R in Systems Biology and Translational Medicine

Gang Chen @ BGI-Shenzhen
chengang@genomics.cn

November 3, 2012

Outline

Next

What is Systems Biology and Translational Medicine?

Systems Biology

The study of the interactions between the components of biological systems, and how these interactions give rise to the function and behavior of that system.
—Wikipedia

Translational Medicine

"...translate the remarkable scientific innovations we are witnessing into health gains ..."
— Elias A. Zerhouni, NEJM 2005

Challenges

Challenges

- Reliable and high-throughout experimental techniques
- High-quality and large-scale sample collection

Challenges

Challenges

- Reliable and high-throughout experimental techniques
- High-quality and large-scale sample collection
- Money

Challenges

Challenges

- Reliable and high-throughout experimental techniques
- High-quality and large-scale sample collection
- Money
- Big and complex data analysis

BIG Biological Data

- One human genome, 3 GB
- 60X depth sequencing of human genome, $3 \times 60 = 180$GB
- A typical study of complex disease need 2K samples in the 1st stage and more than 10K samples for validation.
- $1\text{K} \times 60 \times 3\text{GB} = 180\text{TB}$

BIG Biological Data

- One human genome, 3 GB
- 60X depth sequencing of human genome, $3 \times 60 = 180$GB
- A typical study of complex disease need 2K samples in the 1st stage and more than 10K samples for validation.
- $1\text{K} \times 60 \times 3\text{GB} = 180\text{TB}$

-

Complex Biological Data

Why?

- Biological system is very complicated
- Biological data is very complicated

- DNA
- RNA
- Protein
- Metabolite
- Microbe
- Virus

Next

Diseases

Disease

- Genetic factor
- Environmental factor

Examples

- Traffic accident
- Psoriasis
- Tumor
- ……

Complex Diseases

Complex Disease

- Genes
- Gene Interactions
- Subtypes
- Environment
- Nation
- ……

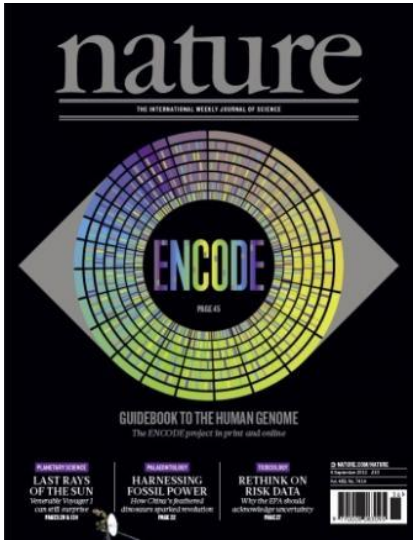
Complex Diseases

Complex Disease

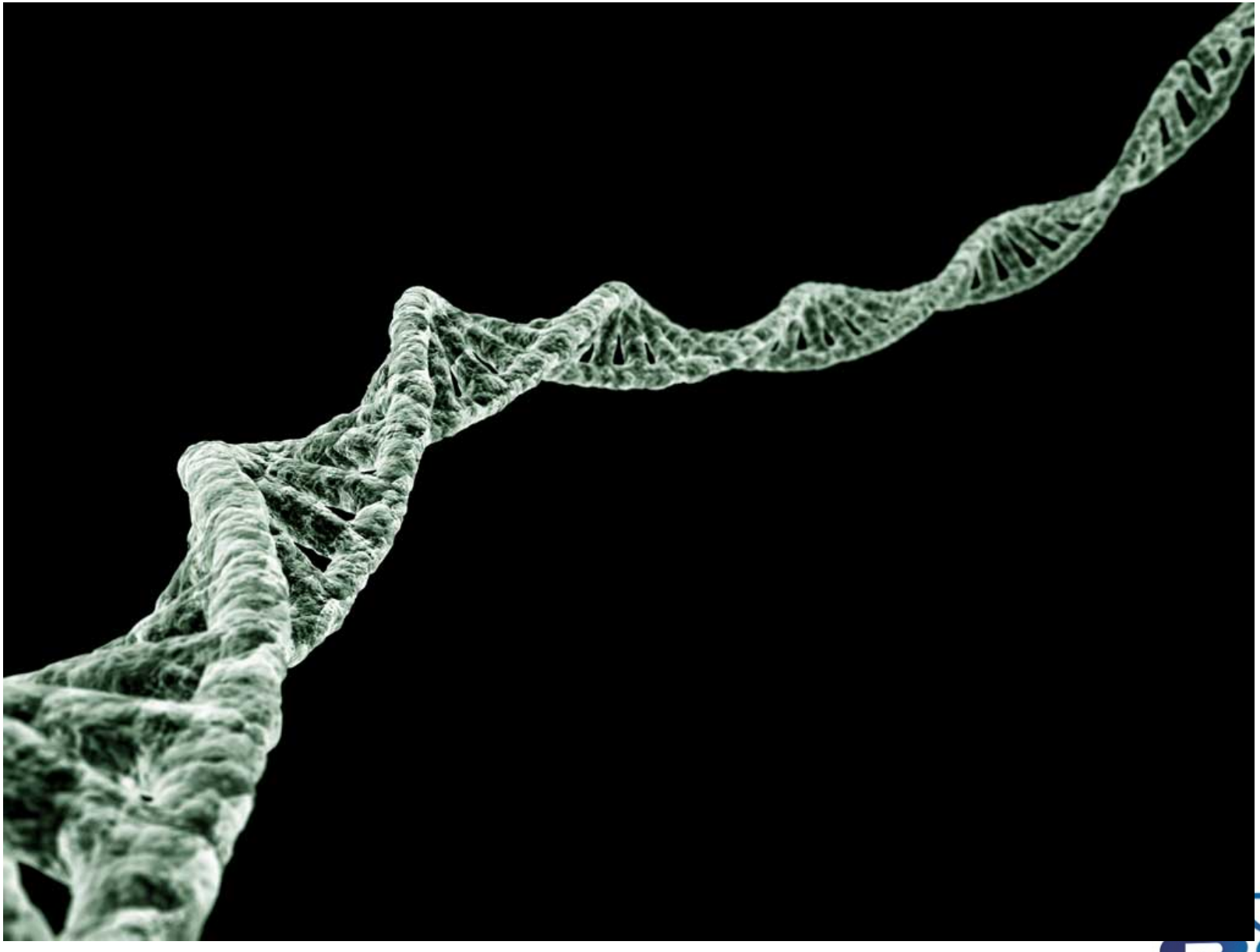
- Genes
- Gene Interactions
- Subtypes
- Environment
- Nation
- ……

You are not an real scientist!

You are not an real scientist!



Could you please help me to find the key I lost in the mountain?

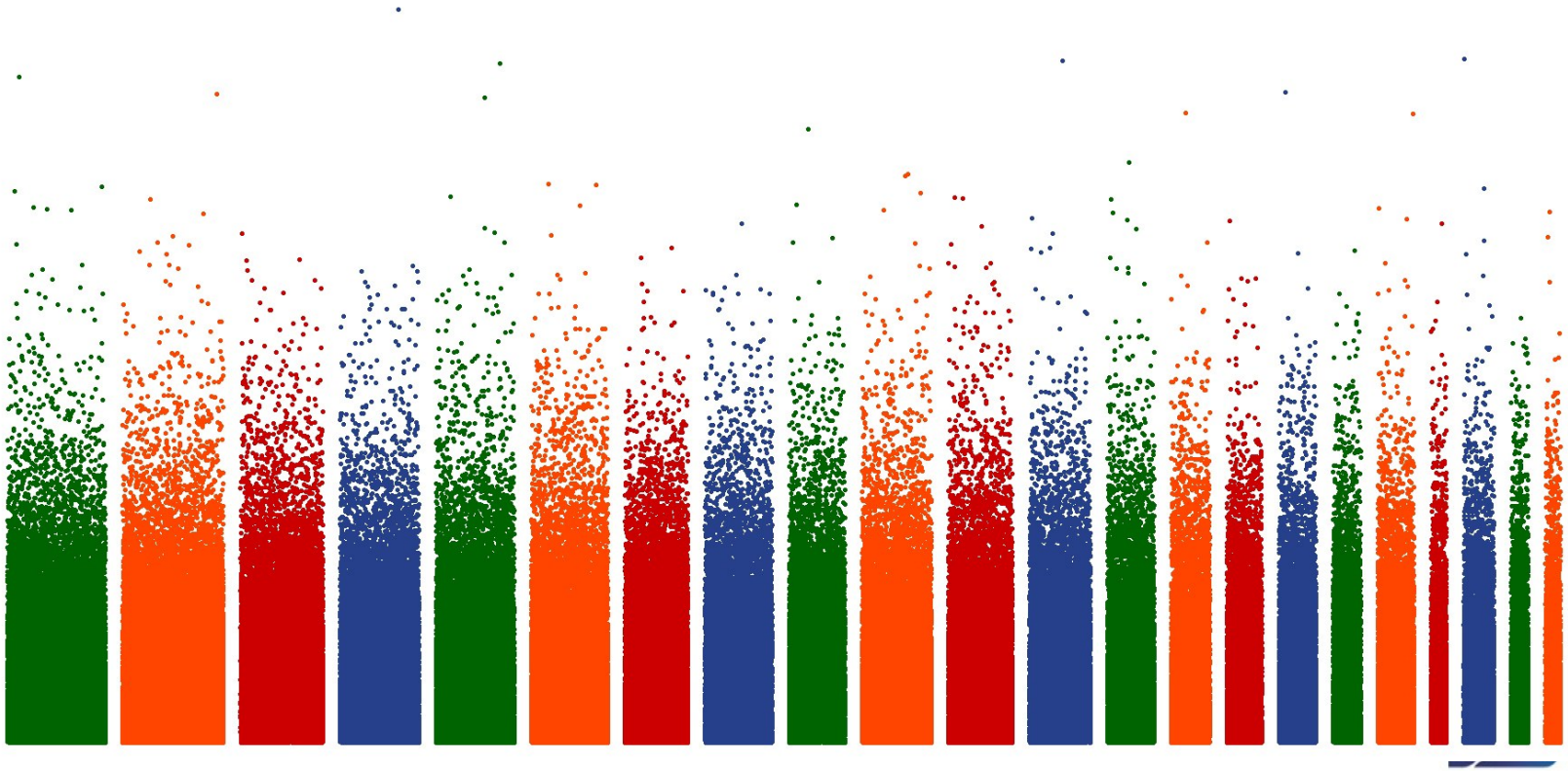


Where is the disease related mutation?

Genome-Wide Association Study

GWAS

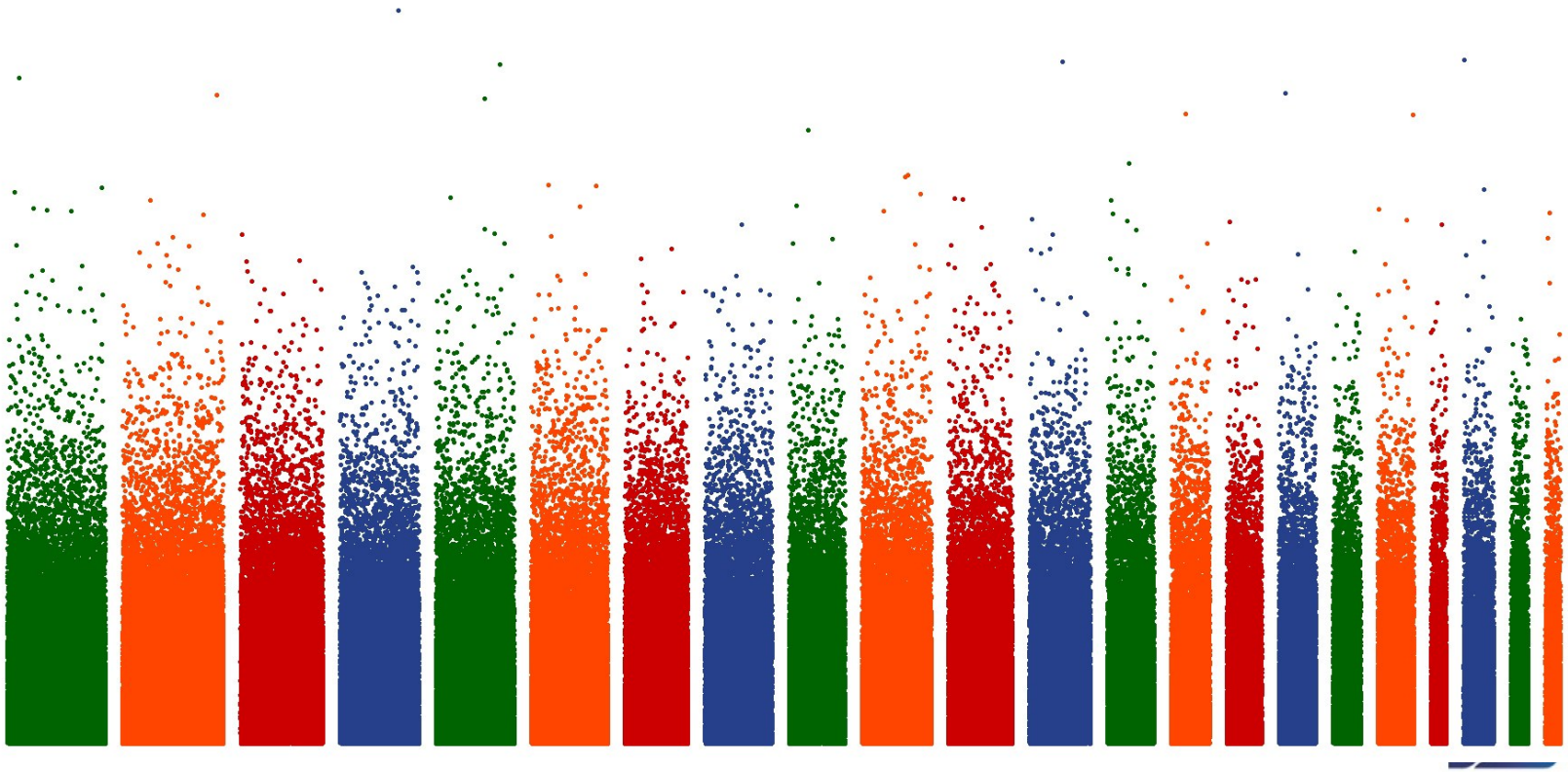
Check the differences of genotype of each loci in genome between patient and healthy people.



Genome-Wide Association Study

GWAS

Check the differences of genotype of each loci in genome between patient and healthy people.



Nature Genetics → Nature GWAS

Next

Why do we want to use R?

- Rapid implementation
- Active community, COS and CRAN
- Statistics is the fundamental of big data analysis
- Data visualization, ggplot2!
- Integration with other programming language, Rcpp
- Widely supported by commercial products

Why do we want to use R?

- Rapid implementation
- Active community, COS and CRAN
- Statistics is the fundamental of big data analysis
- Data visualization, ggplot2!
- Integration with other programming language, Rcpp
- Widely supported by commercial products
- We love R!

Why do we have to use R?

- Most new methods are implemented as R packages
- Bioconductor
- Free and Free
- Training and learning material

Why do we have to use R?

- Most new methods are implemented as R packages
- Bioconductor
- Free and Free
- Training and learning material
- We cannot live without R!

How do we use R?

R in disease study

- Statistical Analysis
- Classification and regression for disease diagnosing and prediction
- Feature selection for disease gene identification
- Visualization for publication and customers

What we have done?

R in BGI

- Thousands of researchers are using R
- R is installed on almost every computer and server
- Most plots in our world-class publications are generated by R
- A novel multi-loci GWAS framework for complex diseases
- A framework for various biological data integration
- An auto report-generation system for personal genome sequencing service

What we have done?

R in BGI

- Thousands of researchers are using R
- R is installed on almost every computer and server
- Most plots in our world-class publications are generated by R
- A novel multi-loci GWAS framework for complex diseases
- A framework for various biological data integration
- An auto report-generation system for personal genome sequencing service Knitr!
- ……

华大基因 BGI

Problems

Problems

- Where is R expert?
- R is a statistical tool or a programming language for data analysis?
- How to use R to handle TB data? Hadoop?
- Interfaces for biological data: sequences, interaction, model ...
- Algorithms for extreme high dimension data
- ……

Expectation

Future

- R for big data
- Packages for new biological problems
- Soft-engineering methodology and tool-chain for R

Expectation

Future

- R for big data
- Packages for new biological problems
- Soft-engineering methodology and tool-chain for R
- R Geek

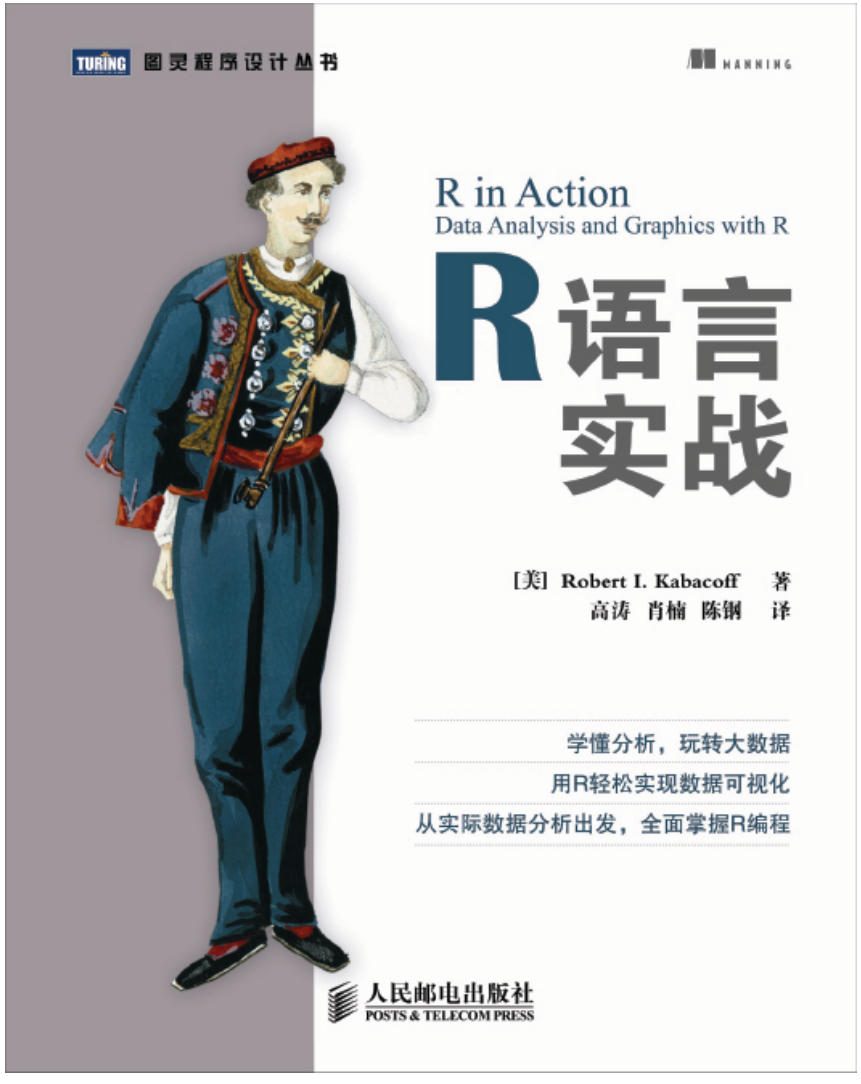
Next

Conferences

Conferences

- ISCB-Asia
  Dec 17 - 19, Shenzhen
- AYRCOB
  Dec 20 - 21, Shenzhen
  http://2012.ayrcob.org
- 7th ICG & Bio-IT APAC
  Nov 28 - Dec 1, Hong Kong

Book


TURING 图灵程序设计丛书
MANNING
R in Action
Data Analysis and Graphics with R
R语言实战
[美] Robert I. Kabacoff 著
高涛 肖楠 陈钢 译
学懂分析，玩转大数据
用R轻松实现数据可视化
从实际数据分析出发，全面掌握R编程
人民邮电出版社
POSTS & TELECOM PRESS

Q & A

Thanks

We are seeking data guru and collaboration!

chengangcs@weibo
chengang@genomics.cn

http://gossipcoder.com

华大基因 BGI