Starting Genetic Imaging Analyses with SOLAR-Eclipse

Peter Kochunov Ph.D.
pkochunov@gmail.com

CBIR
UNIVERSITY OF MARYLAND
CENTER FOR BRAIN IMAGING RESEARCH

UNIVERSITY of MARYLAND
SCHOOL OF MEDICINE

TEXAS BIOMEDICAL
RESEARCH INSTITUTE
FORMERLY THE SOUTHWEST FOUNDATION FOR BIOMEDICAL RESEARCH

To participate in the demo

- Please register and get HCP account
- http://hcpx-demo.humanconnectome.org
- The univariate demo
  - **http://tinyurl.com/mvyxcmq**
  - or
  - https://hcpx-demo.humanconnectome.org/app/template/UnivariatePolygenic.vm
- The bivariate demo
  - **http://tinyurl.com/ltqf8sj**
  - or
  - https://hcpx-demo.humanconnectome.org/app/template/GeneticCorrelation.vm

Introduction

- What is SOLAR-Eclipse
- Downloading and installing SOLAR-Eclipse
- Creating a solar analysis directory
  - Pedigree file
  - Phenotype file
- Common analyses types
  - Heritability
  - Genetic Correlations
    - Get this file: www.mdbrain.org/personalpages/peter/lpa.tar

SOLAR-Eclipse

- Extension of SOLAR for imaging genetics
- Developed for multiplatform (pc/mac/linux)
  - –Genetic analysis of discreet and continuous traits
  - –Supports All Common Genetic Analyses for Continuous and Discrete traits.
    - Heritability
    - Genetic Correlation
    - Quantitative trait Linkage
    - GWAS analyses in related and unrelated samples
  - –Supports uni-and-multivariate analyses
  - –Supports discrete and continuous covariates

Main Strength – Genetic Analysis in Family and Twin Samples

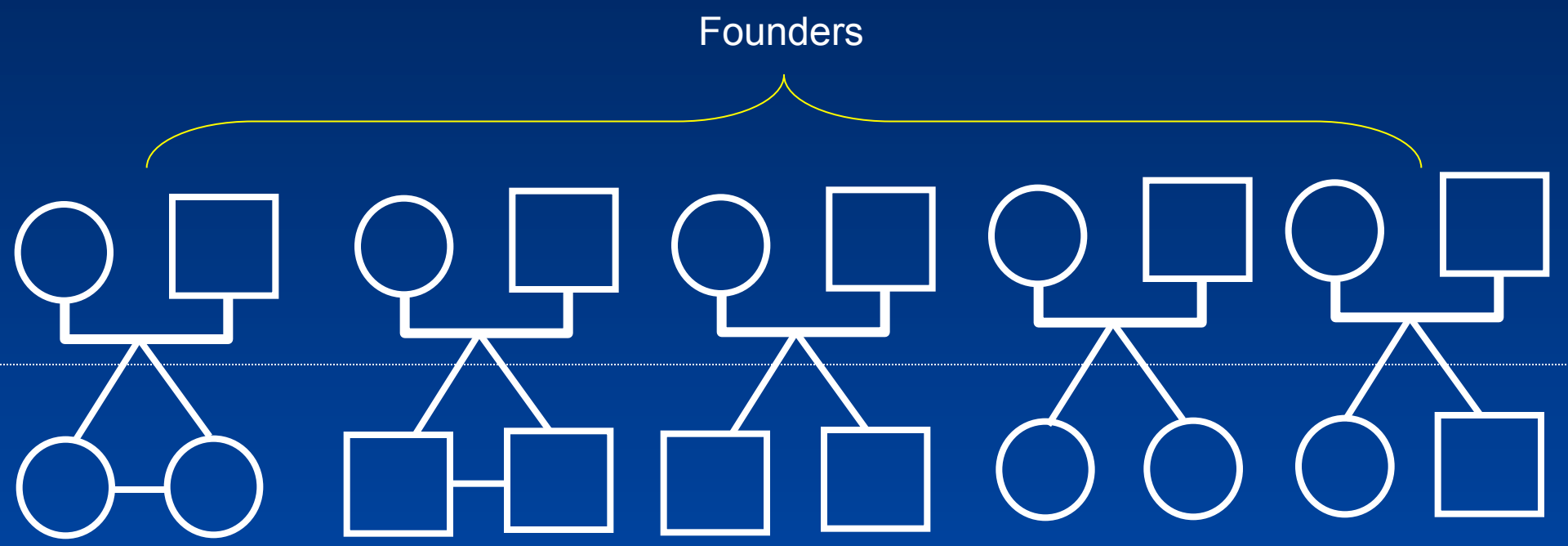
Downloading/Installing SOLAR

- Get it from NITRC website
  - http://www.nitrc.org/projects/se_linux/
  - Use the linux version for most of the features
    - Latest Apple/PC versions are compiled at request
- Email your user name to get the registration code.
  - solar@txbiomedgenetics.org
  - This code will work on any machine with the same user name
  - Registering for a specific domain is also available
- Manual/Instructional videos at http://www.mdbrain.org/solareclipse/

Develop Analysis Plan

- Background
  - Lipo-Protein A (LAP) stimulates lipid metabolism.
  - Specific polymorphisms may influence LPA1/2 levels measured in blood
- Questions
  - Are LPA1/2 concentrations in blood influenced by additive genetic effects?
  - To what extent is white matter integrity measured through DTI-FA determined by additive genetic effects?
  - What extent are the same genetic factors influencing both white matter integrity and LPA levels?
- Approach
  - Measure DTI-FA and [LPA1/2] in blood in N=60/50 MZ/DZ twin pairs.
  - Measure heritability of average FA values
  - Measure heritability of [LPA]
  - Calculate pleiotropy between two traits!

Our Twin Pedigree



Making a pedigree

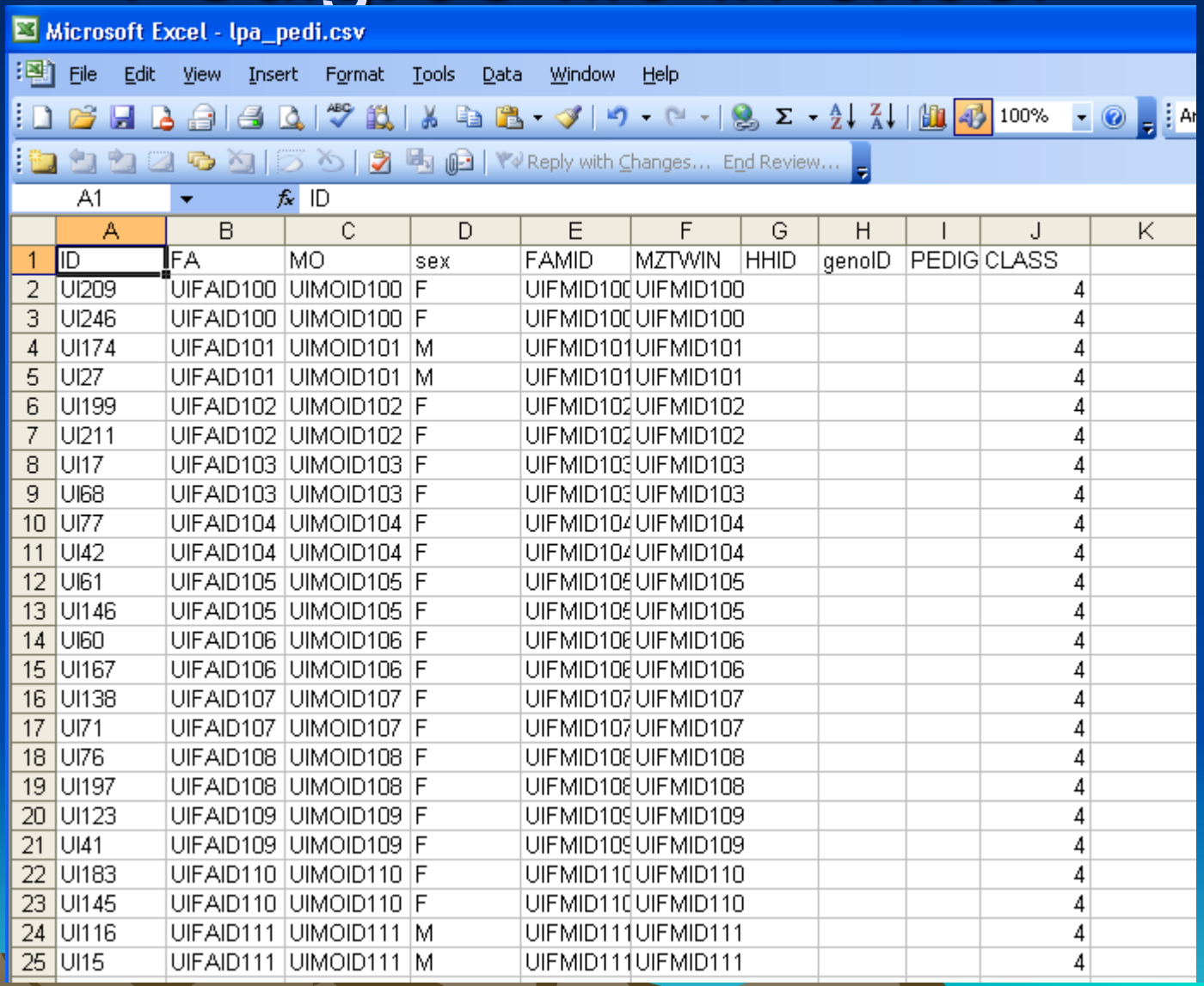
- A very important step.
- Pedigree contains “genetic information” based on relatives.
  - Expressed as kinship matrix
  - MZ twins share 100% genetic variance
  - DZ twins share 50% of genetic variance
- SOLAR pedigree must have “founders”
  - Founders are the basis of the pedigree and are assumed unrelated
  - Founder only pedigrees can be used for GWAS analysis of unrelated individuals.

Actually is a very simple process

ID	FA	MO	sex	FAMID	MZTWIN	HHID	genoID	PEDIGREE NUMBER	CLAS
UI209	UIFAID100	UIMOID100	F	UIFMID100	UIFMID100				
UI246	UIFAID100	UIMOID100	F	UIFMID100	UIFMID100				

- Best done in Excel
- Start coding one subject per line.
- ID – subject id (your choice)
- FA/MO are IDs for parents
  - will have to be included as founders even if they don't have phenotypes
- FAMID – optional family id
- MZTWIN – identifier for MZ twins. A twin pair is defined as subjects with same identifier.

Pedigree file in excel



	A	B	C	D	E	F	G	H	I	J	K
1	ID	FA	MO	sex	FAMID	MZTWIN	HHID	genoID	PEDIG	CLASS	
2	UI209	UIFAID100	UIMOID100	F	UIFMID100	UIFMID100				4	
3	UI246	UIFAID100	UIMOID100	F	UIFMID100	UIFMID100				4	
4	UI174	UIFAID101	UIMOID101	M	UIFMID101	UIFMID101				4	
5	UI27	UIFAID101	UIMOID101	M	UIFMID101	UIFMID101				4	
6	UI199	UIFAID102	UIMOID102	F	UIFMID102	UIFMID102				4	
7	UI211	UIFAID102	UIMOID102	F	UIFMID102	UIFMID102				4	
8	UI17	UIFAID103	UIMOID103	F	UIFMID103	UIFMID103				4	
9	UI68	UIFAID103	UIMOID103	F	UIFMID103	UIFMID103				4	
10	UI77	UIFAID104	UIMOID104	F	UIFMID104	UIFMID104				4	
11	UI42	UIFAID104	UIMOID104	F	UIFMID104	UIFMID104				4	
12	UI61	UIFAID105	UIMOID105	F	UIFMID105	UIFMID105				4	
13	UI146	UIFAID105	UIMOID105	F	UIFMID105	UIFMID105				4	
14	UI60	UIFAID106	UIMOID106	F	UIFMID106	UIFMID106				4	
15	UI167	UIFAID106	UIMOID106	F	UIFMID106	UIFMID106				4	
16	UI138	UIFAID107	UIMOID107	F	UIFMID107	UIFMID107				4	
17	UI71	UIFAID107	UIMOID107	F	UIFMID107	UIFMID107				4	
18	UI76	UIFAID108	UIMOID108	F	UIFMID108	UIFMID108				4	
19	UI197	UIFAID108	UIMOID108	F	UIFMID108	UIFMID108				4	
20	UI123	UIFAID109	UIMOID109	F	UIFMID109	UIFMID109				4	
21	UI41	UIFAID109	UIMOID109	F	UIFMID109	UIFMID109				4	
22	UI183	UIFAID110	UIMOID110	F	UIFMID110	UIFMID110				4	
23	UI145	UIFAID110	UIMOID110	F	UIFMID110	UIFMID110				4	
24	UI116	UIFAID111	UIMOID111	M	UIFMID111	UIFMID111				4	
25	UI15	UIFAID111	UIMOID111	M	UIFMID111	UIFMID111				4	

Lpa_pedi.csv. CSV stands for comma separated value file format

Continued

ID	FA	MO	sex	FAMID	MZTWIN	HHID	genoID	PEDIGREE NUMBER	CLASS
UI209	UIFAID100	UIMOID100	F	UIFMID100	UIFMID100				4
UI246	UIFAID100	UIMOID100	F	UIFMID100	UIFMID100				4

- HHID
  - Household id. Use for study of household effects
  - Use it if you have twins living in different households
- genoID
  - Is not commonly used
- Pedigree Number
  - Is not commonly used in humans
- CLASS
  - Important for mega-genetic analysis. Covariates and other normalizations are performed per class
  - If you plan to combine multiple studies assign them non-overlapping class valus

Code founders like this

ID	FA	MO	sex	FAMID	MZTWIN	HHID	genoID	PEDIGREE	CLASS
UIFAID1			M	UIFMID1					4
UIFAID2			M	UIFMID2					4

- Founders are identified as “orphaned” subjects.
  - Make sure you have the right gender
- Save the file in the .csv format.
  - On the mac use “Windows CSV” option
  - Otherwise, “^L” is used to identify end of line

Fire up solar-eclipse

```
[peterk@medusa LPA]$
[peterk@medusa LPA]$
[peterk@medusa LPA]$
[peterk@medusa LPA]$ solar
```

Solar command starts the tcl shell with R-like interface

```
SOLAR Eclipse version 7.5.3 (Experimental), last updated on May 13, 2014
Copyright (c) 1995-2014 Texas Biomedical Research Institute
Enter help for help, exit to exit, doc to browse documentation.

solar> ls
```

Solar inherits all bash commands, so “ls” gives listing of directory

```
lpa_pedi.csv  lpa_pheno.csv
solar>
```

Lpa_pedi is the file we just created. Lpa_pheno doesn’t exist yet!

heron : peterk | 10.0.4.76 : | login1 : petr | test_medx : bash | 10.0.4.76 : petr

Load your pedigree

```
[peterk@medusa LPA]$
[peterk@medusa LPA]$
[peterk@medusa LPA]$
[peterk@medusa LPA]$ solar

SOLAR Eclipse version 7.5.3 (Experimental), last updated on May 13, 2014
Copyright (c) 1995-2014 Texas Biomedical Research Institute
Enter help for help, exit to exit, doc to browse documentation.

solar> ls
lpa_pedi.csv  lpa_pheno.csv  phenotypes.info
solar> load pedi lpa_pedi.csv
Loading pedigree data from the file lpa_pedi.csv ...
solar> ls
house.gz  lpa_pedi.csv  lpa_pheno.csv  pedigree.info  pedindex.cde  pedindex.out  phenotypes.info  phi2.gz
solar>
```

"Load pedi" command with the name of the file

Several new files will be created

heron : peterk | 10.0.4.76 : | login1 : petr | test_medx : bash | 10.0.4.76 : petr

SOLAR creates several internal files that re-arrange pedigree in computation friendly format

Hint – phi2.gz is the kinship matrix in text format that can be used for other packages

Obtaining phenotypes

- DTI-FA values we calculated using ENIGMA-DTI pipeline
  - http://enigma.ini.usc.edu/protocols/dti-protocols/
  - Average whole brain FA values were used
- LPA 1 and 2 levels were obtain from blood
  - 12 hour fasting
  - Measured using Cardio IQ™ Ion Mobility assay
    - Unknown to me and hence presumed very accurate

Lets make a phenotype file

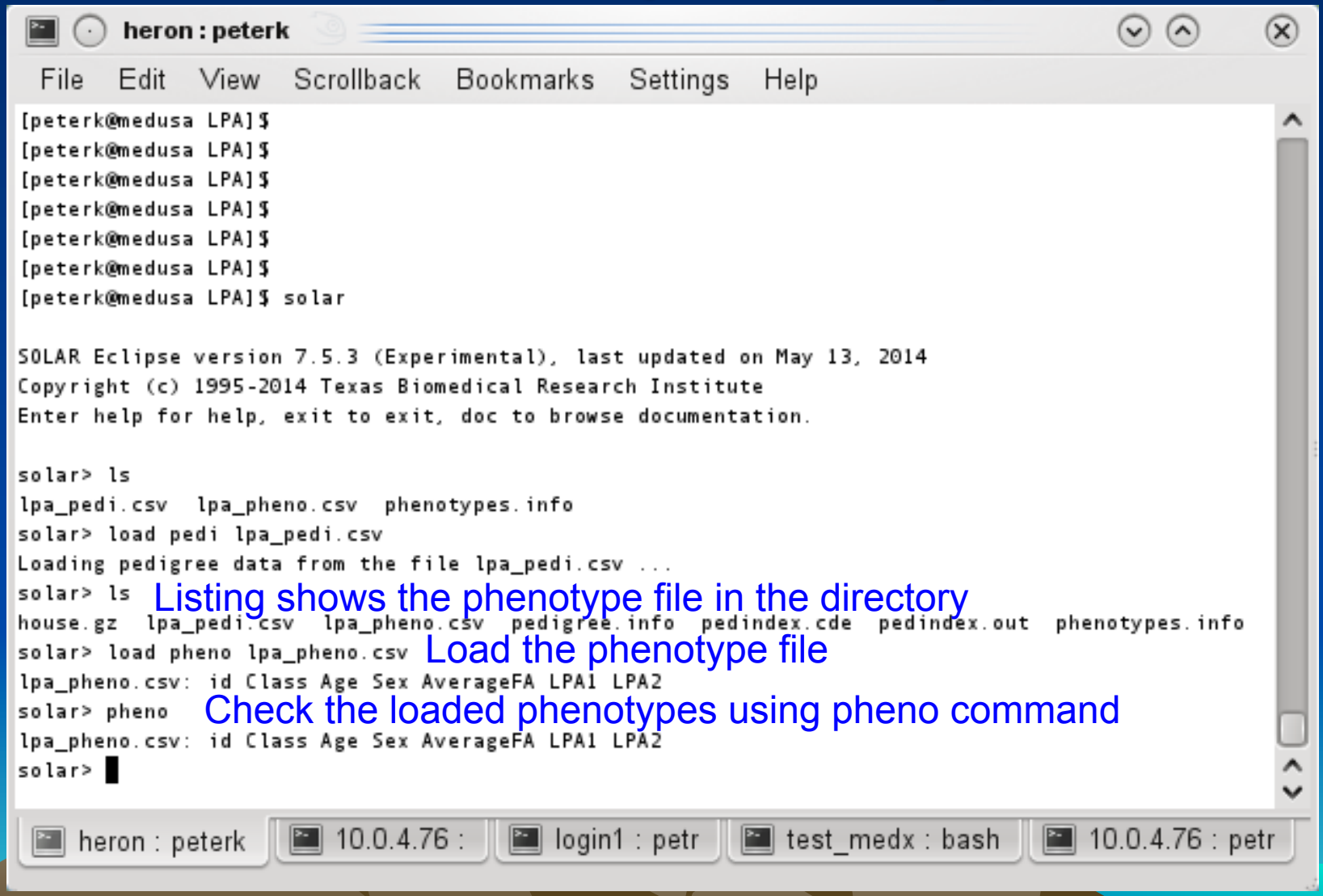
- Phenotypes are stored in a simple csv format
- Subject ID should match IDs in pedigree file
- Multiple phenotypes files can be loaded at once and merged in memory
  - Convenient for storing fixed factors such as sex and age for the entire pedigree
- Load phenotype command is
  - “load pheno file_name.csv”
  - http://www.mdbrain.org//solareclipse/solar_commd.html
  - Type “pheno” to check that phenotypes were loaded

Phenotype file

- Let put average FA values measured from DTI
- Blood-levels of two LPA proteins LPA1 and LPA2
- Add sex and gender
- Simple CSV format.

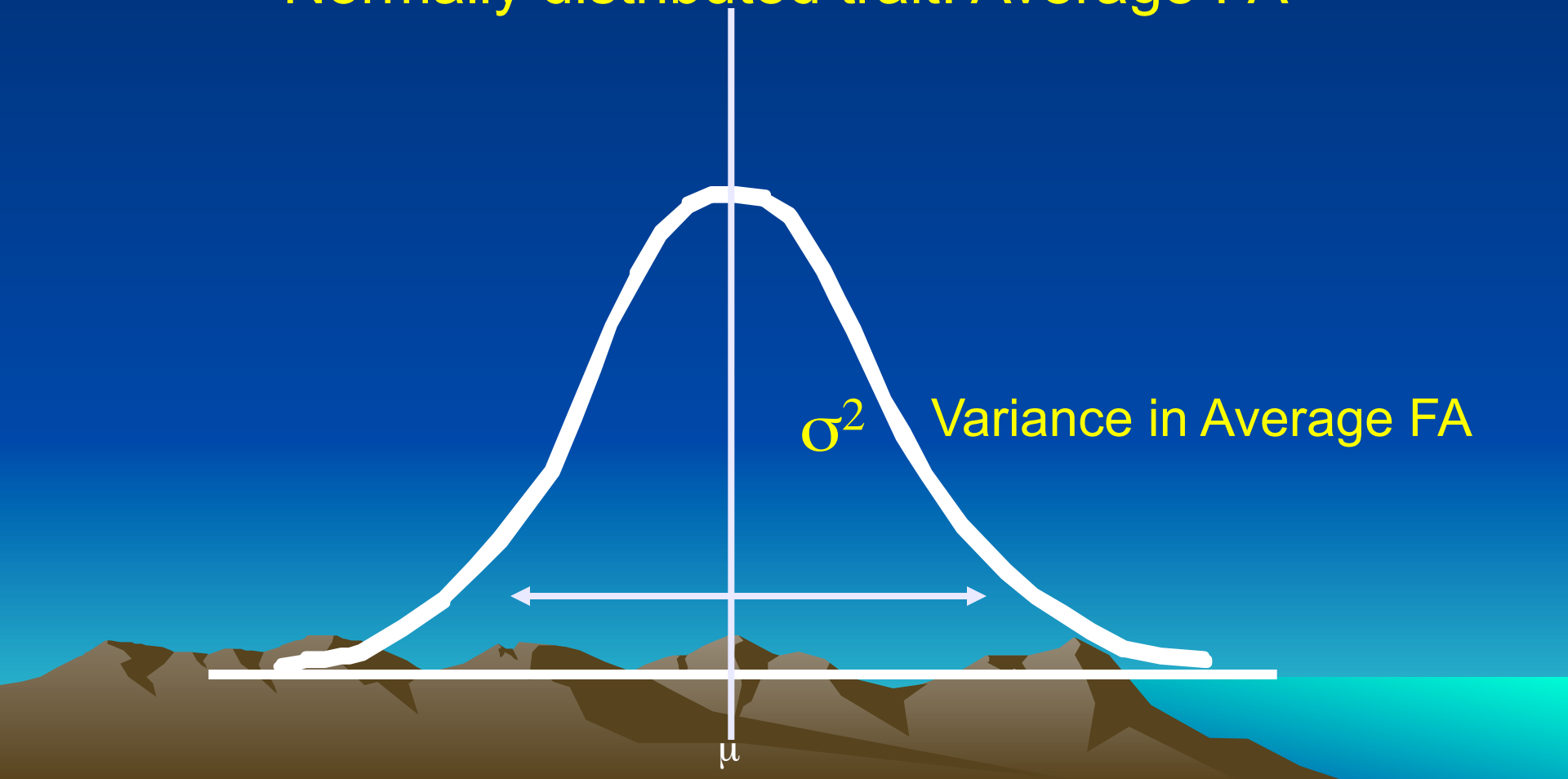
id	Age	Sex	AverageFA	LPA1	LPA2
UI101	38	F	0.489995	6.27615	1.191184
UI102	38	F	0.486044	5.73763	1.05002
UI103	45	F	0.475386	6.42397	1.13037
UI10	35	F	0.455976	5.40412	1.11781
UI104	19	F	0.493612	6.20166	1.01562
UI105	56	F	0.457284	6.03059	1.14993
UI106	35	F	0.471036	1.96892	1.05707
UI107	36	F	0.498711	6.31608	1.17105
UI108	32	F	0.485737	6.12398	1.07816

Load the phenotype file



Calculating heritability

- What is heritability?



Variance Decomposition

$$\sigma_p^2 = \sigma_g^2 + \sigma_e^2$$

σ_p^2 = Total phenotypic variance

σ_g^2 = Variance due to genetic sources

σ_e^2 = Variance due to environmental source

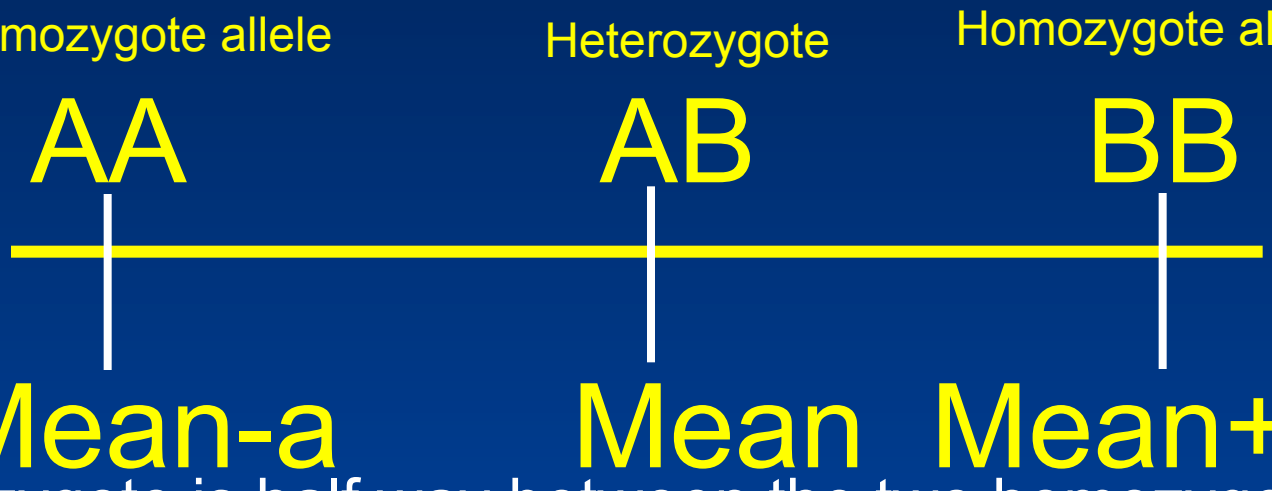
Genetic variance is due to

$$\sigma_g^2 = \sigma_a^2 + \sigma_d^2$$

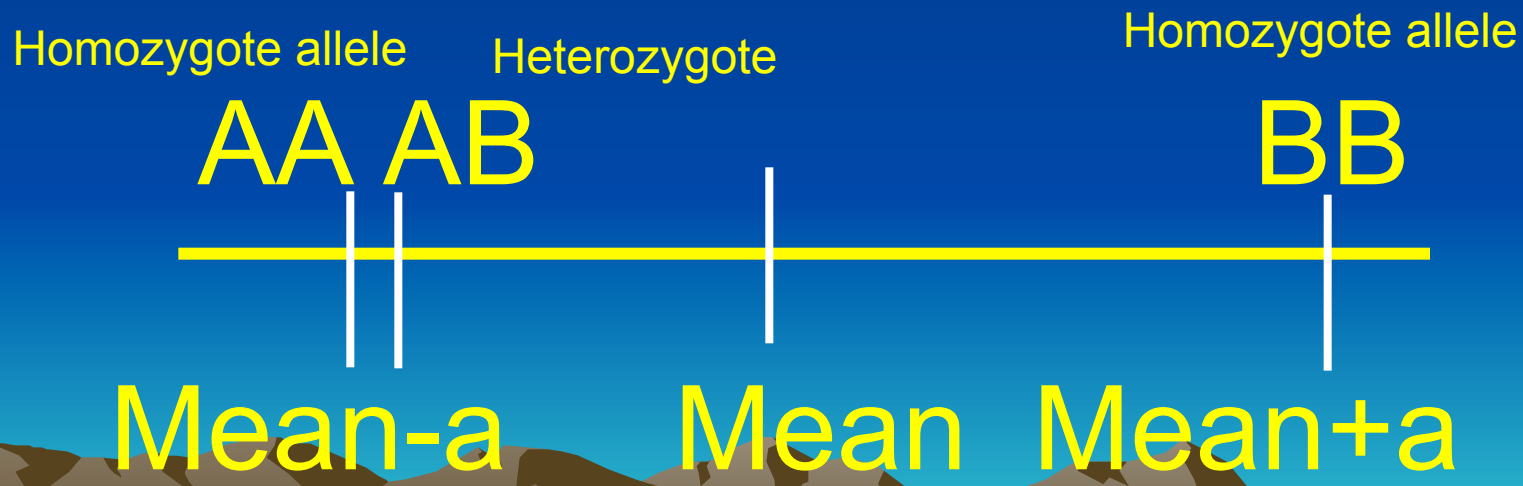
σ_a^2 = Additive genetic variance

σ_d^2 = Dominance variance

Additive genetic effect



heterozygote is half way between the two homozygotes,



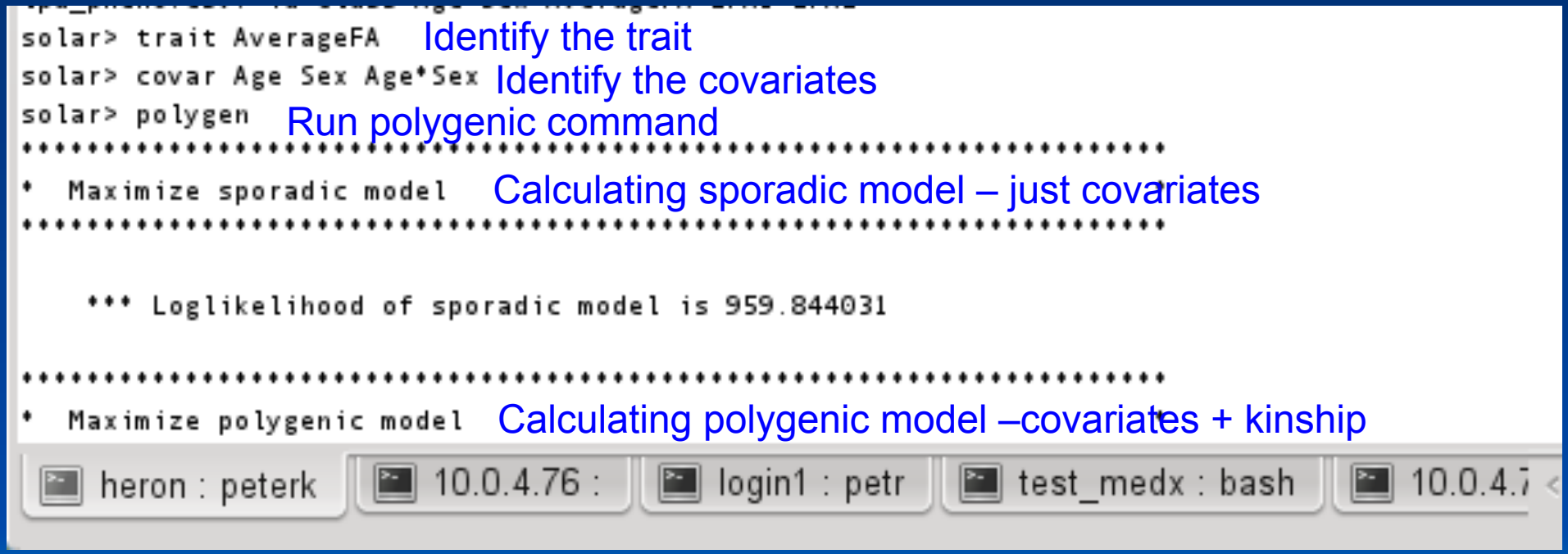
Dominant genetic effect: A dominates B

Additive genetic (narrow sense) heritability

$$h^2 = \frac{\sigma_a^2}{\sigma_p^2}$$

Fraction of the total variance in a trait explained by additive genetic variance

How to calculate it?



Polygenic function tests two GLM models based on their log-likelihood values

Sporadic model assumes no genetic effects (unitary kinship matrix)

Polygenic model assumes genetic effects (non-unitary kinship matrix)

Output

heron : peterk

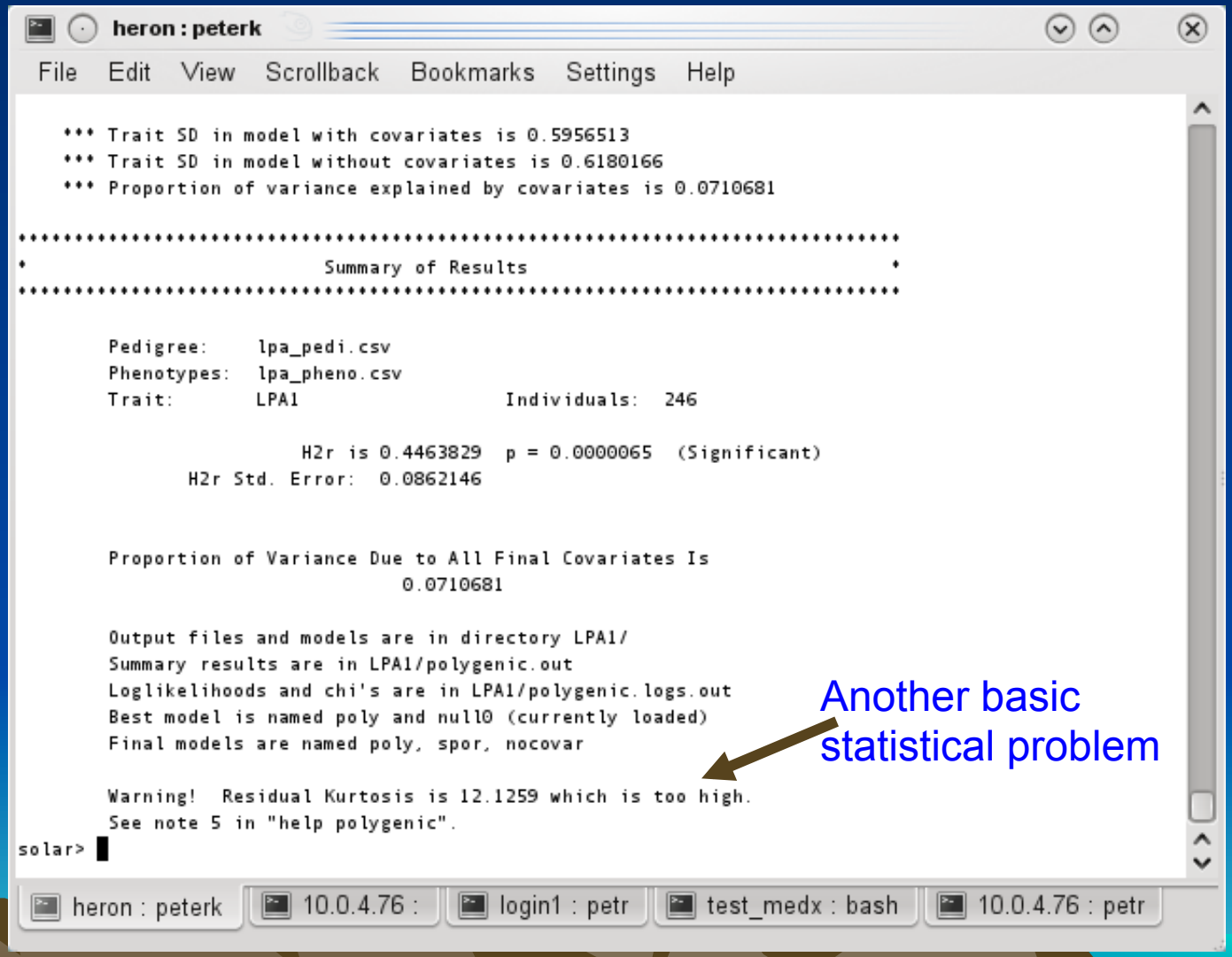
File Edit View Scrollback Bookmarks Settings Help

```
*                          Summary of Results                                  *
**************************************************************************************

        Pedigree:     lpa_pedi.csv
        Phenotypes:   lpa_pheno.csv
        Trait:        AverageFA              Individuals:   246

        WARNING!  Estimated Trait Standard Deviation is 0.01248
        When Trait SD is below 0.5, results are sometimes incorrect
        Multiplying trait by a factor such as 80.2 is recommended

                        H2r is 0.8318344   p = 6.066895e-20   (Significant)
                H2r Std. Error:   0.0346385


        Proportion of Variance Due to All Final Covariates Is
                                0.1987657

        Output files and models are in directory AverageFA/
        Summary results are in AverageFA/polygenic.out
        Loglikelihoods and chi's are in AverageFA/polygenic.logs.out
        Best model is named poly and null0 (currently loaded)
        Final models are named poly, spor, nocovar

        Residual Kurtosis is -0.0206, within normal range
solar>
```

Hmm, we have a basic statistical problem

heron : peterk | 10.0.4.76 : | login1 : petr | test_medx : bash | 10.0.4.7

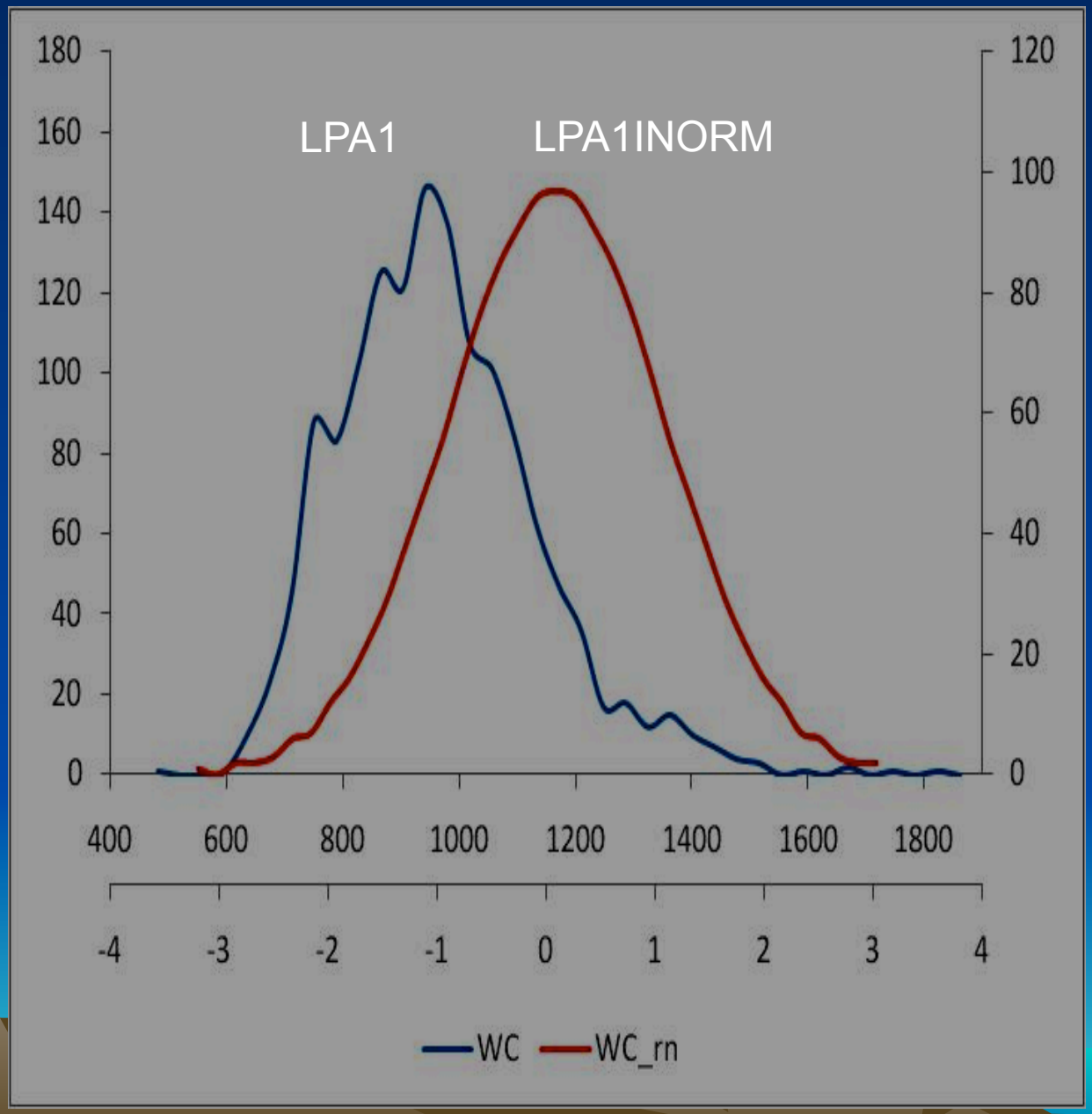
Do the same to LPA1



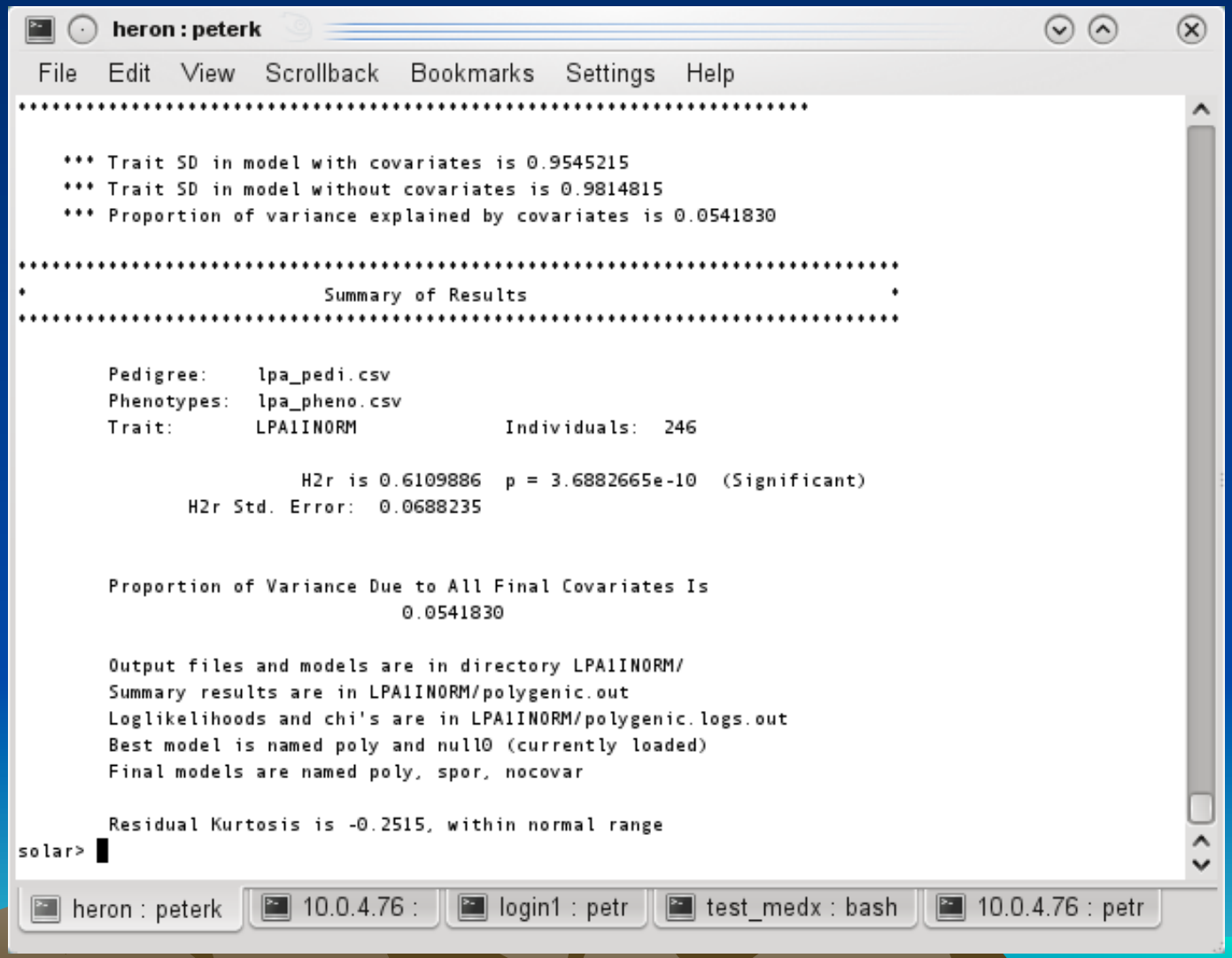
What to do about it?

- Both problems infer lack of normal distribution
- Normality of the distribution has to be enforced for validity of measurements
- Inorm function in solar does it
  - AverageFAINOR = inorm_AverageFA
  - LPA1INORM = inorm_LPA1
- Lets repeat!

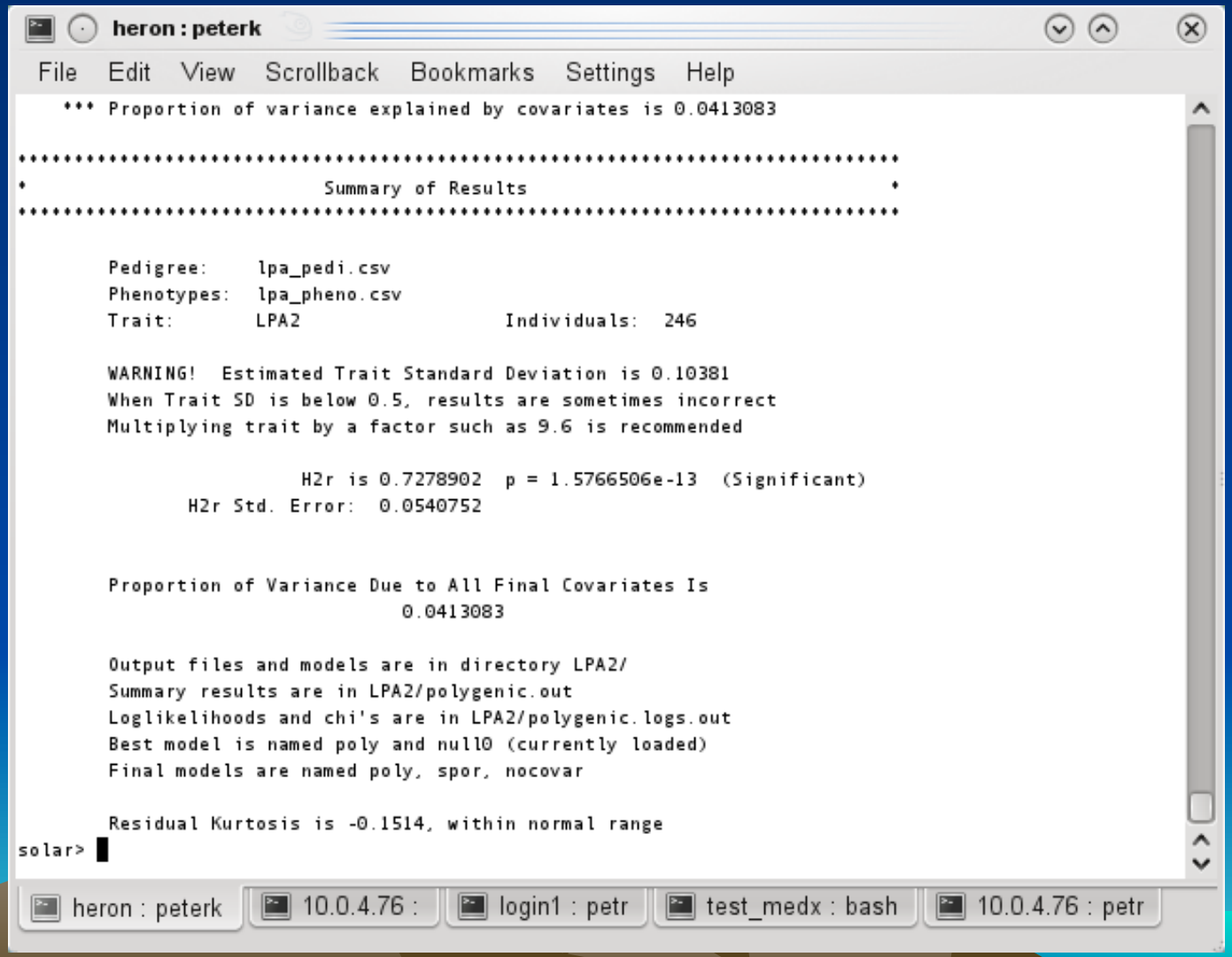
Forcing normal distribution via inverse Gaussian transform.



A-ha, inorm did the trick!



Do the same for LPA2



What do we have

Additive genetic variance explains

- 82% of FA variance
- 61% of the variance in LPA1 protein
- 72% of the variance in LPA2 protein

Now the important part.

- Is that variance shared between traits?
  - This would suggest common genetic or environmental cause

Perform genetic correlation

- Use genetic correlation (ρ_G)

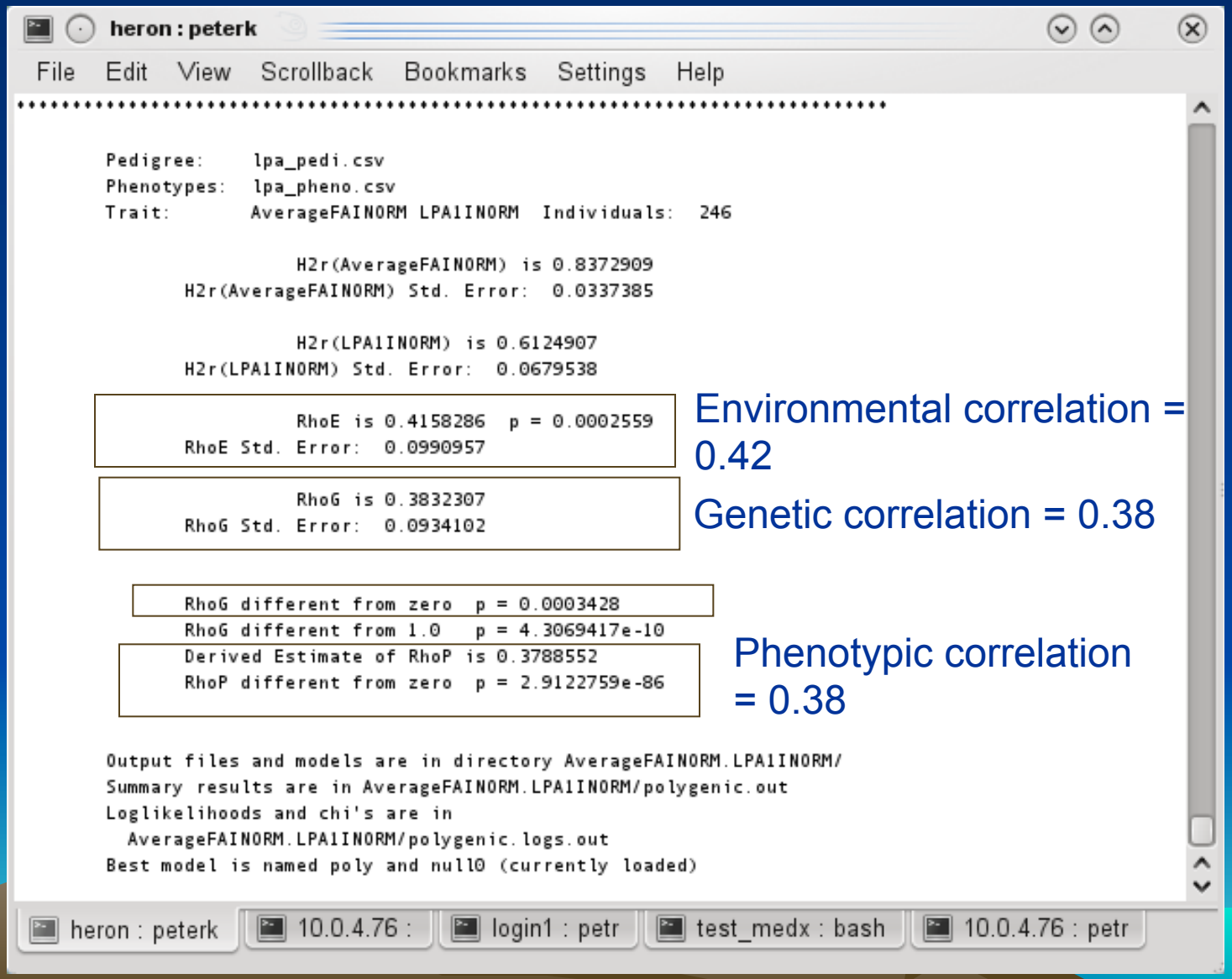
$$r = \sqrt{h_A^2}\sqrt{h_B^2} \cdot \rho_G + \sqrt{1 - h_A^2}\sqrt{1 - h_B^2} \cdot \rho_E$$

  - Pearson's r decomposed into ρ_G and ρ_E
  - ρ_G is the proportion of variability due to shared genetic effects
- To calculate degree of shared genetic variance

Multivariate Analysis in Solar

- Simply use more then one trait
  - trait AverageFAINORM LPA1INORM
  - covar age^1,2#sex
    - Covaries effects of age, age^2, sex and age by sex
  - Polygen –testrhoe –testrhog –testrhop
    - –testrhoe calculates signficance of genetic correlation
    - –testrhog calculates signficance of enviormental correlation
    - –testrhop calculate signficance of the phenotypic (combined) correlation
      - Useful for calculating correlations in family samples

Results for FA and LPA1

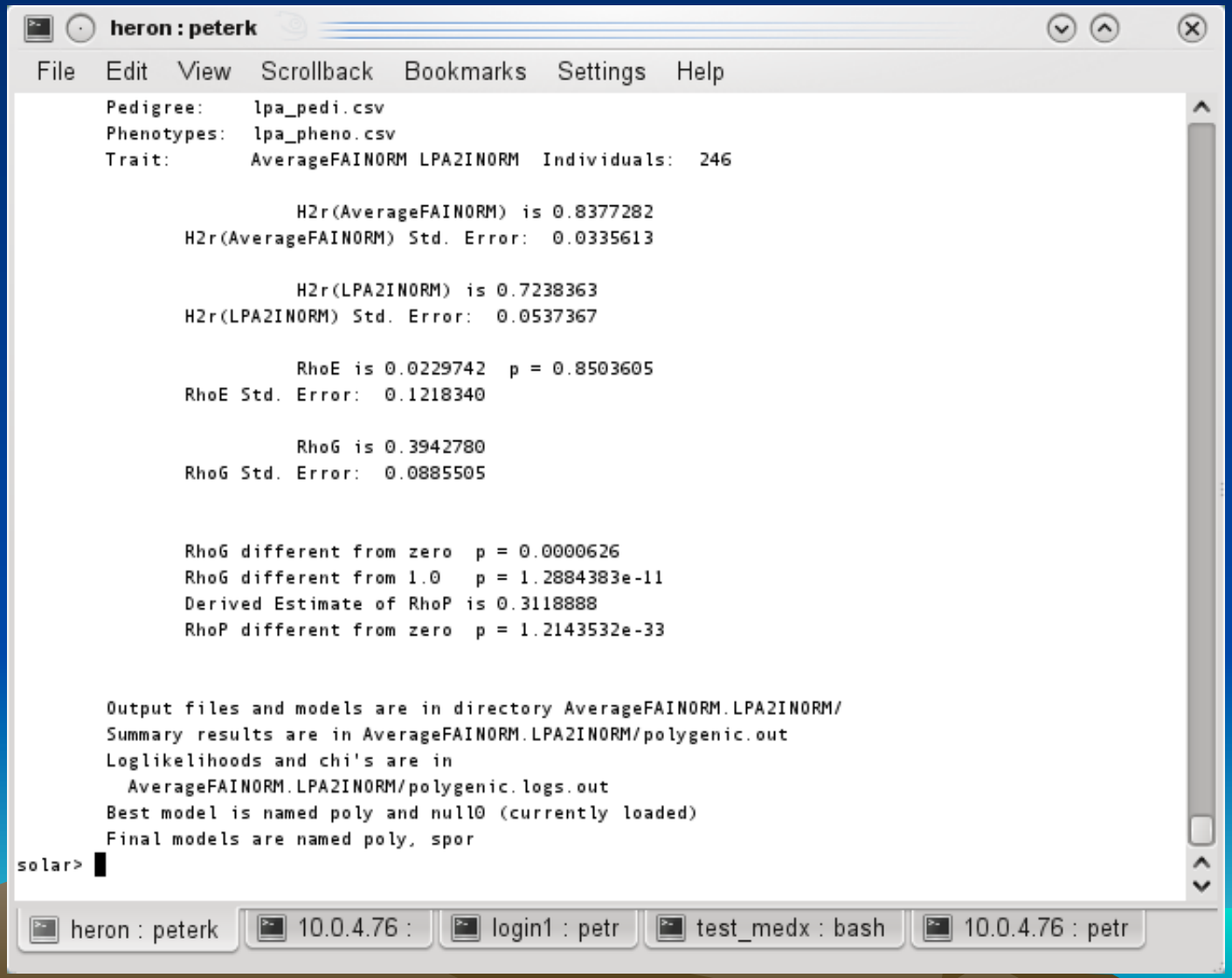


Environmental correlation = 0.42

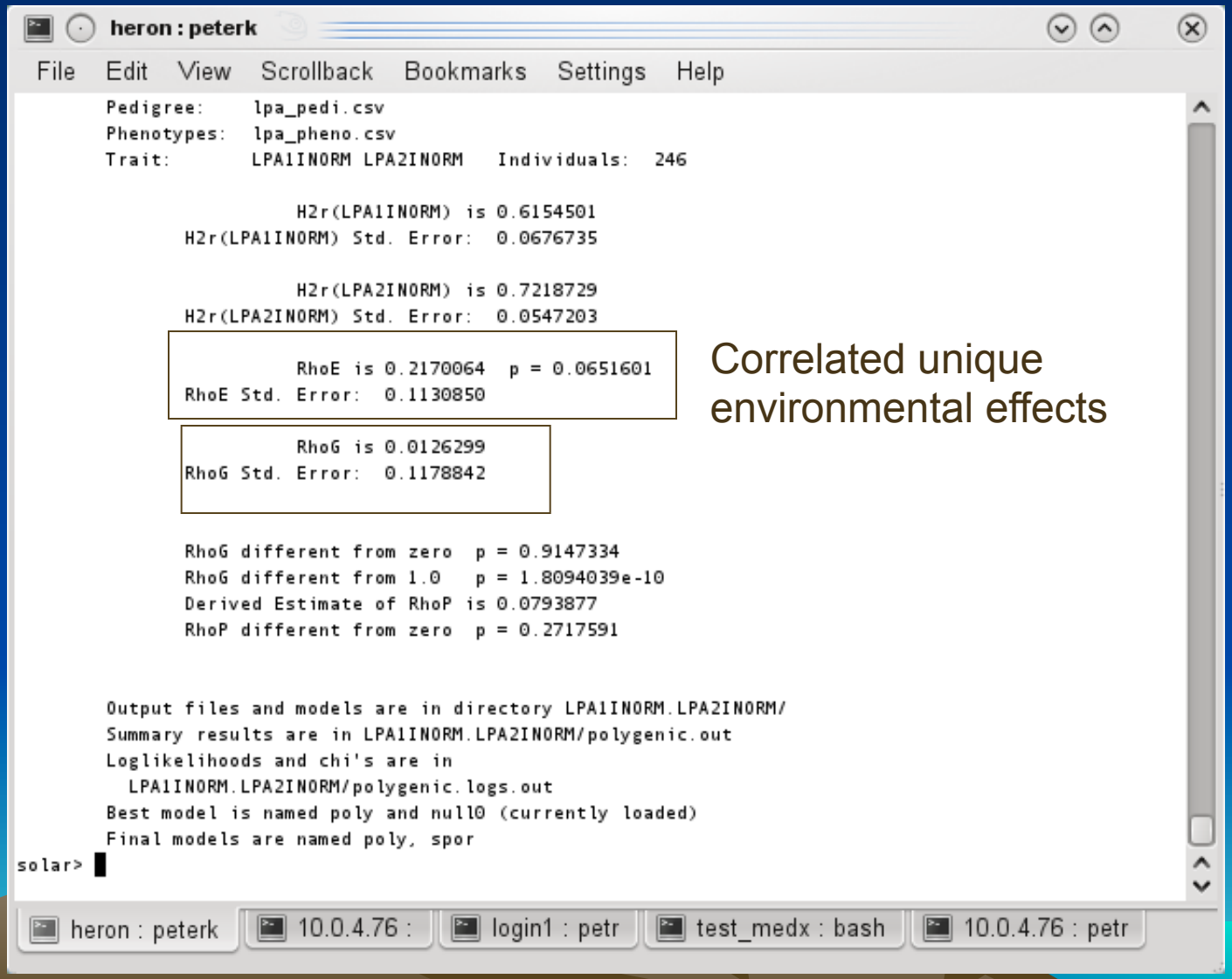
Genetic correlation = 0.38

Phenotypic correlation = 0.38

Results for FA and LPA2



What about LPA1 and LPA2



Correlated unique environmental effects

What have we learned?

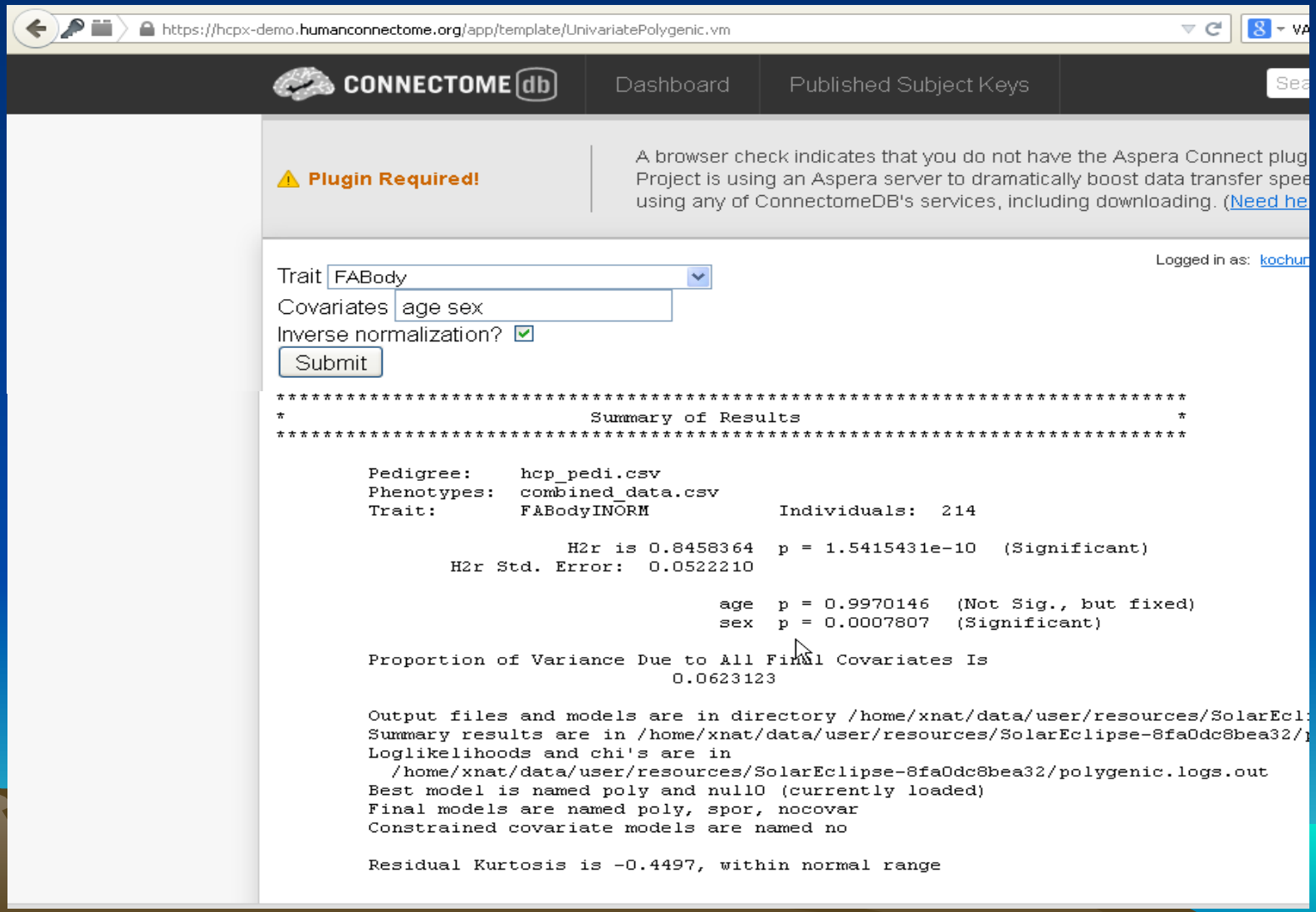
- FA has shared genetic variance with blood volumes of two lipid messenger/receptors
- FA shows shared environmental variance
  - Caused by factors like stress or diet
- The two LPA molecules are genetically independent
  - Encoded by different genes
  - But show some common environmental variance

Contact (pkochunov@gmail.com)

- HCP researchers are running SOLAR-Eclipse as a part of their analysis service
- https://hcpx-demo.humanconnectome.org/app/template/UnivariatePolygenic.vm
  https://hcpx-demo.humanconnectome.org/app/template/GeneticCorrelation.vm
- Visit SOLAR-Eclipse workshop at Imaging Genetic Conference
  - http://www.imaginggenetics.uci.edu/
  - 1/19-20, 2015/ Irvine California
  - To learn how to use SE
    - Voxel-wise genetic analysis
    - Mega and Meta genetic analysis

Human Connectome Web SE

https://hcpx-demo.humanconnectome.org/app/template/UnivariatePolygenic.vm



Human Connectome Web SE

Acknowledgements

- Collaborative effort of
  - John Blangero
  - Charles Peters
  - David Glahn
  - Tom Nichols
  - Bennett Landman
  - Neda Jahanshad
  - Paul Thompson
- Supported by EB015611, P01HL045522, R37MH059490, R01MH078111, R01MH07708143 and R01MH083824