CC BY SA

The GenABEL project for statistical genomics

Yurii Aulchenko

[YuriiA consulting (NL) | ICG SB RAS (RU) | CPHS UoE (UK) | @YuriiAulchenko]

for the GenABEL project contributors

[@GenAproj | www.GemABEL.org]

GenABEL
STATISTICAL GENOMICS

Outline

- Statistical genomics
  - A short history
  - Current state
  - Summary

GenABEL
STATISTICAL GENOMICS

Why are we different? Why do certain people get a disease?

What are the mechanisms underlying these differences?

How genetic variation controls the phenotype?

Statistical genomics

Feature 3

Sample 1

	qt1	qt2	rs10	rs18	rs29	rs65	rs73
1	-0.58	4.46	0	0	0	1	0
2	0.80	0.32	0	0	NA	NA	0
3	-0.52	3.26	0	0	NA	1	0
4	-1.55	888.00	0	0	NA	NA	0
5	0.25	5.70	0	1	0	2	0
6	0.15	4.65	0	0	0	0	0
7	-0.56	4.64	1	1	1	2	0
8	NA	5.77	0	0	0	2	NA
9	-2.26	0.71	0	1	0	2	0
10	-1.32	3.26	0	0	0	2	0

GenABEL
STATISTICAL GENOMICS

Statistical genomics

Feature 3

Sample 1

	qt1	qt2
1	-0.58	4.46
2	0.80	6.32
3	-0.52	3.26
4	-1.55	888.00
5	0.25	5.70
6	0.15	4.65
7	-0.56	4.64
8	NA	5.77
9	-2.26	0.71
10	-1.32	3.26

	rs10	rs18	rs29	rs65	rs73
1	0	0	0	1	0
2	0	0	NA	NA	0
3	0	0	NA	1	0
4	0	0	NA	NA	0
5	0	1	0	2	0
6	0	0	0	0	0
7	1	1	1	2	0
8	0	0	0	2	NA
9	0	1	0	2	0
10	0	0	0	2	0

Traits/ phenotypes

?

Genotypes

GenABEL
STATISTICAL GENOMICS

Genome-wide association scanning (GWAS)

lm(qt1 ~ rs10)

	qt1	qt2
1	-0.58	4.46
2	0.80	6.32
3	-0.52	3.26
4	-1.55	888.00
5	0.25	5.70
6	0.15	4.65
7	-0.56	4.64
8	NA	5.77
9	-2.26	0.71
10	-1.32	3.26

Traits/ phenotypes

?

	rs10	rs18	rs29	rs65	rs73
1	0	0	0	1	0
2	0	0	NA	NA	0
3	0	0	NA	1	0
4	0	0	NA	NA	0
5	0	1	0	2	0
6	0	0	0	0	0
7	1	1	1	2	0
8	0	0	0	2	NA
9	0	1	0	2	0
10	0	0	0	2	0

Genotypes

GenABEL
STATISTICAL GENOMICS

Genome-wide association scanning (GWAS)

	qt1	qt2
1	-0.58	4.46
2	0.80	6.32
3	-0.52	3.26
4	-1.55	888.00
5	0.25	5.70
6	0.15	4.65
7	-0.56	4.64
8	NA	5.77
9	-2.26	0.71
10	-1.32	3.26

Traits/ phenotypes

?

	rs10	rs18	rs29	rs65	rs73
1	0	0	0	1	0
2	0	0	NA	NA	0
3	0	0	NA	1	0
4	0	0	NA	NA	0
5	0	1	0	2	0
6	0	0	0	0	0
7	1	1	1	2	0
8	0	0	0	2	NA
9	0	1	0	2	0
10	0	0	0	2	0

Genotypes

GenABEL
STATISTICAL GENOMICS

Genome-wide association scanning (GWAS)

	qt1	qt2
1	-0.58	4.46
2	0.80	6.32
3	-0.52	3.26
4	-1.55	888.00
5	0.25	5.70
6	0.15	4.65
7	-0.56	4.64
8	NA	5.77
9	-2.26	0.71
10	-1.32	3.26

Traits/ phenotypes

?

	rs10	rs18	rs29	rs65	rs73
1	0	0	0	1	0
2	0	0	NA	NA	0
3	0	0	NA	1	0
4	0	0	NA	NA	0
5	0	1	0	2	0
6	0	0	0	0	0
7	1	1	1	2	0
8	0	0	0	2	NA
9	0	1	0	2	0
10	0	0	0	2	0

Genotypes

GenABEL
STATISTICAL GENOMICS

Genome-wide association scanning (GWAS)

	qt1	qt2
1	-0.58	4.46
2	0.80	6.32
3	-0.52	3.26
4	-1.55	888.00
5	0.25	5.70
6	0.15	4.65
7	-0.56	4.64
8	NA	5.77
9	-2.26	0.71
10	-1.32	3.26

Traits/ phenotypes

?

	rs10	rs18	rs29	rs65	rs73
1	0	0	0	1	0
2	0	0	NA	NA	0
3	0	0	NA	1	0
4	0	0	NA	NA	0
5	0	1	0	2	0
6	0	0	0	0	0
7	1	1	1	2	0
8	0	0	0	2	NA
9	0	1	0	2	0
10	0	0	0	2	0

Genotypes

GenABEL
STATISTICAL GENOMICS

Genome-wide association scanning (GWAS)

	qt1	qt2
1	-0.58	4.46
2	0.80	6.32
3	-0.52	3.26
4	-1.55	888.00
5	0.25	5.70
6	0.15	4.65
7	-0.56	4.64
8	NA	5.77
9	-2.26	0.71
10	-1.32	3.26

Traits/ phenotypes

?

	rs10	rs18	rs29	rs65	rs73
1	0	0	0	1	0
2	0	0	NA	NA	0
3	0	0	NA	1	0
4	0	0	NA	NA	0
5	0	1	0	2	0
6	0	0	0	0	0
7	1	1	1	2	0
8	0	0	0	2	NA
9	0	1	0	2	0
10	0	0	0	2	0

Genotypes

GenABEL
STATISTICAL GENOMICS

Genome-wide association scanning (GWAS)

Few

100,000-40,000,000,000

1,000-100,000

	qt1	qt2
1	-0.58	4.46
2	0.80	6.32
3	-0.52	3.26
4	-1.55	888.00
5	0.25	5.70
6	0.15	4.65
7	-0.56	4.64
8	NA	5.77
9	-2.26	0.71
10	-1.32	3.26

	rs10	rs18	rs29	rs65	rs73
1	0	0	0	1	0
2	0	0	NA	NA	0
3	0	0	NA	1	0
4	0	0	NA	NA	0
5	0	1	0	2	0
6	0	0	0	0	0
7	1	1	1	2	0
8	0	0	0	2	NA
9	0	1	0	2	0
10	0	0	0	2	0

Traits/ phenotypes

?

Genotypes

GenABEL
STATISTICAL GENOMICS

Scanning through "omics" space

100-100,000

1,000-100,000

	qt1	qt2
1	-0.58	4.46
2	0.80	6.32
3	-0.52	3.26
4	-1.55	888.00
5	0.25	5.70
6	0.15	4.65
7	-0.56	4.64
8	NA	5.77
9	-2.26	0.71
10	-1.32	3.26

Traits/ phenotypes

?

100,000-40,000,000,000

	rs10	rs18	rs29	rs65	rs73
1	0	0	0	1	0
2	0	0	NA	NA	0
3	0	0	NA	1	0
4	0	0	NA	NA	0
5	0	1	0	2	0
6	0	0	0	0	0
7	1	1	1	2	0
8	0	0	0	2	NA
9	0	1	0	2	0
10	0	0	0	2	0

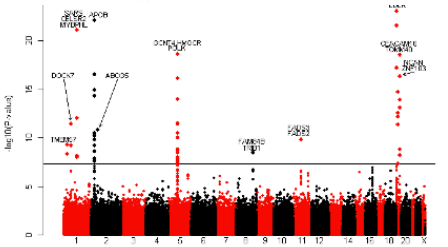
Genotypes

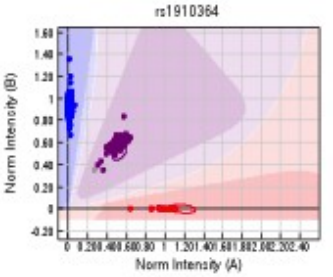
GenABEL
STATISTICAL GENOMICS

Statistical genomics: what is so special?

- Rules governing genes & experimental design: analysis methodology and results visualization
- Technological inputs: data formats, quality control, analysis methods
- Analysis is computationally challenging (and IO demanding)

$$y = X\beta + Zu + \epsilon$$





GenABEL
STATISTICAL GENOMICS

Analysis scenarios

- Classic GWAS scenario
  - One trait – one genetic marker at a time
  - Correlations between phenotypes – mixed models
- Emerging scenarios
  - One trait – multiple genetic markers
  - Multiple traits – single / multiple markers

GenABEL
STATISTICAL GENOMICS

Outline

- Statistical genomics
- **A short history**
- Current state
- Summary

A short history

Package

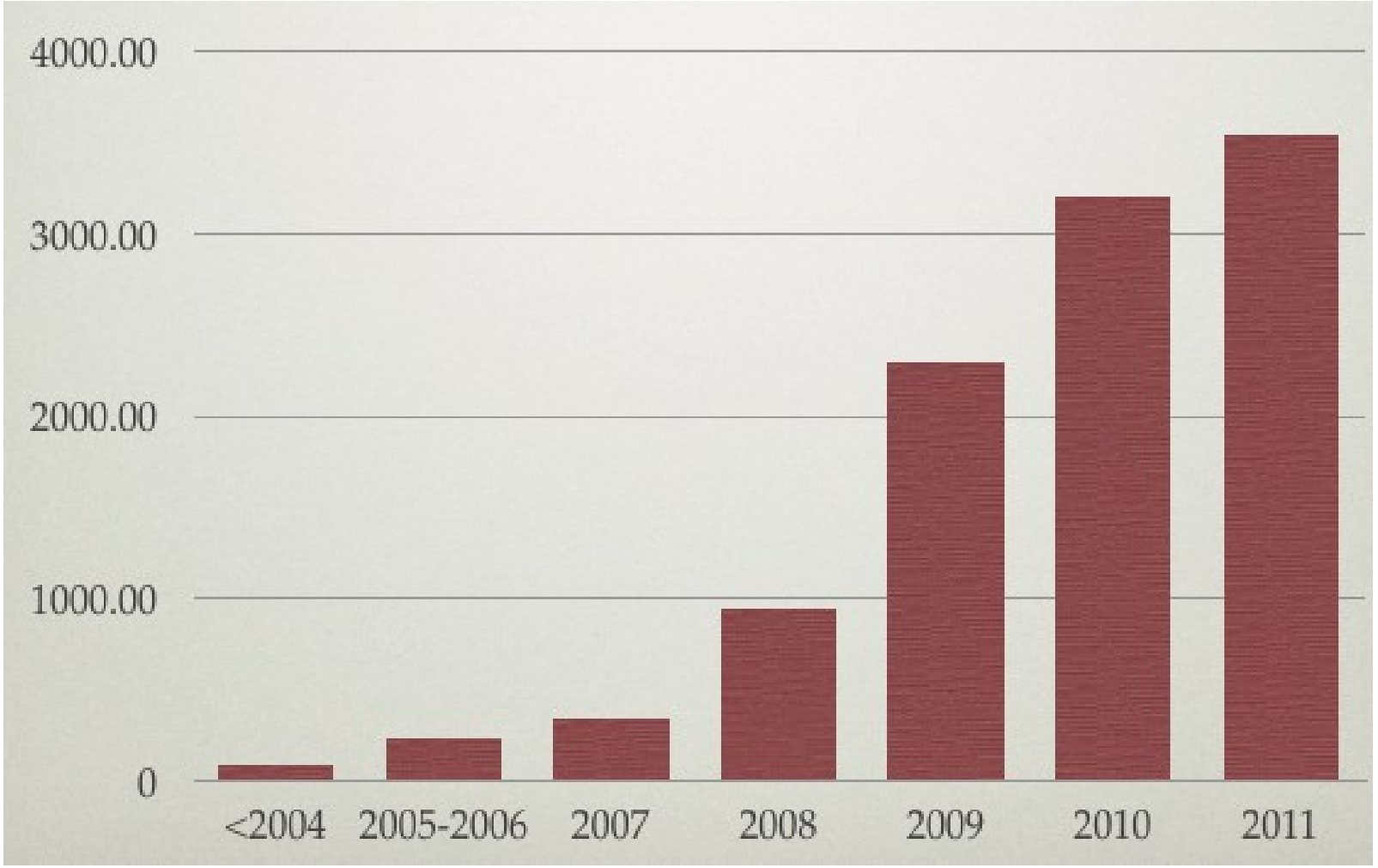
Paper

2006	2007	2008	2009	2010	2011	2012	2013...
GenA	GenA						

GenABEL package

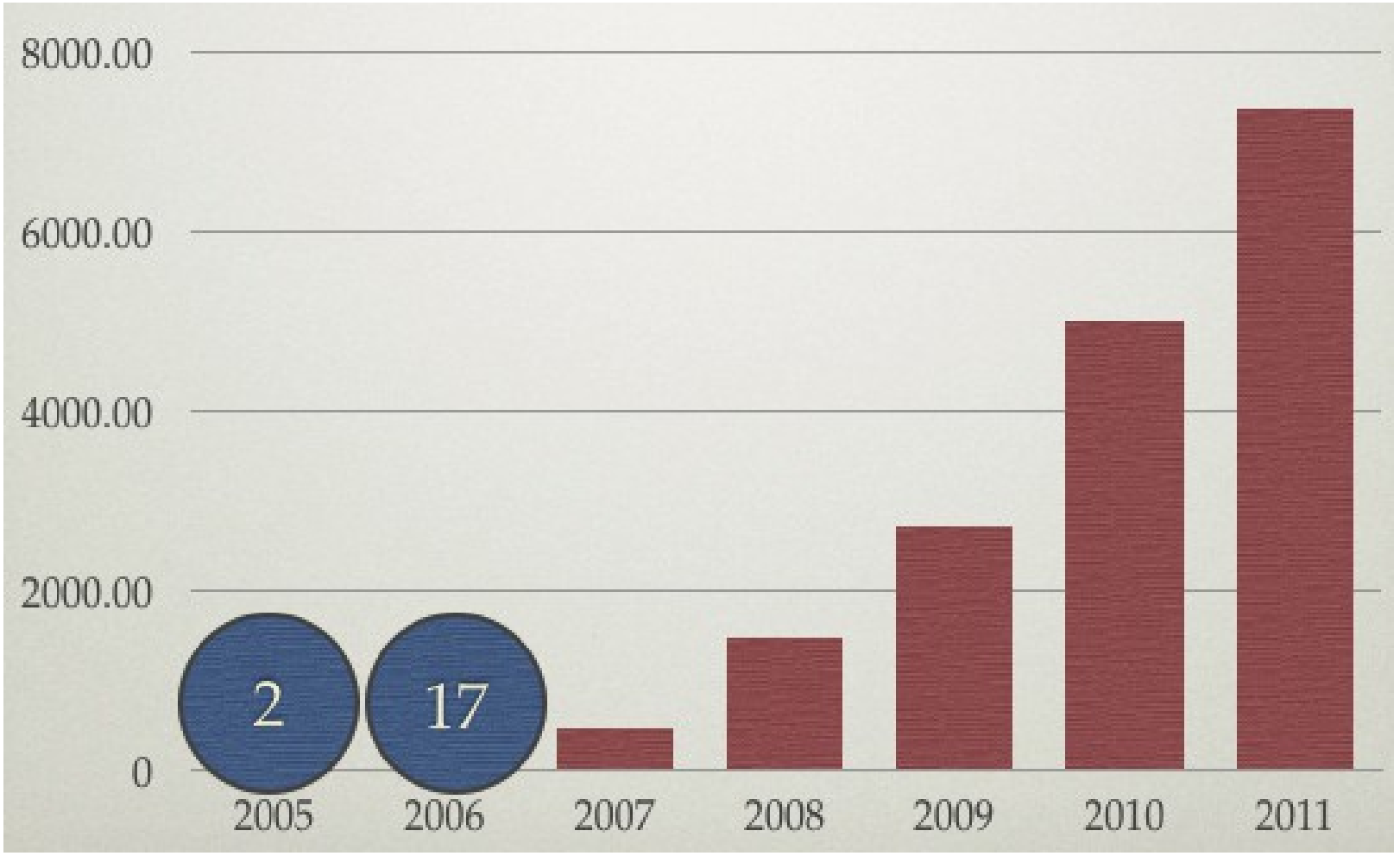
GenABEL
STATISTICAL GENOMICS

GWAS PUBLICATIONS



GenABEL
STATISTICAL GENOMICS

LOCI IDENTIFIED IN GWAS

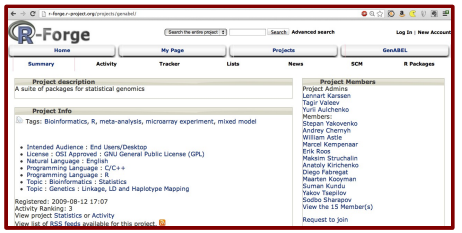


GenABEL
STATISTICAL GENOMICS

A short history

Package

Paper



Data Analysts Captivated by R's Power

MetA ProbA ParallA MixA

2006	2007	2008	2009	2010	2011	2012	2013...

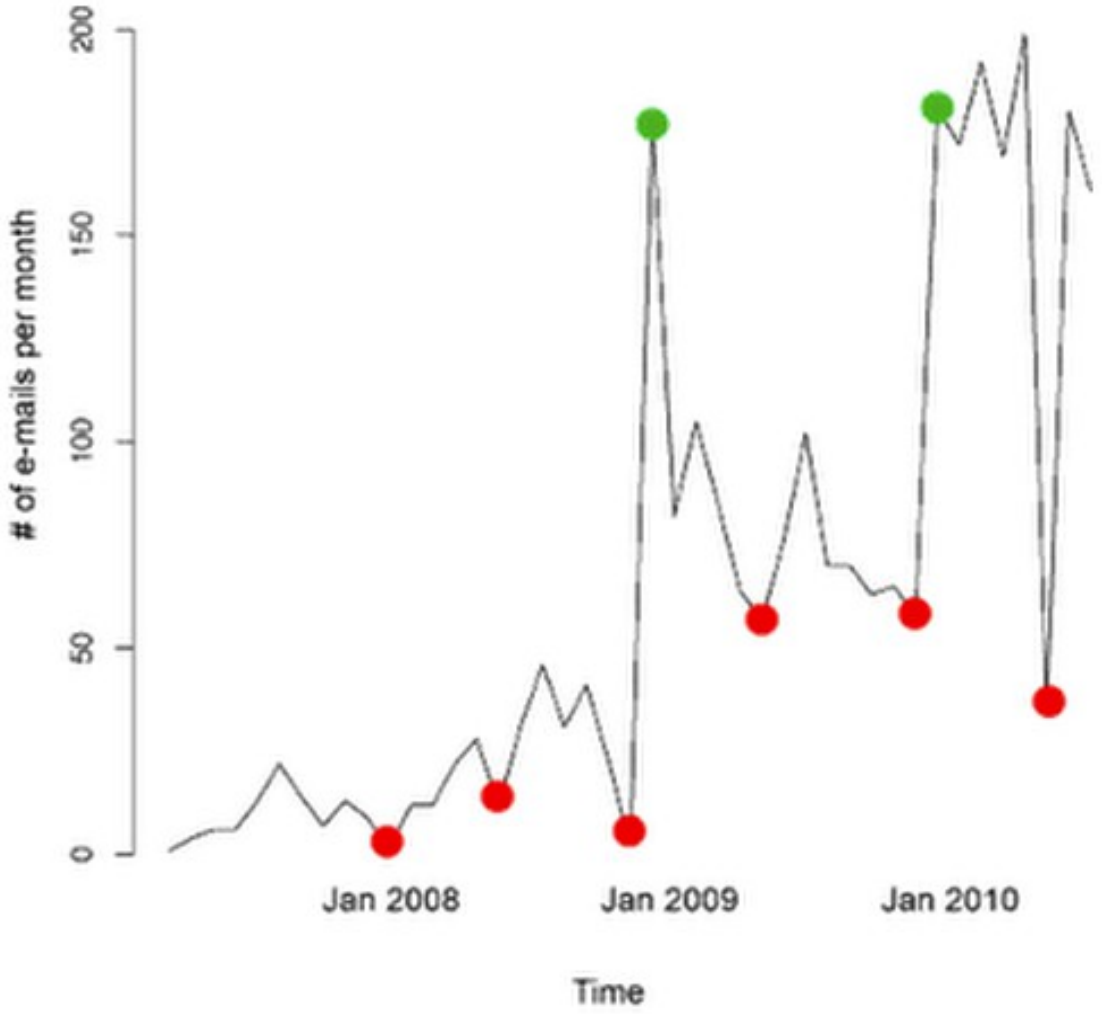
GenA GenA DatA ParallA ProbA

GenABEL package

GenABEL suite

GenABEL
STATISTICAL GENOMICS

Turning point



GenABEL
STATISTICAL GENOMICS

The GenABEL project

Mission: to provide a framework for development of statistical genomics methodology

Vision: collaboration, transparency and free exchange of code, ideas, and data is a key to agile and robust methodology development

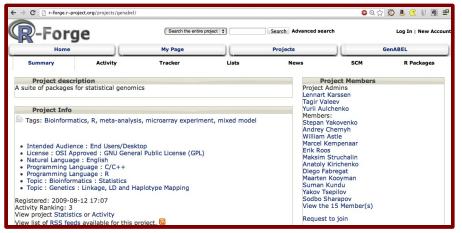
Strategy: community-based and driven methodology discussion, development, implementation, dissemination, maintenance, ***and application***

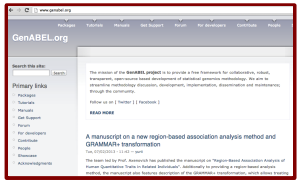
GenABEL
STATISTICAL GENOMICS

A short history

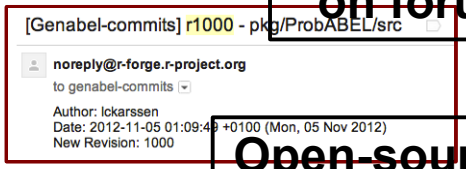
Package

Paper





1000 posts on forum



Open-source tutorial

Data Analysts Captivated by R's Power

MetA ProbA ParallA MixA PredictA PredictA GenA

2006 | 2007 | 2008 | 2009 | 2010 | 2011 | 2012 | 2013...

GenA GenA DatA ParallA ProbA VariA VariA GenA OmicA

GenABEL package

GenABEL suite

GenABEL project

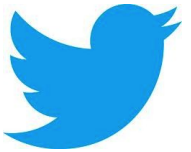
GenABEL
STATISTICAL GENOMICS

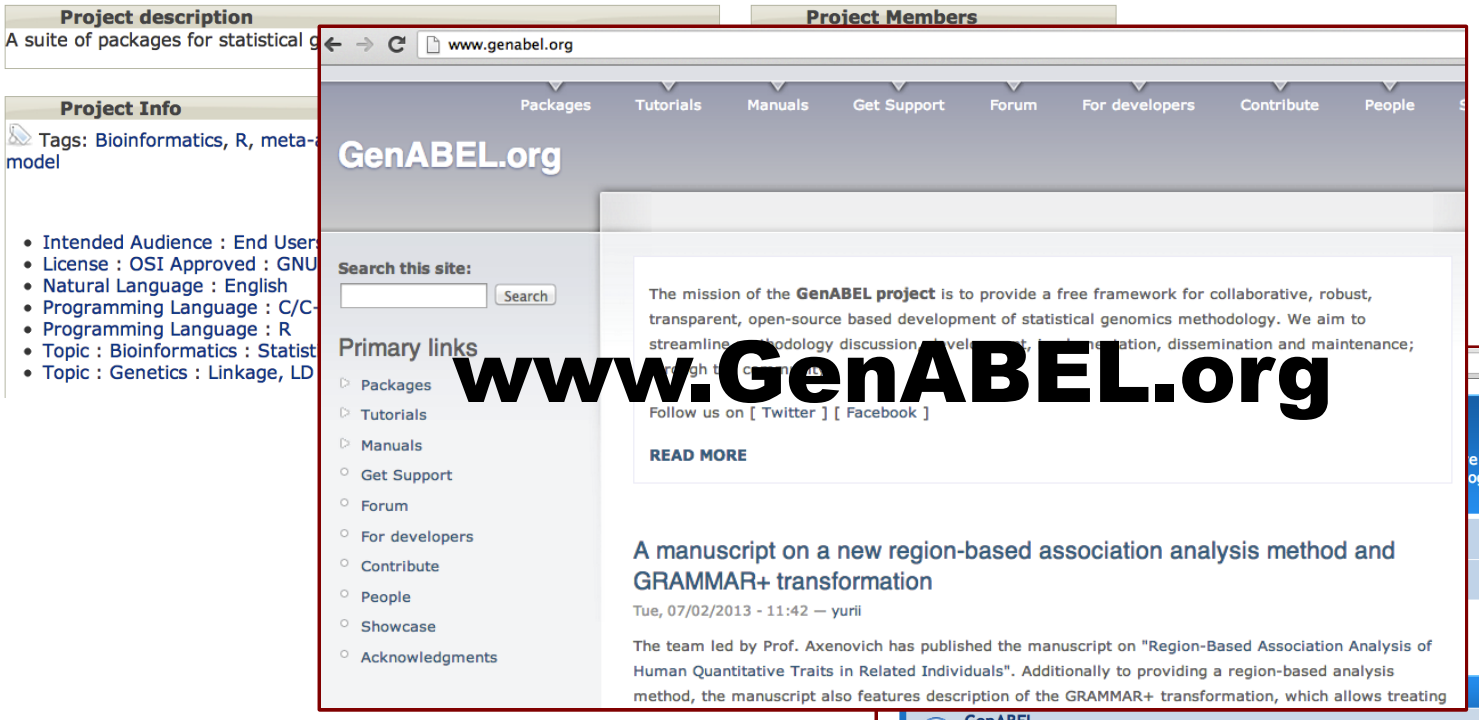
Outline

- Statistical genomics
- A short history
- **Current state**
- Summary

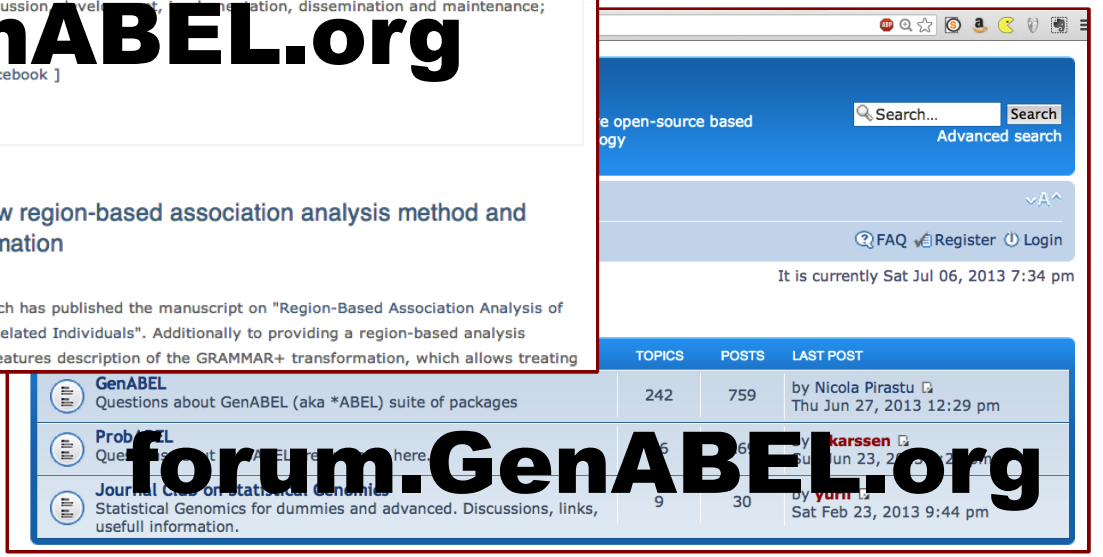
Infrastructure

GenABEL @ R-Froge





www.GenABEL.org



forum.GenABEL.org

GenABEL
STATISTICAL GENOMICS

Project in numbers

Code of 9 packages	
Language	# kLines of code
R	19
C++	19
C	17
Other	2
Rnw/Roxy	20

Estimated
12 man-years
$1,500,000

People	
Developers	15 (5)
Forum	430 (71)

Communications	
Devel-list	>700 posts
Forum	>1000 posts

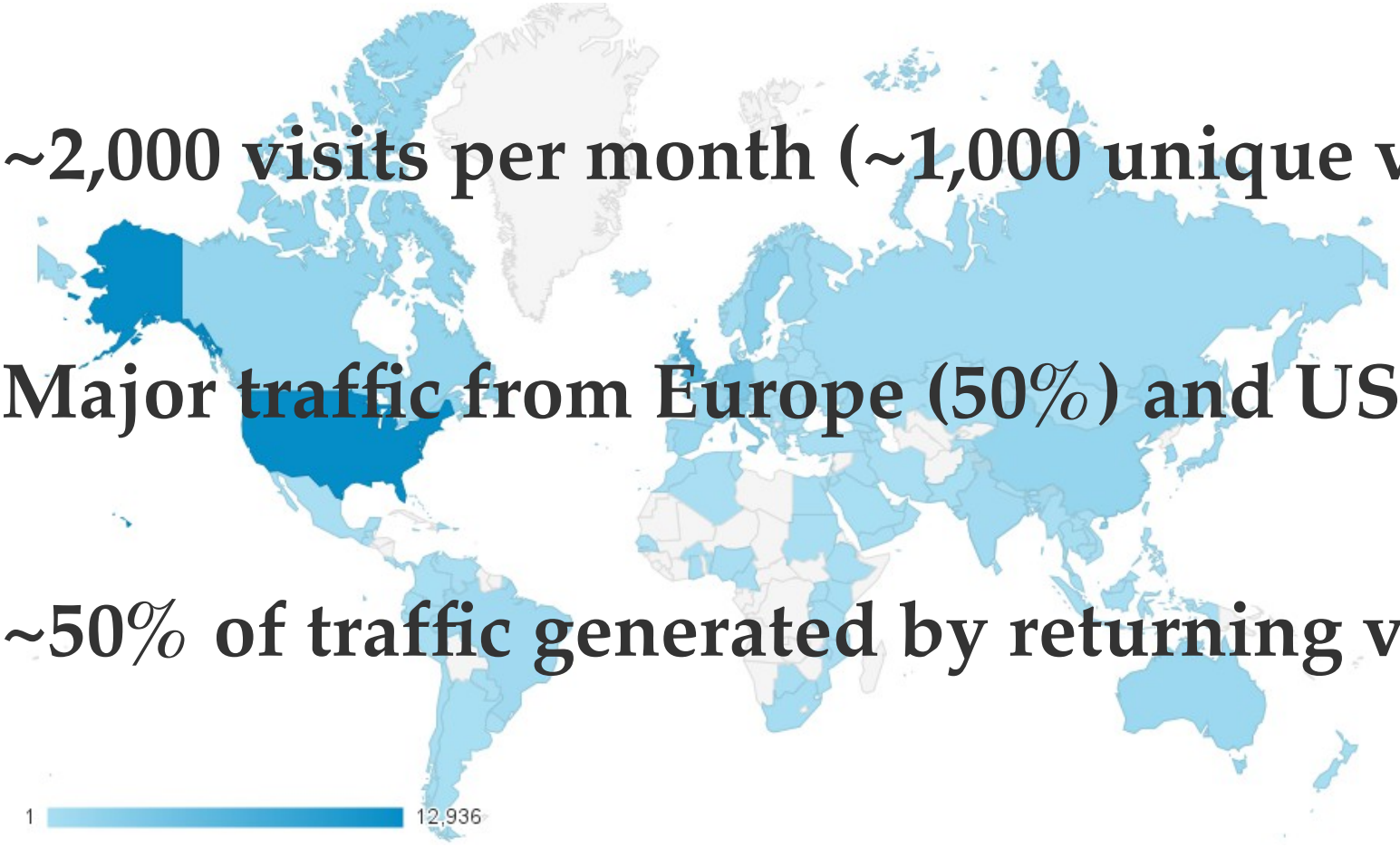
Documentation	
Manuals	>200 pages
Tutorials	>250 pages
Videos	~10 min

Publications	
Total	7 (4)
# citations	>700 (>500)

GenABEL
STATISTICAL GENOMICS

www.GenABEL.org

- ~2,000 visits per month (~1,000 unique visitors)
- Major traffic from Europe (50%) and US (25%)
- ~50% of traffic generated by returning visitors



GenABEL
STATISTICAL GENOMICS

GenABEL-package

Genome-wide analysis of association between directly typed SNPs and quantitative, binary and time-till-event outcomes

Type of analysis	# functions
Data manipulations	~40
Quality control and descriptives	~10
Analysis	~30
Graphics & data presentation	~5
Total	**391**

Highlights:

- Converters between different data formats
- Powerful QC organized around the check.marker() function
- A line of mixed-models based tools for correction for population stratification

GenABEL
STATISTICAL GENOMICS

Other R-packages

GWAS analyses

- *VariABEL (5):* tools for "environmental sensitivity" vGWAS
- *MixABEL (12):* advanced mixed models for GWAS

Post-GWAS

- *MetABEL (7):* meta-analysis of GWAS results
- *PredictABEL (111):* assessment of (genetic) risk prediction models

Support

- *DatABEL (72):* out-of-RAM large matrices storage and access
- *ParallABEL (52):* parallelization algorithms for GWAS

GenABEL
STATISTICAL GENOMICS

Non-R packages

ProbABEL: GWAS of imputed data (quantitative, binary, time-till-event traits; regression and mixed models)

Filevector: C++ base for the DatABEL-package, facilitating out-of-core computations on large matrices

OmicABEL: rapid mixed-model based GWAS especially for multiple trait ("omics") analysis.

GenABEL
STATISTICAL GENOMICS

Outline

- Statistical genomics
- A short history
- Current state
- **Summary**

GenABEL
STATISTICAL GENOMICS

Summary

- GenABEL is problem-centered project aiming towards agile development of statistical genomics methodology
- The GenABEL suite consist of 9 packages implementing close to 1,000 functions facilitating analyses of polymorphic genomes
- GenABEL suite is widely used for GWAS analyses of human, farm, pet animal, and plant data
- The project runs on enthusiasm and spare time of several people (and $10 a month from "YuriiA consulting")

GenABEL
STATISTICAL GENOMICS

Difficulties we face

Core functionality

- The project would gain from re-design and added "core" functionality (e.g. regarding access to different data formats; parallelization)
- This gain is on the project, and not individual developer's level

Coordination and communication

- Coordination takes time
- It may take a while before problems well-known for developers get through to the end-user (anyone willing to become our PR officer?)

GenABEL
STATISTICAL GENOMICS

Vacant roles

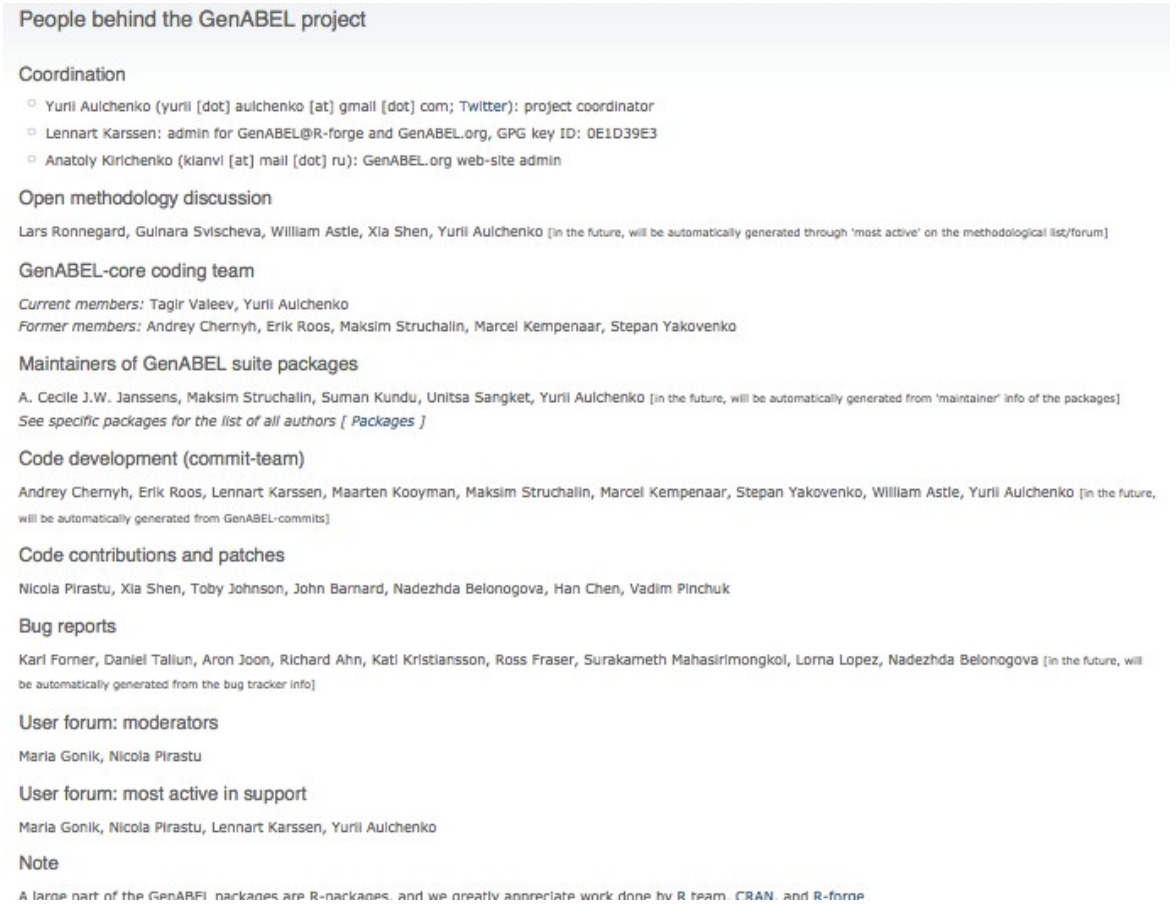
- Project Coordinator
- Lead Developer(s)
- Public Relations Officer

GenABEL
STATISTICAL GENOMICS

Current support for the project

GenABEL
STATISTICAL GENOMICS

Acknowledgements



People behind the GenABEL project

Coordination

- Yurii Aulchenko (yurii [dot] aulchenko [at] gmail [dot] com; Twitter): project coordinator
- Lennart Karssen: admin for GenABEL@R-forge and GenABEL.org, GPG key ID: 0E1D39E3
- Anatoly Kirichenko (kianvi [at] mail [dot] ru): GenABEL.org web-site admin

Open methodology discussion

Lars Ronnegard, Gulnara Svischeva, William Astle, Xia Shen, Yurii Aulchenko [in the future, will be automatically generated through 'most active' on the methodological list/forum]

GenABEL-core coding team

Current members: Tagir Valeev, Yurii Aulchenko
Former members: Andrey Chernyh, Erik Roos, Maksim Struchalin, Marcel Kempenaar, Stepan Yakovenko

Maintainers of GenABEL suite packages

A. Cecile J.W. Janssens, Maksim Struchalin, Suman Kundu, Unitsa Sangket, Yurii Aulchenko [in the future, will be automatically generated from 'maintainer' info of the packages]
See specific packages for the list of all authors [Packages]

Code development (commit-team)

Andrey Chernyh, Erik Roos, Lennart Karssen, Maarten Kooyman, Maksim Struchalin, Marcel Kempenaar, Stepan Yakovenko, William Astle, Yurii Aulchenko [in the future, will be automatically generated from GenABEL-commits]

Code contributions and patches

Nicola Pirastu, Xia Shen, Toby Johnson, John Barnard, Nadezhda Belonogova, Han Chen, Vadim Pinchuk

Bug reports

Karl Forner, Daniel Taliun, Aron Joon, Richard Ahn, Kati Kristiansson, Ross Fraser, Surakameth Mahasirimongkol, Lorna Lopez, Nadezhda Belonogova [in the future, will be automatically generated from the bug tracker info]

User forum: moderators

Maria Gonik, Nicola Pirastu

User forum: most active in support

Maria Gonik, Nicola Pirastu, Lennart Karssen, Yurii Aulchenko

Note

A large part of the GenABEL packages are R-packages, and we greatly appreciate work done by R team, CRAN, and R-forge

CRAN

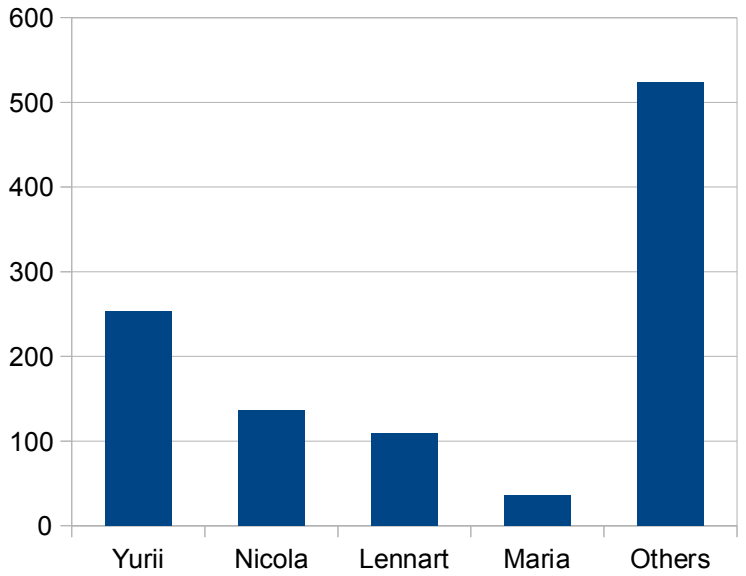
R-Forge

GenABEL
STATISTICAL GENOMICS

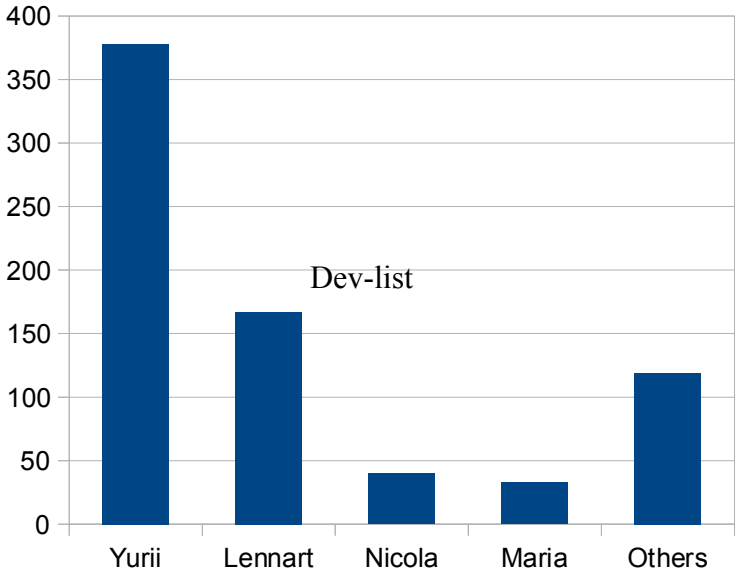
Key people

Lennart Karssen, Nicola Pirastu, Maria Gonik

Forum (431/70) members



Dev-list (45/12) members



GenABEL
STATISTICAL GENOMICS