when $\theta \rightarrow 0$, the NB distribution will be approximately equivalent to Poisson distribution. On the other hand, when the mean is tiny, the difference between the mean and the variance will also be small, even if the dispersion is relatively large. This scenario is frequently observed when the mean is < 1 in empirical analysis.

In the OneK1K scRNA-seq data, we observe many genes with equal or very similar means and variances. After looking into the actual distribution, most cases are from low-mean genes. For example, among 24,903 genes in CD4_{NC} cells, 3,609 have equal mean and variance across all 980 individuals (including 539 with all zeros). The non-zero mean estimates for these genes range from $[2.94\text{E-}6, 1.62\text{E-}4]$, and the proportion of individuals with zero expression ranges from $[0.9286, 0.9990]$. For those 3,609 genes, the distribution of gene expression within an individual is a special case. Their counts are either all 0s or all 0s except one 1. These two cases will lead to equal means and variances. However, this does not necessarily mean that the original distribution is Poisson. Here is a very simple simulation to prove it.

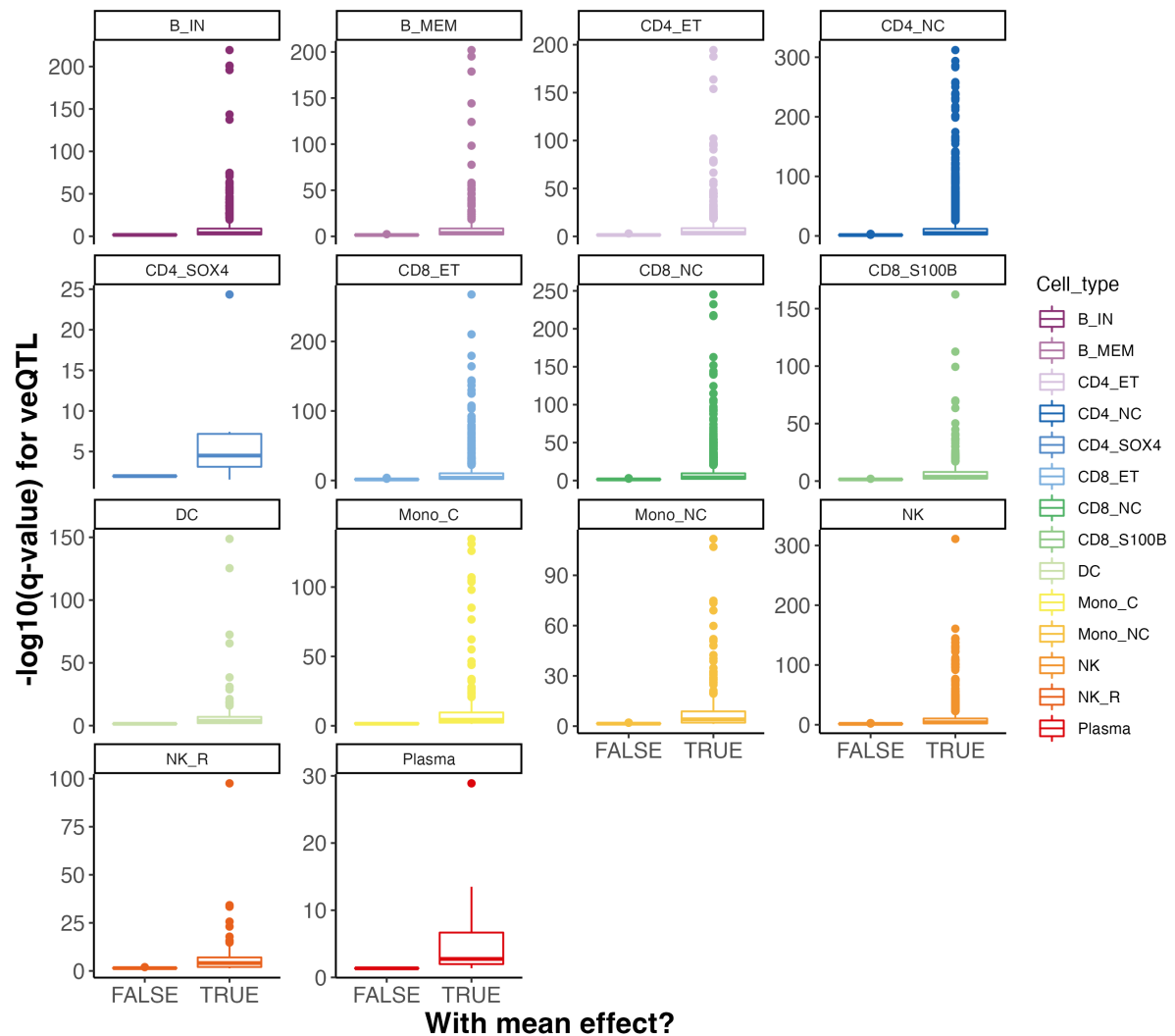
We simulate 10,000 individuals, each with 1,000 cells, and all the within-individual distributions are drawn from $NB \sim (\mu = 0.001, \theta = 1)$. The distribution patterns of these individuals are displayed in detail (**Supplementary Data 10**). We can see that the first four patterns account for ~98.2% of the distribution. Also, when there are only 0 and 1, the variance is no larger than the mean, so the moment estimate for theta is always 0 or negative. This can be easily proven as follows:

Let $Fano\ factor = \frac{\sigma^2}{\mu} = \frac{m}{m+k-1}$, where m is the number of 0s and k is an integer > 1 . This index is always between 0 and 1, so the variance is always less than the mean. Since the expected μ value is small, empirically $m \gg k$, the Fano factor will be very close to 1. Also, Let $\hat{\theta}_{MOE} = \frac{\sigma^2 - \mu}{\mu^2} = \frac{1-k}{k} \cdot \frac{m+k}{m+k-1}$. This index is always negative because of the first component. Also, since $m \gg k$ the second component is close to 1. The final estimate will be only slightly larger than $\frac{1-k}{k}$, and the lower boundary is $-\frac{1}{2}$.



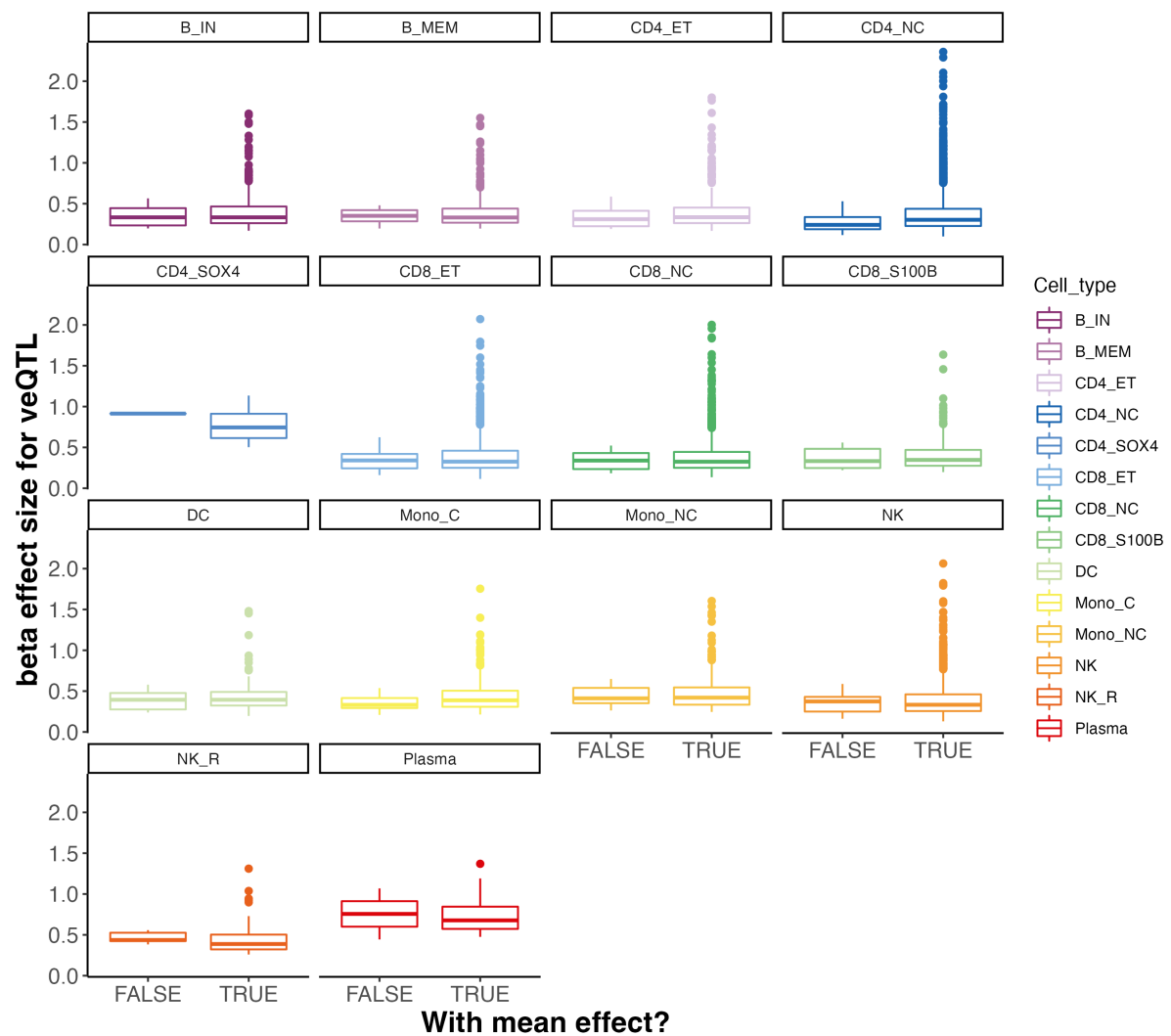
Supplementary Figure 1. The intra-individual mean-variance relationship and overlap between eGenes and vGenes per cell type.

(A) The Spearman's correlation estimates between the intra-individual mean and variance of each gene. The colour of the violin plot denotes the corresponding cell type. The bottom 10% correlation estimates are omitted. All violins have the same maximum width. **(B)** The relationship between intra-individual mean and mean-variance correlation per gene in CD4 naïve cells. **(C)** The relationship between intra-individual mean and proportion of no-expression individuals per gene in CD4 naïve cells. **(D)** The number of eGene, vGene, and vGene that is not identified as eGene in each cell type. **(E)** The relationship between sample size and number of vGenes. **(F)** The relationship between the median number of cells per individual and the number of vGenes. Each dot represents a cell type, and the colour of the dots corresponds to the panel.



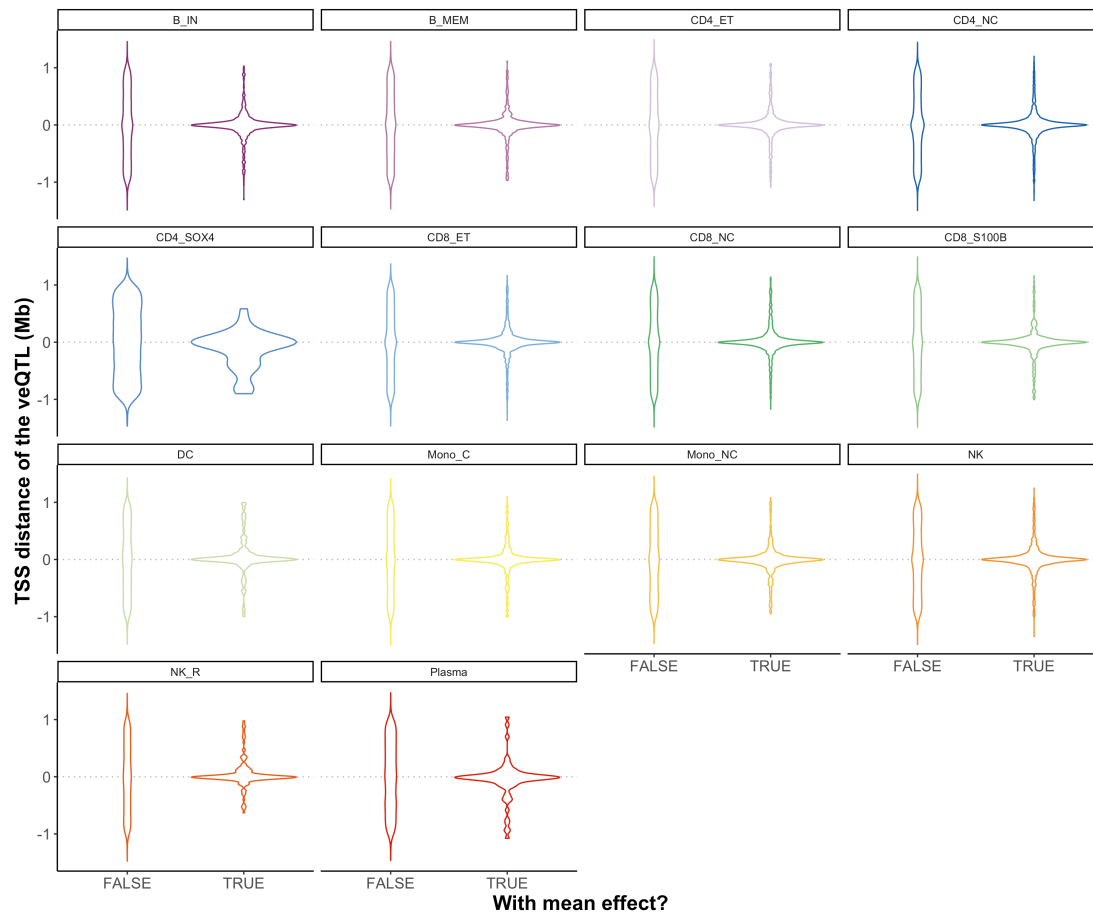
Supplementary Figure 2. Association significant level of top veQTL with or without mean effects.

The boxplot shows the distribution of the significance level of the top veQTL for each vGene, grouped by whether this vGene is also an eGene. TRUE (FALSE) group means the corresponding vGene is (not) also an eGene. The y-axis indicates the $-\log_{10}(\text{q-value})$ for each veQTL. Box plots show the median (centre line), interquartile range (box, 25th–75th percentiles), $1.5 \times \text{IQR}$ (whiskers), and outliers (individual points). The beta effect size is the slope estimate for the top SNP of the vGene in the veQTL association tests. The color of the boxplot corresponds to different cell type.



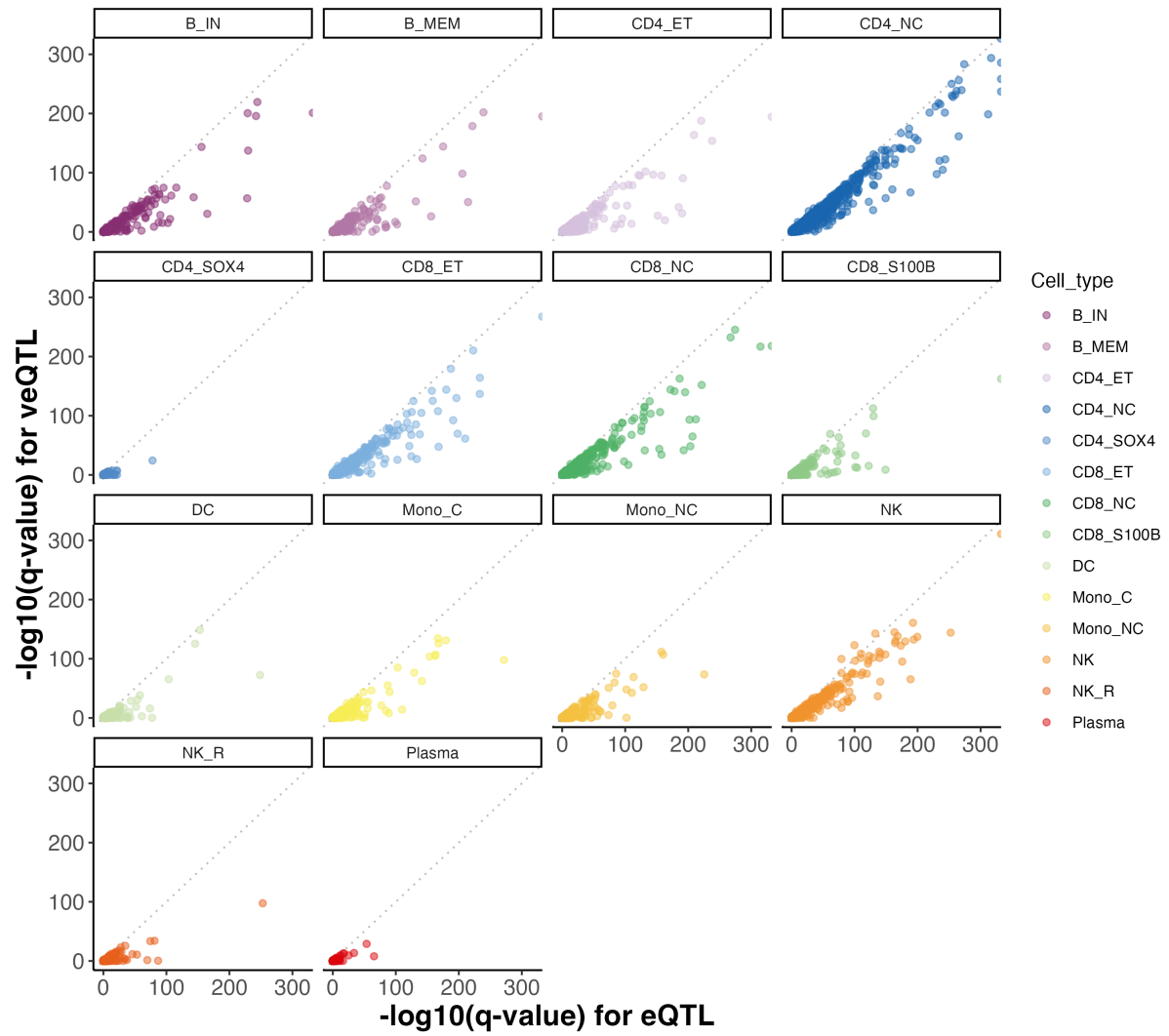
Supplementary Figure 3. Effect size of top veQTL with or without mean effects.

The boxplot shows the distribution of beta effect size of the top veQTL for each vGene, grouped by whether this vGene is also an eGene. Box plots show the median (centre line), interquartile range (box, 25th–75th percentiles), 1.5× IQR (whiskers), and outliers (individual points). The color of the boxplot corresponds to different cell type.



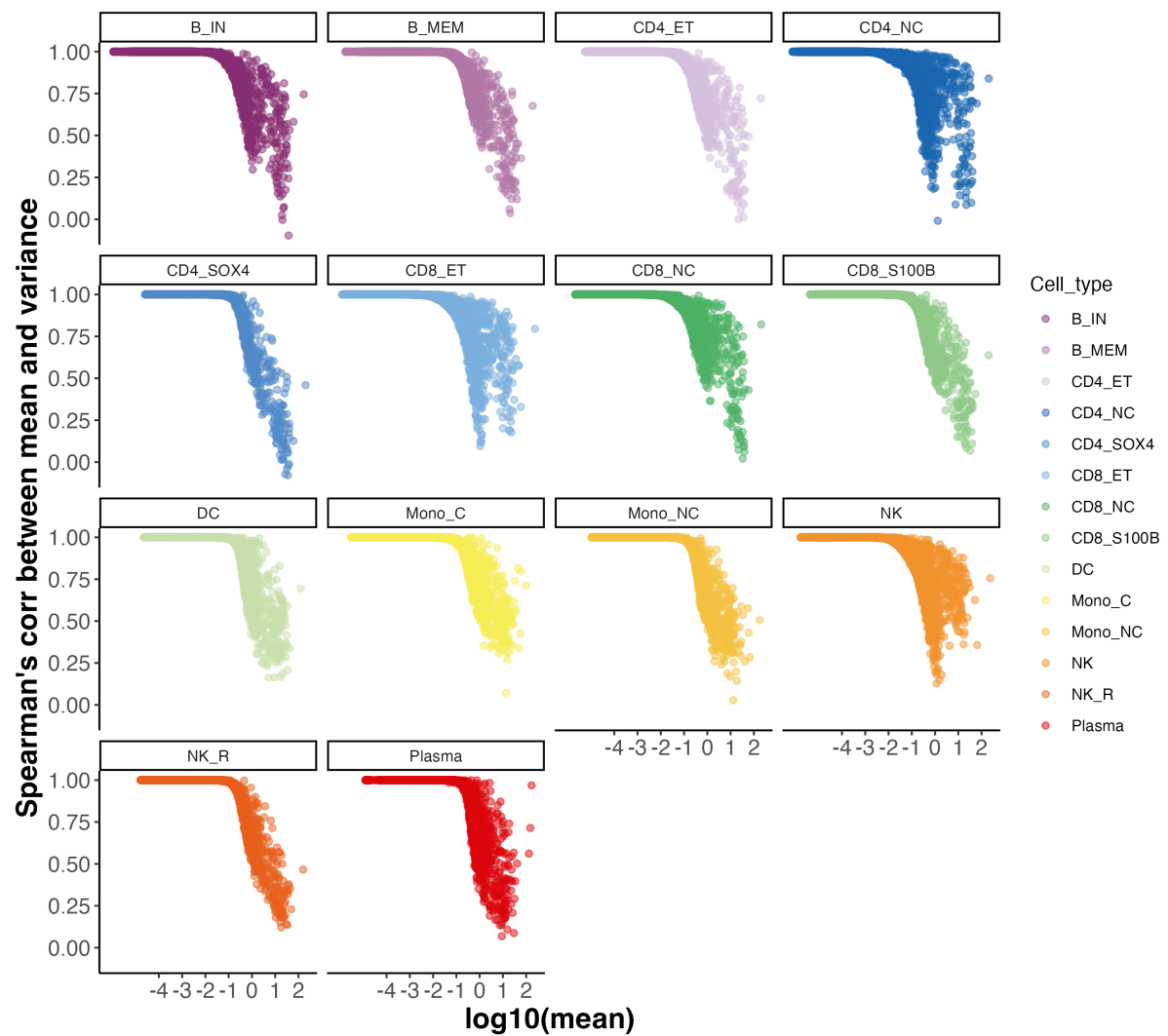
Supplementary Figure 4. TSS distance top veQTL with or without mean effects.

This violin plot compares the distance between the top veQTL to the transcription start site for those vGenes with or without mean effects. The x-axis indicates the two groups of veQTLs. The y-axis indicates the distance by Mb. The color of the boxplot corresponds to different cell type.



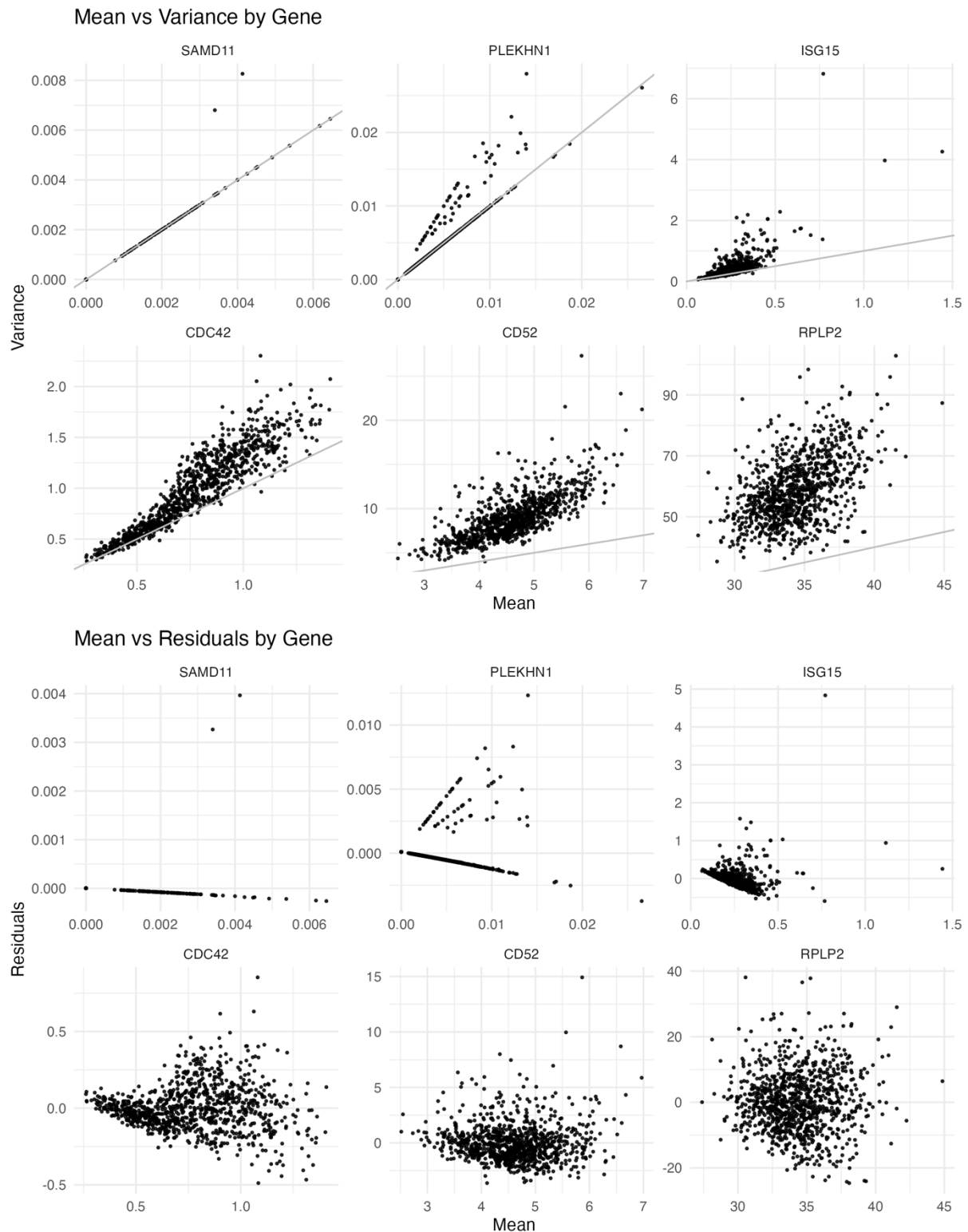
Supplementary Figure 5. Comparison of association significant level between eQTL and veQTL.

The x-axis and y-axis denote the $-\log_{10}(q\text{-value})$ for eQTL and veQTL, respectively. Each dot indicates a SNP-gene pair test, and the colour of the dot indicates the cell type. The grey dashed line denotes the diagonal line of the coordinate panel.



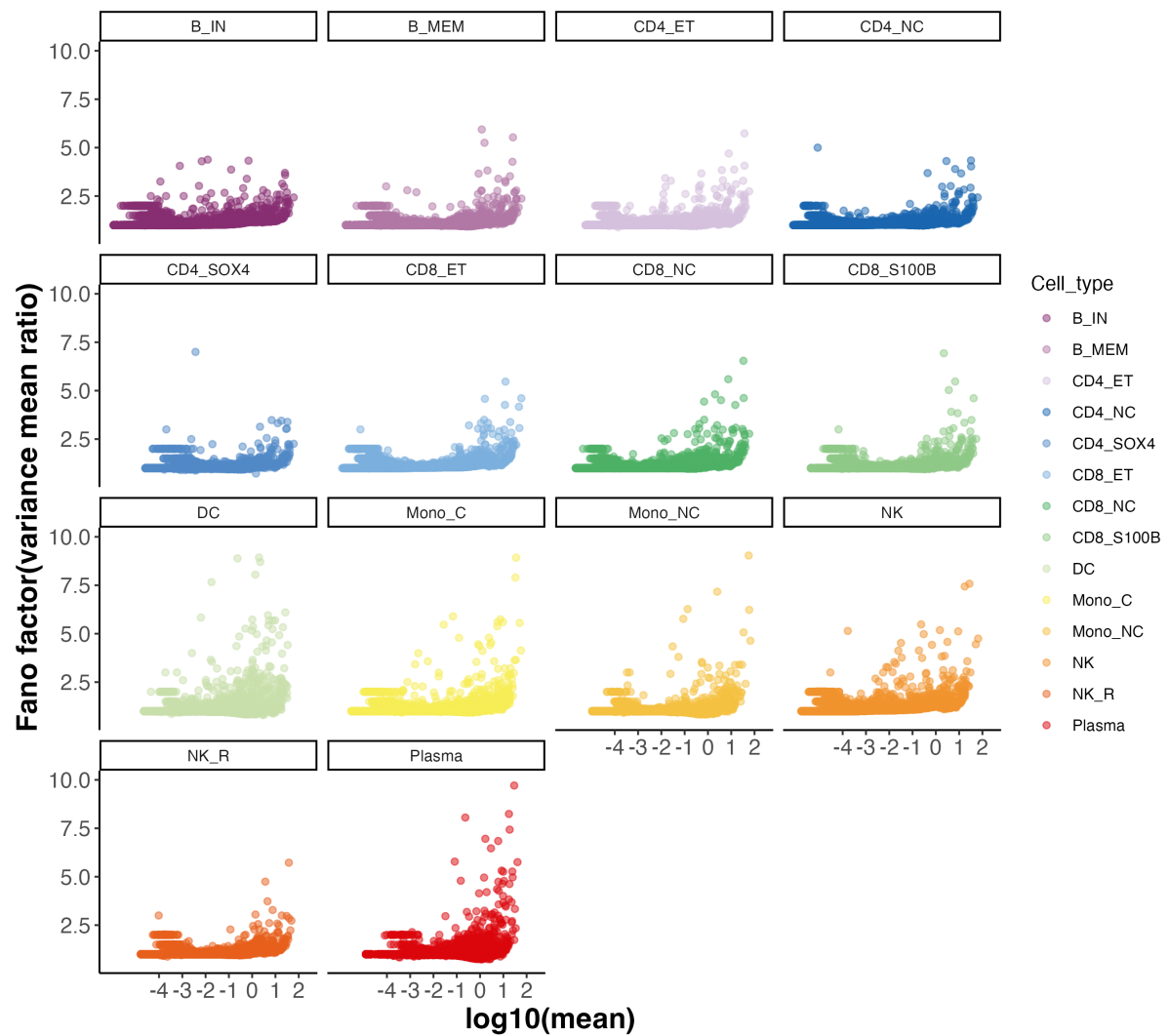
Supplementary Figure 6. The relationship between the intra-individual mean and the mean-variance correlation.

The scatter plot shows the relationship between the intra-individual mean and the Spearman's correlation efficient between intra-individual mean and variance across individuals per gene in 14 cell types. Each dot indicates one gene and there are 24,903 genes in total.



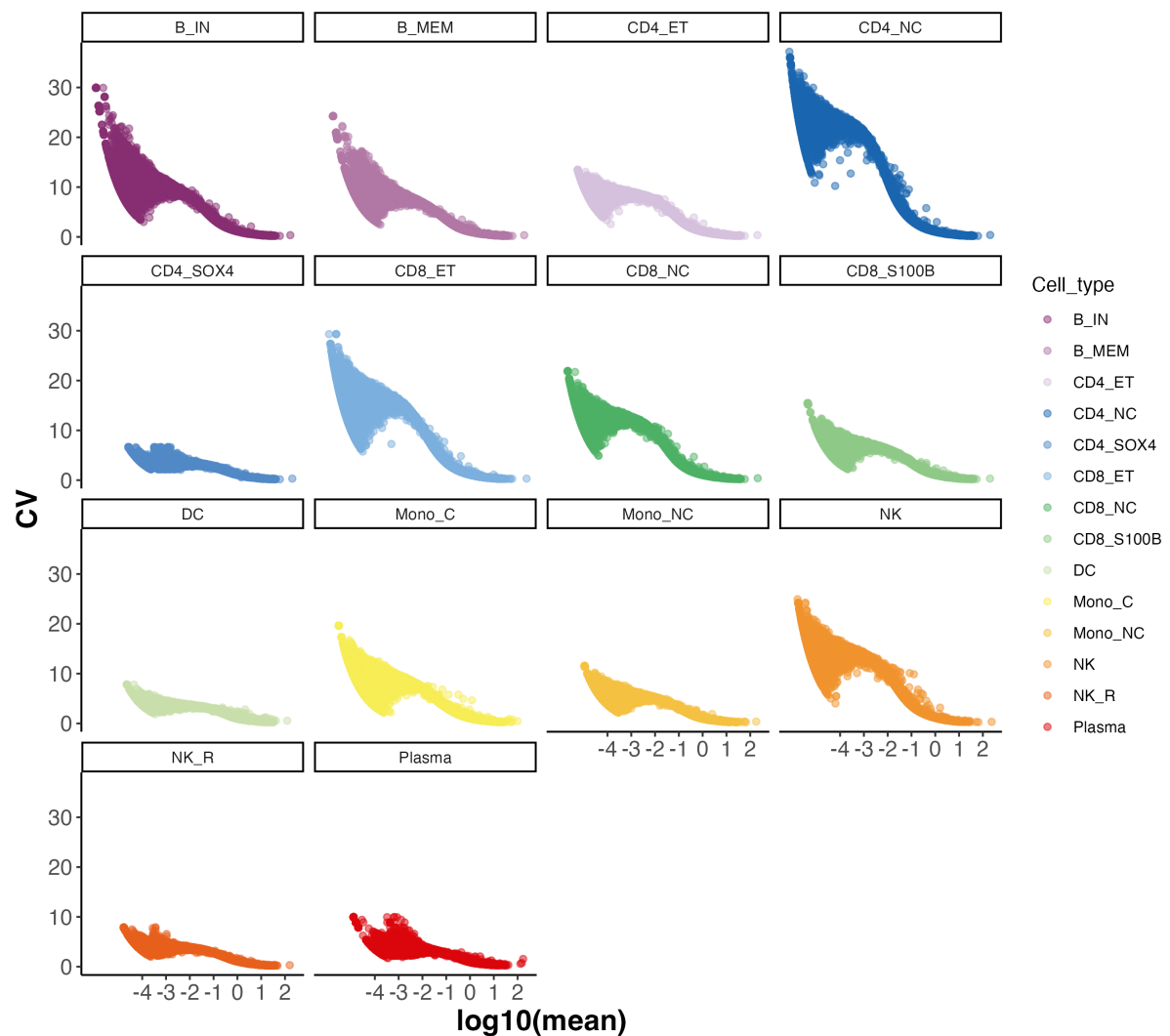
Supplementary Figure 7. The relationship between the intra- mean, variance, and residuals for six example genes.

The scatter plot shows the relationship between the intra-individual mean and variance (upper panel), and between intra-individual mean and residuals from variance regressing out the mean (lower panel). The six genes are selected based on different mean levels in CD4_{NC} cell. Each dot indicates one individual's estimate (N = 980). The majority of genes (19,237/24,903) fit the pattern of *SAMD11* or *PLEKHN1*, rather than the other four genes



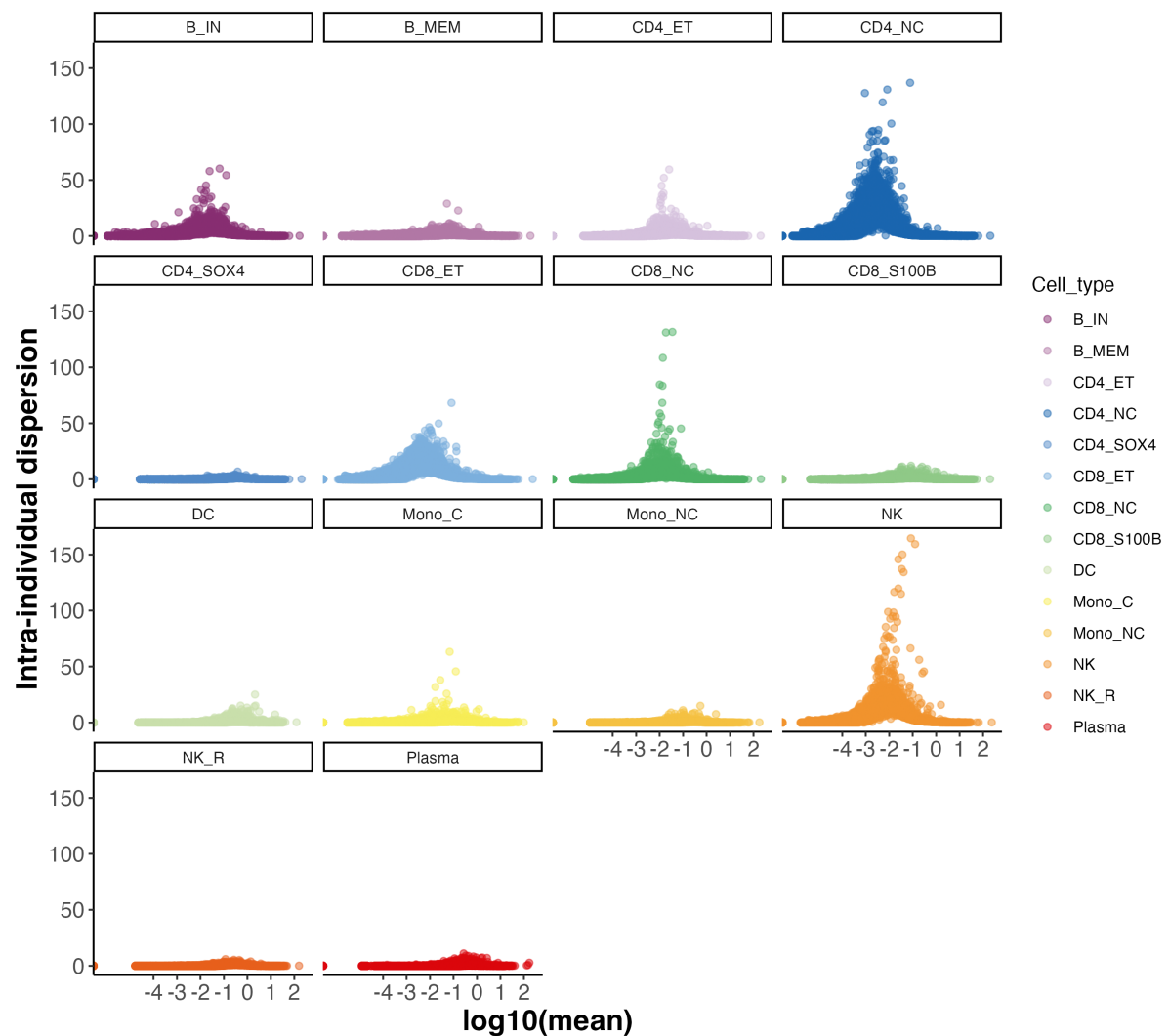
Supplementary Figure 8. The relationship between the mean and the Fano factor (variance mean ratio).

The scatter plot shows the relationship between the intra-individual mean and the proportion of non-expressed individuals for each gene across 14 cell types. The x-axis denotes the $\log_{10}$ transformed mean per gene. The y-axis denotes the Fano factor, which is calculated as variance divided by mean. Each dot represents the mean Fano factor for a gene across individuals and the color corresponds to different cell type. 46 outliers that with Fano factor > 10 were omitted. Genes with all-zero expression do not have a Fano factor.



Supplementary Figure 9. The relationship between the mean and the coefficient of variation (CV).

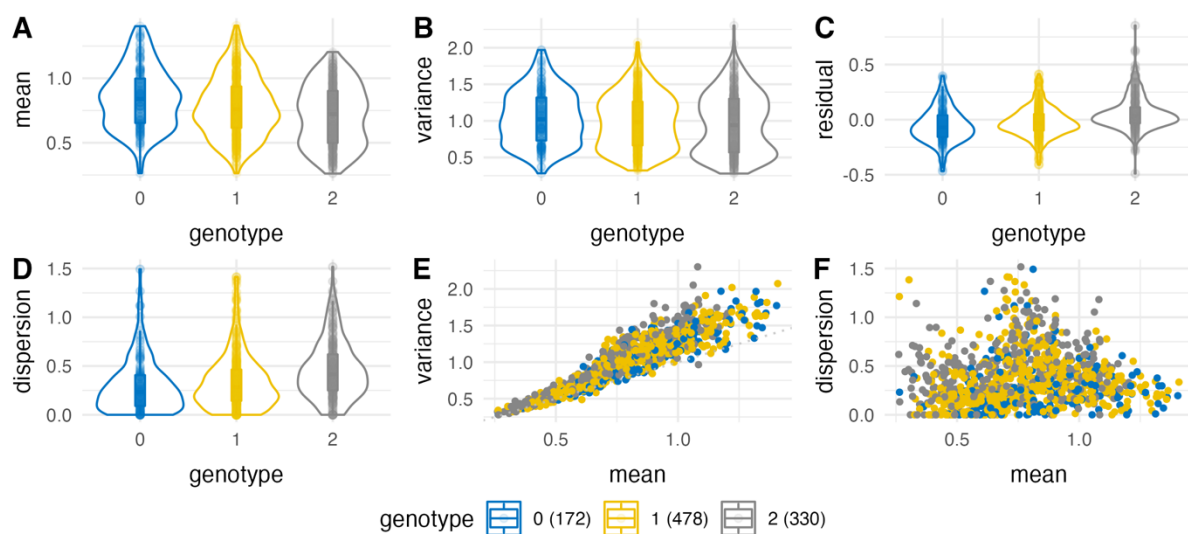
The scatter plot shows the relationship between the intra-individual mean and the proportion of non-expressed individuals for each gene across 14 cell types. The x-axis denotes the log10 transformed mean per gene. The y-axis denotes the CV (coefficient of variation), calculated as the standard deviation divided by the mean for each gene. The number of dots (genes) is corresponding to column “Nr genes” in Table 1. Genes with all-zero expression do not have a CV estimate.



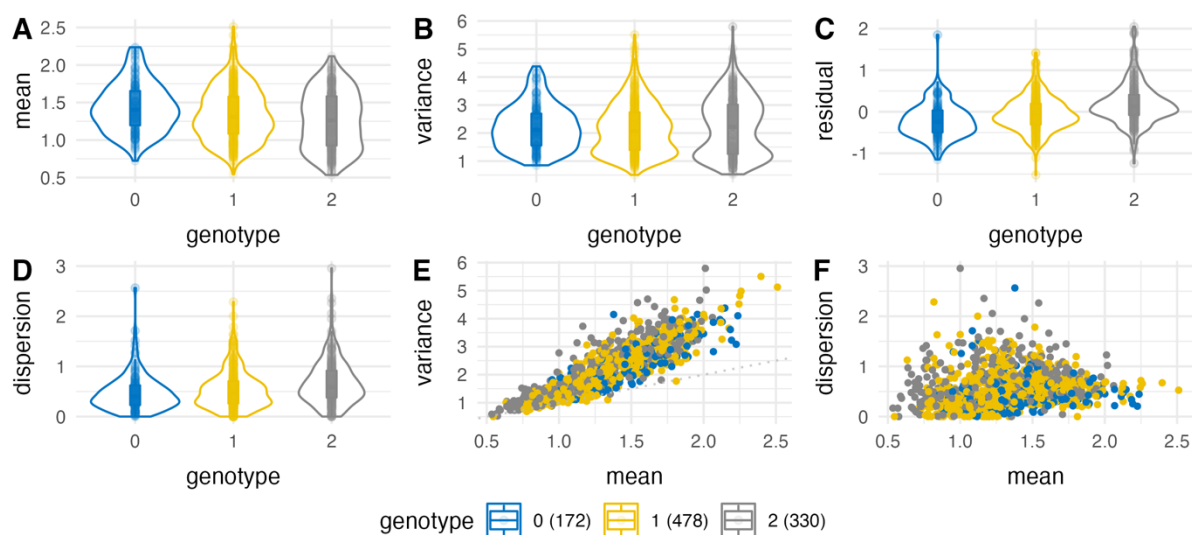
Supplementary Figure 10. The relationship between the mean and the dispersion.

The scatter plot shows the relationship between the intra-individual mean and the intra-individual dispersion for each gene across 14 cell types. The x-axis denotes the $\log_{10}$ transformed mean per gene. The y-axis denotes the intra-individual estimates of dispersion per gene, calculated using the Cox-Reid adjusted MLE method. The number of dots (genes) is corresponding to column "Nr genes" in Table 1.

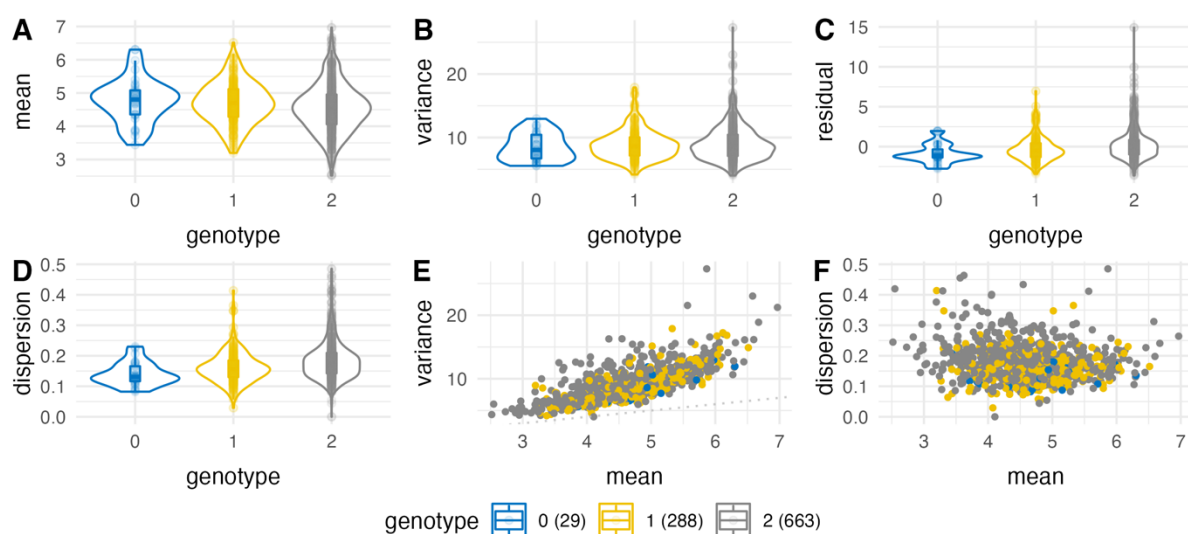
1:22426187_C_A (rs1534949) for *CDC42* in CD4_{NC} cells



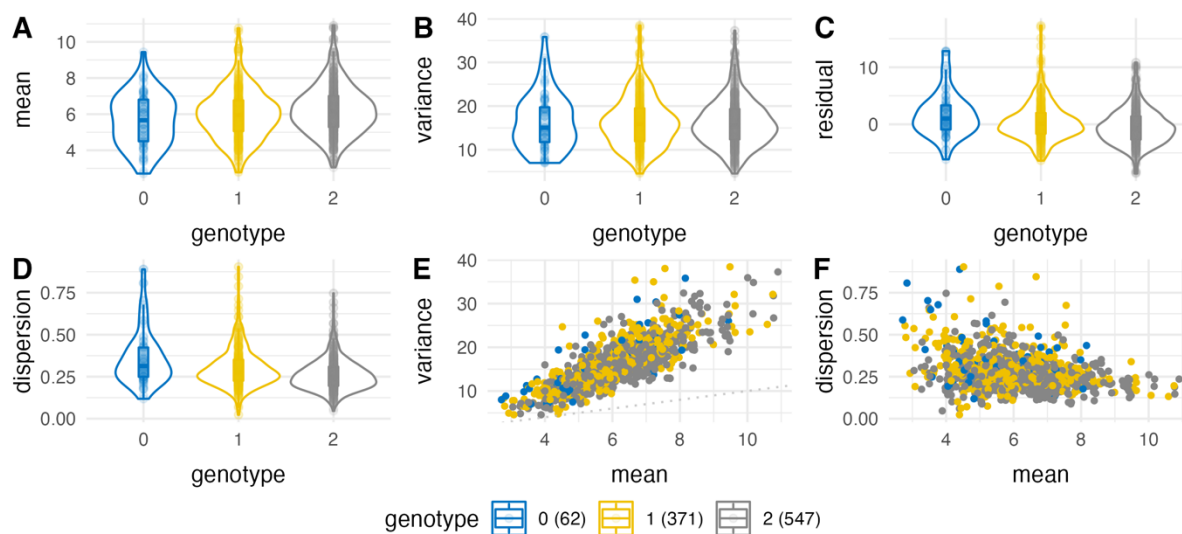
1:22426187_C_A (rs1534949) for *CDC42* in NK cells



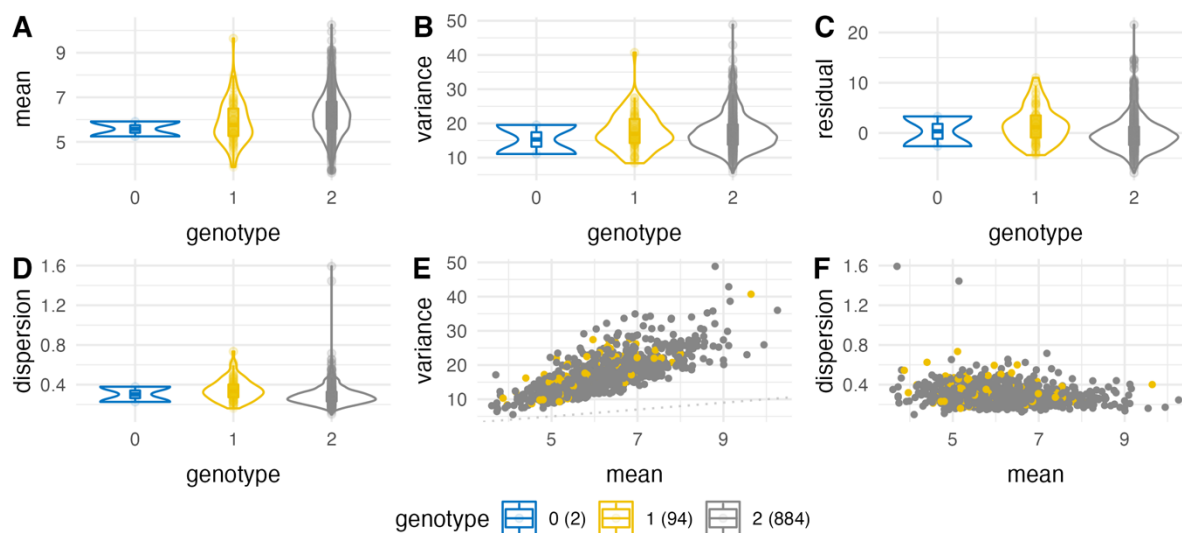
1:26616280_A_G (rs3924324) for *CD52* in CD4_{NC} cells



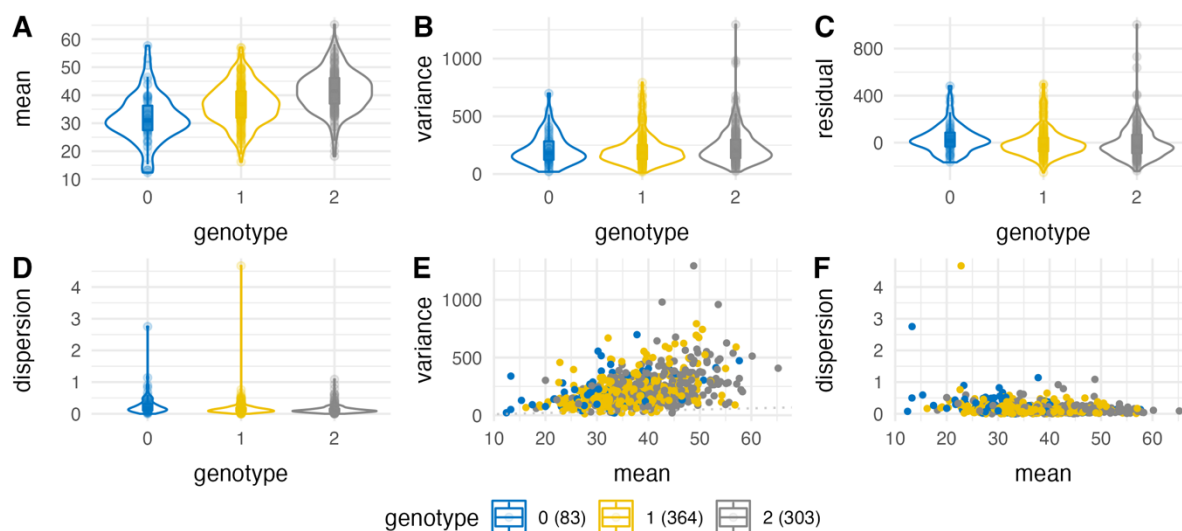
1:26645806_C_T (rs11589222) for *CD52* in CD8_{ET} cells



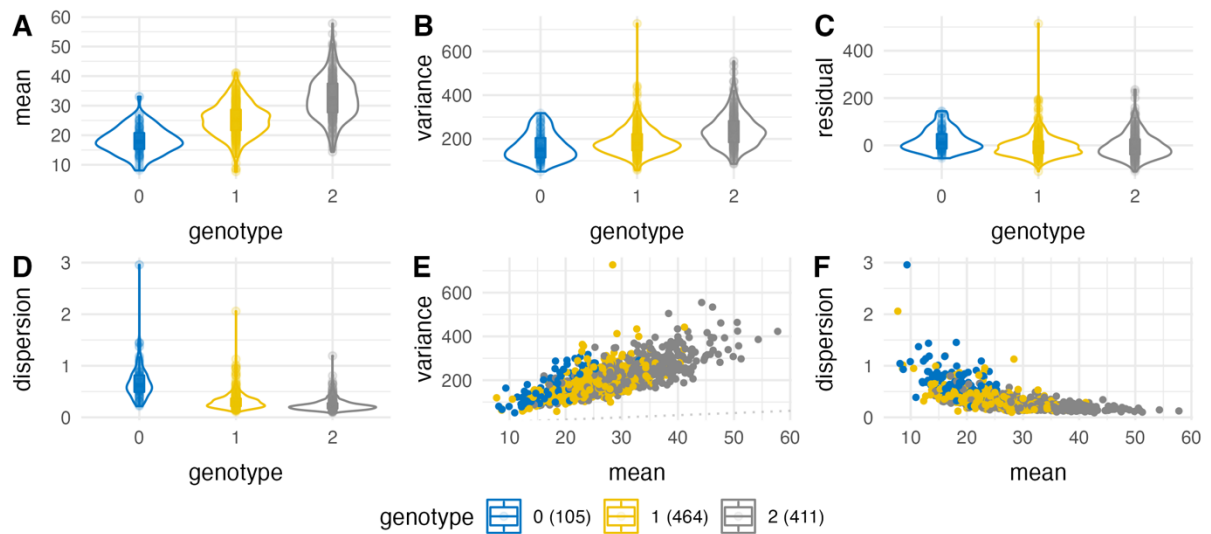
1:153514241_C_A (rs7535476) for *S100A4* in NK cells



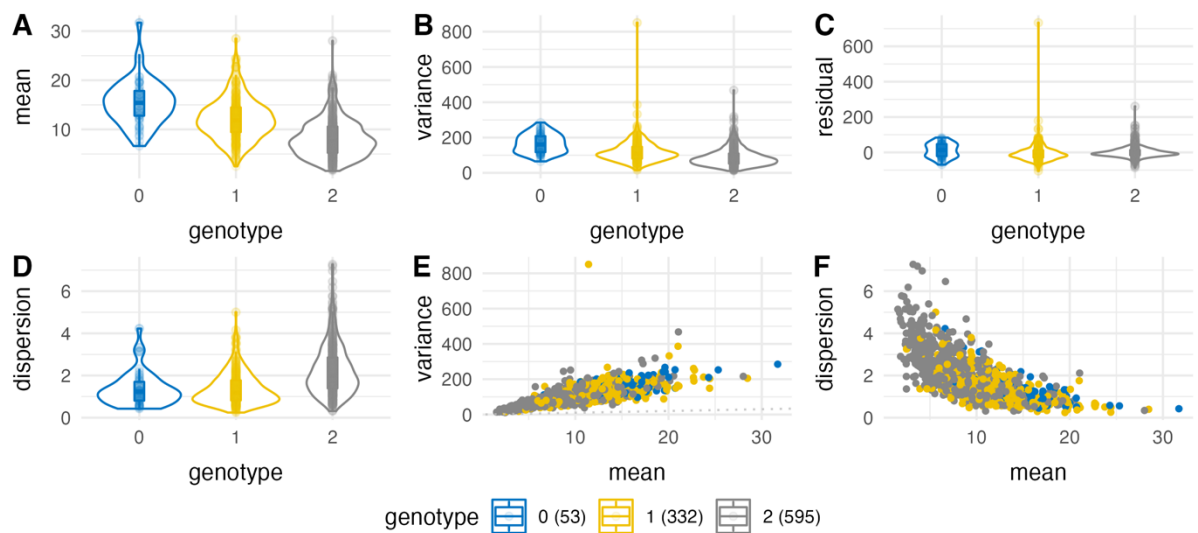
2:85916005_G_A (rs4832181) for *GNLY* in NK_R cells



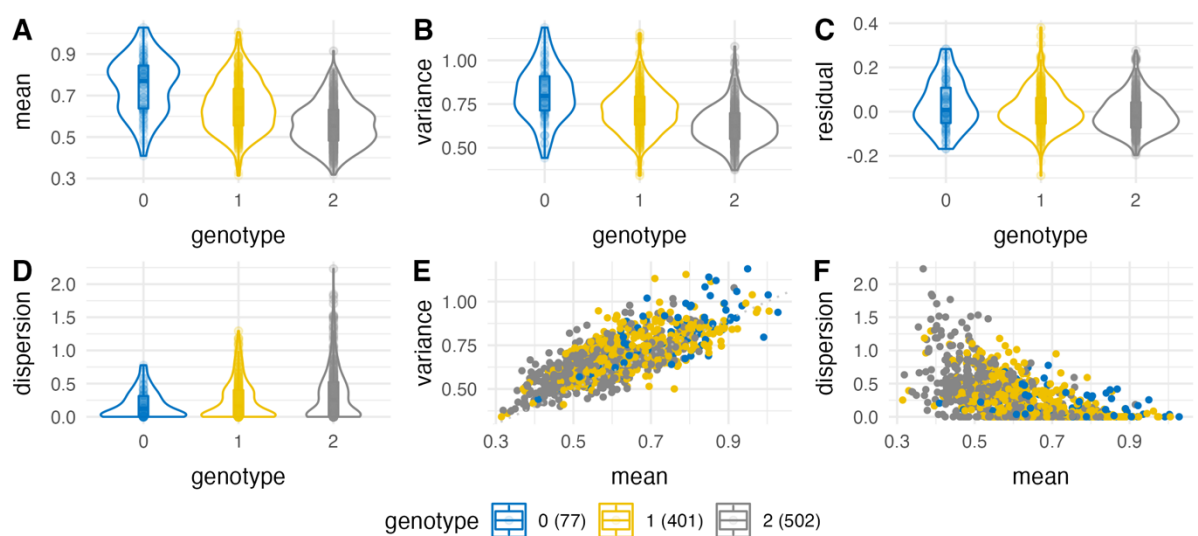
2:85920249_G_T (rs3755007) for *GNLY* in NK cells



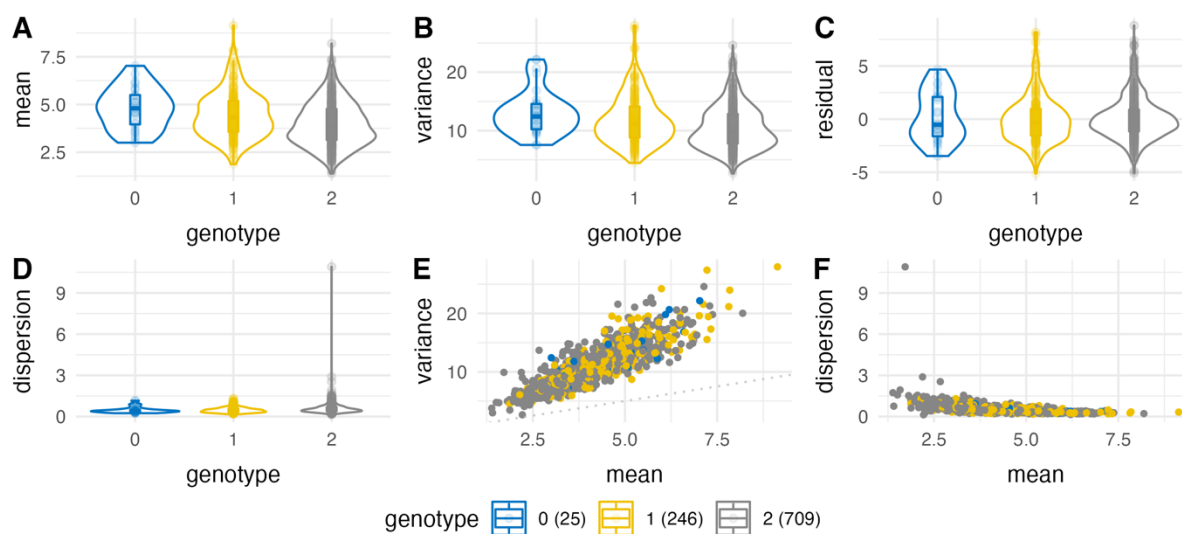
2:85934499_A_C (rs12151621) for *GNLY* in CD8_{ET} cells



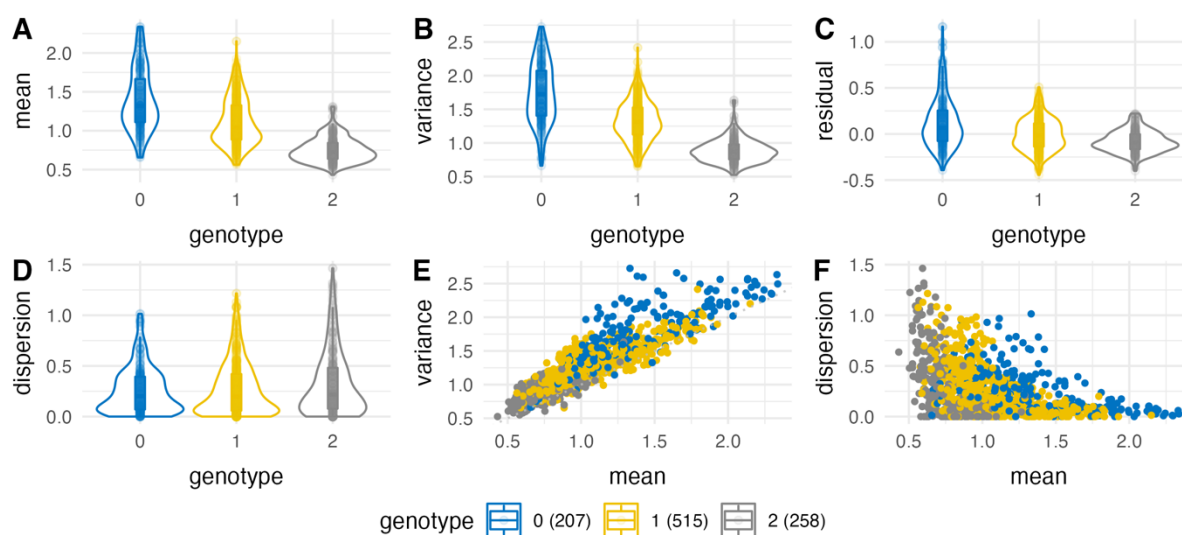
2:127516475_G_C (rs6732878) for *GYPC* in CD4_{NC} cells



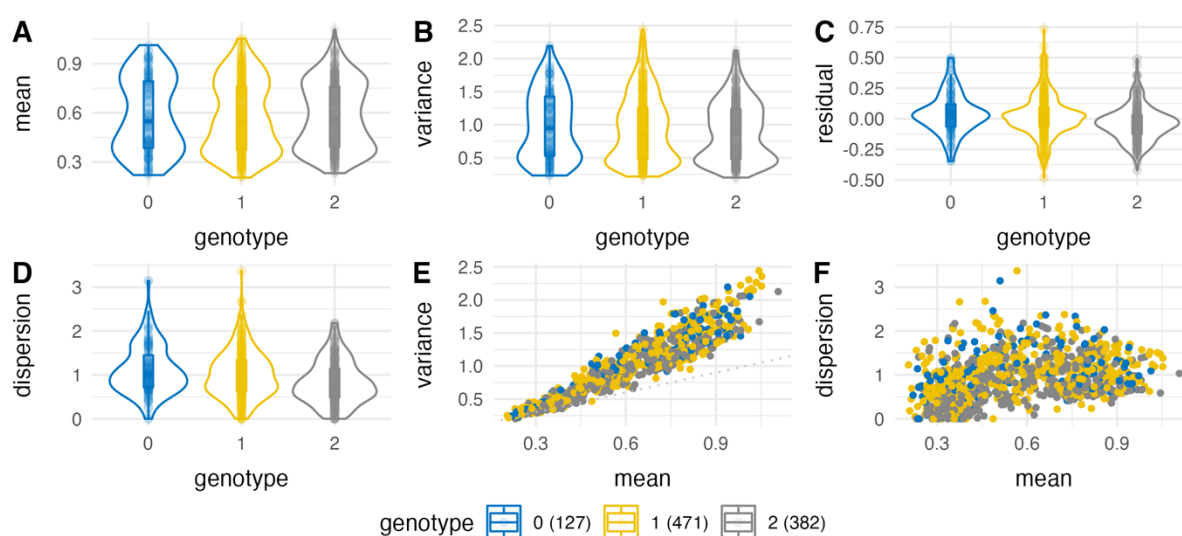
4:15957763_G_A (rs4698429) for *FGFBP2* in NK cells



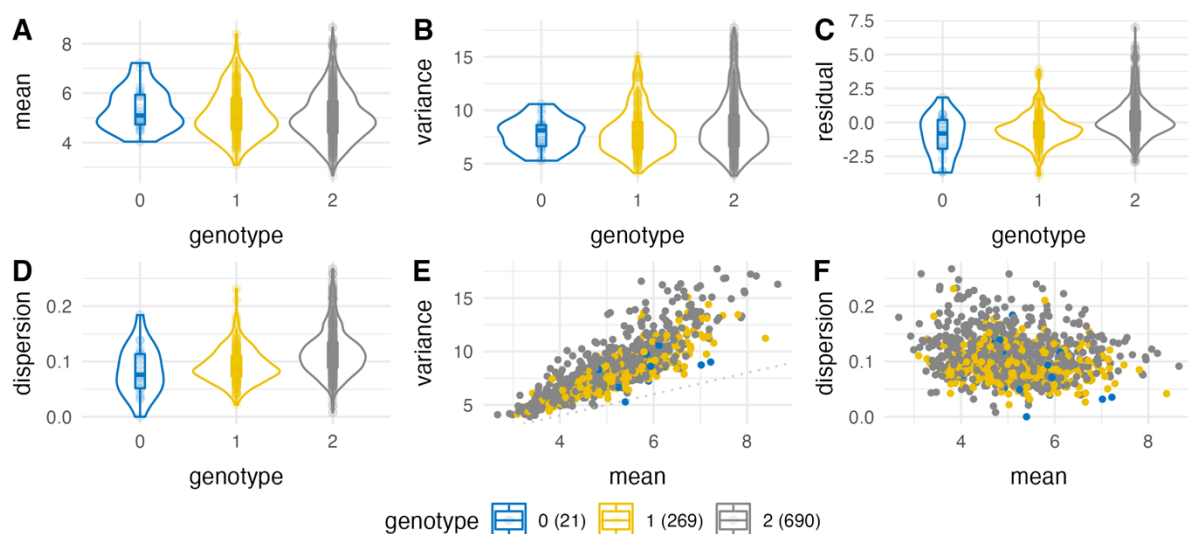
4:119204466_T_C (rs28517808) for *SNHG8* in CD4_{NC} cells



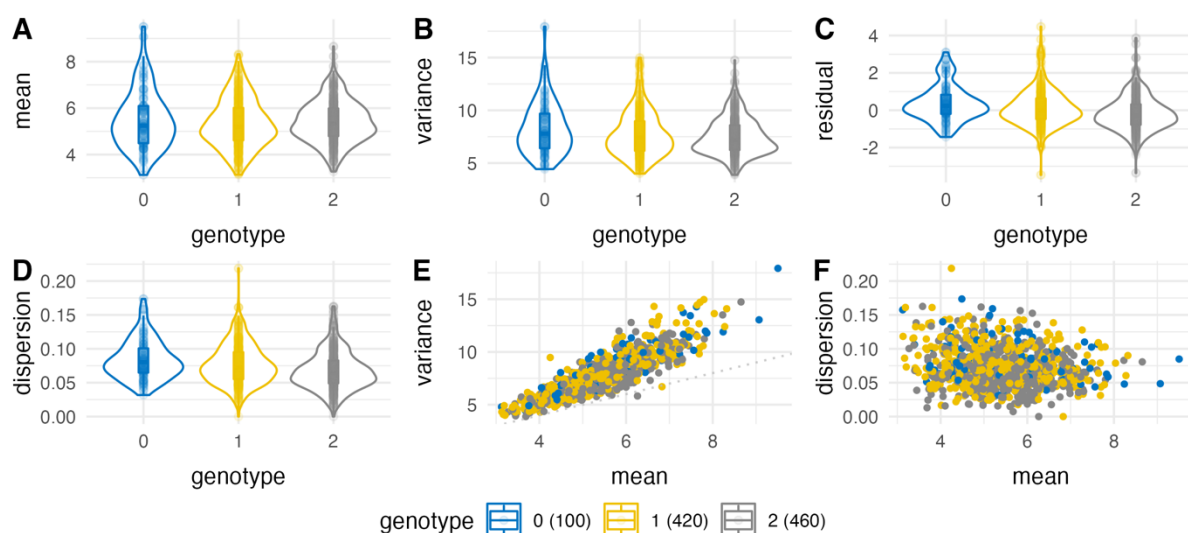
5:178986632_C_A (rs7703730) for *HNRNPH1* in CD4_{NC} cells



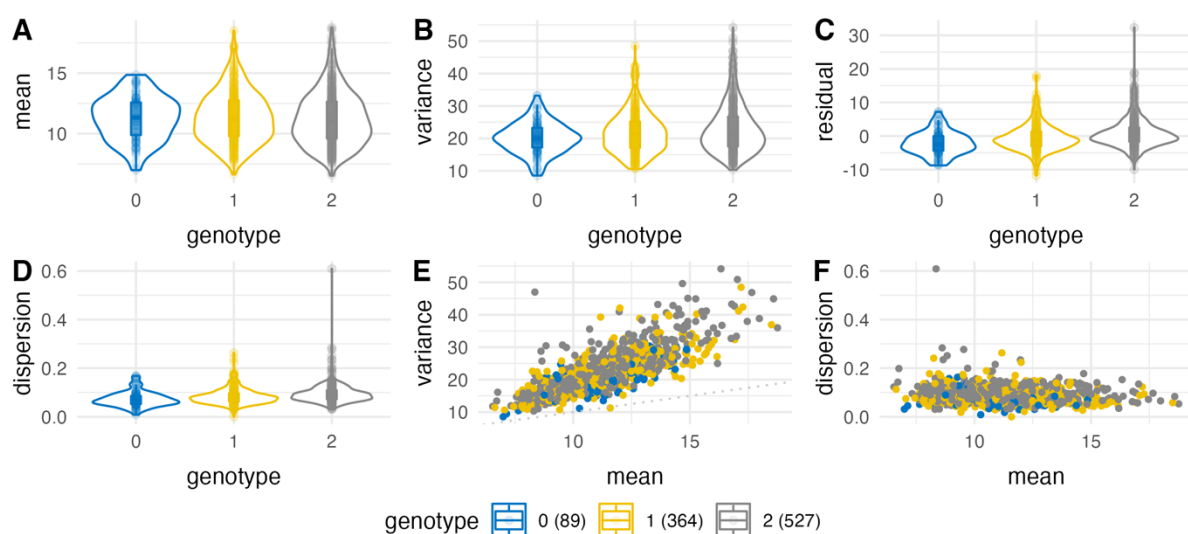
6:29913266_T_G (rs1061156) for *HLA-A* in CD4_{NC} cells



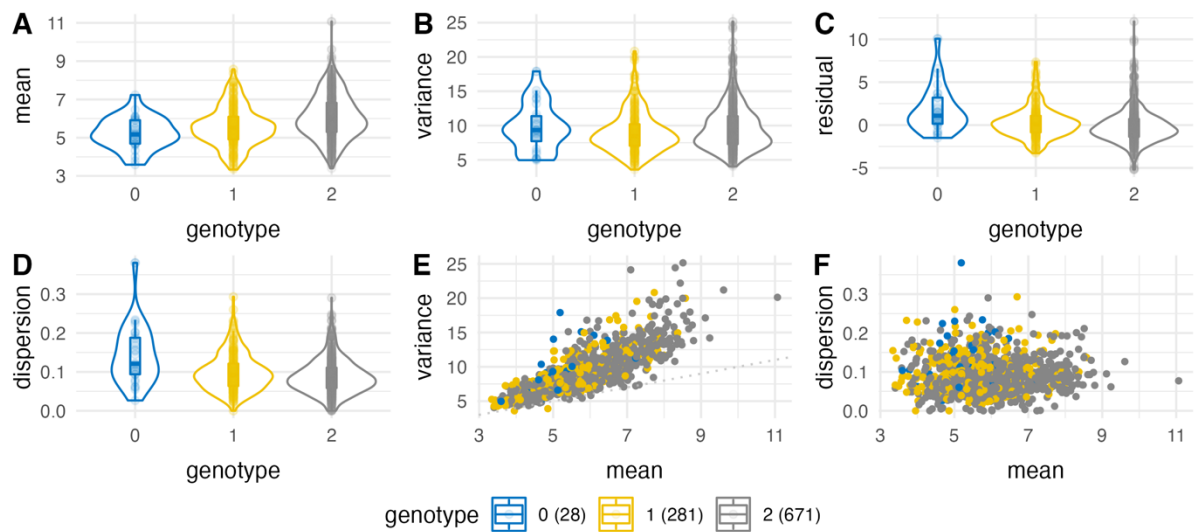
6:31221914_C_T (rs9264219) for *HLA-C* in CD4_{NC} cells



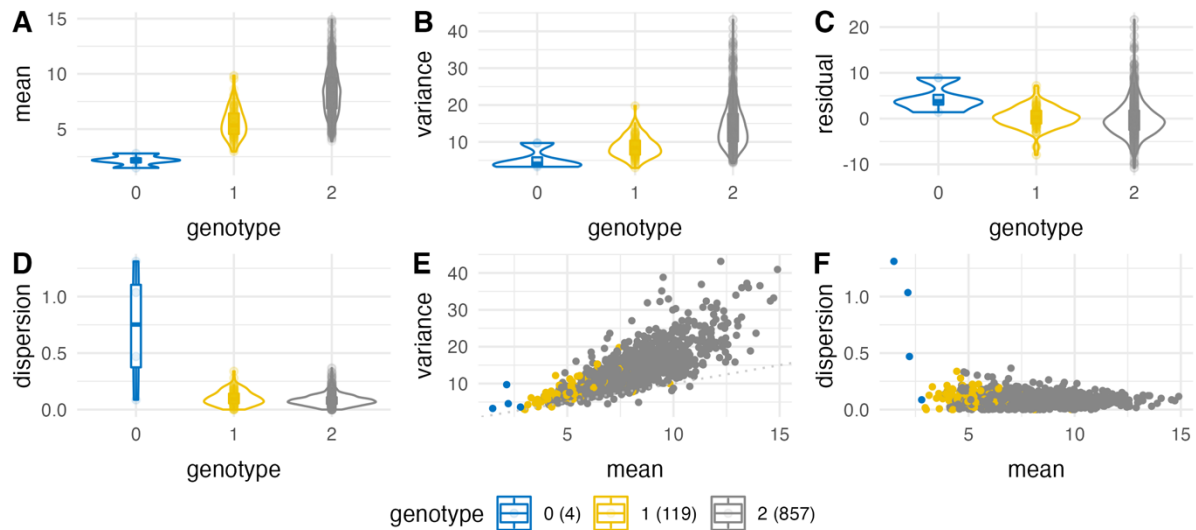
6:31263051_G_A (rs2853926) for *HLA-C* in CD8_{ET} cells



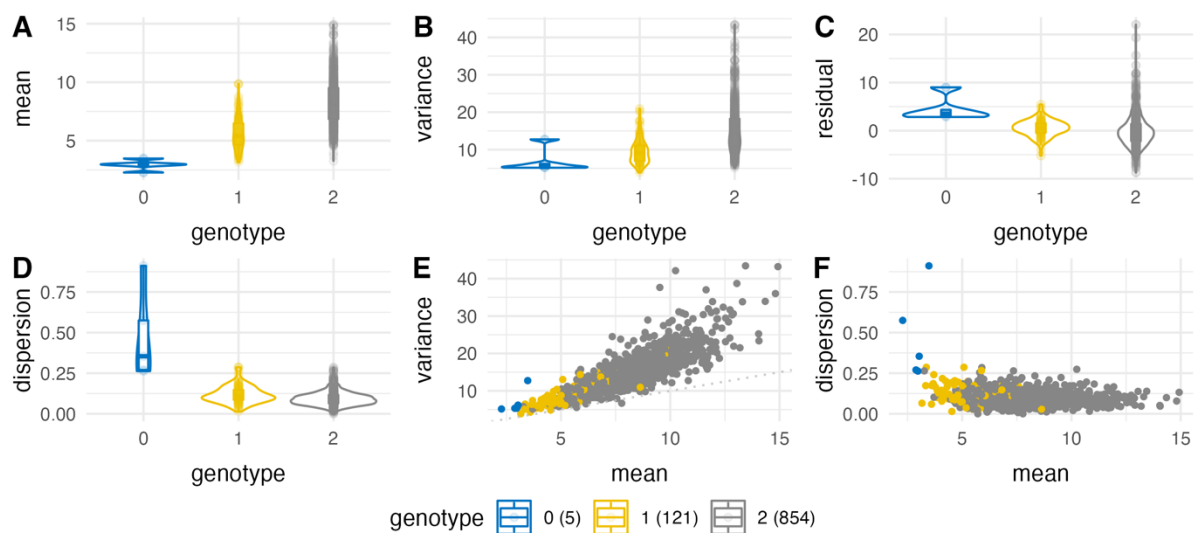
6:31321360_A_G (rs2844585) for *HLA-C* in CD8_{NC} cells



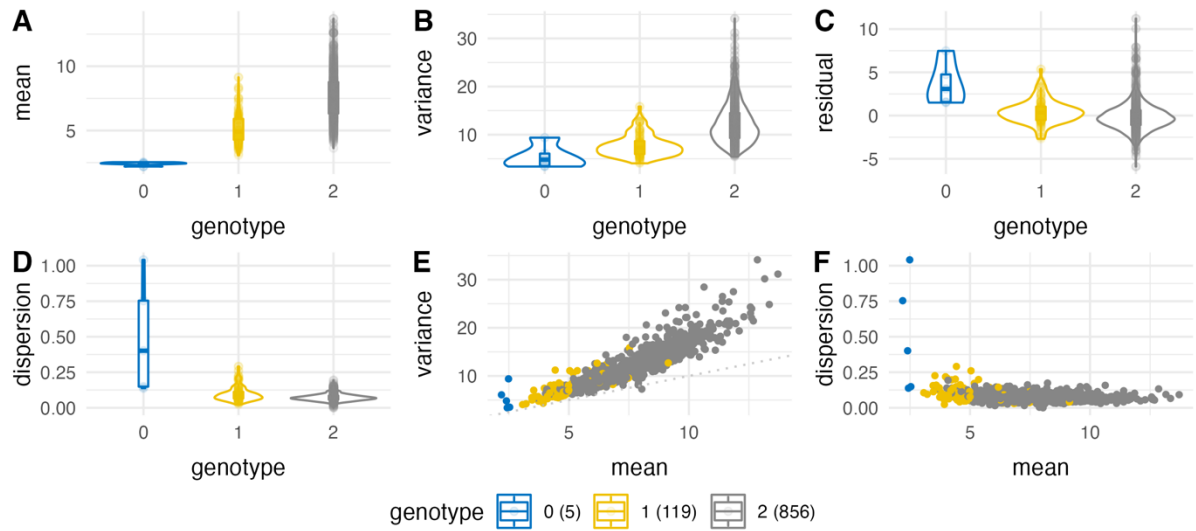
6:31324955_T_C (rs34437781) for *HLA-B* in CD4_{ET} cells



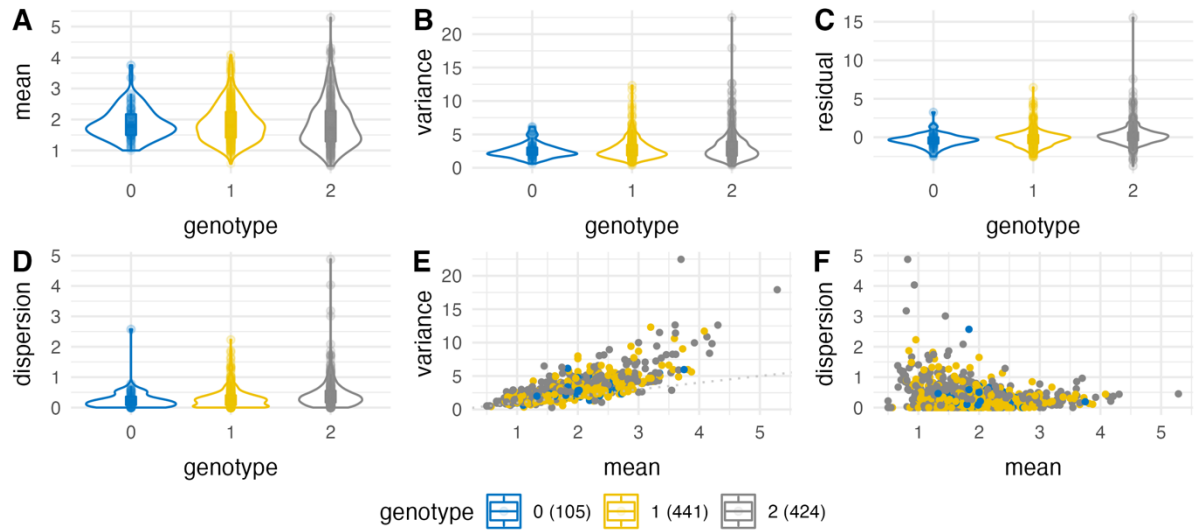
6:31327701_G_T (rs9378249) for *HLA-B* in CD8_{NC} cells



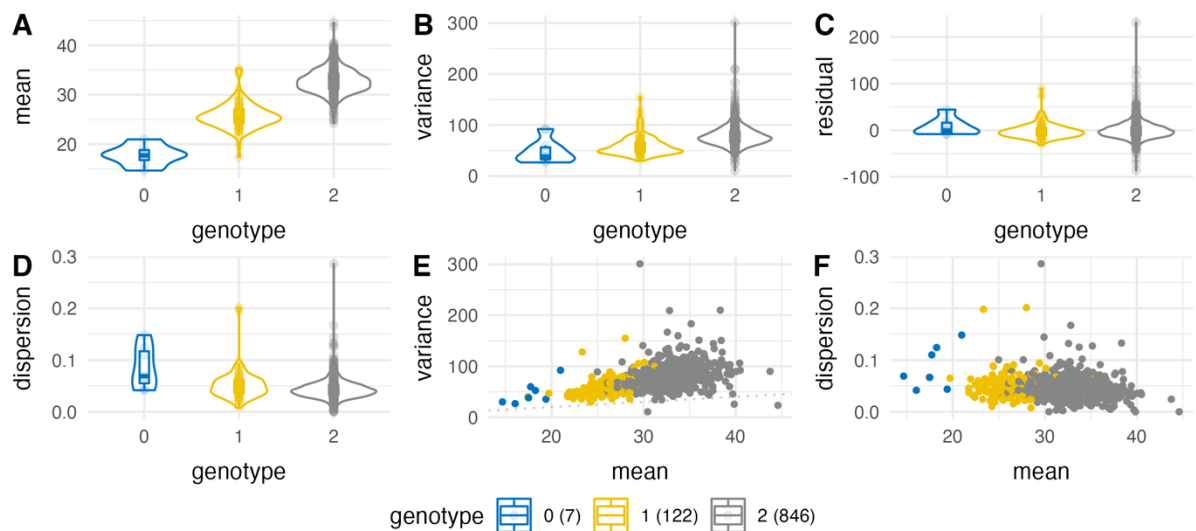
6:31366295_A_C (rs9394070) for *HLA-B* in CD4_{NC} cells



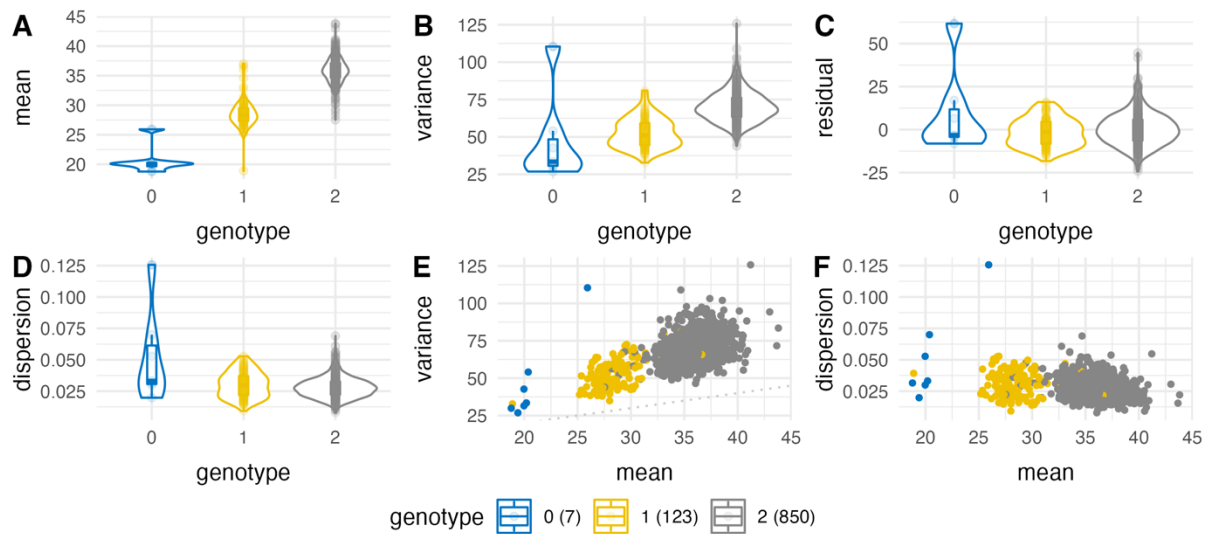
6:32589326_A_C (rs9271503) for *HLA-DQA1* in B_{MEM} cells



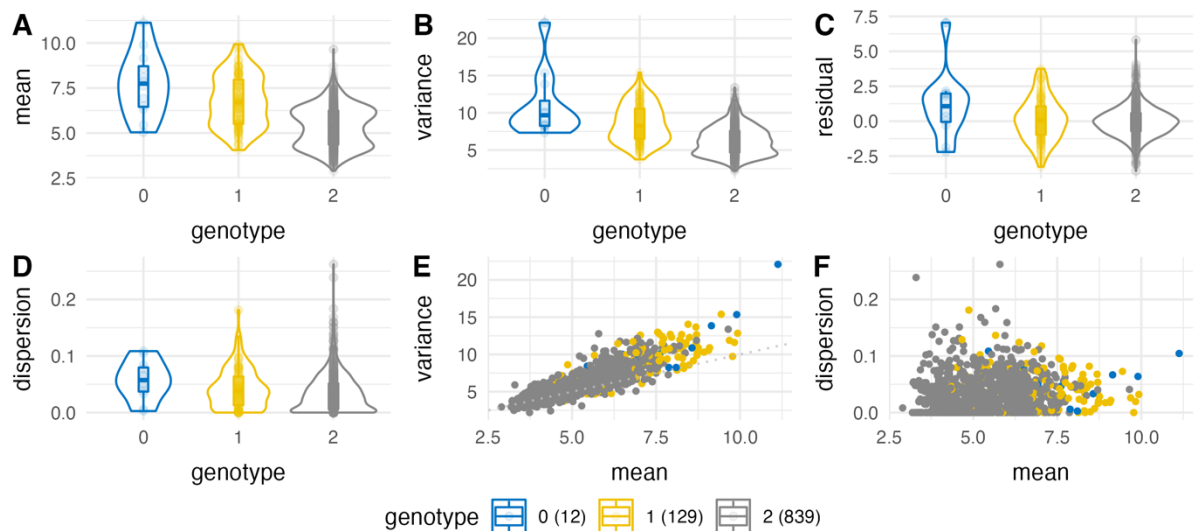
6:33239869_T_C (rs17215231) for *RPS18* in B_{IN} cells



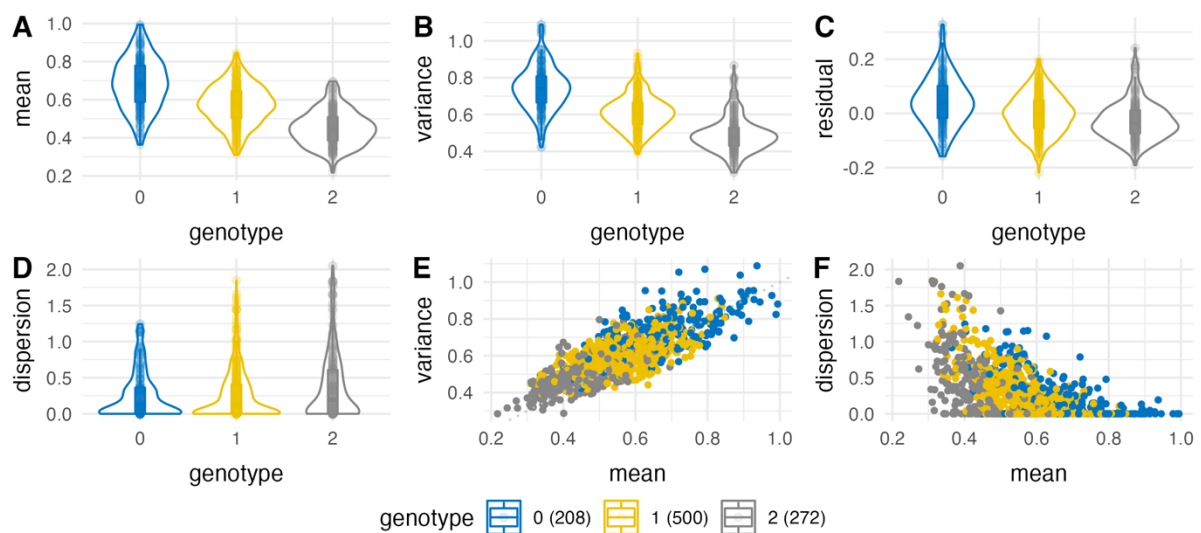
6:33239869_T_C (rs17215231) for *RPS18* in CD4_{NC} cells



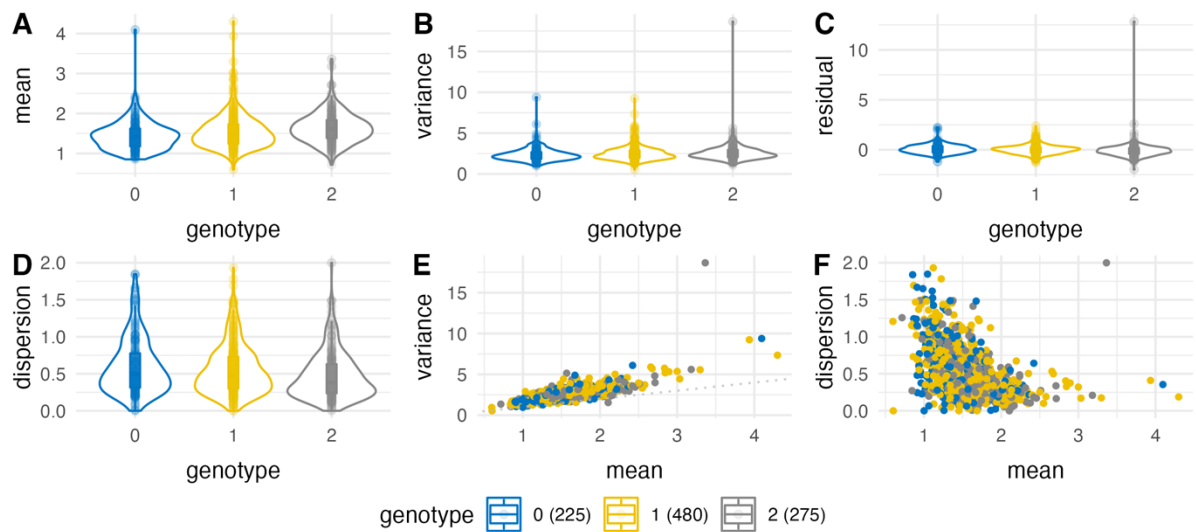
6:34372804_G_A (rs775635) for *RPS10* in CD8_{NC} cells



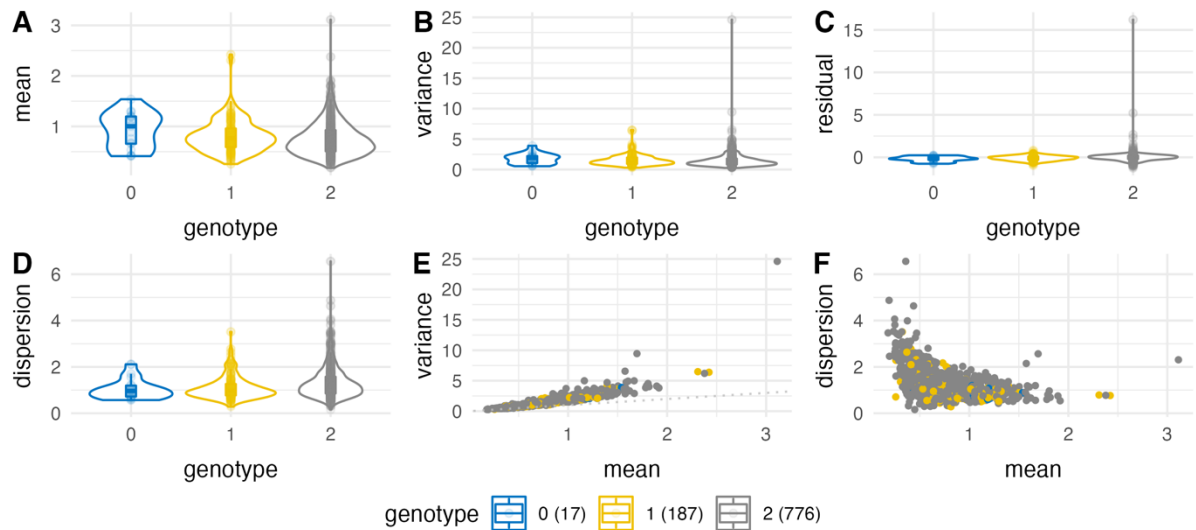
6:167370999_G_A (rs2769346) for *RNASET2* in CD4_{NC} cells



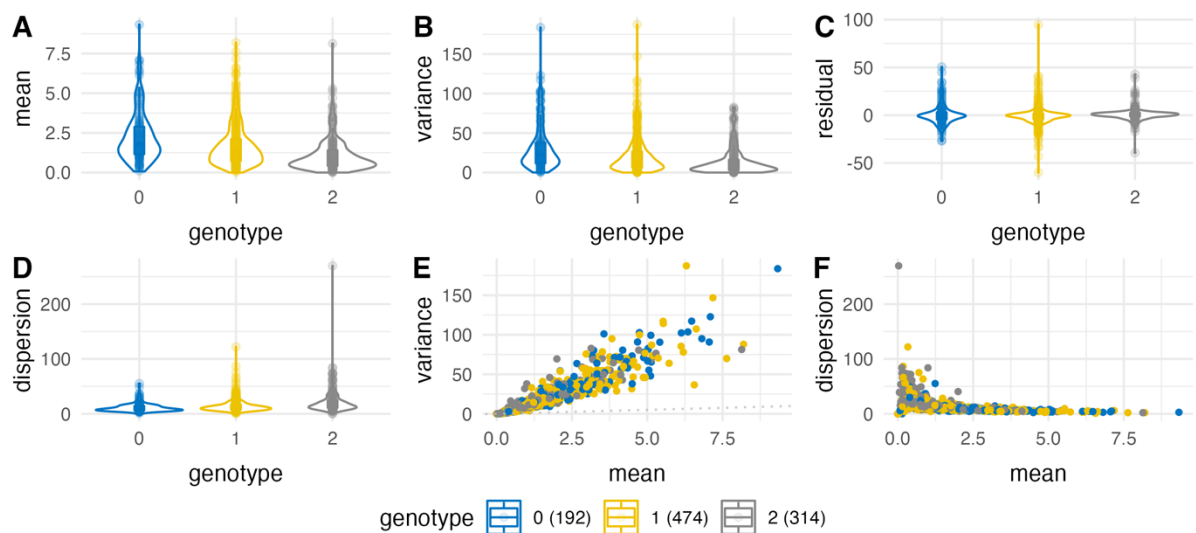
8:144075281_T_C (rs4424237) for *LY6E* in NK cells



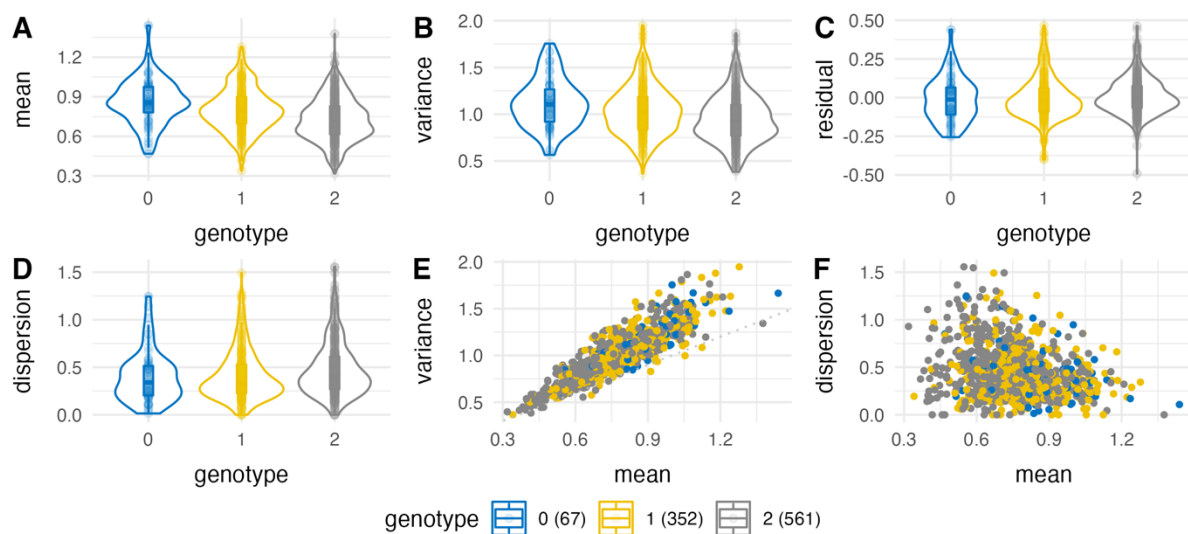
9:75769950_G_C (rs2795112) for *ANXA1* in CD4_{NC} cells



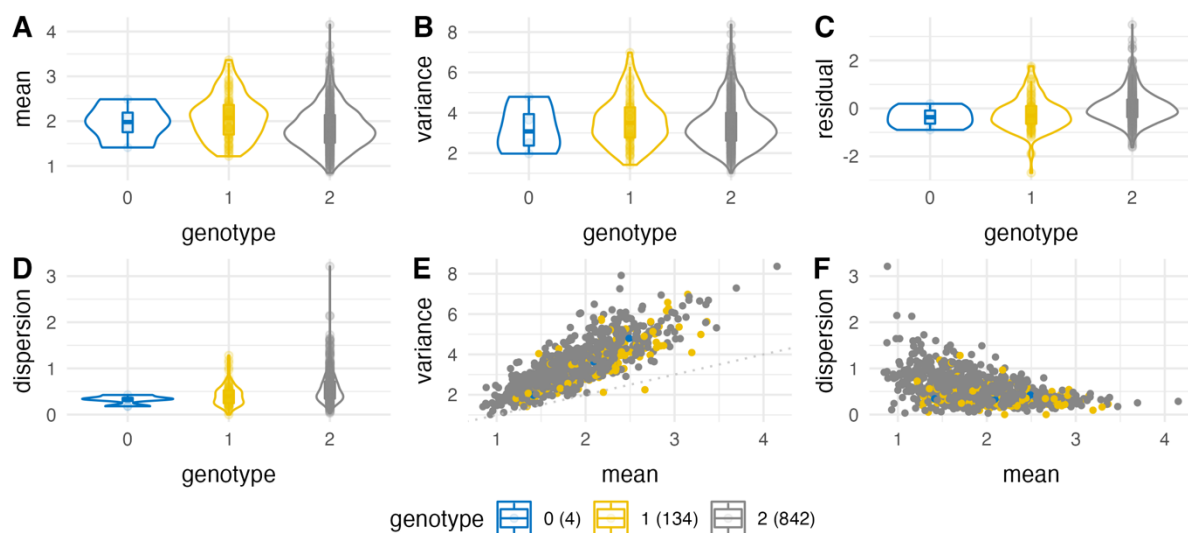
9:139848273_G_A (rs2271869) for *PTGDS* in NK cells



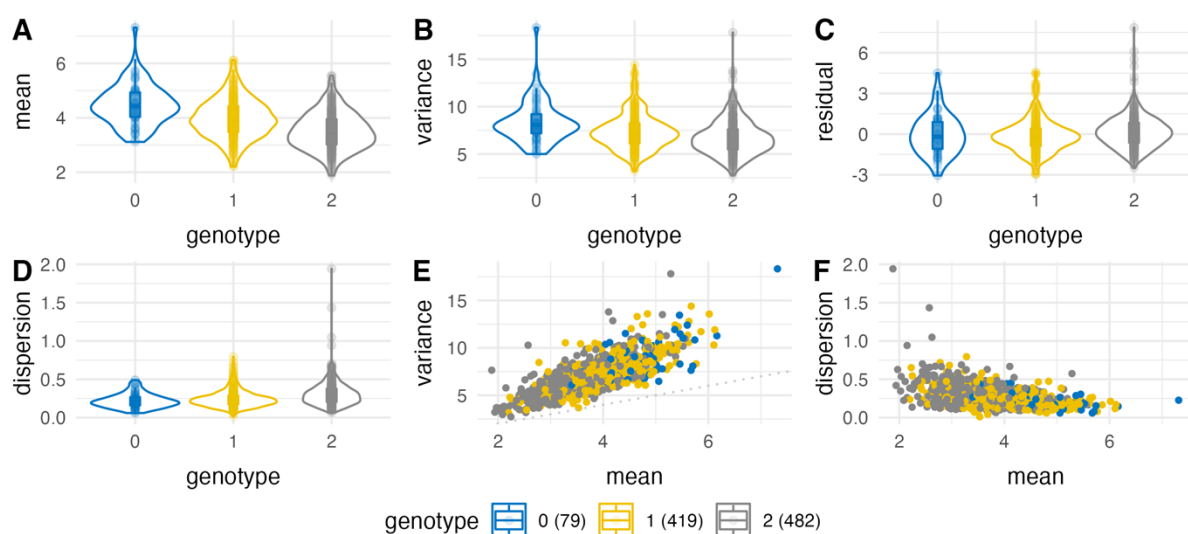
11:303271_A_T (rs6598046) for *IFITM2* in CD4_{NC} cells



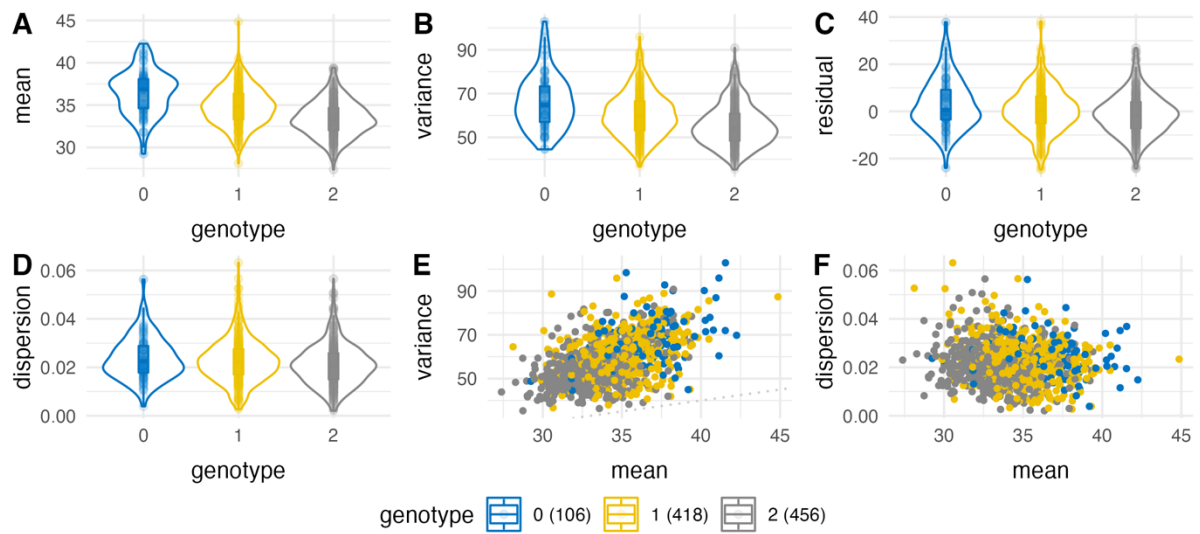
11:307539_A_G (rs11412325) for *IFITM2* in CD8_{ET} cells



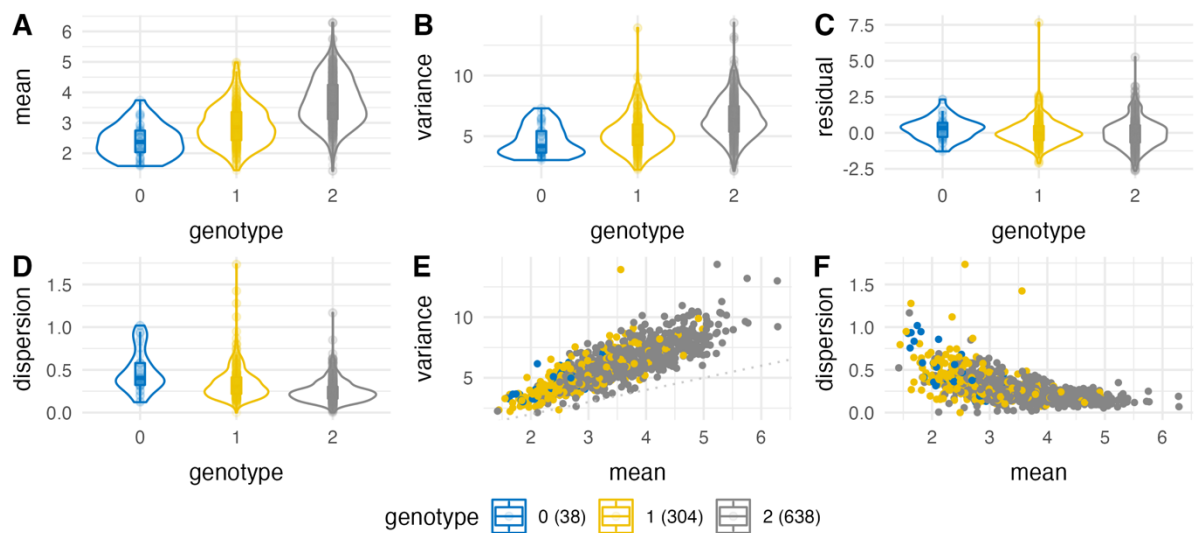
11:349122_C_T (rs7117996) for *IFITM2* in NK cells



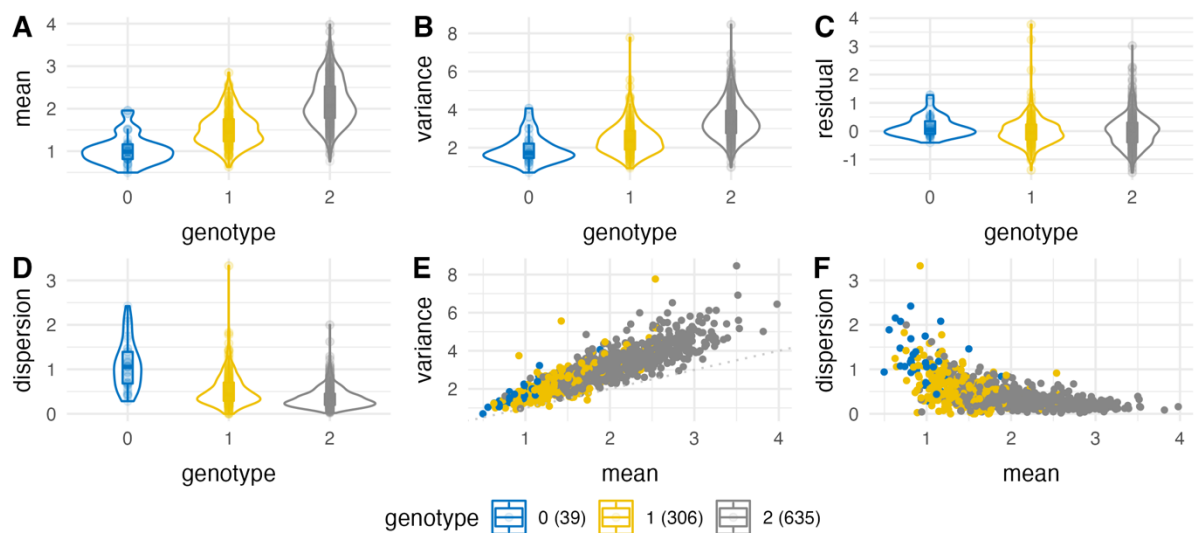
11:802902_T_G (rs28360884) for *RPLP2* in CD4_{NC} cells



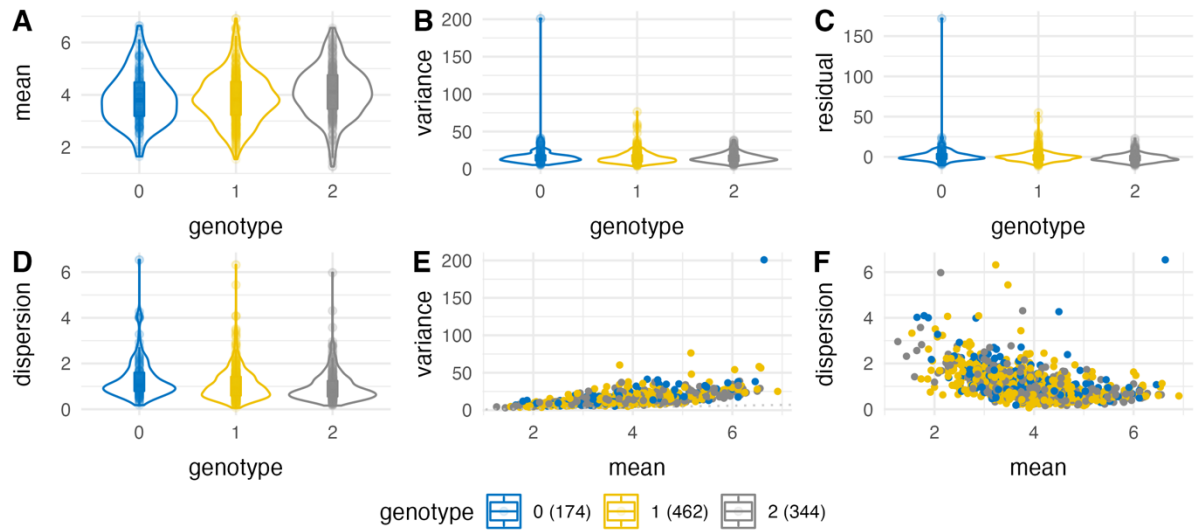
11:65644027_T_C (rs583887) for *CTSW* in NK cells



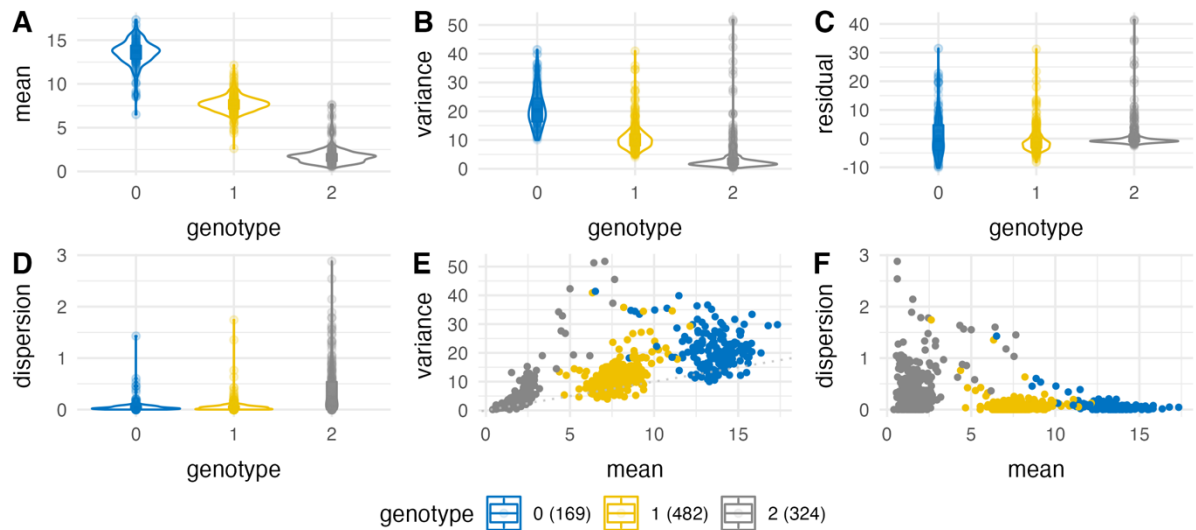
11:65645354_A_G (rs596002) for *CTSW* in CD8_{ET} cells



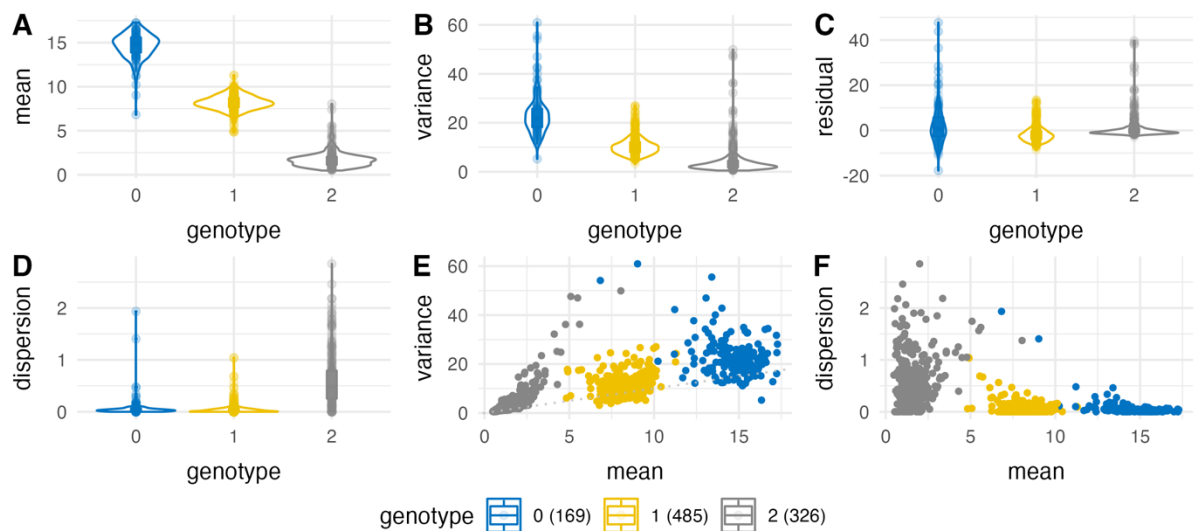
12:9623841_T_G (rs10743738) for *KLRB1* in CD4_{ET} cells



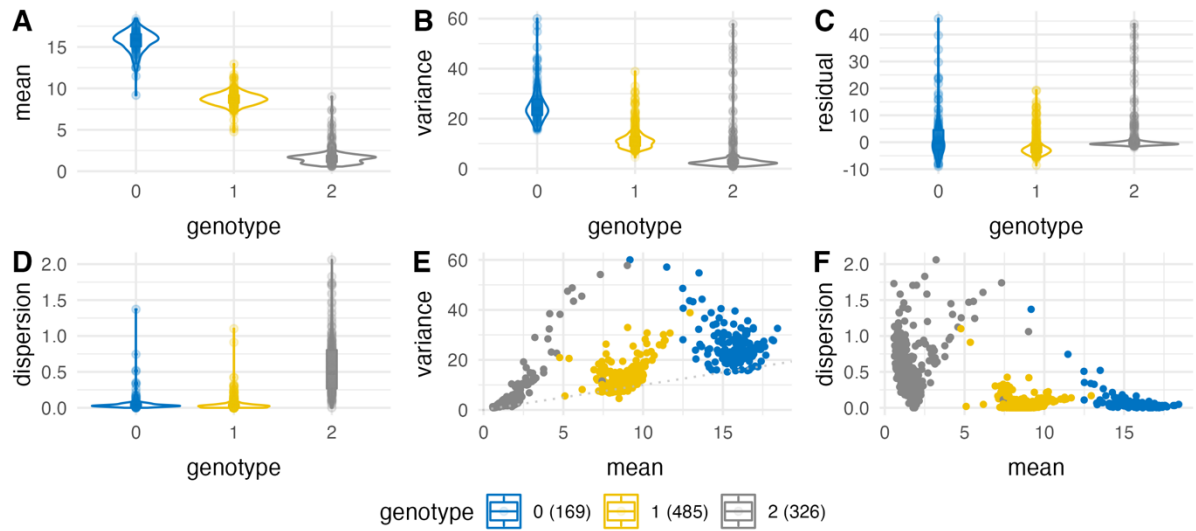
12:56435929_C_G (rs1131017) for *RPS26* in B_{IN} cells



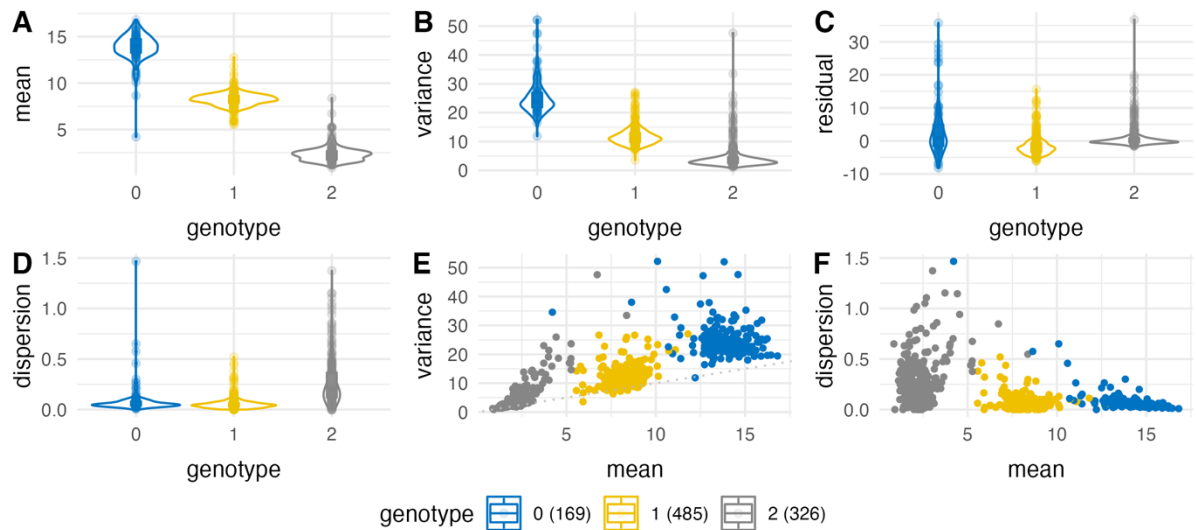
12:56435929_C_G (rs1131017) for *RPS26* in CD4_{ET} cells



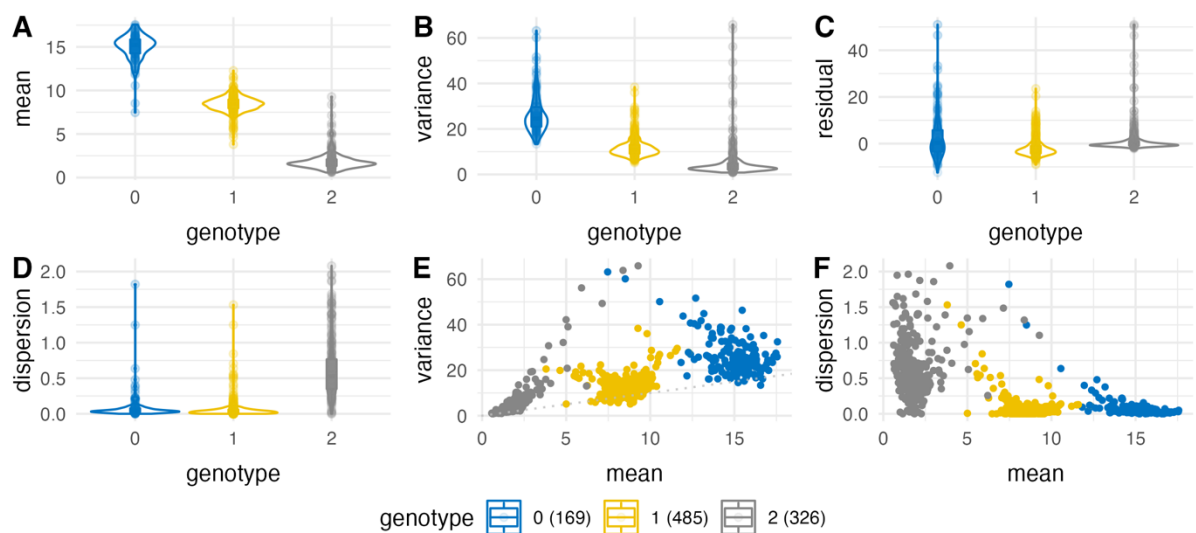
12:56435929_C_G (rs1131017) for *RPS26* in CD4_{NC} cells



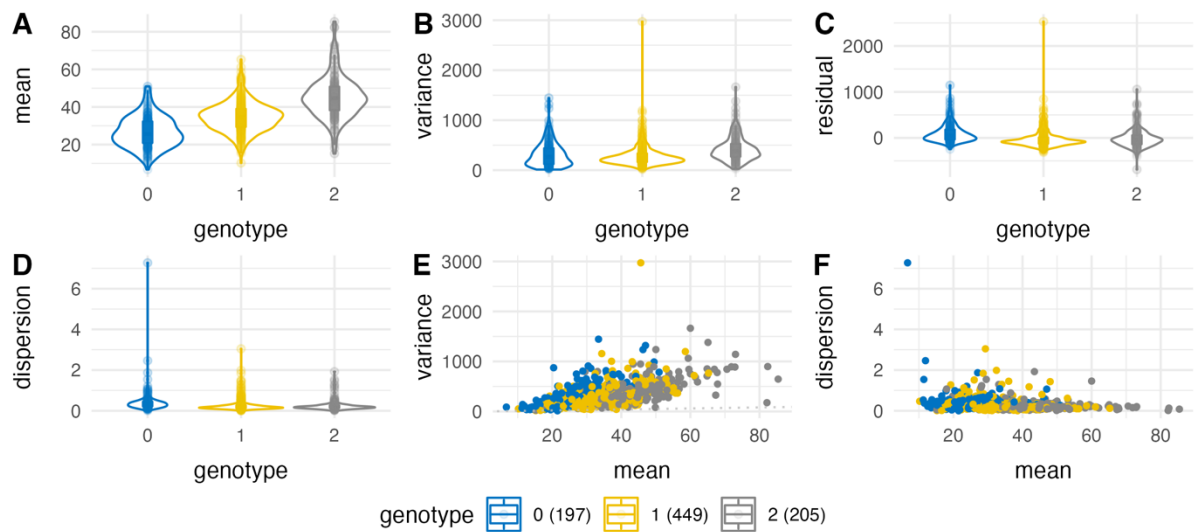
12:56435929_C_G (rs1131017) for *RPS26* in CD8_{ET} cells



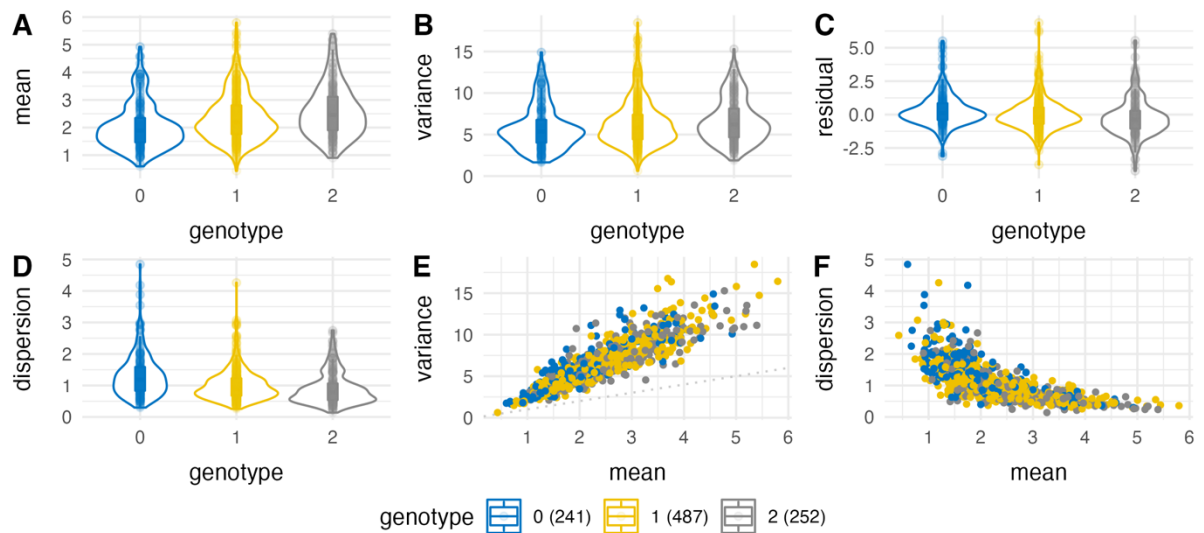
12:56435929_C_G (rs1131017) for *RPS26* in CD8_{NC} cells



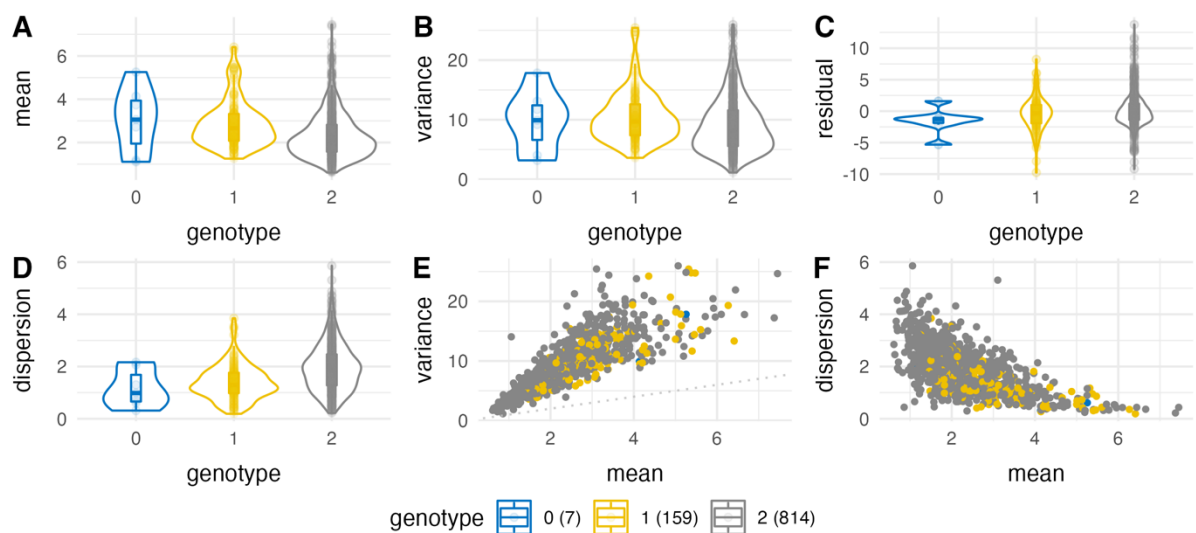
12:69747834_T_C (rs1384) for *LYZ* in Mono_C cells



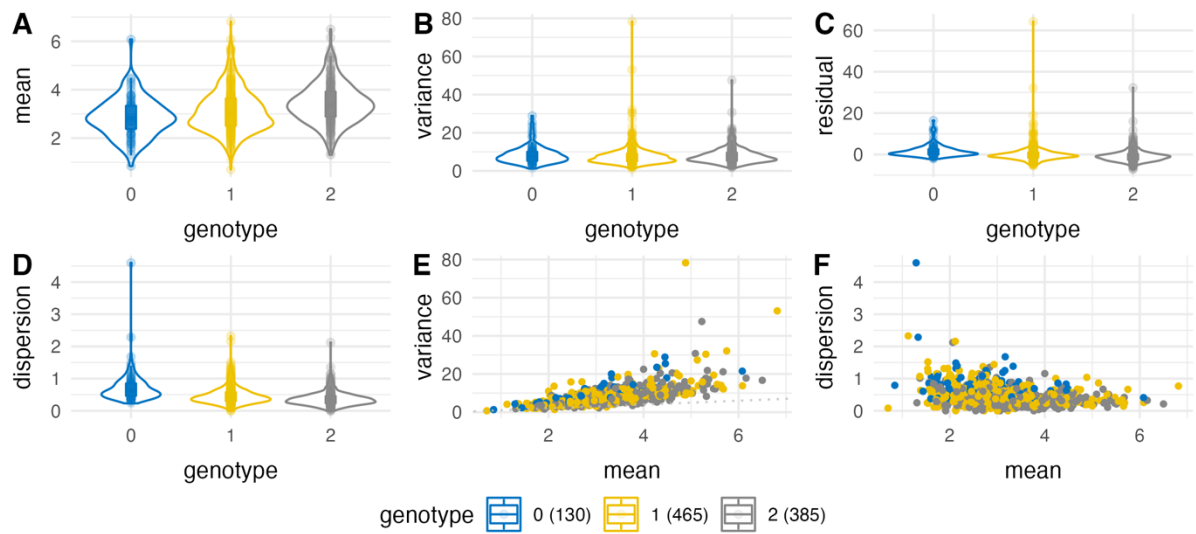
14:25083383_G_A (rs11158812) for *GZMH* in NK cells



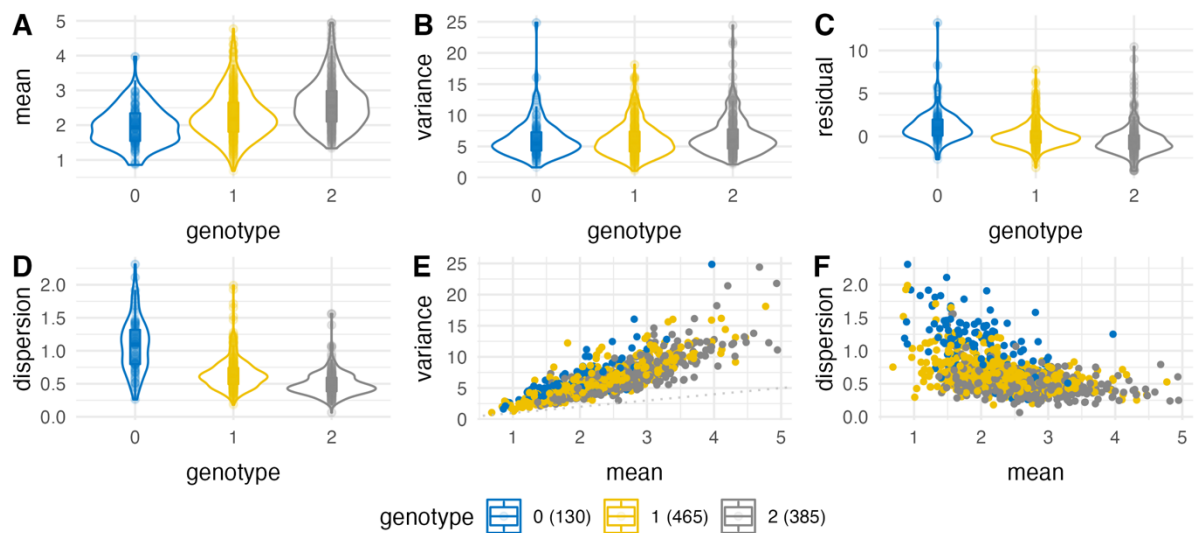
16:3115272_C_T (rs45499297) for *IL32* in NK cells



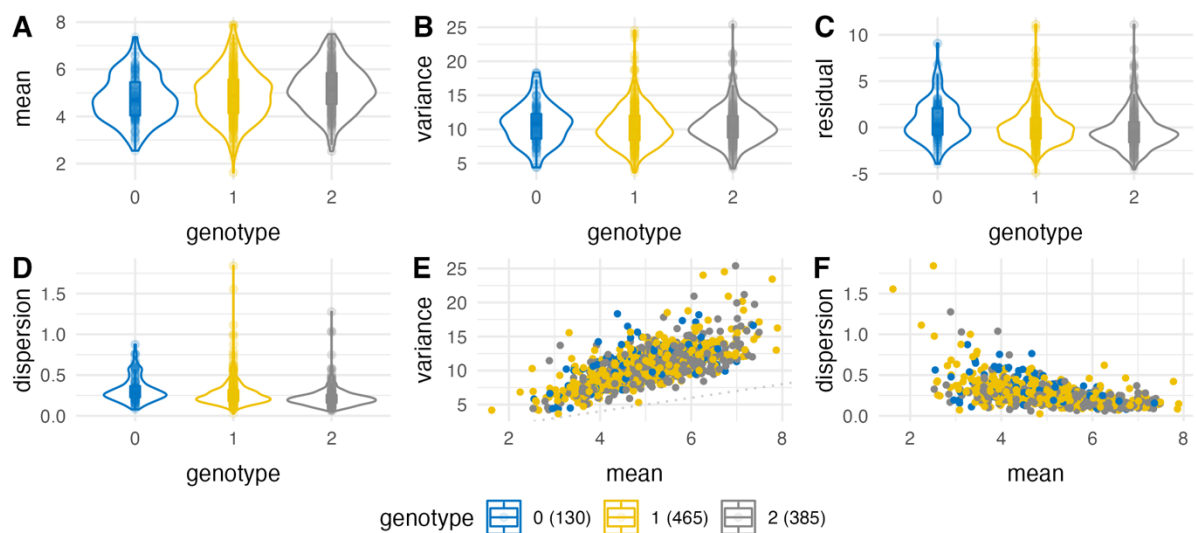
16:3115628_A_C (rs1554999) for *IL32* in CD4_{ET} cells



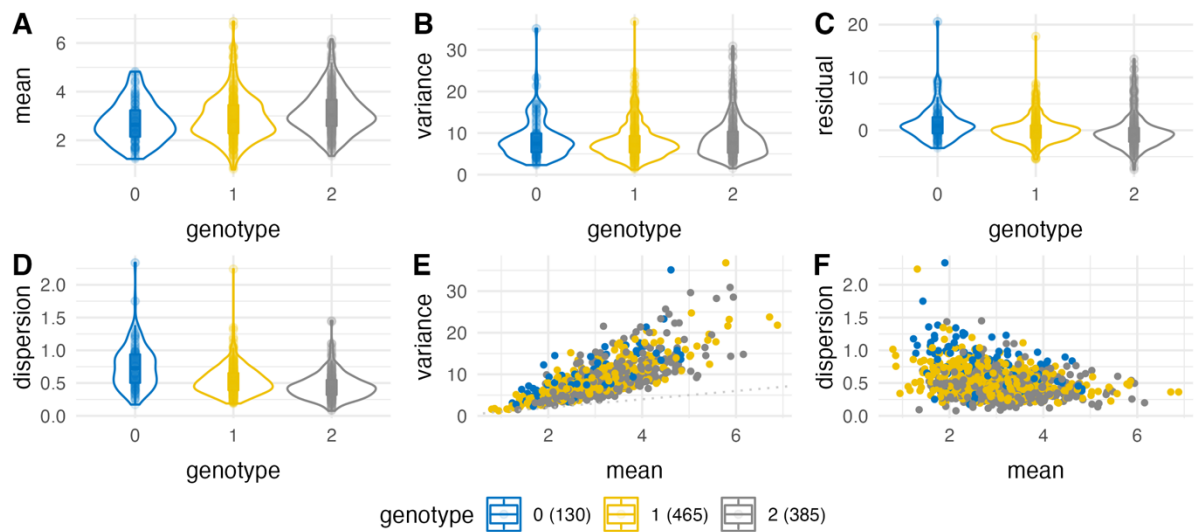
16:3115628_A_C (rs1554999) for *IL32* in CD4_{NC} cells



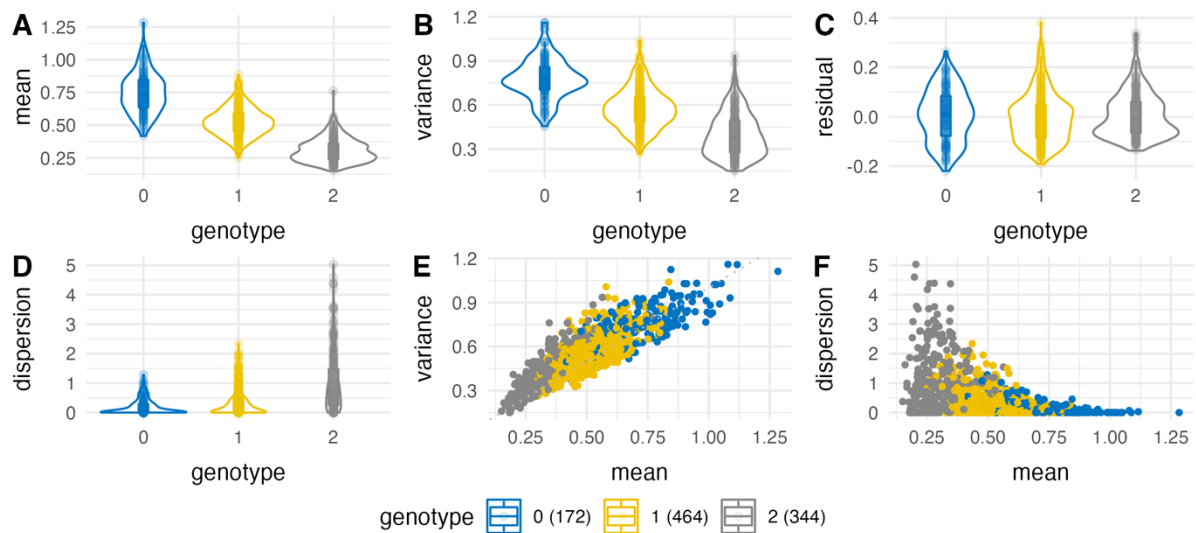
16:3115628_A_C (rs1554999) for *IL32* in CD8_{ET} cells



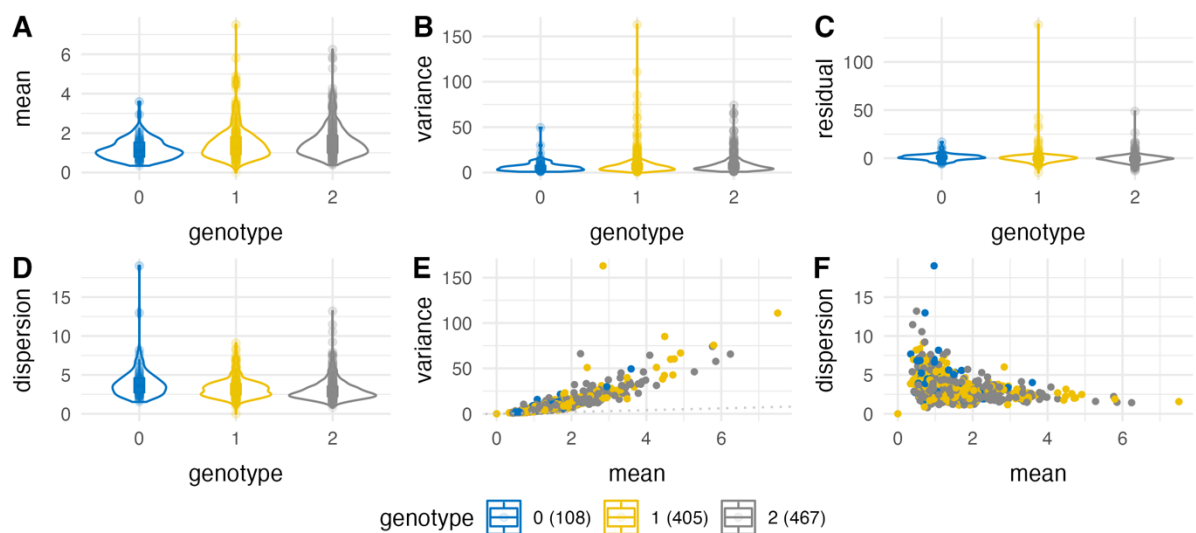
16:3115628_A_C (rs1554999) for *IL32* in CD8_{NC} cells



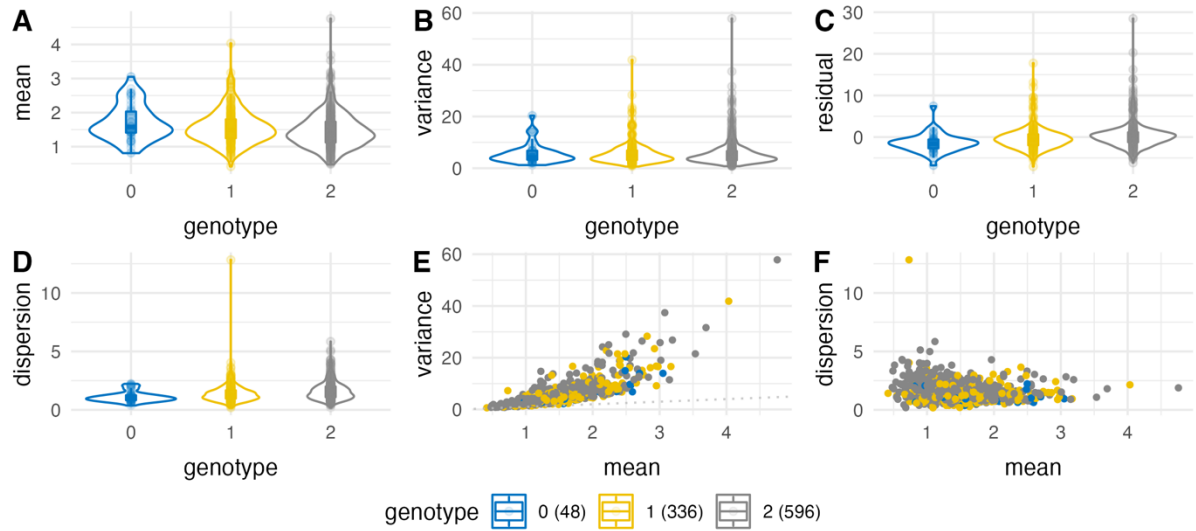
17:7207964_A_C (rs7503161) for *EIF5A* in CD4_{NC} cells



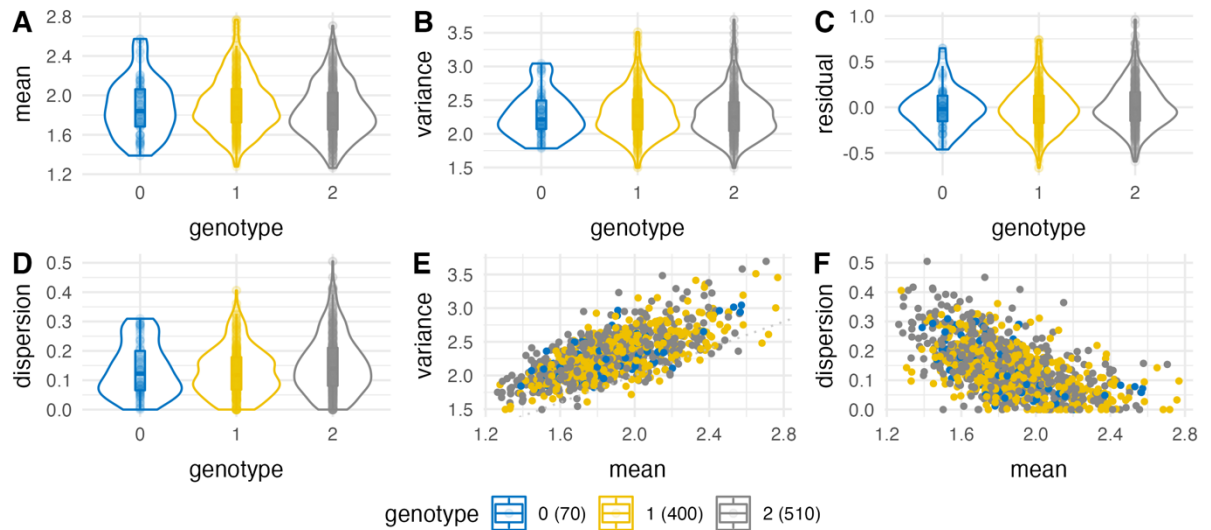
17:34397258_G_A (rs854471) for *CCL3* in NK cells



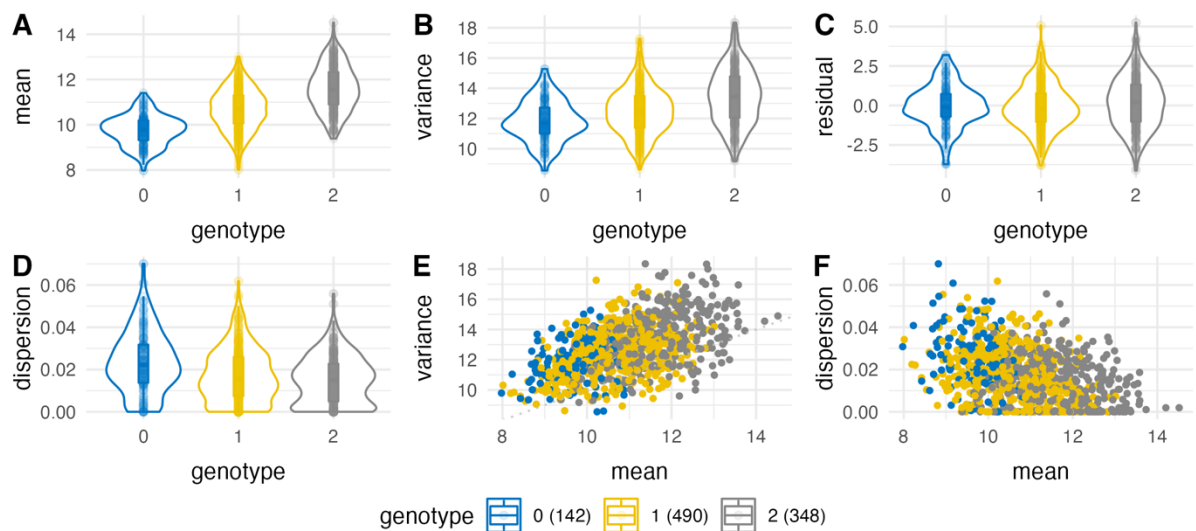
17:34411105_A_G (rs1634490) for *CCL4* in CD8_{ET} cells



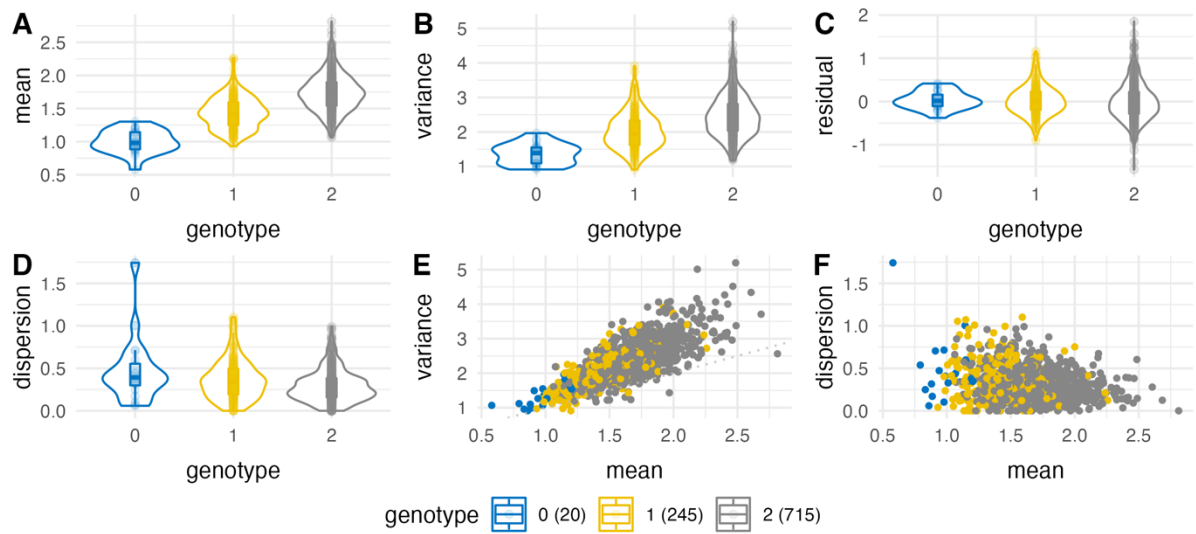
19:35658380_G_T (rs12461097) for *FXYD5* in CD4_{NC} cells



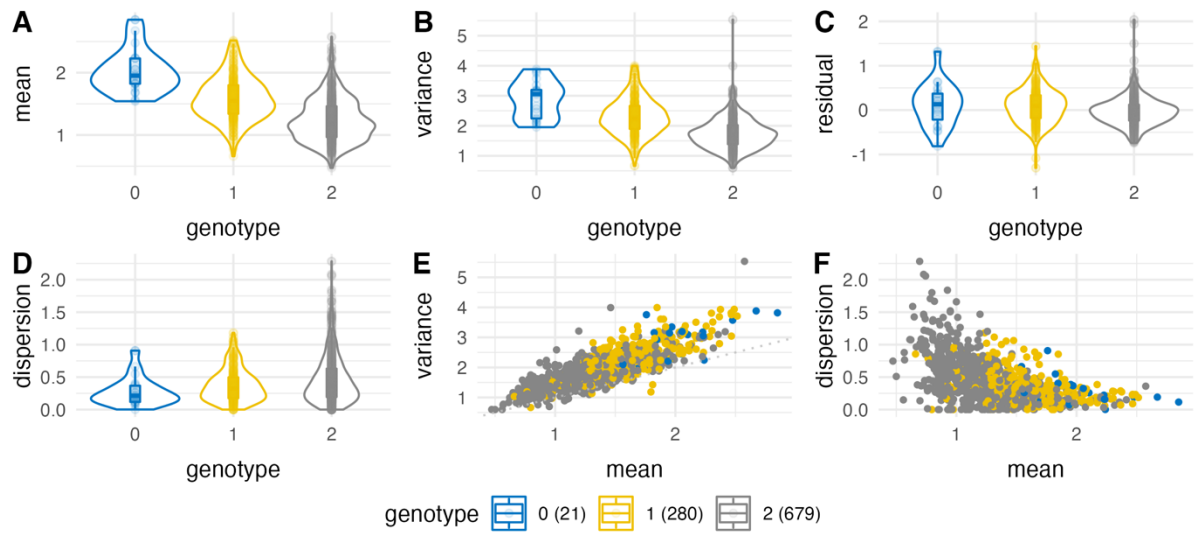
19:54700668_C_T (rs34172242) for *RPS9* in CD4_{NC} cells



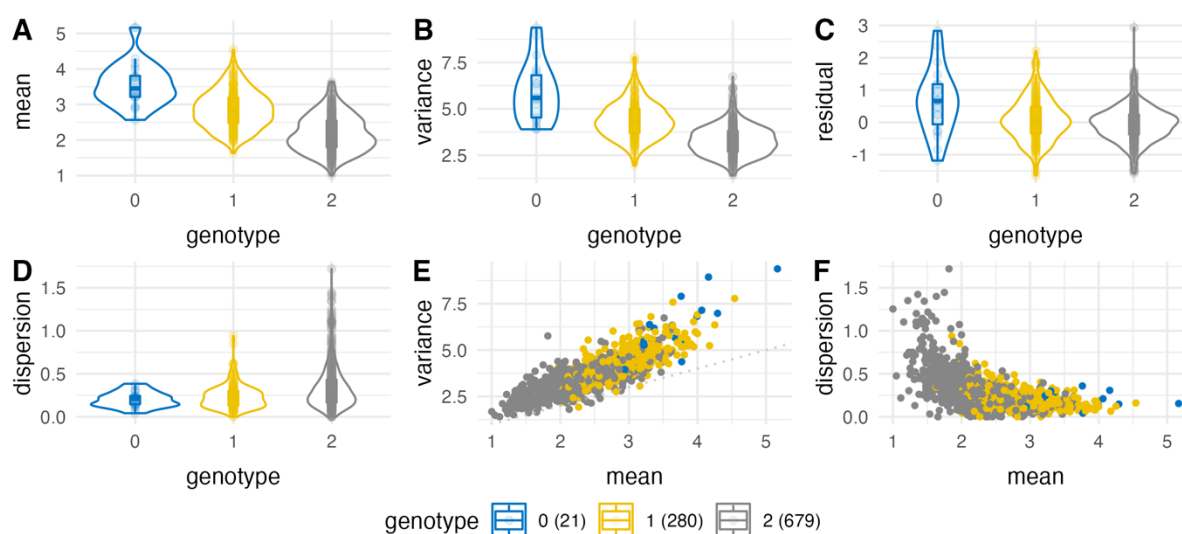
20:62152519_G_C (rs72629024) for *PPDPF* in NK cells



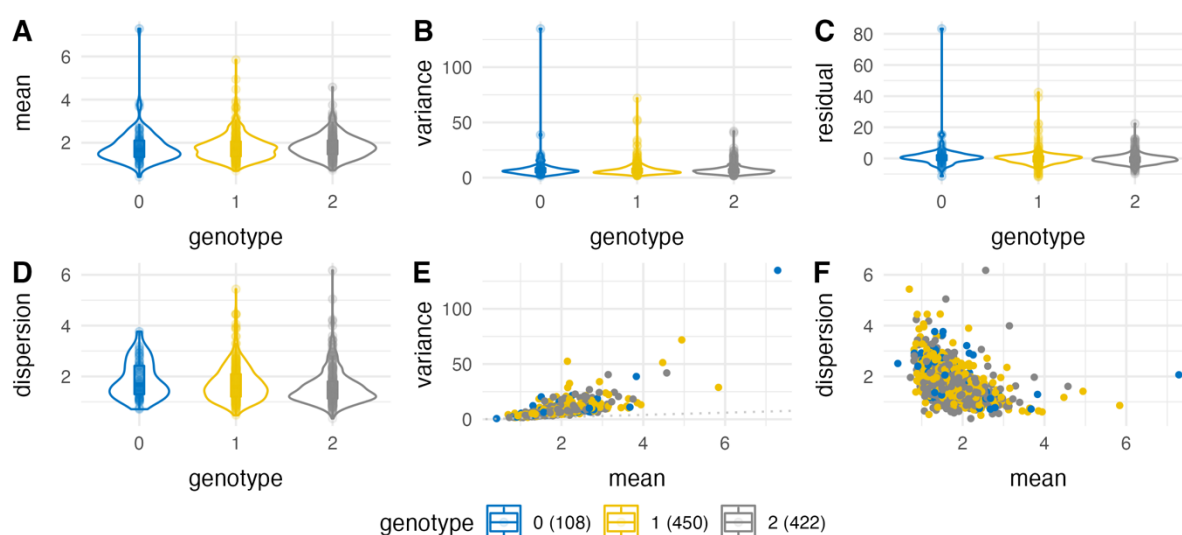
21:46328099_T_C (rs760462) for *ITGB2* in CD8_{ET} cells



21:46328099_T_C (rs760462) for *ITGB2* in NK cells

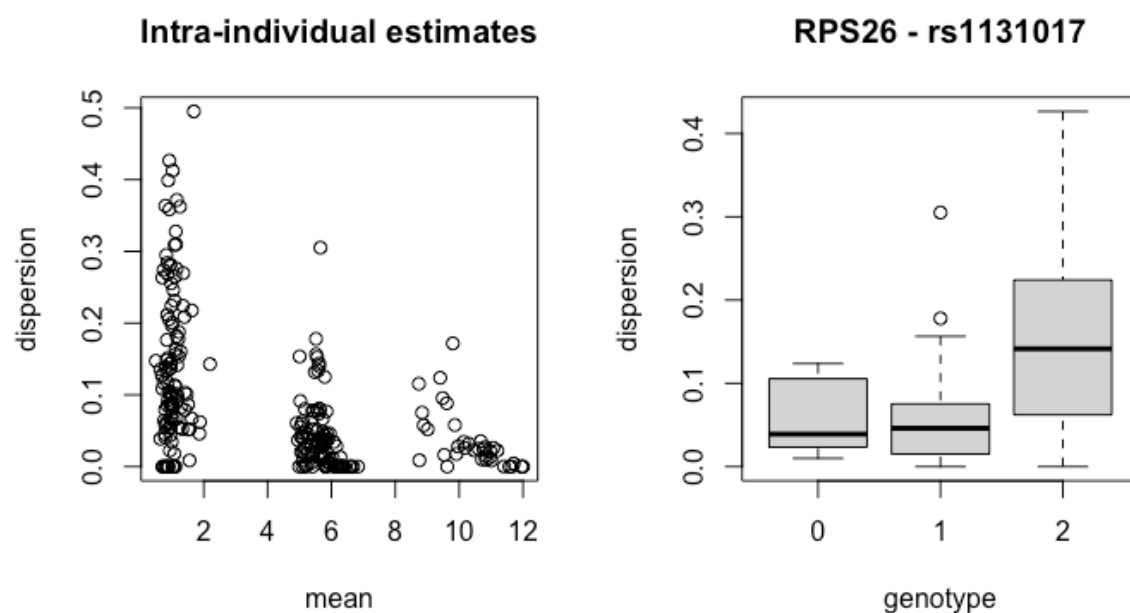


22:38069305_A_G (rs62236671) for *LGALS1* in NK cells

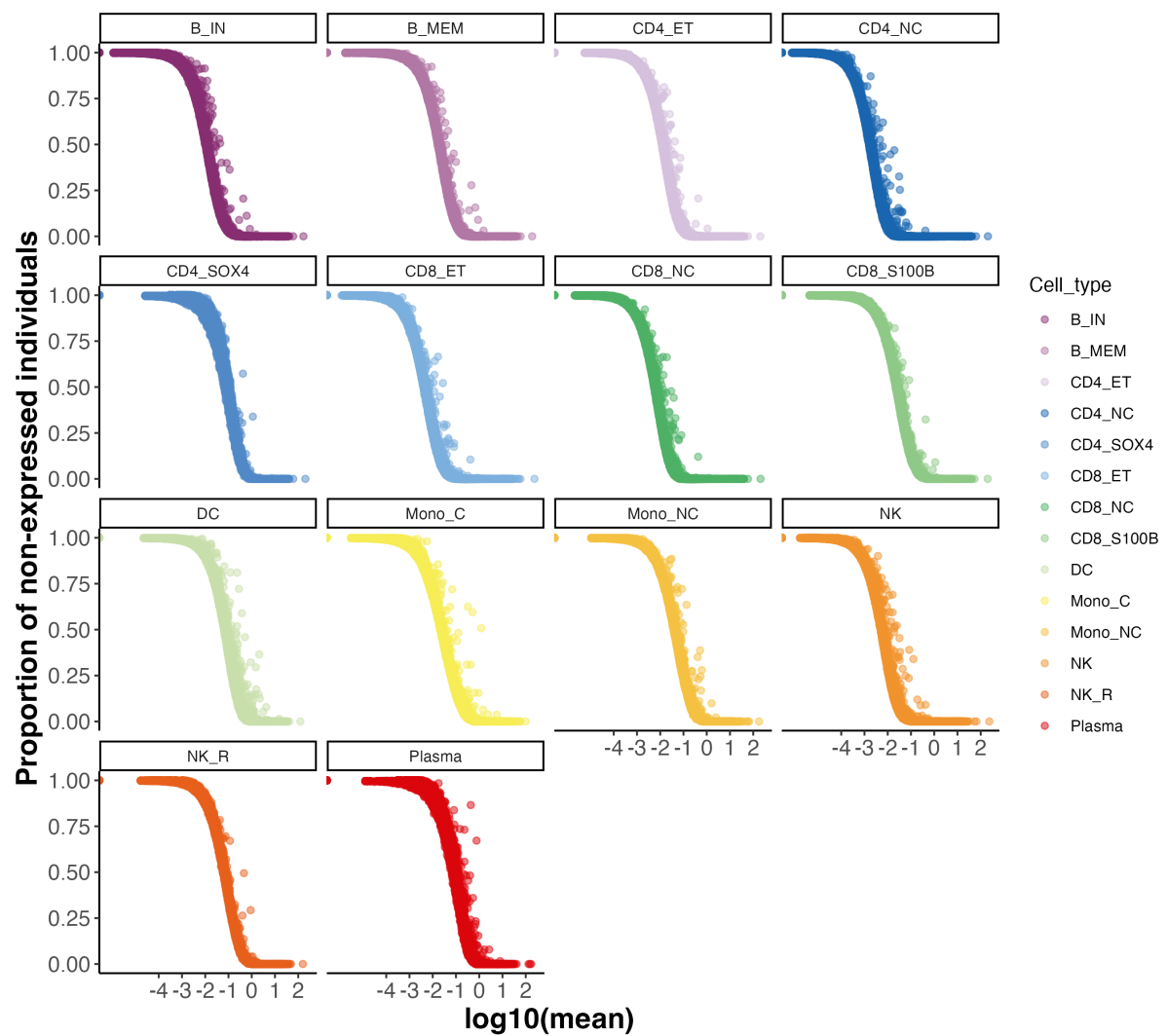


Supplementary Figure 11. Genetic association plot for 55 deQTL-dGene pairs.

(A-D) The violin plots of individual genotype of each deQTL corresponding to the intra-individual mean, variance, residual, dispersion of the dGene expression in certain cell types. Box plots show the median (centre line), interquartile range (box, 25th–75th percentiles), 1.5× IQR (whiskers), and outliers (individual points). The x-axis indicates the genotype (coded as 0, 1, 2 indicating number of alternative allele carried). The number of donors per genotype group is annotated within bracket. The color of the boxplot and dots corresponds to the three genotype groups **(E)** Scatter plot of intra-individual mean against intra-individual variance of expression. **(F)** Scatter plot of intra-individual mean against intra-individual dispersion of expression. The dispersion was estimated by CR-MLE method.

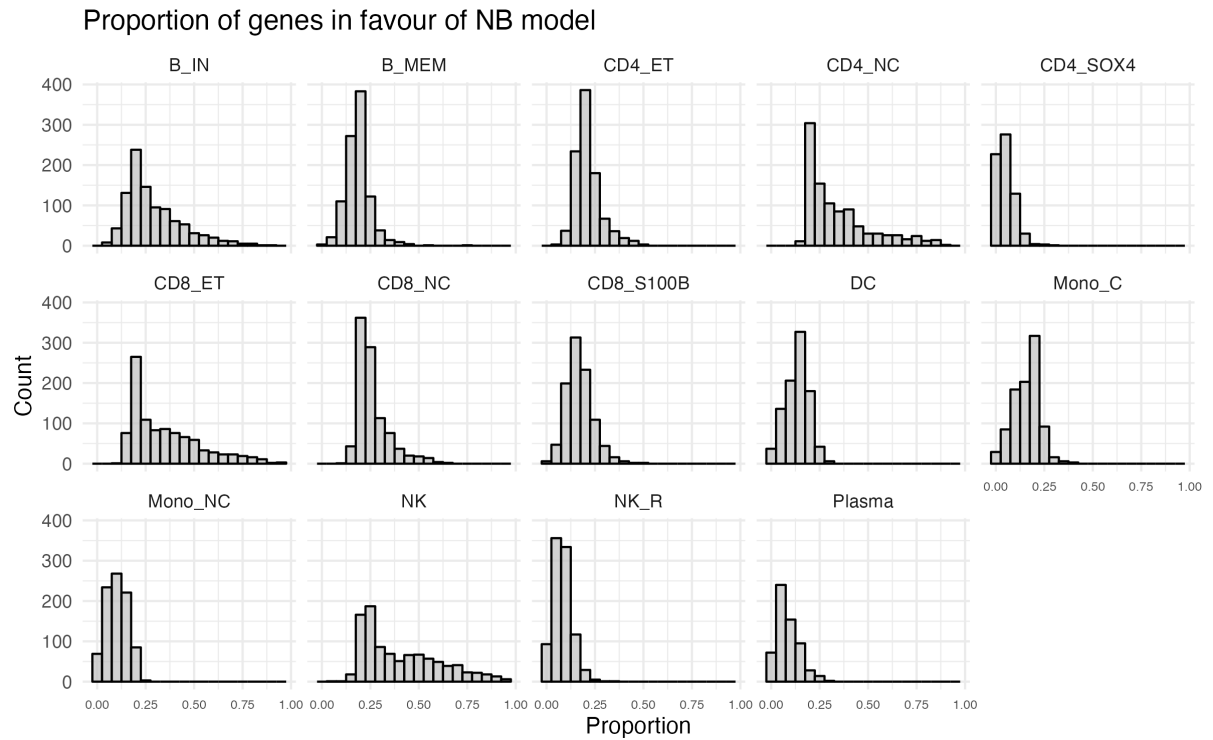


Supplementary Figure 12. Mean-dispersion and genetic association plot for rs1131017-*RPS26* locus in CD4 T cells in Perez et al. The left panel showed the intra-individual estimates of the mean and dispersion for *RPS26* in CD4 T cells in the replication cohort (N = 135). The right panel indicates the association pattern between genotype and intra-individual dispersion estimates. Box plots show the median (centre line), interquartile range (box, 25th–75th percentiles), 1.5× IQR (whiskers), and outliers (individual points).



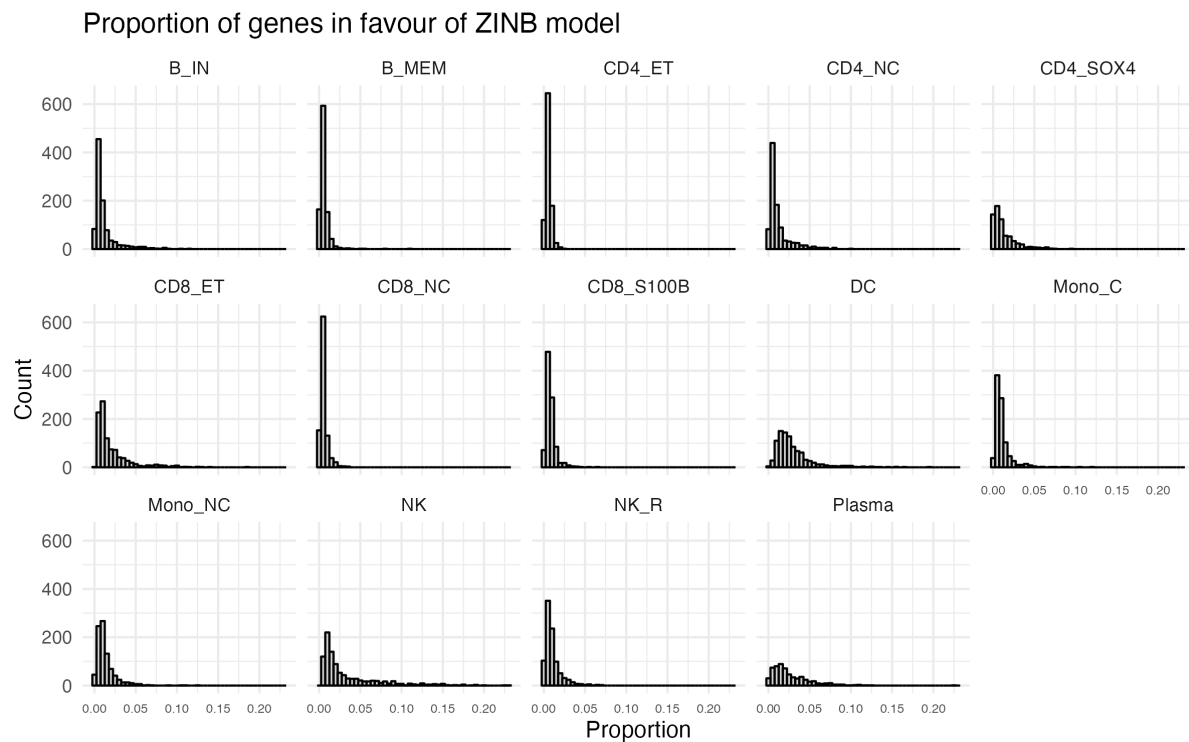
Supplementary Figure 13. The relationship between the mean and the proportion of non-expression individuals.

The scatter plot shows the relationship between the intra-individual mean and the proportion of non-expressed individuals for each gene across 14 cell types. Each dot denotes one gene. The x-axis denotes the log₁₀ transformed mean per gene. The y-axis denotes the proportion of individuals that do not express that gene. The number of dots (genes) is corresponding to column "Nr genes" in Table 1.



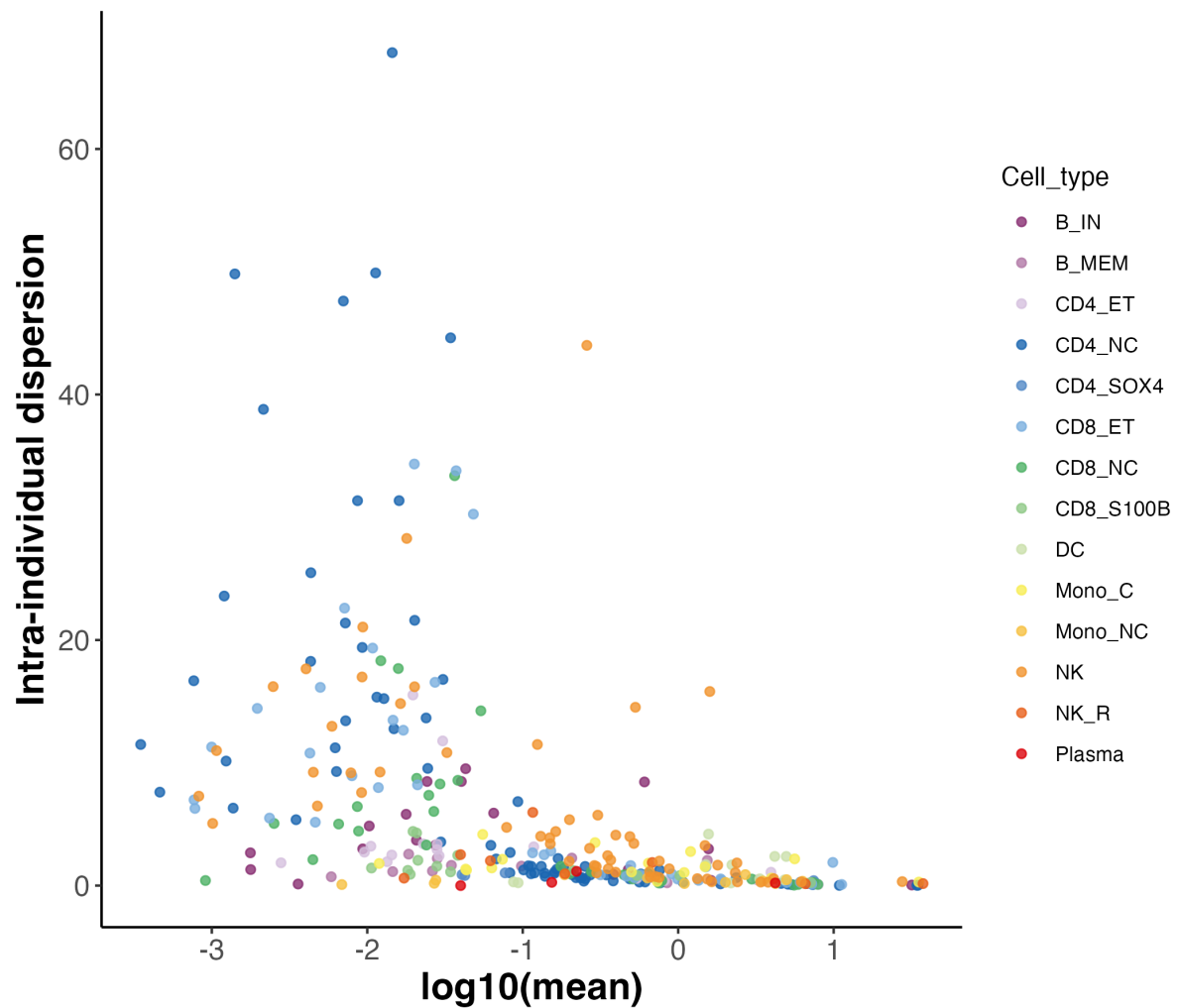
Supplementary Figure 14. Proportion of genes showing nominal evidence favouring the NB model rather than the Poisson model

The histogram summarises how many genes in each individual are more consistent with the Negative Binomial model than with the Poisson model. The nominal evidence is determined by a likelihood ratio test with a P-value < 0.05. Each panel denotes one of the 14 cell types.



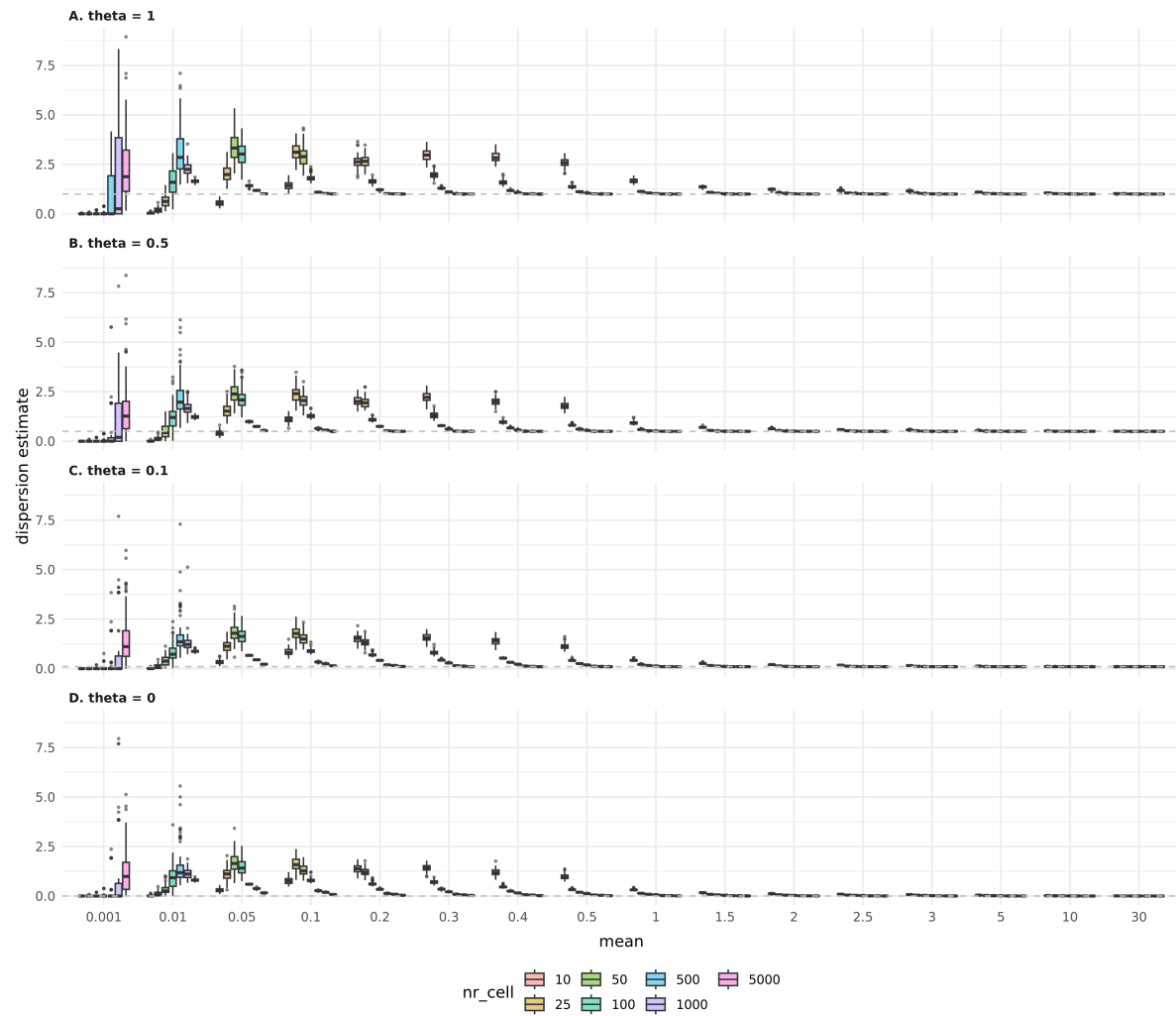
Supplementary Figure 15. Proportion of genes showing nominal evidence favouring the ZINB model rather than the NB model

The histogram summarises how many genes in each individual are in favour of the Zero-Inflated Negative Binomial model rather than the Negative Binomial model. The nominal evidence is determined by a likelihood ratio test with a P-value < 0.05. Each panel denotes one of the 14 cell types.



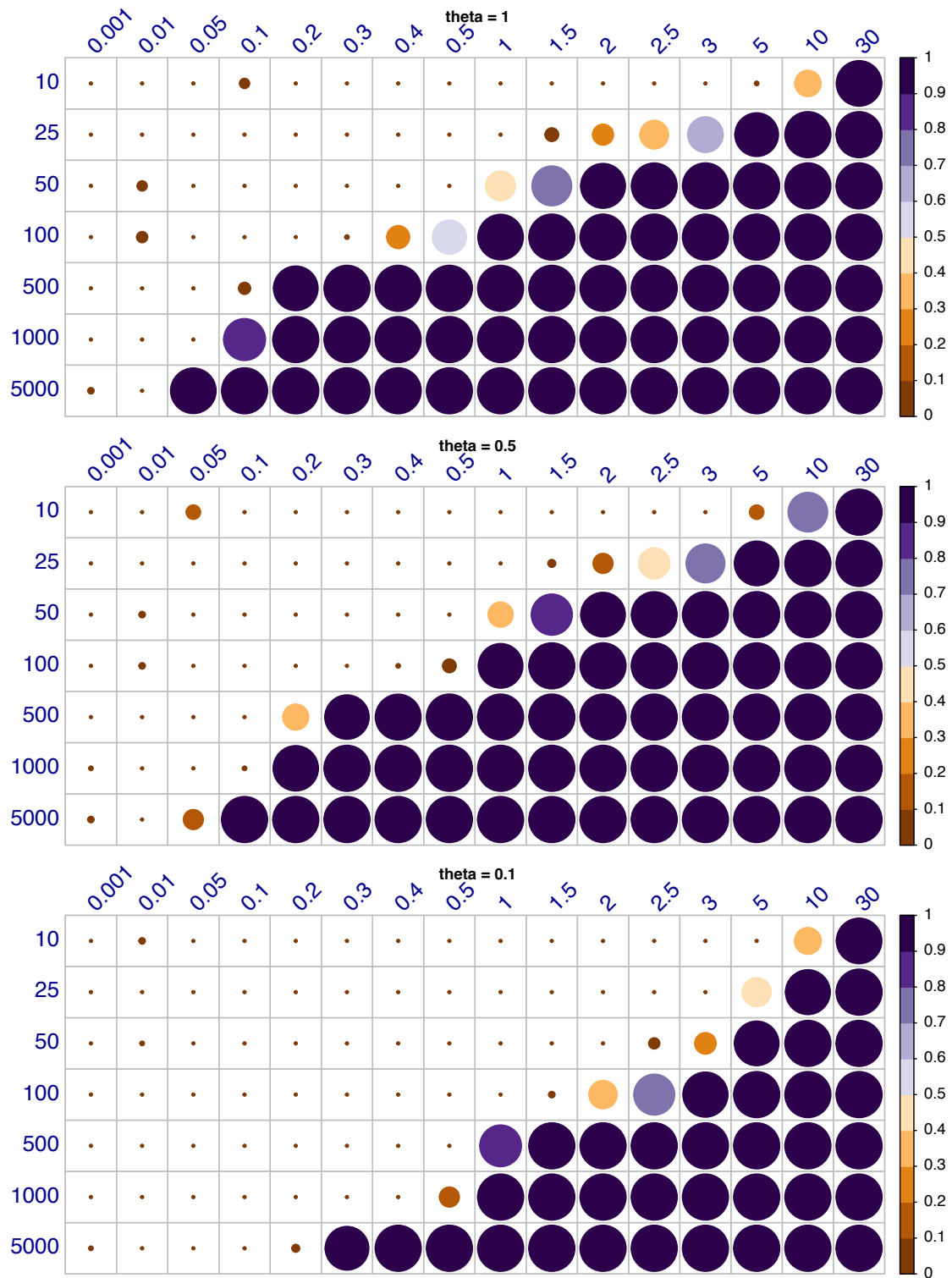
Supplementary Figure 16. Distribution of dispersion estimates for 199 genes with 326 deQTLs in the preliminary test.

This plot indicates the intra-individual estimates of mean and dispersion for 199 genes with 326 deQTL in 14 cell types in the preliminary test before filtering out the inflated genes. The x-axis indicates the log10-transformed intra-individual mean estimates, and the y-axis indicates the intra-individual dispersion estimates using the Cox-Reid MLE method. The colour of the dot indicates the corresponding cell type.



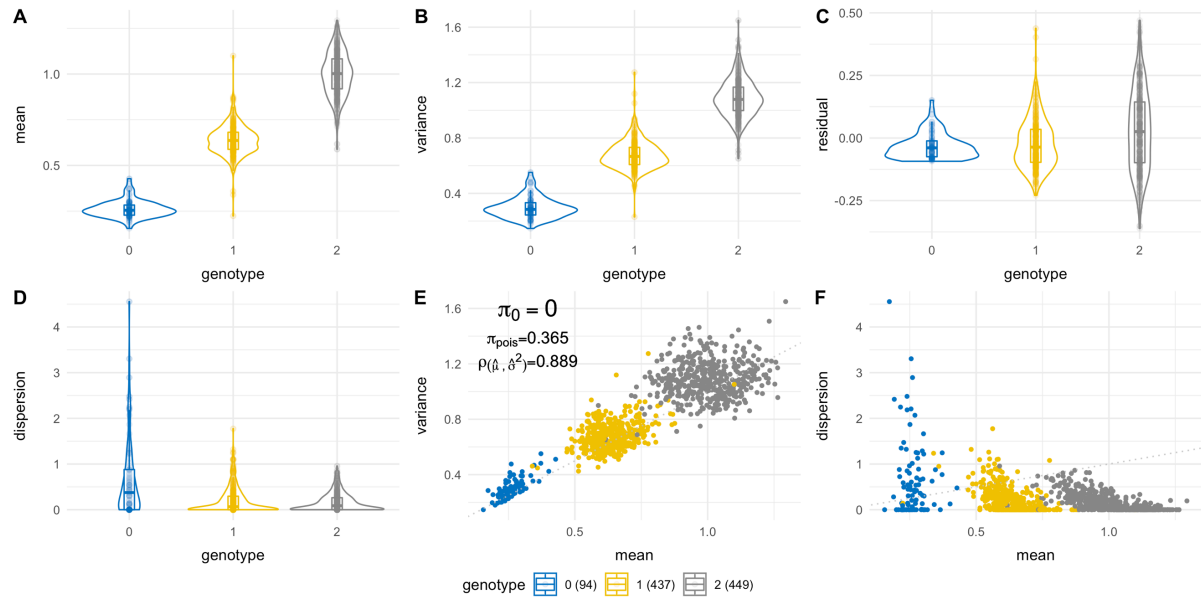
Supplementary Figure 17. Behaviour of CR-MLE dispersion estimates across simulated expression regimes.

Dispersion estimates obtained using the Cox–Reid adjusted maximum likelihood estimator (CR-MLE) under simulated count data. For each simulation, gene expression counts were generated under a negative binomial model with specified mean expression (x-axis) and dispersion parameter (panel label). The y-axis shows the estimated dispersion values obtained from CR-MLE. Colours indicate different numbers of cells per individual (sample size). The horizontal grey dashed line denotes the true dispersion parameter used to generate the data. Results illustrate that dispersion estimates become increasingly inflated and variable at low mean expression levels and small sample sizes, particularly when the true dispersion lies near zero. Box plots show the median (centre line), interquartile range (box, 25th–75th percentiles), 1.5× IQR (whiskers), and outliers (individual points).



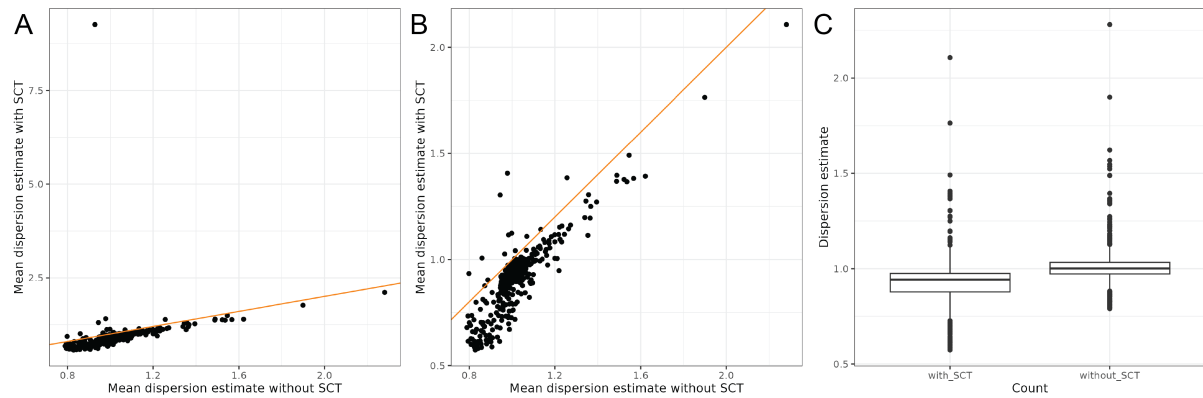
Supplementary Figure 18. The accuracy of dispersion estimates in simulation.

The heatmap plot demonstrates the accuracy of the dispersion estimates in the simulation under different scenarios. The row indicates the number of cells per individual, and the column indicates the mean of intra-individual expression. Each circle indicates the proportion of 100 replicates that the estimate is within $\pm 5\%$ of the true dispersion parameter.



Supplementary Figure 19. Example of inflated estimation of dispersion for *SMDT1* in CD4_{NC} cells.

The mean threshold for unbiased estimation of dispersion in CD4_{NC} cells is 0.3. Although the mean of intra-individual mean for all individuals passed the threshold (0.3), the mean in the genotype 0 group (G/G of rs133379) did not, thus might suffer from inflated estimates. Box plots show the median (centre line), interquartile range (box, 25th–75th percentiles), 1.5× IQR (whiskers), and outliers (individual points).



Supplementary Figure 20. Mean dispersion estimates per gene before and after SCTransform

(A) mean dispersion estimates for all the genes with mean > 0.3 in CD4 Naïve cells (N = 980). (B) Same to panel A but exclude the outlier gene *MALAT1*. (C) Boxplot of mean dispersion estimates grouped by whether the count has been SCTransformed or not. Box plots show the median (centre line), interquartile range (box, 25th–75th percentiles), 1.5× IQR (whiskers), and outliers (individual points).