

# Analysis of sorghum breeding trials using pedigree information

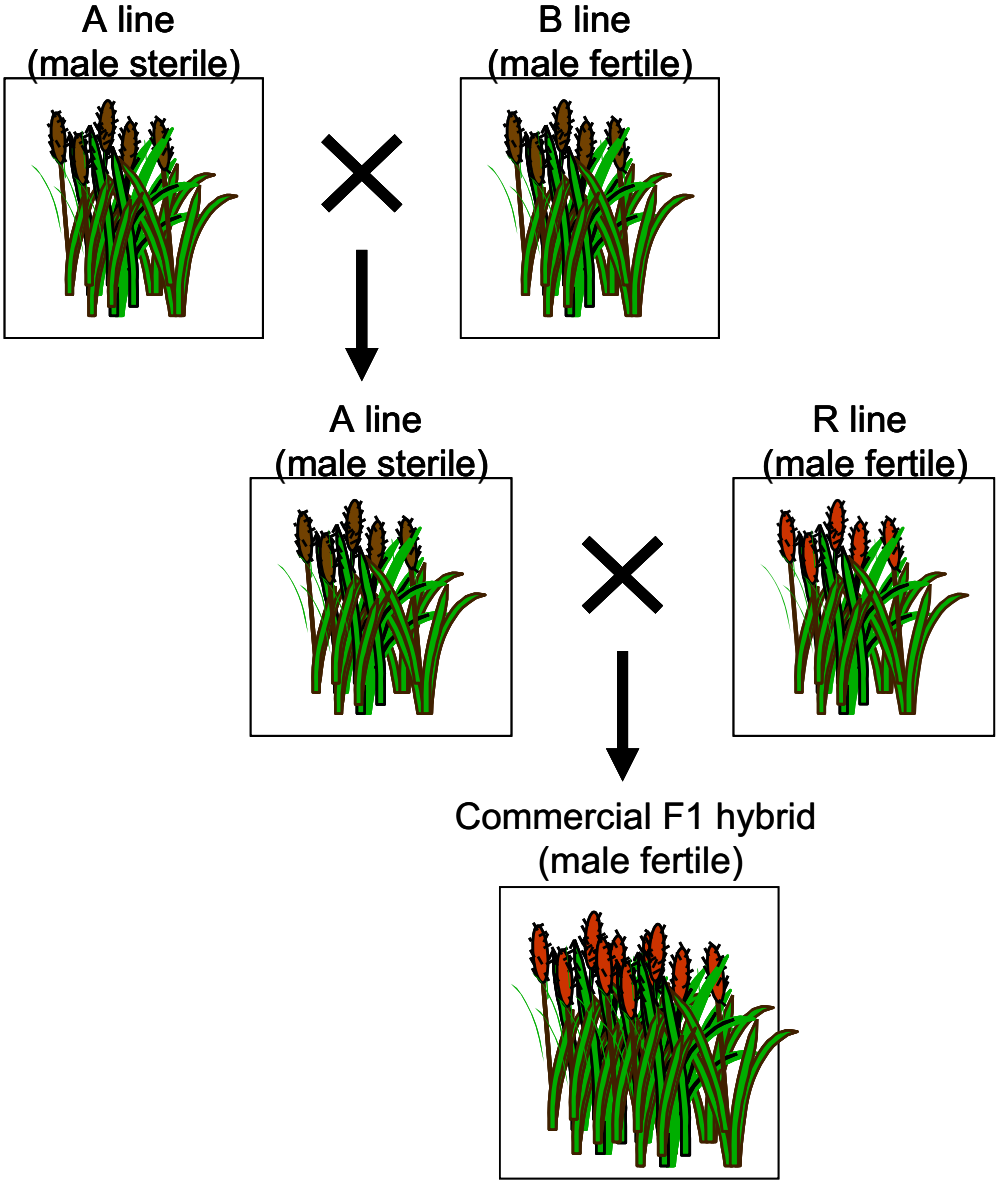
Colleen Hunt  
David Butler  
Brian Cullis

## Hybrid Crops

- Most plants have both male and female parts and therefore fertilize themselves.
- A hybrid is produced by taking the pollen from one plant and pollinating a different plant.
- Hybrid crops are developed because they produce more grain, fruit, or flowers than they would if they were left as an inbred.  
(sorghum, maize, rice, sunflowers)



# Hybrid seed Production



# Sorghum Breeding

- The DEEDI sorghum breeding team produces germplasm
  - to advance to the seed companies to make commercial varieties
  - To include in future crosses
- To assess the germplasm they need to create hybrids and use them in their breeding trials



# Aim

We need to find the very best statistical model to analyse the hybrids in order to assess the parents that will be used for future crosses.



# Mixed Model



The most basic model for fitting yield ( $y$ ) for  $i$  genotypes with  $j$  reps is

1. Fixed site mean
2. Random genotype term with mean 0 and variance,  $\sigma_g^2$  known as the genetic variance.
3. Random error term with mean 0 and variance  $\sigma^2$

$$y_{ij} = \mu + u_i + e_{ij}$$

or

$$y = X\tau + Zu + e$$



# Including the pedigree effects

We partition the genotype effect into three potential parts

## 1. Additive effects

- How the parents perform on average

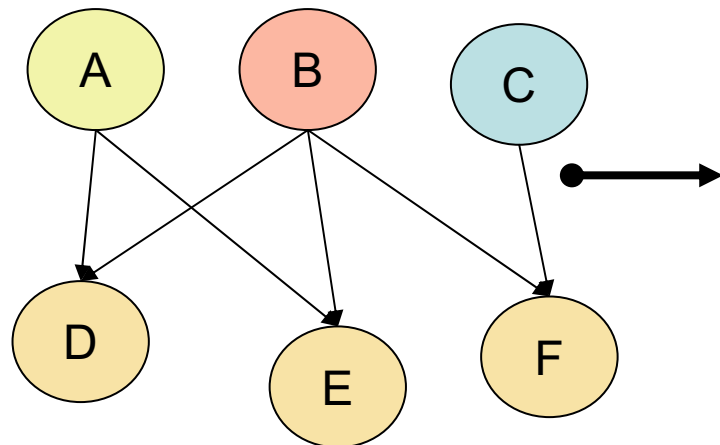
## 2. Dominance effects

- How the hybrid performs in comparison to the combination of parents performance

## 3. Residual genetic effects

## Additive effects

- Additive effects are obtained by including a relationship matrix (known as A) into our design matrix Z.
- This A matrix gives the model information on the relatedness of each genotype in the trial
- For example: parents A, B and C produced offspring D, E and F



	A	B	C	D	E	F
A	1	0	0	0.5	0.5	0
B		1	0	0.5	0.5	0.5
C			1	0	0	0.5
D				1	0.5	0.25
E					1	0.25
F						1



## Information for the A matrix

- In our sorghum trials we have planted D, E and F but we want to know about A, B and C
- We can create a pedigree file that tells us about all the hybrids in the trial, their parents, their grand-parents, their great grand-parents and so on
- We are fortunate with sorghum that we have all this information inside the PBMASS database.
- We can trace the ancestry back as far as around 20 generations

## The pedigree file

- For the simple A, B, C, D, E and F example our pedigree file would look something like this
- To get our A matrix needed for our model we need to do this

```
asreml.Ainverse (pedigreefile.data,...)
```

Geno	Female Parent	Male Parent
A	0	0
B	0	0
C	0	0
D	A	B
E	A	B
F	B	C

# Dominance



A hybrid is considered to be high dominant if its predicted value is greater than the combination of its parents additive effects

Predicted hybrid effects and additive effects

	A1	A1
B1	-3.66	-3.34
B2	4.18	-1.89

Dominance effects

	A1	A2
B1	-2.65	0.18
B2	2.90	-0.65

We have random effects, they are not so neat and tidy so we need to account for any extra residual genetic effects



## The D matrix

- The same pedigree file is used to create the dominance D matrix
- It is calculated using a monte carlo simulation approach
- `asreml.monte (pedigreefile.data, ...)`

# Example Data



- The trial has 471 genotypes
  - This includes 455 test genotypes and 16 commercial entries
  - The 455 test genotypes are made up from combinations of 5 male parents and 163 female parents
  - The 163 females are F4, F5, F6, F7, F8, F9, FF with some females having multiple generations in the trial
- The pedigree file has 890 entries.
- After fitting the model we will get
  - 890 additive effects
  - 630 dominance effects (see later)
  - 890 residual genetic effects (with parents given values of 0)

We re-write our original model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\tau} + \mathbf{Z}_g\mathbf{g} + \mathbf{Z}_u\mathbf{u} + \mathbf{e}$$

Where  $\mathbf{g}$  contains the genotype terms and  $\mathbf{u}$  is the rest.

Now we partition  $\mathbf{g}$  into the three genetic terms

$$\mathbf{y} = \mathbf{X}\boldsymbol{\tau} + \mathbf{Z}_a\mathbf{a} + \mathbf{Z}_d\mathbf{d} + \mathbf{Z}_l\mathbf{l} + \mathbf{Z}_u\mathbf{u} + \mathbf{e}$$


Where  $\mathbf{a}$  contains the additive terms,  $\mathbf{d}$  is dominance and  $\mathbf{l}$  are the residual genetic effects



$\mathbf{Z}_a$  and  $\mathbf{Z}_l$  are equal to  $[\mathbf{0} \ \mathbf{Z}_g]$

To better deal with parental lines that have zero dominance we remove these from the D matrix to make it easier to calculate D inverse

$\mathbf{Z}_d$  is the design matrix for the non - zero dominance lines


$$\mathbf{a} \sim N(\mathbf{0}, \sigma_a^2 \mathbf{A})$$

$$\mathbf{d} \sim N(\mathbf{0}, \sigma_d^2 \mathbf{D})$$

$$\mathbf{l} \sim N(\mathbf{0}, \sigma_l^2 \mathbf{I})$$



## In terms of asreml

```
asreml (yield~1,  
        random=~ped (Geno) +ped (DGeno) +ide (Geno) ,  
        ginverse=list (Geno=aytf.ginv, DGeno=aytf.dinv) ,  
        .....)
```



## Results

- We fit this model to 6 sites of data
- Using a REMLRT we can see a vast improvement when fitting a Pedigree Dominance model

	REMLlogl for Genotype model	REML logl for Pedigree Dominance Model	REMLLRT	P-value
Biloela	-186.89	-142.71	88.36	<0.001
Dalby Box	-219.79	-142.39	154.8	<0.001
Hermitage	-217.87	-194.24	47.26	<0.001
Kilcummin	-43.47	-26.86	33.22	<0.001
Liverpool Plains	-271.48	-230.11	82.74	<0.001
Springsure	-274.84	-243.87	61.94	<0.001

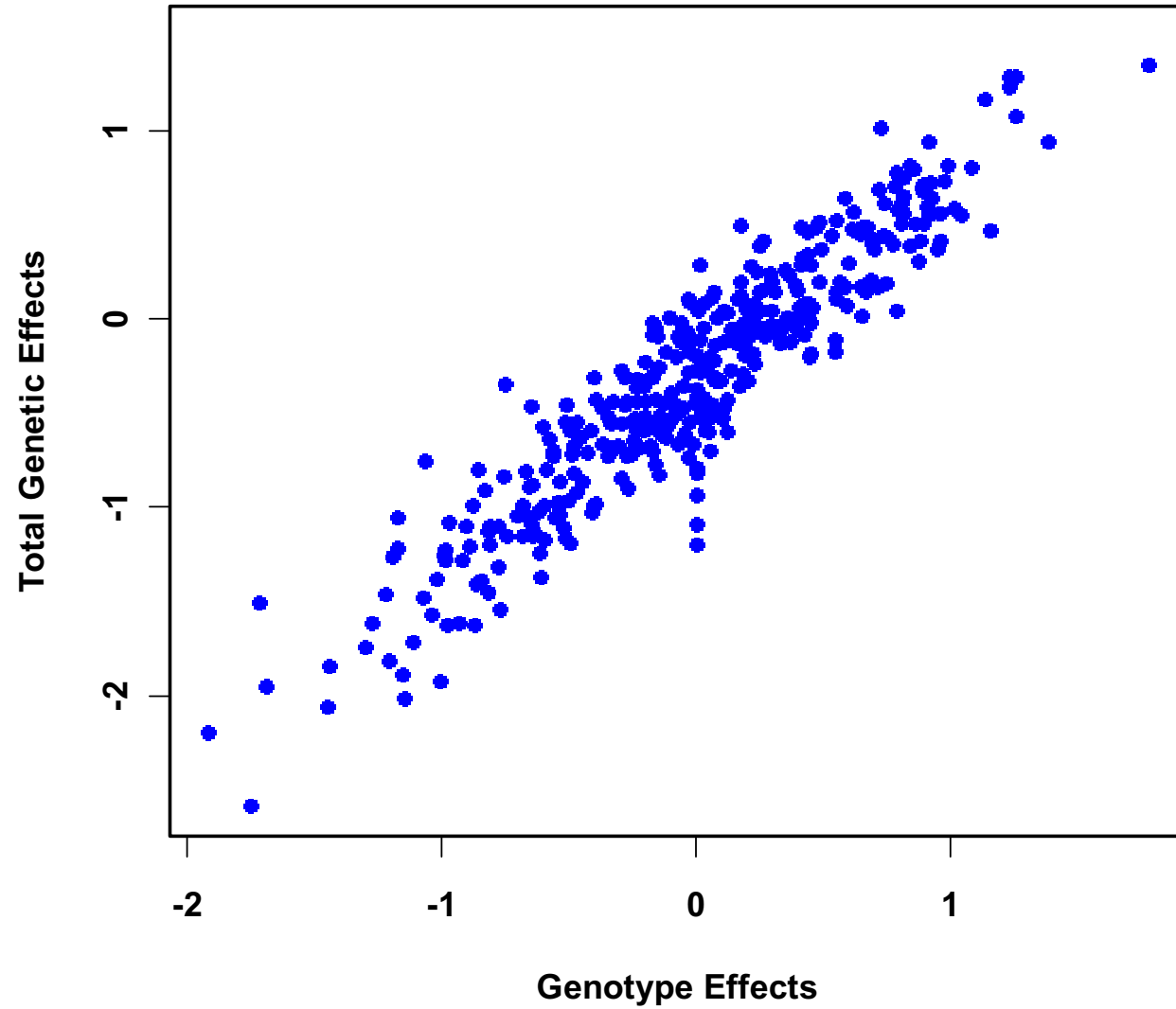
# Genetic Variances



	<b>mean t/ha</b>	<b>Geno Variance</b>	<b>Additive</b>	<b>Dominance</b>	<b>Residual Genetic</b>
Biloela	4.27	0.278	0.422	0.015	0.039
Dalby Box	6.93	0.43	0.512	0.139	0.000
Hermitage	10.51	0.582	0.756	0.164	0.165
Kilcummin	3.01	0.181	0.193	0.000	0.062
Liverpool Plains	9.9	1.053	0.880	0.207	0.257
Springsure	4.4	0.125	0.174	0.266	0.000

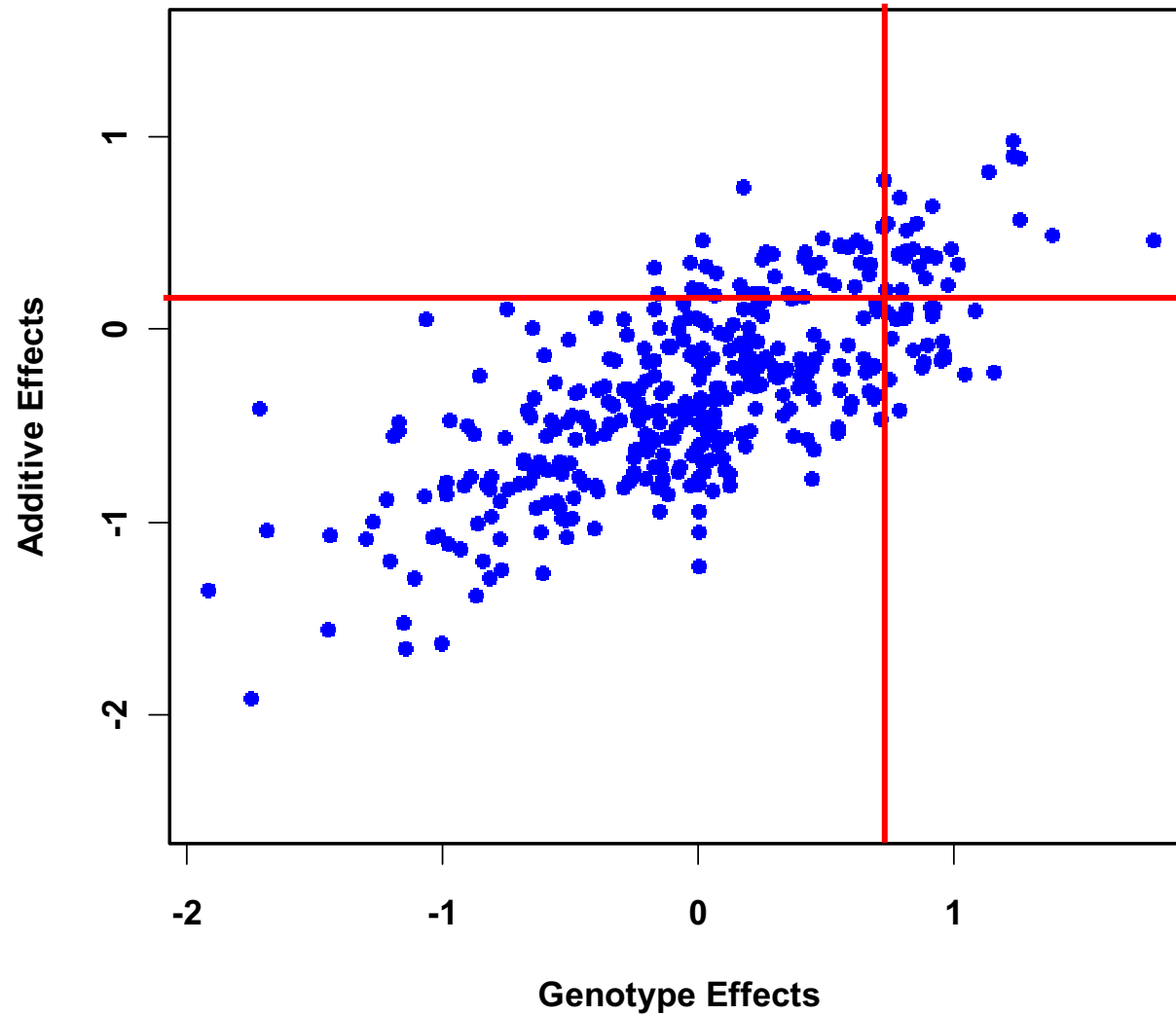
# How different are the blups?

Hermitage



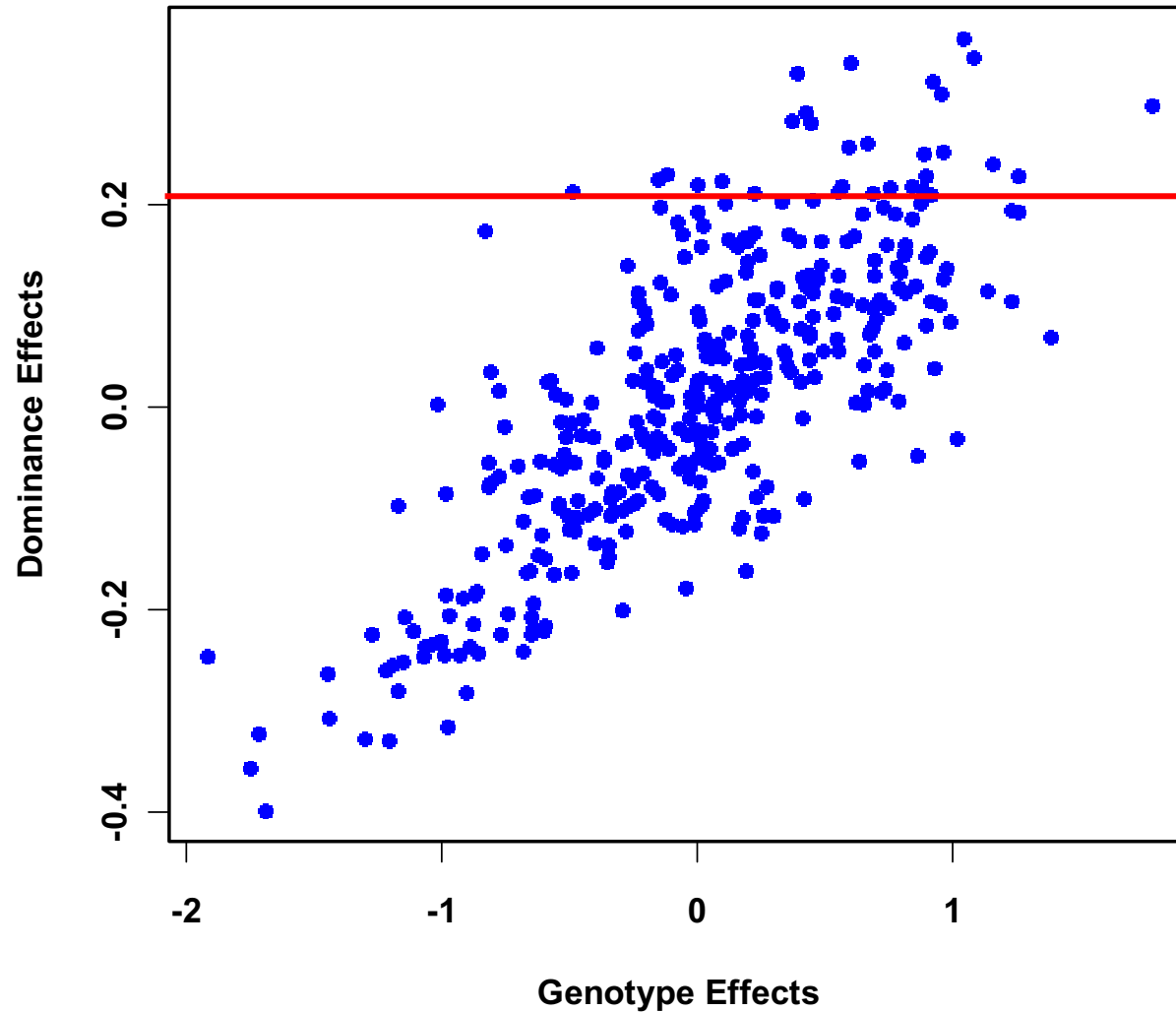
# Additive effects

Breeders  
select the best  
parents from  
their additive  
effects

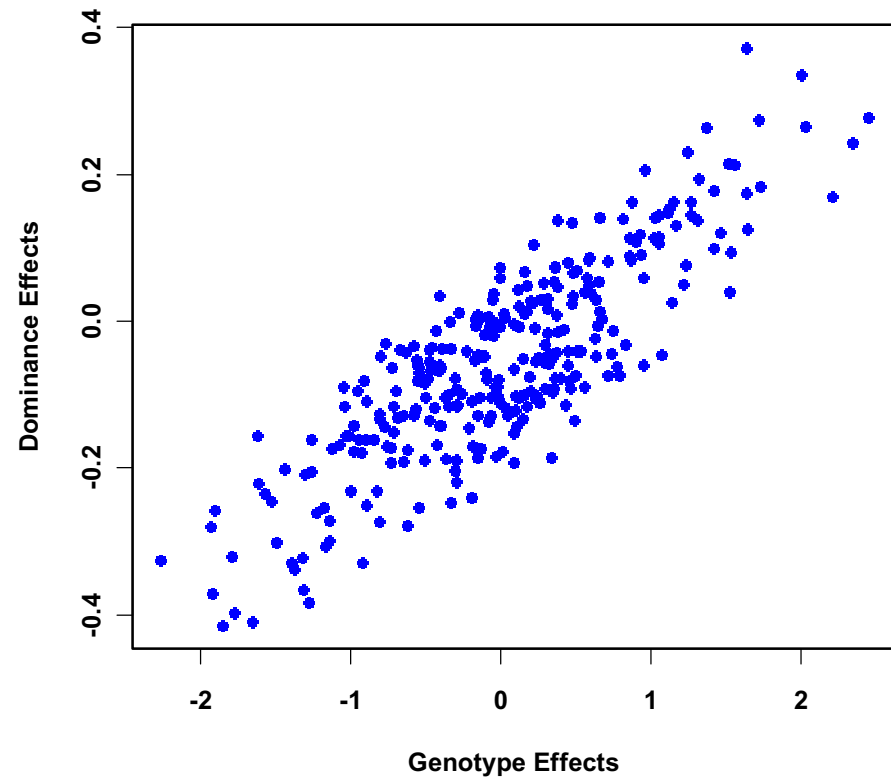
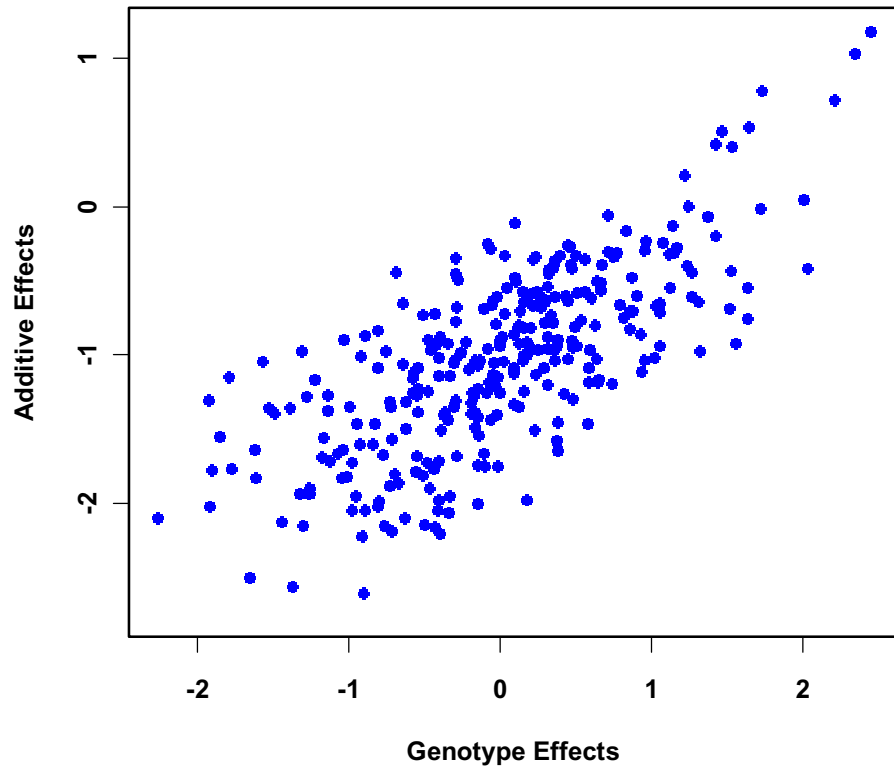


# Dominance effects

A line with high dominance effect performs better than the combination of parental effects



# Liverpool Plains





## Multi-Environment Analysis

- We can see from the individual analyses that each site has a different amount of additive, dominance and residual genetic effects.
- We need to account for possible genotype by trial interactions for each of the three genetic partitions.
- Site correlations will differ for each genetic part.



## In terms of asreml

### Single site

```
asreml (yield~1,  
        random=~ped (Geno) +ped (DGeno) +ide (Geno) ,  
        ginverse=list (Geno=aytf.ginv, DGeno=aytf.dinv) ,  
        .....)
```

### MET

```
asreml (yield~site,  
        random=~diag (site) :ped (Geno) +  
                diag (site) :ped (DGeno) +diag (site) :ide (Geno) ,  
        ginverse=list (Geno=aytf.ginv, DGeno=aytf.dinv) ,  
        .....)
```

# Comparing MET models



	REMLlogl
diag(site):Geno	-1173.77
fa(site,1):Geno	-1071.71
fa(site,2):Geno	-1029.18
diag(site):ped(Geno)+ diag(site):ide(Geno)	-967.33
fa(site,1):ped(Geno)+ fa(site,1):ide(Geno)	-883.47
diag(site):ped(Geno)+ diag(site):ped(Dgeno)+ diag(site):ide(Geno)	-961.65
fa(site,1):ped(Geno)+ fa(site,1):ped(Dgeno)+ fa(site,1):ide(Geno)	Watch this space!

## Conclusion

- For hybrid crops always analyse using the pedigree information

