

Using Pedigree Founders Only Or Using All Pedigree Members?

- A Comparison of SNP-Expression
Regression Results

**Young Ju Suh¹, Hye-Soon Lee^{2,3},
Franak Batliwalla², Wentian Li²**

*1. Department of Preventive Medicine, College of Medicine, **Ewha Womans University**, Seoul, Korea. 2. The Robert S. Boas Center for Genomics and Human Genetics, **Feinstein Institute for Medical Research**, North Shore LIJ Health System, Manhasset, NY, USA. 3. Department of Internal Medicine, **Hanyang University Medical College**, Seoul, Korea.*

GOALS OF OUR STUDY

Three different approaches for SNP-expression association analysis for expression quantitative trait (eQT) when there is a family-specific environmental effect, or a genetic effect unrelated to the SNP being tested:

- 1. Use uncorrelated founders only.**
- 2. Use all pedigree members without any treatment of the correlated samples.**
- 3. Use all pedigree members for a linear mixed model.**

We aim at examining whether the result (R^2 , p-values, etc.) strongly, or only weakly, depend on which approach is adopted.

ABSTRACT

We observed that SNP-expression linear regression analyses by **using founders only**, and by **using all pedigree members**, may lead to different significance of association signal as well as different conclusions. Regression analyses based all pedigree members can be carried out with a linear mixed model, either assuming **one random effect** (in intercept) or **two** (in both intercept and slope). Using linear mixed model may greatly improve the data-fitting result (e.g. much higher R^2 value), but it may not improve the significance. These observations lead to the recommendation of further confirmation studies for any claimed SNP-expression association based on one dataset.

METHODS

- **Single Nucleotide Polymorphisms (SNPs)**

Use a reduced number SNP's = 2,263:

- * To be on autosomal chromosomes.
- * To have all three genotypes present in all 194 samples.
- * To have the minor allele frequency > 0.1 .

METHODS

- **Expression Quantitative Trait (eQT)**

Focus on limited number of eQT's from the sources:

- * The **27 eQT's** selected for having the strongest *cis* signal in genome-wide linkage analysis (listed in Cheung et al. paper).
- * **3 eQT's** that are of interests to our investigation of rheumatoid arthritis, two at HLA-DRB1 and one at PTPN22.
- * From a genome-wide association analysis of $2,263 \times 3554 = 8,042,702$ linear regression models on the 56 pedigree founders, we selected **47 eQT** which has at least one SNP to have $p\text{-value} < 4.4 \times 10^{-6} = 0.01/2,263$ (significance level of 0.01 with a Bonferroni correction).

METHODS

- **Subjects of The Study**

Case 1: 56 uncorrelated founders only.

Case 2: 194 all pedigree members.

METHODS

- Linear Regression (LR) And Linear Mixed Model (MM)

Fit a regression model for $y = \log_2(\text{expression})$ as a function of $x = 0, 1, 2$ copies of the minor allele:

LR: $y_j = a + bx_j + \varepsilon_j$

MM1: $y_{ij} = a + \varepsilon_i + bx_{ij} + \varepsilon_{ij} = a + bx_{ij} + \varepsilon_i \cdot 1 + \varepsilon_{ij}$

MM2: $y_{ij} = a + \varepsilon_i + (b + \delta_i)x_{ij} + \varepsilon_{ij} = a + bx_{ij} + \varepsilon_i \cdot 1 + \delta_i \cdot X_{ij} + \varepsilon_{ij}$

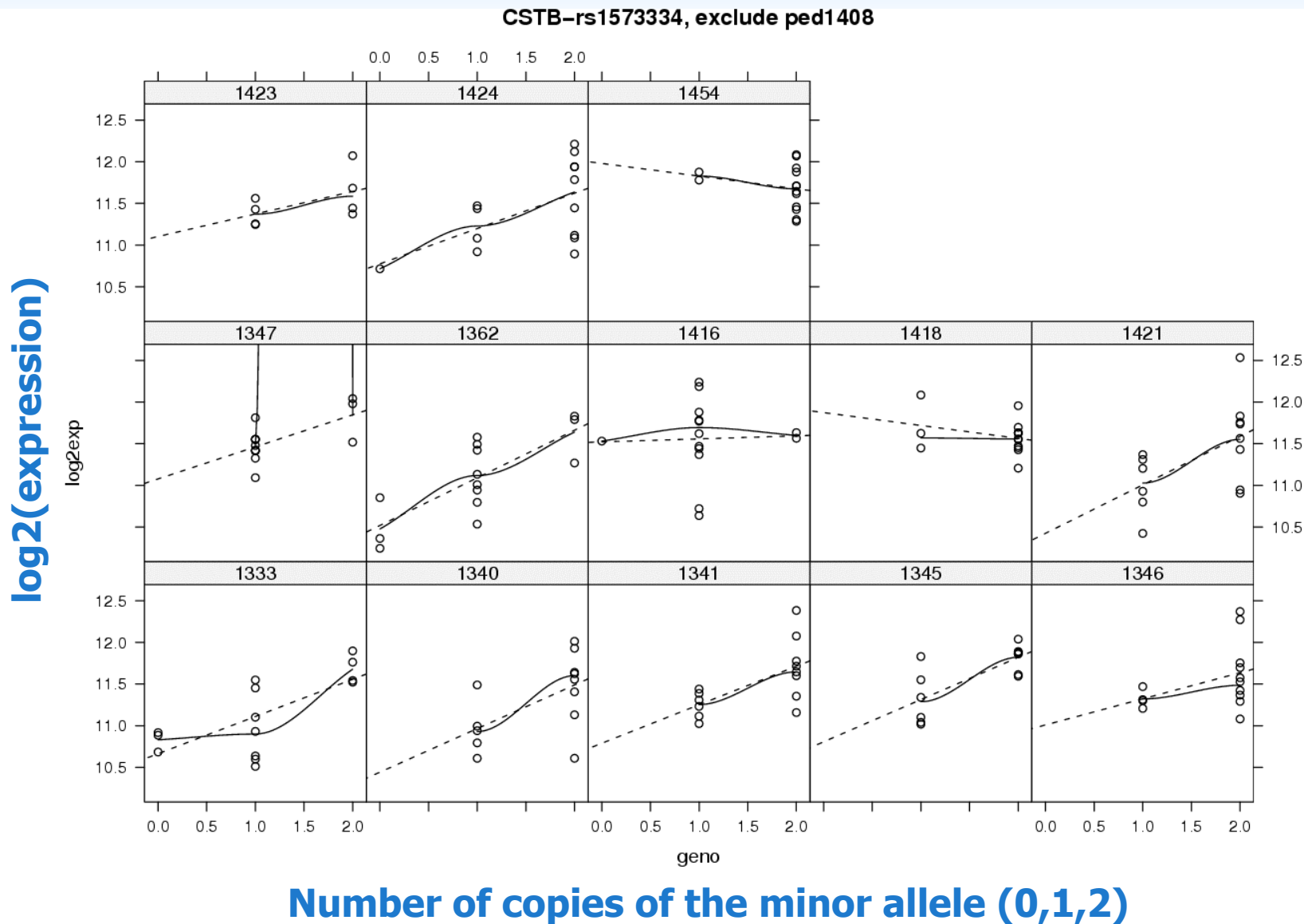
where MM1 and 2 are linear mixed models with a random effect **on the intercept** a , and with a random effects **on both intercept** a **and slope** b , respectively (i is the pedigree index and j is the sample index.)

RESULTS

- **The Strongest Association Signal in CSTB-rs157334 Pair**

The CSTB-rs157334 pair provides the strongest association signal (for the 56-sample dataset) for all 27 eQT's listed in [Cheung et al. 2005].

Figure 1. SNP-expression regression analysis for each pedigree for the CSTB-rs157334 pair.



RESULTS

- **Comparisons**

1. To compare LR on 56-sample set (Case 1) and LR on 194-sample set (Case 2).
2. To compare LR and MM models for the 194-sample set (within Case 2).

Table 1. Results for various regression models (LR, MM1, MM2) for SNP-expression pairs with the most significant p-values on the 56-sample dataset for the 27 eQT's, and for the 47 eQT's from our own regression analysis, plus two eQT's of interests to the study of rheumatoid arthritis.

expression-SNP	N=56		N=194					
	LR		LR		MM1		MM2	
	R^2	p -value	R^2	p -value	R^2	p -value	R^2	p -value
CSTB-rs157334	.43,.40	3.3×10^{-8}	.26,.23	8.3×10^{-10}	.44,.40	7.9×10^{-13}	.50,.48	5.3×10^{-6}
CSTB-rs1999811	.32,.30	4.5×10^{-6}	.02,.02	0.043	.34,.33	0.0072	.34,.33	0.0072
HSD17B12-rs1334334	.35,.29	1.8×10^{-6}	.21,.18	3.2×10^{-11}	.47,.44	1.9×10^{-6}	.49,.47	0.00045
DDX17-rs243404	.32,.31	7.4×10^{-6}	.03,.03	0.012	.33,.33	0.036	.33,.33	0.036
HLADRB2-rs1395579	.36,.35	9.4×10^{-6}	.09,.08	8.4×10^{-5}	.55,.48	0.042	.59,.53	0.12
IFRD1-rs1079549	.44,.39	2.5×10^{-8}	.02,.02	0.027	.33,.32	0.0054	.33,.32	0.0053
MT1H-rs1423508	.43,.40	5.1×10^{-8}	.002,.002	0.52	.33,.32	0.40	.33,.32	0.40
CEBPD-rs1363062	.39,.29	2.8×10^{-7}	.049,.036	0.0022	.43,.41	0.0051	.46,.45	0.019
PTPN22-rs984068	.21,.21	0.00042	.002,.005	0.52	.48,.49	0.84	.52,.51	0.84
HLADRB1-rs941838	.31,.28	5.9×10^{-5}	.006,.011	0.30	.40,.43	0.55	.40,.43	0.55



Two R^2 values are presented: one for $y = \log_2(\text{expression})$ and another for $y = \text{expression}$. The p -value is for testing $b = 0$.

RESULTS

- The mixed models lead to a better data-fitting performance with R^2 . The better performance also leads to smaller AIC and BIC values for the MM models, similar to what was observed in [Yu et al., 2006].
- The CSTB-rs157334 and HSD17B12-rs1334334 pairs are perhaps the only SNP-expression pairs whose significant association is confirmed in the larger dataset (194 samples); association in other pairs are *not* confirmed.

DISCUSSION

- This observation is somewhat surprising, because there are several reasons to believe that the 194-sample dataset would reproduce the conclusion from the 56-sample dataset with higher significance:
 - 1.** The extra samples in 194-sample dataset are offsprings of those in the 56-sample dataset, so if there is a **relativeness in their eQT's**, it will only **reinforce the association signal**.
 - 2.** With a **larger sample size** in the 194-sample dataset, we should expect a smaller p-value.

DISCUSSION

- However, out of the 27 SNP-eQT pairs from [Cheung et al. 2005], only 4 could be considered “interesting” (significant and better p-values in the 194-sample dataset); and out of the 47 SNP-eQT pairs selected by our genome-wide run, the only two interesting pairs are already in the above-mentioned list.

CONCLUSION

- We need to be cautious in claiming SNP-eQTL association when a significant result is only observed in one small dataset.
- Pedigree members from pedigrees whose founders are used as the first round analysis may have already provided a dataset for confirmation.