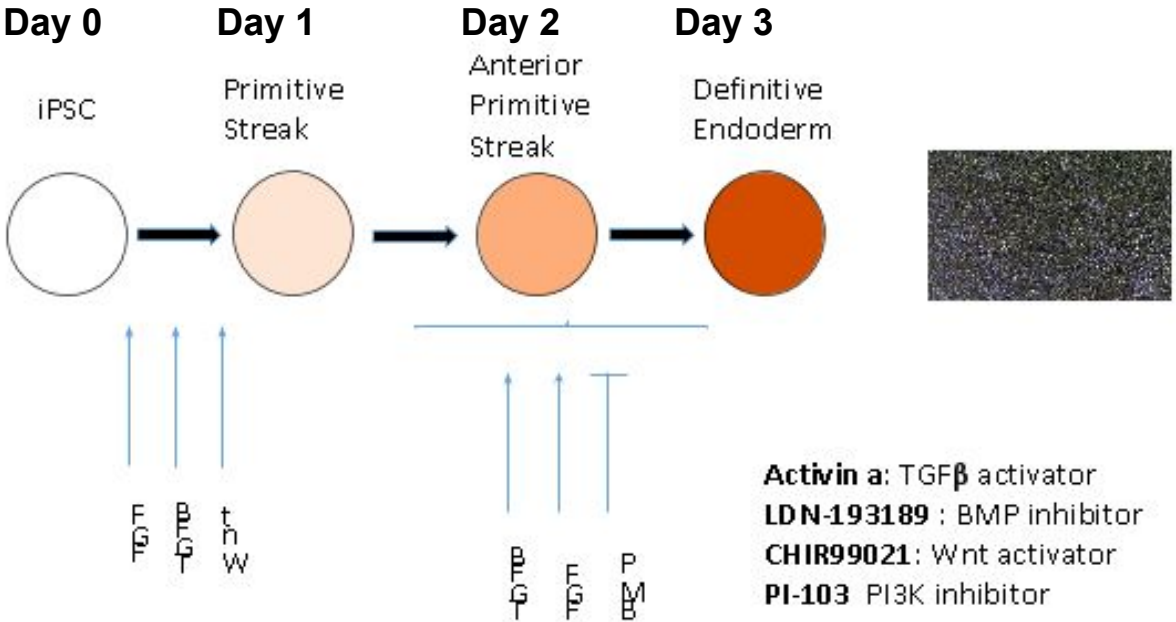
Analysis of RNA-seq data from time series experiments

Joyce Hsiao and Lauren Blake
Gilad and Lynch lab meeting
July 6, 2016

iPSC differentiation into endoderm: experimental design

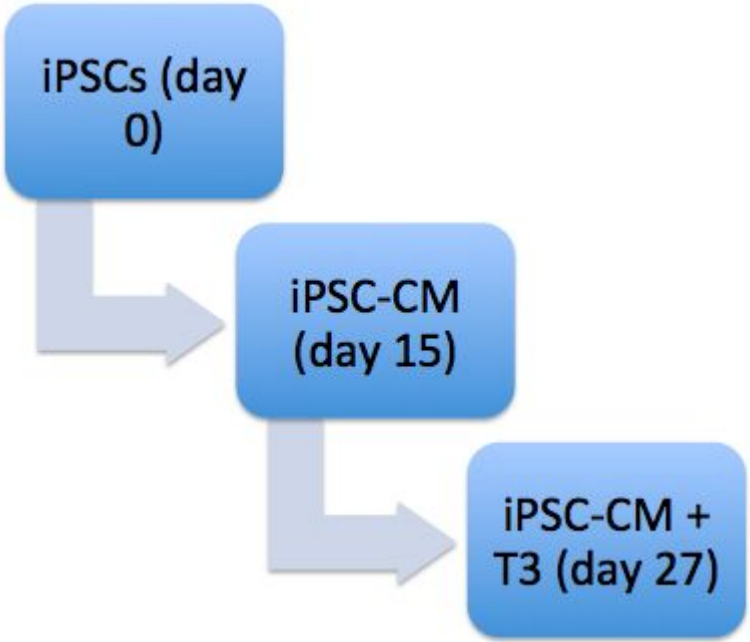
4 chimpanzees, all replicated once

6 Humans, 2 are replicated once

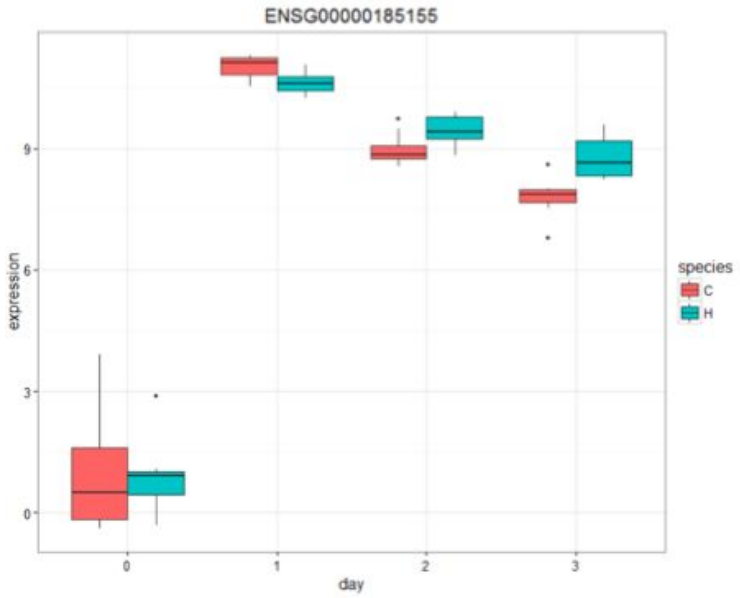


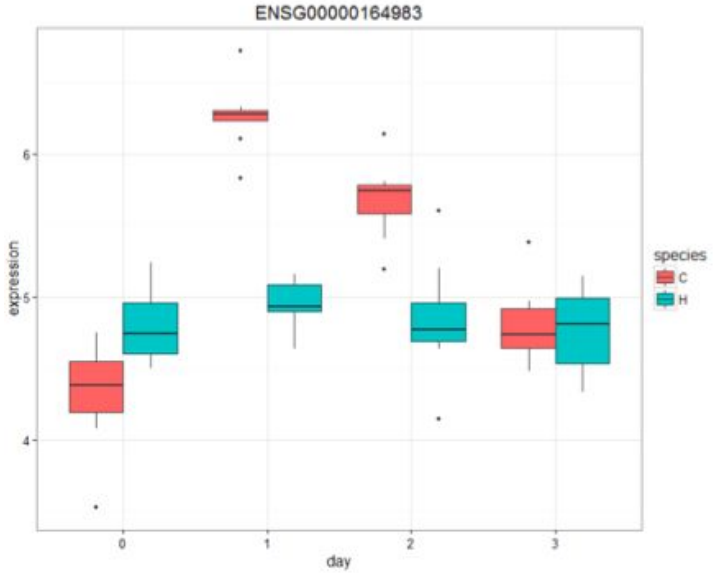
Sammy Thomas

Characterizing gene expression patterns of human and chimpanzee iPSC-CM: Experimental design



Motivating questions: Local and global trends





Sammy Thomas

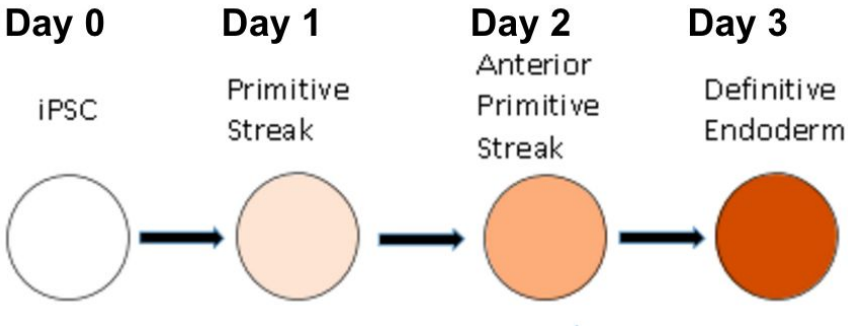
Modelling time series data

Concerns

1. Repeated measurements
    a. Correlation between time points: specific to individual?
    b. Patterns of change over time: linear or polynomial? Discrete or continuous?
2. Fixed versus random effect
3. Specific to our design
    a. Often small number of individuals, multiple replicates per individuals, and small number of time points

Fixed versus random effects: inference

- Fixed effect: the selected timepoints or conditions are fixed
- Random effect: the selected timepoints or conditions are a random subset of a larger population, therefore the inference can be extended beyond the timepoints or conditions in the current study



Are the repeated measurements a random subset of the developmental process?

Fixed versus random effects: correlation

- Random effect allows the modelling of correlation between timepoints or between replicates of the same individuals

Three possible correlation structures between replicates/timepoints

Unstructured

$$\begin{bmatrix} \sigma_1^2 & \sigma_{12} & \sigma_{13} & \sigma_{14} \\ \sigma_{12} & \sigma_2^2 & \sigma_{23} & \sigma_{24} \\ \sigma_{13} & \sigma_{23} & \sigma_3^2 & \sigma_{34} \\ \sigma_{14} & \sigma_{24} & \sigma_{34} & \sigma_4^2 \end{bmatrix}$$

1. Correlations between all time points are different.
2. # of parameters = t(t+1)/2

Compound symmetry

$$\begin{bmatrix} \sigma^2+\sigma_1^2 & \sigma_1^2 & \sigma_1^2 & \sigma_1^2 \\ \sigma_1^2 & \sigma^2+\sigma_1^2 & \sigma_1^2 & \sigma_1^2 \\ \sigma_1^2 & \sigma_1^2 & \sigma^2+\sigma_1^2 & \sigma_1^2 \\ \sigma_1^2 & \sigma_1^2 & \sigma_1^2 & \sigma^2+\sigma_1^2 \end{bmatrix}$$

1. Same correlation between all time points
2. Parameters = 2

Autoregressive (AR1)

$$\sigma^2 \begin{bmatrix} 1 & \rho & \rho^2 & \rho^3 \\ \rho & 1 & \rho & \rho^2 \\ \rho^2 & \rho & 1 & \rho \\ \rho^3 & \rho^2 & \rho & 1 \end{bmatrix}$$

1. Exponential increase in correlation over time.
2. Parameters = 2

Analysis strategy

Example: Human vs. Chimp X 3 Timepoints

1. Evaluate both global and local trend
2. Start with fixed effect model, evaluate residual assumptions, add random effects
3. Compare mixed models: Bayesian Information Criteria (BIC)

All can be done in limma, except for the model comparison!

Starting model: fixed effects

Log2 expression = species + day + species X day

```
designMatrix <- model.matrix(~ species + day + species*day)
fit <- lmFit(yy, design = design)
contrastMatrix <- makeContrasts(day1 - day2, day1 - day3,
                                day2 - day3, levels = designMatrix)
fit_contrast <- contrasts.fit(fit, contrastMatrix)
```

Global trends: significant interaction of species and day

Local trends: contrast test

Fixed effects + individual random

```
dupcor <- duplicateCorrelation(y, design, block = block)
fit_cor <- lmFit(y, designMatrix, block = block, correlation = dupcor$consensus)
```

For every gene, block matrix describes the sample relatedness

	Human1, T1	Human1, T2	Human1, T3	Chimp1, T1	Chimp1, T2	Chimp1, T3
Human1, time1	1	1	1	0	0	0
Human1, time2	1	1	1	0	0	0
Human1, time3	1	1	1	0	0	0
Chimp1, time1	0	0	0	1	1	1
Chimp2, time2	0	0	0	1	1	1
Chimp3, tim3	0	0	0	1	1	1

Fixed effects + time random

```
dupcor <- duplicateCorrelation(y, design, block = block)
fit_cor <- lmFit(y, designMatrix, block = block, correlation = dupcor$consensus)
```

For every gene, block matrix describes the sample relatedness

	Human1, T1	Human1, T2	Human1, T3	Chimp1, T1	Chimp1, T2	Chimp1, T3
Human1, time1	1	0	0	1	0	0
Human1, time2	0	1	0	0	1	0
Human1, time3	0	0	1	0	0	1
Chimp1, time1	1	0	0	1	0	0
Chimp2, time2	0	1	0	0	1	0
Chimp3, tim3	0	0	1	0	0	1

Model selection

- Evaluate residuals: are the residuals correlated with predictor variable, such as time?
- Model fitness criteria such as Bayesian Information Criteria

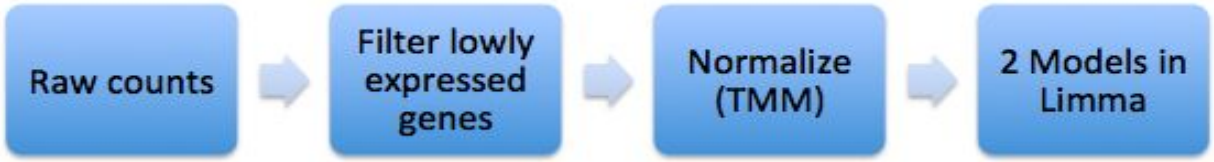
Pipeline for fitting models

Example data: time series RNA-seq data from yeast (Leong et al 2014)

library(“fission”) on Bioconductor

Strain	Time points (minutes)	Biological replicates
WT and del of aft1	0, 15, 30, 60, 120, 180	3/strain

Processing



2 models:

1) Log2 Expression = Strain + Time + Strain*Time
2) Log2 Expression = Strain + Time + Strain*Time + Individual (random)

Fixed effect model

1) **Design matrix**

```
design_all <- model.matrix(~ strain + minute + strain*minute, data = strains)
```

2) **Voom for gene expression**

3) **LMfit**

4) **Diagnostics**

Checking for homoskedastity

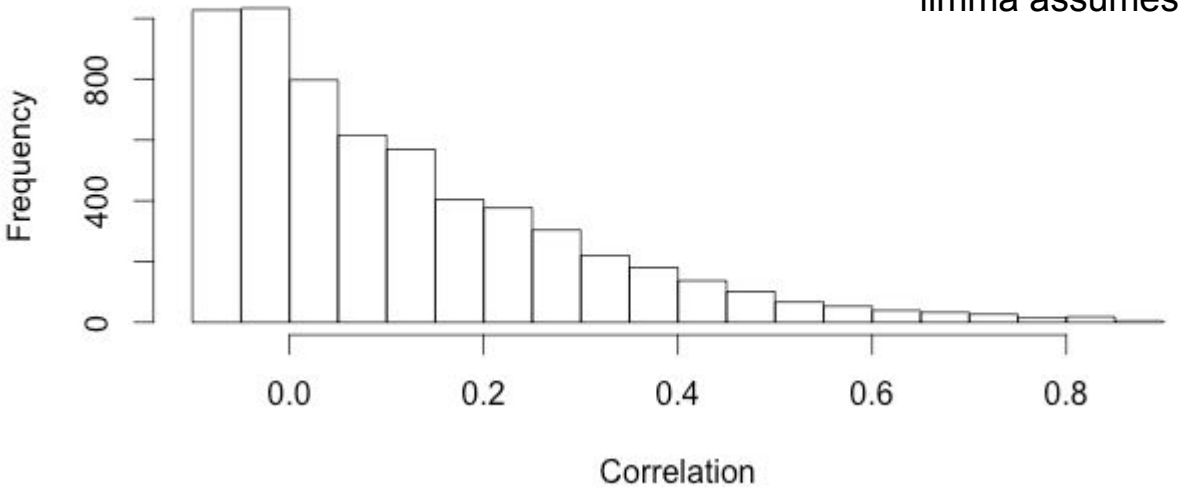
There is an assumption of linear models that the variance is the same for all x’s. Can check that there’s random scatter aka no fanning with a residual plot.

Cor(residuals, values for explanatory variable)

Correlations between replicates (repeated measurements)

cpm.voom <- voom(design_all, normalize.method="none", plot=T)

corfit <- duplicateCorrelation(cpm.voom$E, design_all, block=strains$replicate)



limma assumes same correlation across genes

cor = 0.08

Re-run voom

Re-run voom with blocking and correlation specified ()

```
fit1 <- lmFit(cpm.voom$E, design_all,
block=strains$replicate, correlation=
corfit$consensus)
```

Modified t-statistic and degrees of freedom

Comparison of the 2 models

No random variable

```
##                   logFC    AveExpr         t      P.Value  adj.P.Val
## SPCC70.08c     1.3905768  2.4174953  5.225025 1.852789e-05 0.09264082
## SPNCRNA.1457  -1.1654550  3.3659116 -5.032770 3.075214e-05 0.09264082
## SPBC2F12.09c   1.7088389  1.5162066  4.232133 2.542675e-04 0.44744302
## SPNCRNA.184    2.4995601 -1.3277111  4.096002 3.633343e-04 0.44744302
## SPBCPT2R1.08c -1.9911625  0.1929998 -4.087691 3.713220e-04 0.44744302
## SPCC1235.02    0.3053335  6.8080988  3.730911 9.383816e-04 0.90783243
## SPCC1672.03c   0.4401719  5.8348015  3.685501 1.054743e-03 0.90783243
## SPAC21E11.04   0.5627842  3.6957647  3.594411 1.332242e-03 0.99737734
## SPAP8A3.12c   -0.2954698  6.8119612 -3.506295 1.667857e-03 0.99737734
## SPNCRNA.863    1.2725884  2.1707604  3.484012 1.764985e-03 0.99737734
```

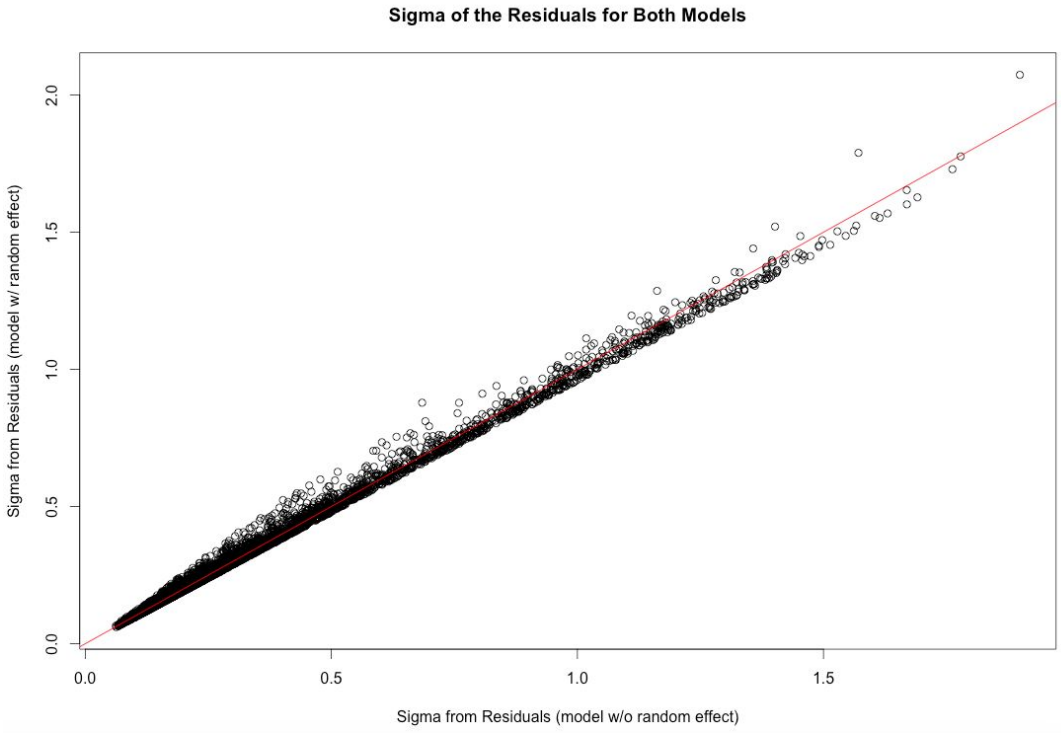
With Random Variable

```
##                   logFC    AveExpr         t      P.Value  adj.P.Val
## SPCC70.08c     1.3905768  2.4174953  5.475298 9.622458e-06 0.05797531
## SPNCRNA.1457  -1.1654550  3.3659116 -5.080355 2.717013e-05 0.08185003
## SPBCPT2R1.08c -1.9911625  0.1929998 -4.539349 1.133413e-04 0.22762719
## SPBC2F12.09c   1.7088389  1.5162066  4.378320 1.732892e-04 0.26101685
## SPNCRNA.184    2.4995601 -1.3277111  4.165733 3.029777e-04 0.36508807
## SPCC1235.02    0.3053335  6.8080988  3.932766 5.567882e-04 0.55910814
```

The SD of the residuals in model w/o random effect > with random effect

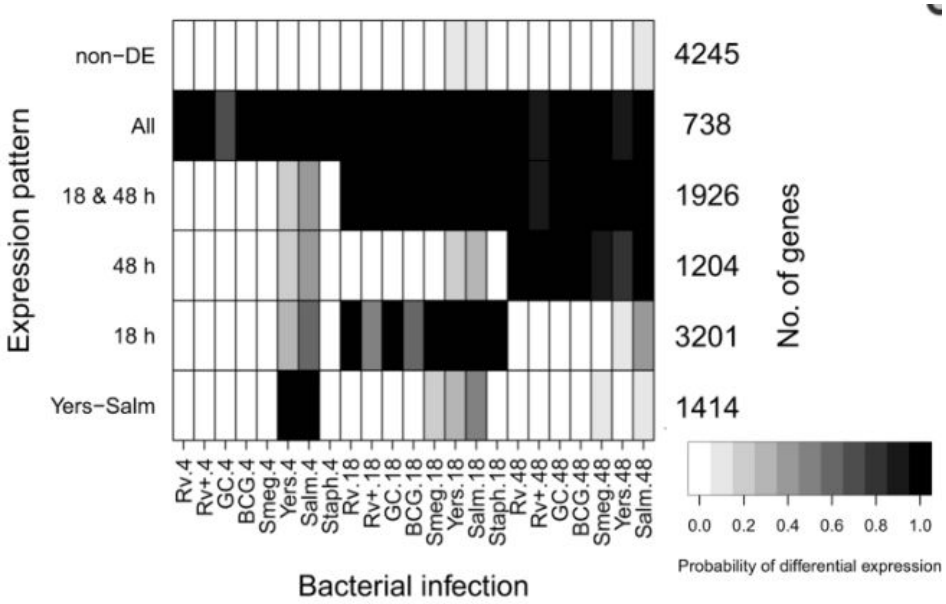
fit <- lmFit(voom_gene_expression_matrix, design_matrix)

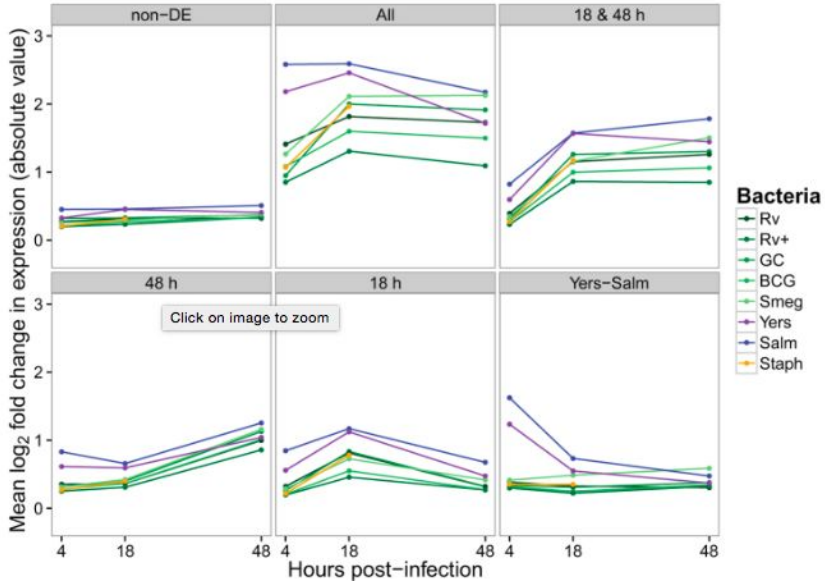
no_random_residuals <- fit$sigma



Clustering: showing common patterns in different genes

Cormotif





Blischak et al. 2015

Summary

- Assess global and local trends in time series experiments
- Use *limma* to perform significant testing and diagnostics
- Visualize change patterns across time

How to pick a covariance structure?

can be described in a fairly intuitive manner, though as we'll see they can be very similar to one another.

Structure	Description	# of Parameters	{i,j}th element
AR(1)	Autoregressive(1)	2	$\sigma_{ij} = \sigma^2 \rho^{\lvert i-j \rvert}$
CS	Compound Symmetry	2	$\sigma_{ij} = \sigma_1 + \sigma^2 1(i = j)$
UN	Unstructured	t(t+1)/2	$\sigma_{ij} = \sigma_{ij}$
TOEP	Toeplitz	t	$\sigma_{ij} = \sigma_{\lvert i-j \rvert+1}$
VC	Variance Components	q	$\sigma_{ij} = \sigma_k^2 1(i = j)$ and i corresponds to the kth effect
ARH(1)	Heterogeneous AR(1)	t+1	$\sigma_{ij} = \sigma_i \sigma_j \rho^{\lvert i-j \rvert}$
CSH	Heterogeneous CS	t+1	$\sigma_{ij} = \sigma_i \sigma_j [\rho 1(i \neq j) + 1(i = j)]$
TOEPH	Heterogeneous TOEP	2t-1	$\sigma_{ij} = \sigma_i \sigma_j \rho_{\lvert i-j \rvert}$