Salambo Bioinformatics Platform

Part #2: Case Study

Andrey Ziyatdinov - UGCD, Sant Pau
Angel Martinez-Perez - UGCD, Sant Pau

About this presentation

- We will talk about current results computed by SalamboR
  - Heritabilities and correlations for GAIT2 traits
- Focus on things we **all** (mathematicians + biologists) care about
  - Input: data quality control and data cleaning
  - Analysis: transformations, covariates, polygenic model
  - Output: interpretations of output values and p-values

The only code snippet in this talk

R

```
library(salamboR)

load("set.RData")

# solarPolygenic
mod <- solarPolygenic(traits="FVIIIc", covlist=c("AGE", "SEX"), set = set)

tab <- mod$df

mod2 <- solarPolygenic(traits="logFVIIIc", covlist=c("AGE", "SEX"), set = set, sceen = TRUE)

tab2 <- mod2$df

# solarCorrelations
cor <- solarCorrelations(traits = c("FVIIIc", "APCR"), set = set)
tab3 <- cor$df

# solarAssoc
A <- solarAssoc(traits="APCR", snps="FVLeiden", set = set)
tab4 <- A$df
```

Introduction

SalamboR tool, GAIT2 traits, 25 publications

SalamboR Tool

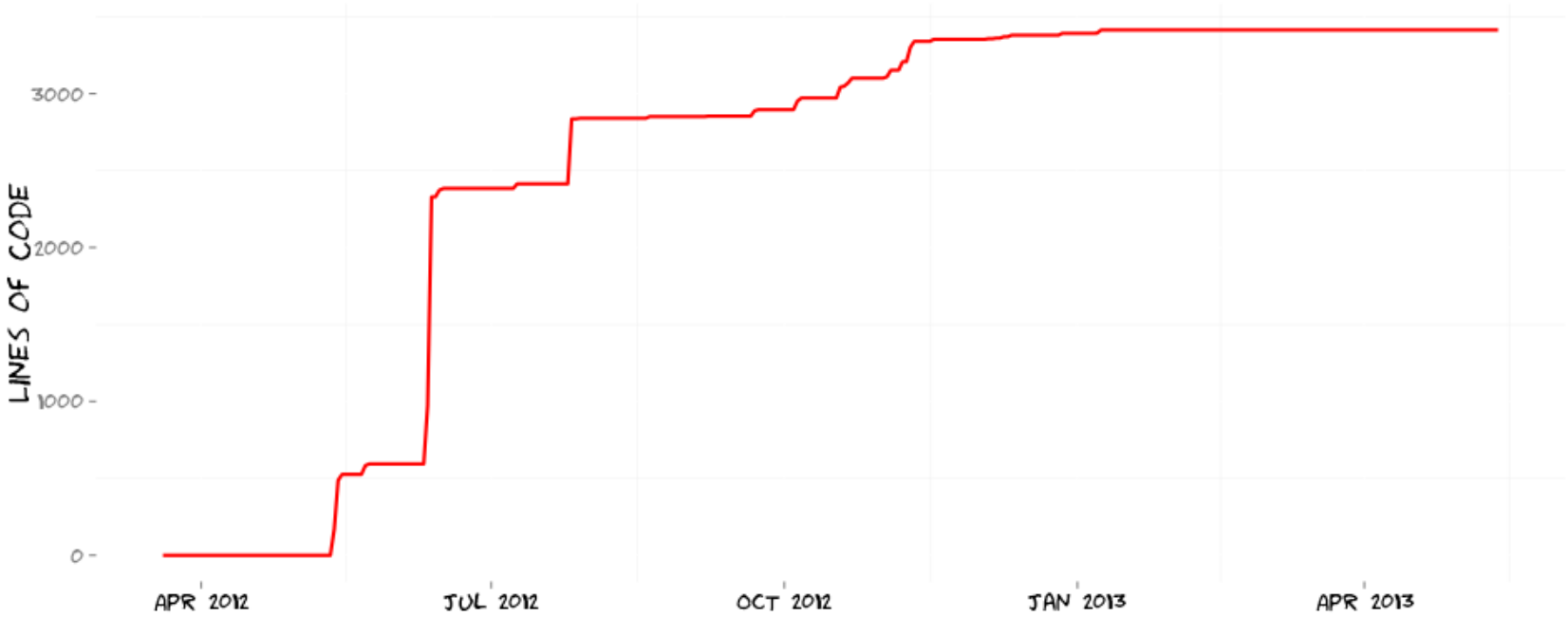
- **Salambo** is the bioinformatics platform consisting of:
  - MySQL db, Salambo Manager, salamboDB for data management
  - SalamboR for statistical analysis
  - Salambo Miner, Salambo Enrichment
- **R** is the language for statistical computing
  - Interface to external tools (SOLAR, pedchek, etc)
  - Graphics
  - Reproducible reports

Statistical Analysis

- Polygenic model
  - Inheritance of a phenotypic characteristic (trait) is attributed to two or more genes
  - Also includes interactions with the environment (multifactorial inheritance)
  - Methods based on variance component analysis
  - `solarPolygenic` function
- Family-based study
  - Extended pedigree families
  - Favorable for linkage based analysis
  - `drawgeneSet` function

Volume of code in SalamboR

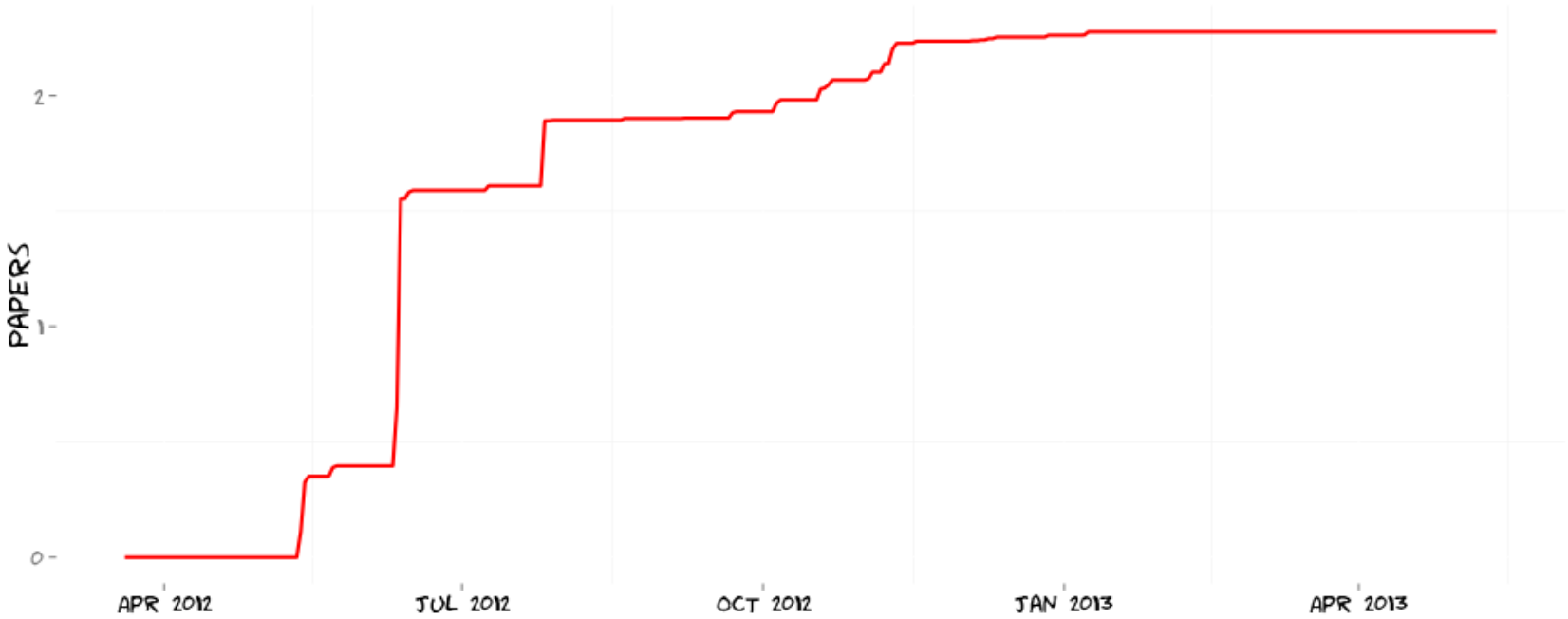
lines of codes



data source: svn log from http://neurochem.sisbio.recerca.upc.edu/papers/santpau/R/SalamboR/

Volume of code in SalamboR

1 paper = 1500 words = 1500 lines of code



data source: svn log from http://neurochem.sisbio.recerca.upc.edu/papers/santpau/R/SalamboR/

Data preparation

GAIT2 Traits, Data Cleaning, Transformations

GAIT2 Traits

Group	Traits	No. Traits
D	AT, VT, Throm, CVI, RheArt, Psori, Asthma, AllerRhi, Hypertension, Malignancy, AIThyroid, Adermatitis	12
TEG	CT_intem, CT_intemuk, CT_natem, CT_tiftem, CFT_intem, CFT_intemuk, CFT_natem, CFT_tiftem, alpha_intem, alpha_intemuk, alpha_natem, alpha_tiftem, MCF_t_intem, MCF_t_intemuk, MCF_t_natem, MCF_t_tiftem, MCF_intem, MCF_intemuk, MCF_natem, MCF_tiftem, TGTlagtime, TGTETP, TGTPeak, Lclisis, LclisisTM	25
PFA	PFAadp, PFAepin	2
TE	COX1, COX2, mPGES1, TXAS, TxB2, PGE2ion, PGE2lps, PLA2G4A, TxB2_PTES, TxB2_PCTES, PGE2ion_LE, PGE2ion_MO, PGE2lps_LE, PGE2lps_MO	14
LC	e5LOX, FLAP, LTA4H, LTC4S, LTB4, LTB4ion, e5LOX_NE, FLAP_NE, LTB4ion_NE, LTB4ion_MO	10
GTF	TMTHG, THRGEN, RTMTH, FIBc, FVIIc, FVIIIc, FIXc, FXIc, FXIIc, APCR, APTT, ATIIIf, D_Dimer, GAB2, GACA, GAFS, GAPS, MAB2, MACA, MAFS, MAPS, Pcam, PSfree, PSfunc, PSt, PT, RUSSELL, TT	28
HOMO	CYS_HPLC, HCY_HPLC, MET_HPLC, SAM_HPLC, SAH_HPLC, SAM_SAH	6
HEMO	PTES, PCTES, LE, NE, LI, MOt	6

"Cleaning data has a steep learning curve and high difficulty level period. Implementing good procedures for data munging is 80% of the job."

jasonpbecke
Forum at news.ycombinator.com

Results

Heritabilities, Correlations

Protocol

- Pre-compute transformations of traits
- Compute univariate models with screening option
  - List of significant covariates
  - Betas
  - Heritabilities
- Compute phenotypic correlations (the effect we observe)
- Compute environmental and genetic correlations (the effect we derive from variance component models)

Protocol for computing univariate models

Output: List of covariates, Betas, Heritability

- Univariate model: trait ~ covariate 1 + covariate 2 + ...
- Trait is transformed (if necessary) to approach the normal distribution
- Use of covariates to explain the variation attributed to the environment
- Screening of covariates is applied to avoid attributing the variance not related to the covariates

List of covariates:

- The Big Four: AGE, SEX, CONTRA, SMOKING
- Additional: AINEs, antiAgreg
- To be added: BMI, Physical Exercise
- Other trait-specific

Heritability

Trait VT

	N	missings	Nused	h2r	h2r.se	h2r.P	covlist
VT	1114	50	935	0.67	0.16	1.60e-06	AGE

- **N**: total number of samples
- **missings**: number of missing samples of trait
- **Nused**: number of sample used by `SOLAR` (excluded missings of both trait and covariates)
- **h2r**: estimated Heritability
- **h2r.se**: standard error of estimation
- **h2r.P**: p-value of the significance
- **covlist**: list of covariates in the final model (screened)

Covariates

Trait VT

	AGE	AGE.se	AGE.P	SEX	SEX.se	SEX.P	contraception	contraception.se	contraception.P
VT	-0.02	0.00	5.15e-10	0.00	0.00	1.36e-01	0.00	0.00	1.68e-01

	AINEs	AINEs.se	AINEs.P	antiAgreg	antiAgreg.se	antiAgreg.P	smoking	smoking.se	smoking.P	var4cov
VT	0.00	0.00	7.85e-01	0.00	0.00	9.26e-01	0.00	0.00	8.92e-01	

- **AGE**: estimation of coefficient beta for `AGE` covariate
- **AGE.se**: standard error of estimation
- **AGE.P**: p-value of the significance
- ...
- **var4cov**: variance captured by model (case of continuous traits)

If screening is set, all covariates except `AGE` will be removed from the final model (alpha 0.1).

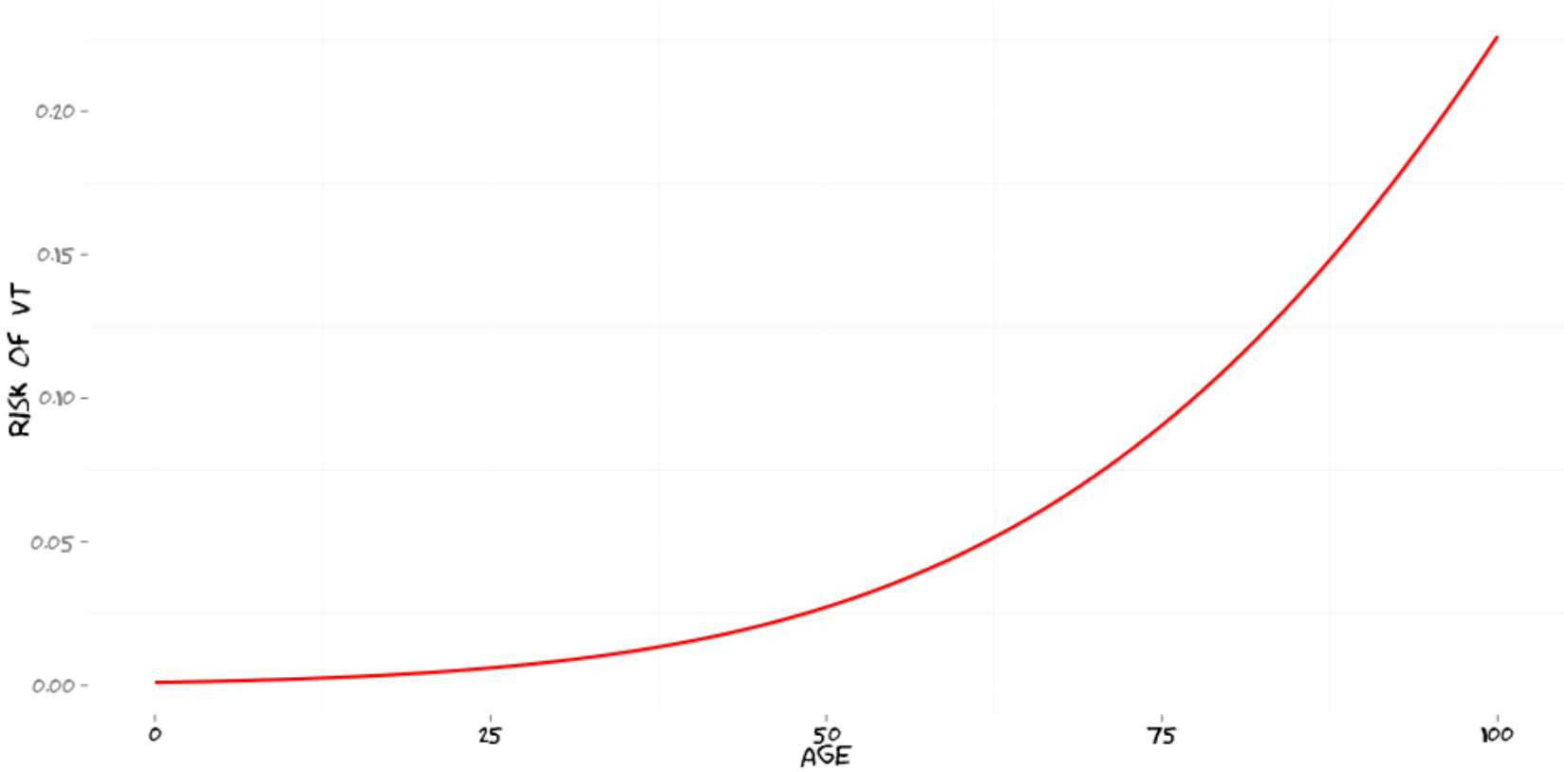
Relation VT ~ AGE

- Trait VT is binary (0 or 1, Male or Female)
- `SOLAR` transforms binary traits (probit)
- Coefficient beta is estimated between **transformed VT** and AGE

Rule (for the current settings)

- Negative beta: the greater AGE the greater risk of VT

Risk of VT ~ AGE



Protocol for computing genetic and environmental correlations

- Trait is transformed (if necessary) to approach the normal distribution
- Known environmental variance: covariates defined by univariate models with screening option
- Bivariate model: trait1 + trait2 ~ covariate 1 + covariate 2 + ...
- Unknown environmental and genetic correlations will be computed by variance component model

When calculations of the correlations failed:

- Signs of errors: correlation value is 1 or -1, p-value is extremely small like 1e-115
- Reasons (actually unknown): missing data, low heritability of traits, numerical problems in `SOLAR`
- Solution (manual): One can attempt different combinations of the covariates

Genetic and Environmental Correlations

TGT group of traits

trait1	trait2	Nused	N.both	N.affected	h2r.trait1	h2rSE.trait1	h2r.trait2	h2rSE.trait2	h2r.min
TGTlagtime	VT	935	919	84	0.63	0.06	0.69	0.16	0.63
TGTPeak	VT	935	919	84	0.70	0.05	0.68	0.15	0.68
Lclisis	VT	935	920	84	0.50	0.01	0.50	0.02	0.50

- **Nused**: number of samples used by `SOLAR`
- **N.both**: number of samples where both traits have data
- **N.affected**: number of samples where trait2 (disease) has data
- **h2r.trait1**: estimated heritability for trait1
- **h2r.trait2**: estimated heritability for trait2
- **h2r.min**: Minimum value of the heritabilities

Genetic and Environmental Correlations

TGT group of traits

trait1	trait2	rhoG	rhoGse	pvalG0	pvalG1	rhoE	rhoEse	pvalE
TGTlagtime	VT	0.16	0.16	3.22e-01	1.80e-06	0.23	0.30	4.55e-01
TGTPeak	VT	0.37	0.14	9.11e-03	1.34e-05	0.32	0.28	2.40e-01
Lclisis	VT	0.53	0.03	3.36e-06	1.00e+00	0.83	0.04	1.74e-06

- **rhoG**: estimation of genetic correlation
- **rhoG.se**: standard error of estimation
- **pvalG0**: p-value of the test whether some genes are shared
- **pvalG1**: p-value of the test whether not all genes are shared
- **rhoE**: estimation of envirinmental correlation
- **rhoE.se**: standard error of estimation
- **pvalE**: p-value of the significance

Was it hard to compute'em all?

- Round 0 @ 06/06/2012
  - Received the list of traits and diseases
- Round 1 @ 15/06/2012
  - First attempt to calculate correlations for a small group of traits (TG group)
- Round 2 @ 27/07/2012
  - Standard covariates
  - A lot of missing correlations and inconsistent results
- Round 3 @ 30/10/2012
  - No covariates
  - Inormal transformations of traits
  - Still many missing correlations
- Round 4 @ nowdays
  - All correlations computed (except some few rare ones)
  - Defined the transformations of traits
  - Added new covariates: AINEs, ANTIagreg
  - Defined the protocol of calculations

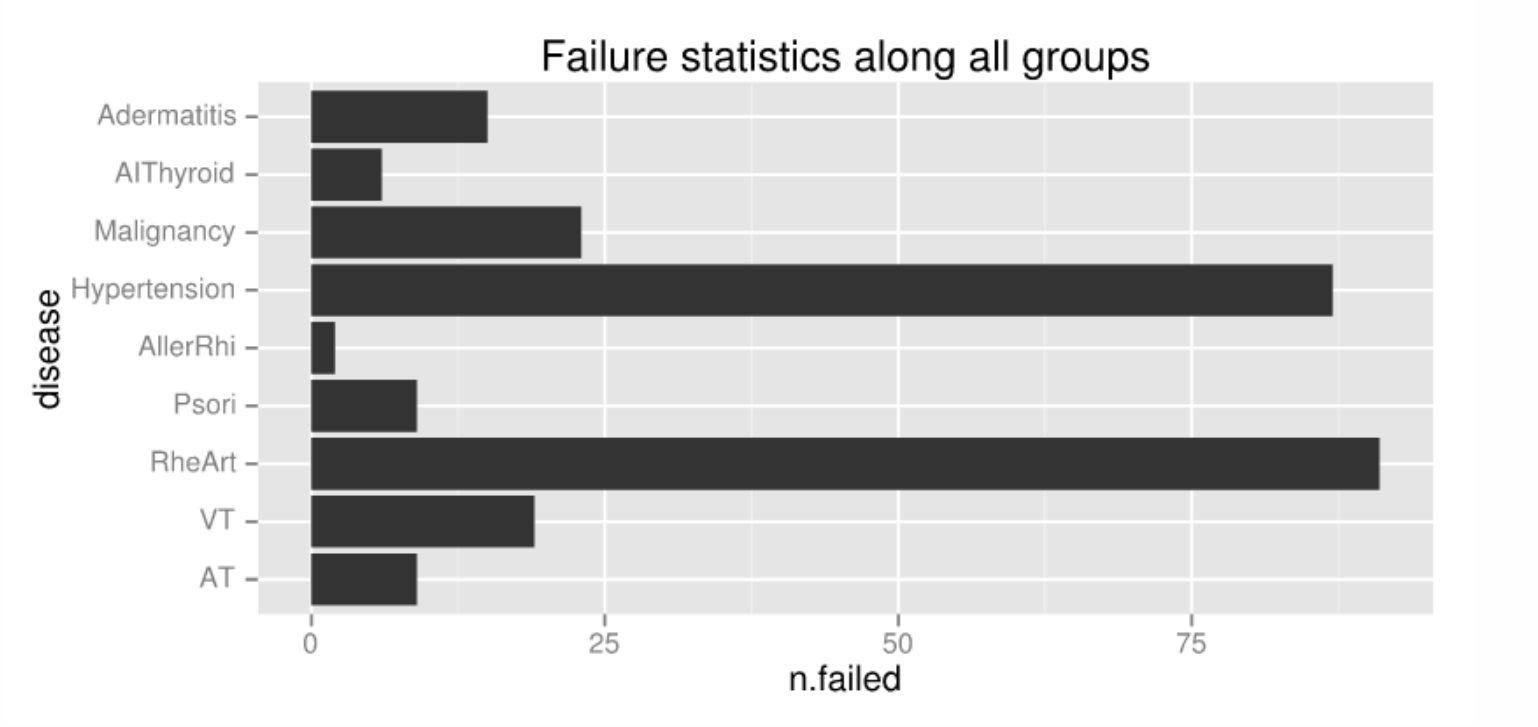
Was it hard to compute'em all?

Statistics by groups of traits

Group of Traits	# diseases	# traits	# corr.	# corr. failed	# corr. computed	# corr. computed, %
TEG	12	25	300	60	240	80 %
PFA	12	2	24	5	19	79.2 %
TE	12	14	168	60	108	64.3 %
LC	12	10	120	26	94	78.3 %
GTF	12	28	336	83	253	75.3 %
HOMO	12	6	72	15	57	79.2 %
HEMO	12	6	72	12	60	83.3 %

Was it hard to compute'em all?

Statistics of failures on diseases



Conclusions

The way to obtain fruitful results

- The biologists need:
  - Results the mathematicians are confident about
  - Clear explanation of variables
  - Excel tables
  - Interpretation of the current results to build new hypotheses
  - Option to re-run analyses
  - ~~Endless patience with the mathematicians~~
- The mathematicians need:
  - Knowledge of the biological/clinical background
  - ~~Infinite patience with `SOLAR`~~

Thank You!

SOLAR energy is our future!

Andrey Ziyatdinov - UGCD, Sant Pau
Angel Martinez-Perez - UGCD, Sant Pau