Importance of data science in the field of pharmacogenomics

Shigeo Kamitsuji, PhD
Statistical Genetics Analysis Division
StaGen Co., Ltd.

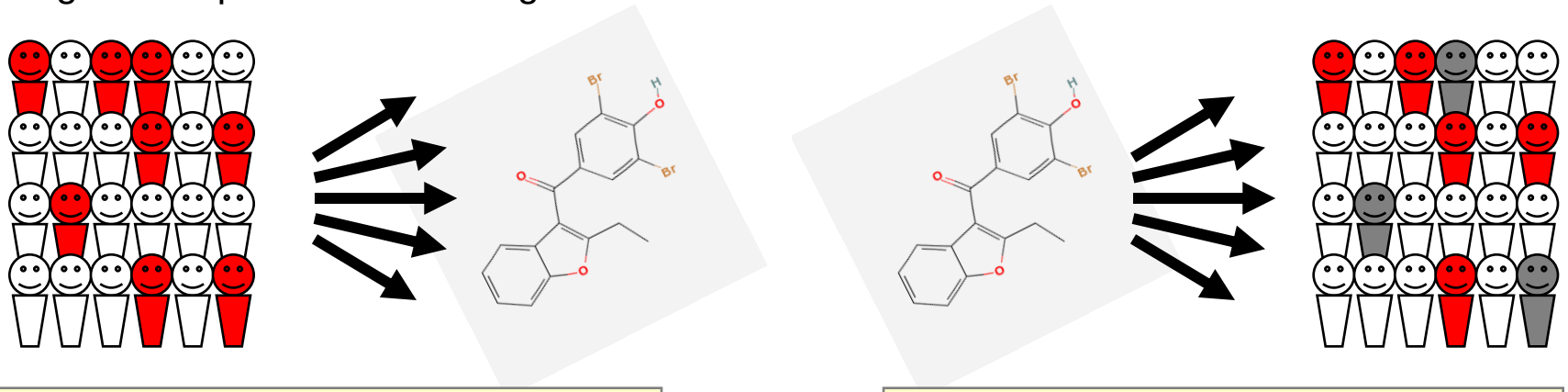
Pharmacogenomics and Personalized medicine

PharmacoGenomiCS (PGx)

Effects of drugs are studied based on individual genomic sequences

Personalized medicine (order-made medicine)

Personalized medicine is a new concept in medical treatment that implies the delivery of drugs to the patient based on genomic variation.



Suitable medicine for all patents

Suitable medicine for each patient

Association study using the genome data

Candidate gene-based association study

- the approach for evaluatinng the association between the phenotype and a specific genome sequence selected based previous study and medical aspects.

Genome-Wide Association Study (GWAS)

Approach for identifying the genetic variation assoicated with the phenotype from whole genome in human

- All SNPs loci in human have been identified by HapMap project (2002).
- 500K or 1 M SNPs for each person can be observed immediately and inexpensively by using DNA chips.
- Algorithms for implementing the genome-wide association analysis have been developed.

Drug document recommended by FDA in US

drug	disease	association	Document
Herceptin	breast cancer	effect	Work for women whose breast cancer cells carry extra copies of a protein called HER2
Mercaptopurine	childhood leukemia	adverse events	Do not use to patients with TPMT deficiency
Strattera	attention deficit disorder	adverse events	Differences in genomic sequences in CYP2D6 gene cause different probabilities of adverse events
Tarceva	lung cancer	effect	Work for individualwhose lung cancer cell carry EGFR
irinotecan	cancer	adverse events	Association between genetic polymorphism of UGT1A1 gene and severe adverse reactions to irinotecan
BiDil	cardiac failure	effect	For African-American

Statistical Genetics

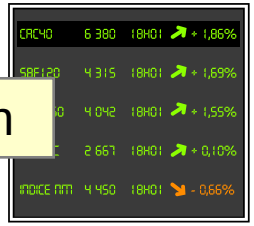
Mathematical statistics

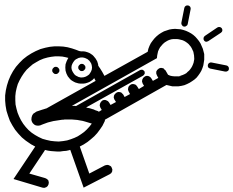
station
area
walks

TOPIX
rate
weather

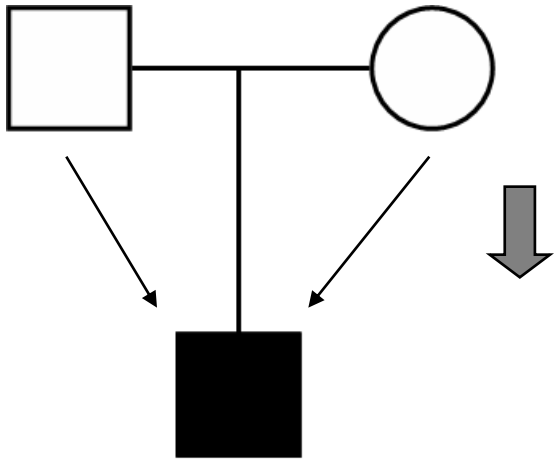
age
drugs
medical data

Modelization





Statistical genetics

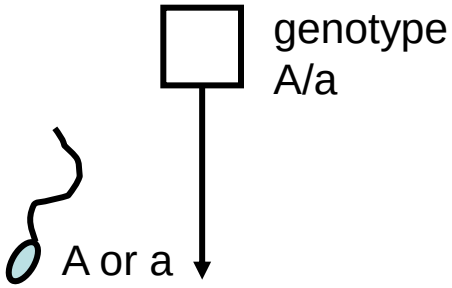


Laws of inheritance
Statistical model is known.

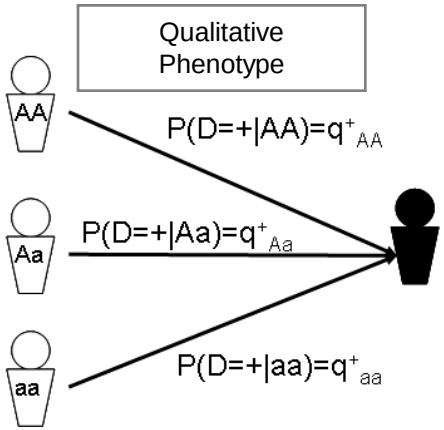
Laws of Inheritance

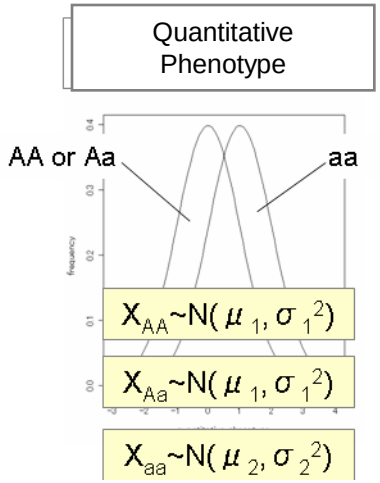
Mendelian inheritance (Mendel's law)

Law of Segregation

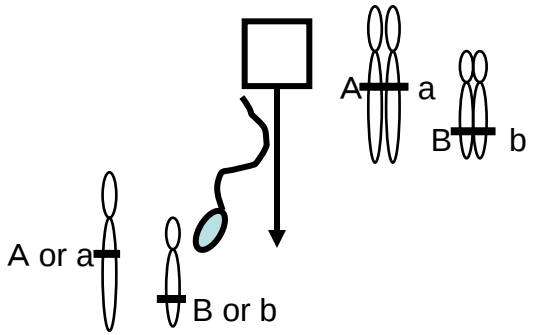


Law of Dominance

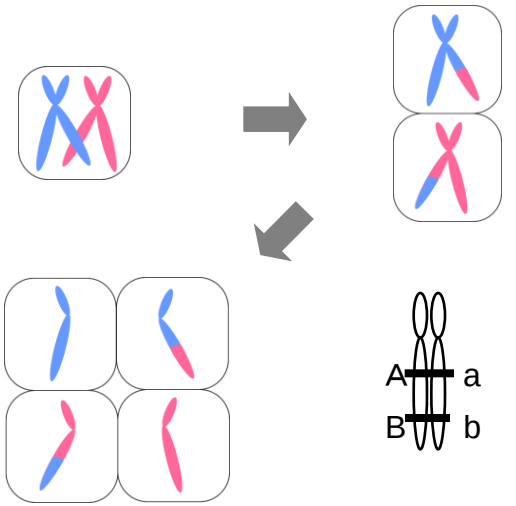




Law of Independent Assortment



Crossover/Recombination



Laws of Inheritance

Mendelian inheritance
+
Recombination

Objects satisfied with laws of inheritance

Allele	
Allele of an offspring are surely inherited from the parents according to the laws of inheritance	
Locus (loci)	**Genotype**
A position on a chromosome two alleles exist	A pair of alleles

Polymorphism		alleles	levels
SNP	(Single Nucleotide Polymorphism locus)	A,T,G,C	2
STRP	(Short Tandem Repeat Polymorphism locus)	number of repeat	≧2
DIP	(insertion/deletion locus)	Ins,Del	2
VNTR	(Variable Numbers of Tandem Repeat locus)	number of repeat	≧2
CNV	(Copy Number Variation)	number of duplication	≧2

Genotype
SNP, STRP, CNV,…

Phenotype
Affected/Non-affected,
quantitative bio-marker,
amount of protein (expression)

Failures to replicate the results of genome association study

ANALYSIS

nature genetics

Systematic meta-analyses of Alzheimer disease genetic association studies: the AlzGene database

/naturegenetics

Lars Bertram[1], Matthew B McQueen[2,3], Kristina Mullin[1], Deborah Blacker[2,4] & Rudolph E Tanzi[1]

Literature searches

Studies included in AlzGene and comparison to other databases. The results presented here are based on a 'data freeze' of the AlzGene database on December 1, 2005, and they cover 789 publications reporting on 802 different polymorphisms in 277 genes (after screening ~23,500 titles and abstracts; **Supplementary Fig. 1** online). Since that time, we have continued our systematic screen of the literature, and as of August 15, 2006, AlzGene included the data of 875 publications (representing 1,055 polymorphisms and 355 genes). To test our ability to capture all of the published genetic association data targeted for AlzGene, we compared the studies we identified to those in two other publicly available databases with a similar focus (HuGENet and GAD). Across ten randomly selected genes, AlzGene identified 112 publications, and HuGENet and GAD list 77 and 20 studies, respectively (**Supplementary Table 1** online).

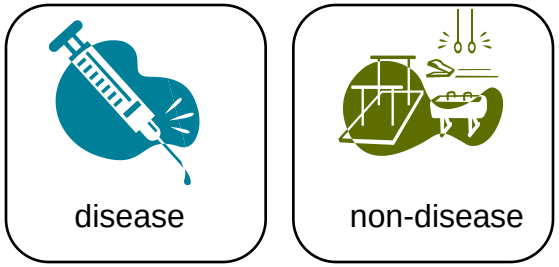
Conclusions

In this study, we have conducted the most comprehensive assessment of currently available data on the genetics of Alzheimer disease and, to the best of our knowledge, of data on any genetically complex disease. Based on the allele distributions of genetic variants with available data in at least three independent case-control samples, we systematically meta-analyzed 127 polymorphisms across 69 different putative Alzheimer disease risk genes, following recently suggested guidelines for the meta-analysis of genetic association data[22–24] and its online curation[8,25]. In addition to *APOE*-ε4 and four other probably ε4-related effects, we discovered 20 polymorphisms in 13 genes that yielded significant allelic summary ORs. Although these ORs were generally modest (showing average 'risk' effects of 1.25 and average 'protective' effects of 0.82), they were very similar to those estimated in previous large-scale meta-analyses across a range of different diseases[5,6,16]. In collaboration

Three key words for SGA in StaGen

Study design

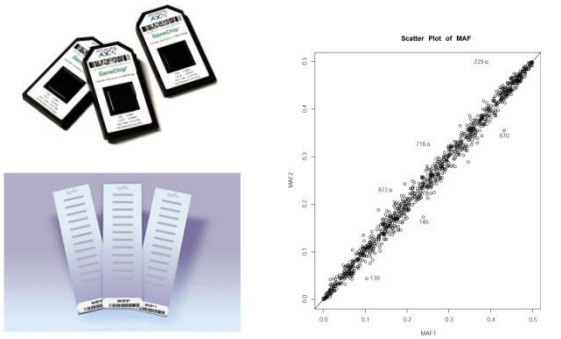
case group control group



How large should the size be to achieve 80% power in case-control studies for a SNP with the odds ratio of 3?

Appropriate study design to avoid incorrect conclusions

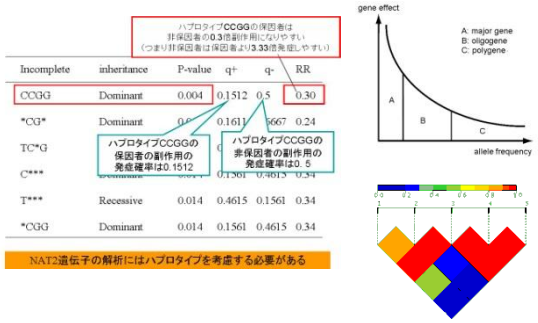
Quality control of genotype data



It is inevitable to remove genotype data which are incorrect or not in accord with the laws of inheritance. Multiple comparison problem in genome-wide SNP analysis should also be considered.

Unreliable information is eliminated and incorrect results are avoided.

Analysis based on laws of inheritance



For obtaining reliable conclusions, genome-wide association study (gwa) and haplotype association study should be based on the laws of inheritance. Multiple comparison problem in gwa is critical.

Analysis based on the laws of inheritance to obtain the results with high reproducibility.

"Sophisticated" Statistical Genetics Analysis＝SGA

Services we provided

Intellectual property Division

- This division transfers technology nurtured at Tokyo Women's medical university into thrid parties.
  - Patents in terms of the relationship between phenotype and genotype

Services of Statistical Genetics Analysis Division

- Training in SGA
- SGA on request and Temporary research personnel service
- Consultation in SGA
- System development

Intellectual property Division

Patnet No.	Registration No.3656952 (Japan)
Project Name	SAA1／Amyloidosis
Title	Method for Detecting Morbidity Risk of AA-Amyloidosis in Rheumatoid Arthritis Patient by Detecting a New SNP in SAA1 Genetic Locus
Abstract	To provide a method capable of detecting morbidity risk of AA-amyloidosis in rheumatoid arthritis using new oligonucleotides probes in the method. SOLUTION: In this method for the detecting SAA1 (serum amyloid A1 protein) gene, the morbidity risk of AA-amyloidosis in rheumatoid arthritis, it is determined whether a base at the -13 of human SAA1 gene is thymine or cytosine. The DNA contains a base sequence GCCACCGTTC CCTGG or a base sequence GCCACCGCTC CCTGG.
Working example	DNA chip etc.

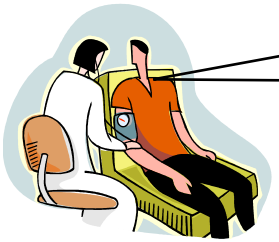
Patnet No.	Registration No.3839836 (Japan)
Project Name	MTX
Title	Method to administer methotrexate (MTX)
Abstract	Method to estimate the effective dose of MTX for the treatment of rheumatoid arthritis. The individual effecctive dose of MTX can be estimated by the analysis of a SNP in MTHFR gene.
Working example	DNA chip etc.

Patents in terms of the relationship between phenotype and genotype

SGA training: Basic course

Fundamental knowledge about PGx

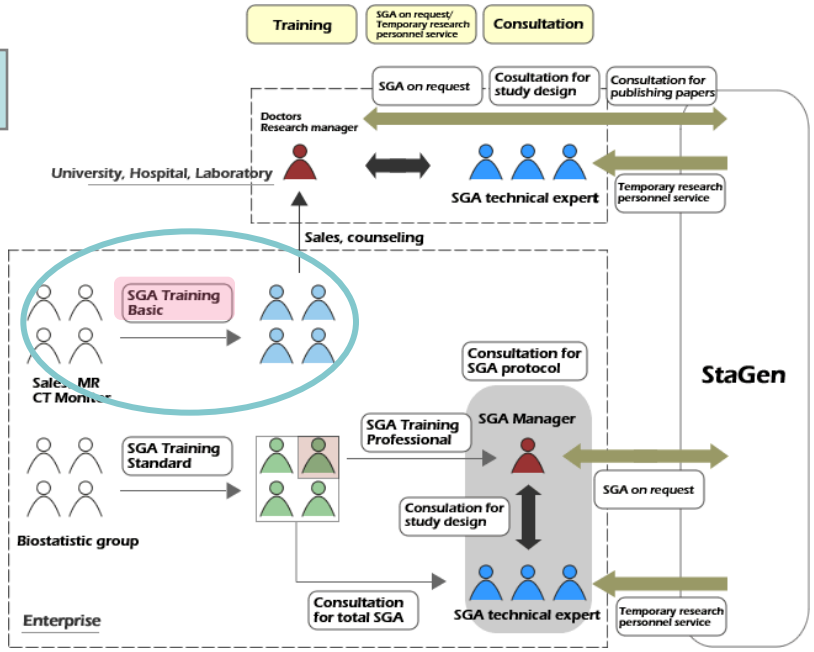
- What is the purpose of PGx?
- How can we extract knowledge from PGx?
- How can the results from PGx be applied?



Genome study may provide you with medicine with higher quality.

How has the evidence for the association between the adverse events and a genetic polymorphism been obtained?

This result was obtained by a GWA study. There is a report that the patients with genotype A/A at locus y in gene *x* have 2.8 times more chance to suffer from the adverse events by the drug.



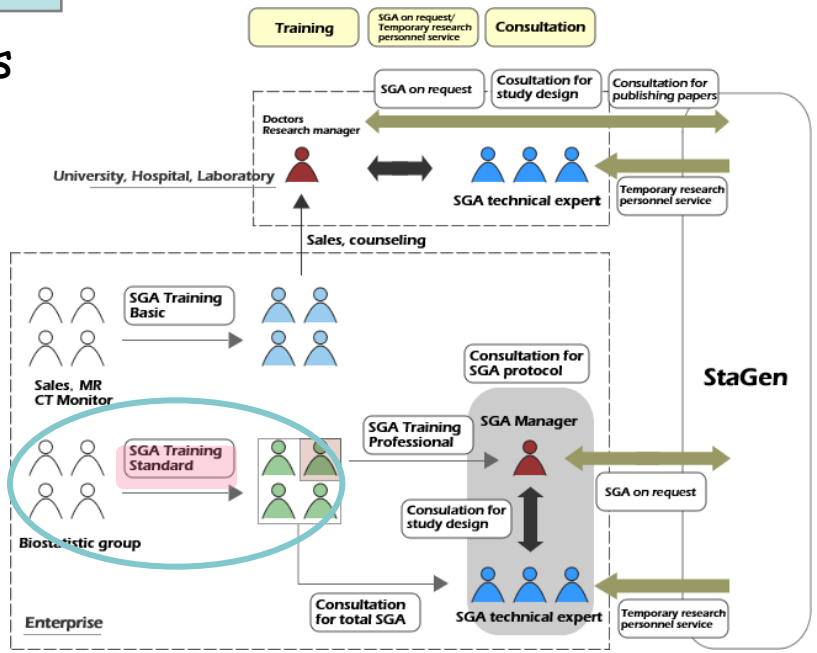
Knowledge about PGx is necessary for various professionals such as drug-developers, medical doctors, nurses, medical representatives (MR), genetic counselors, and pharmacists.

SGA training: Standard course

Acquisition of higher level knowledge about PGx

- Relationship between biological data and laws of inheritance
- Understanding of the mathematical aspects of SGA
- Exercise using SGA tools

+



"Introduction of statistical genetics"

by Naoyuki Kamatani

Concise explanation

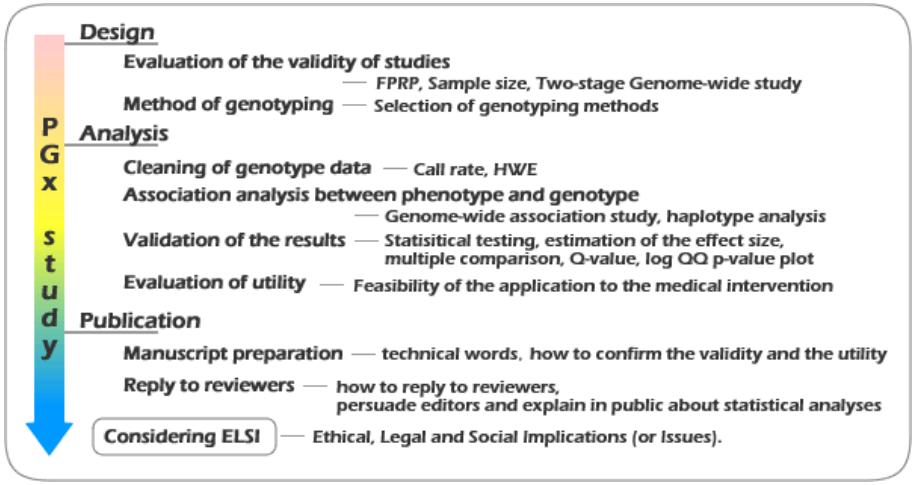
by StaGen

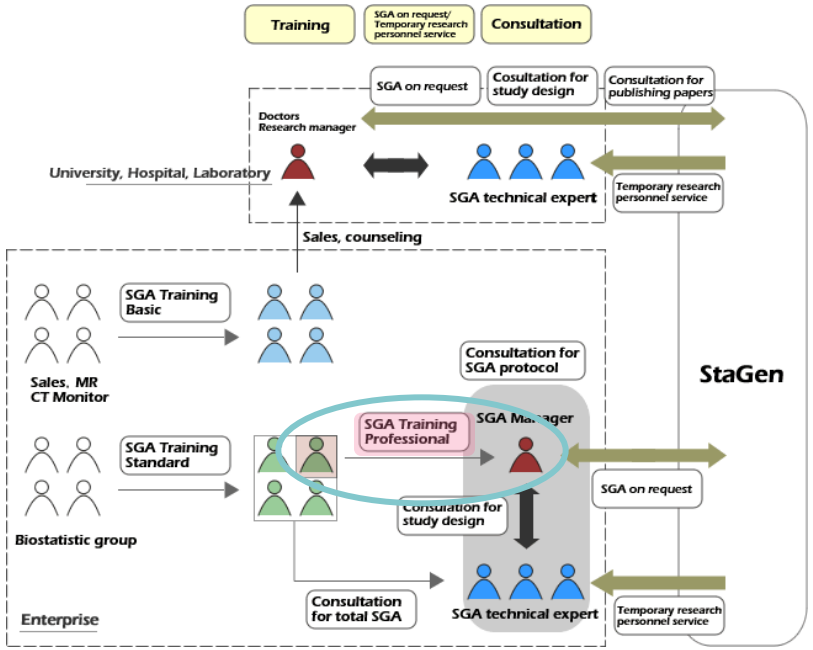
Education of knowledge and techniques in statistical genetics for SGA managers and SGA technicians

SGA training: Professional course

Training course for SGA managers

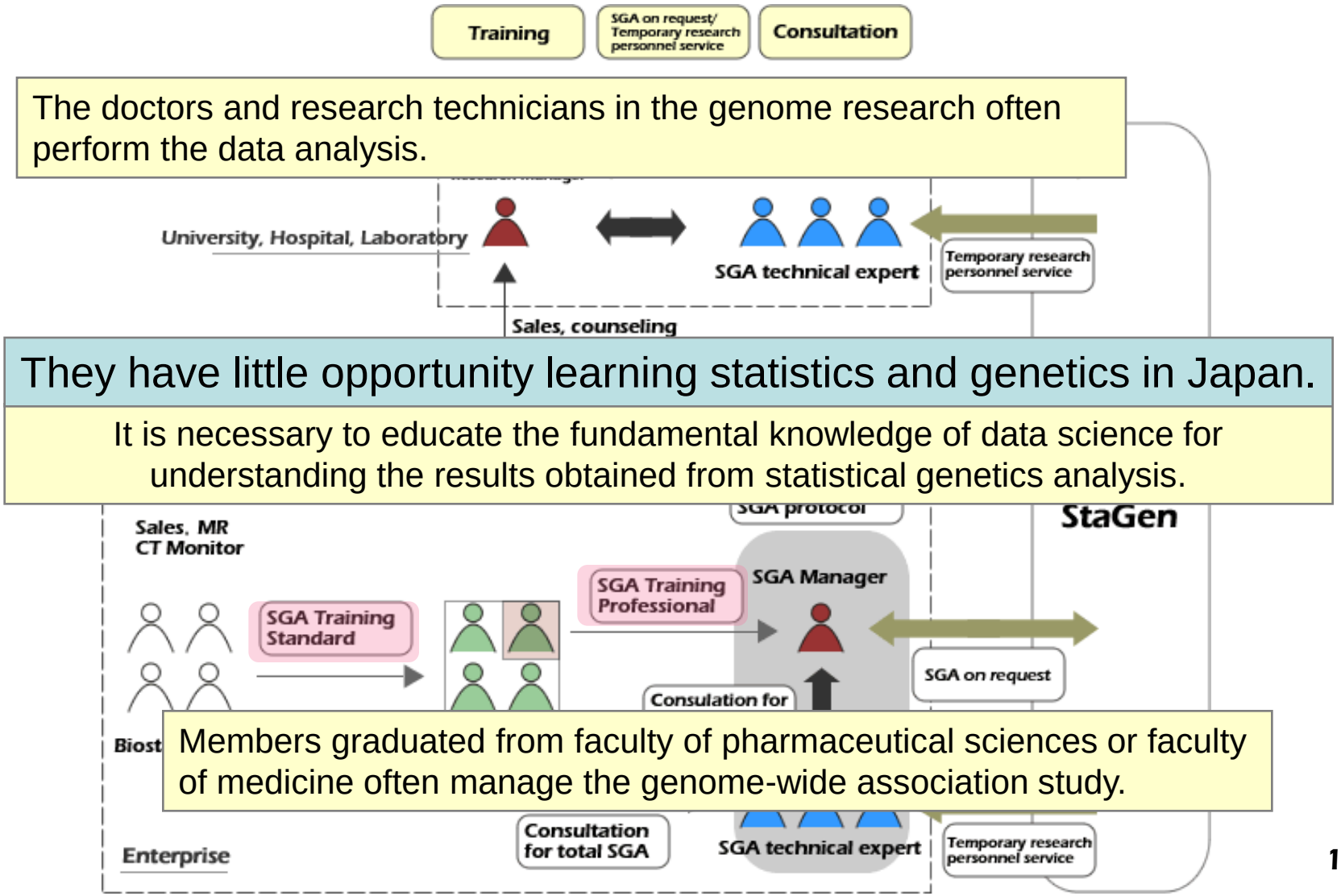
- Optimal designing of PGx studies given various factors
- Training for practical PGx studies using simulation and real samples



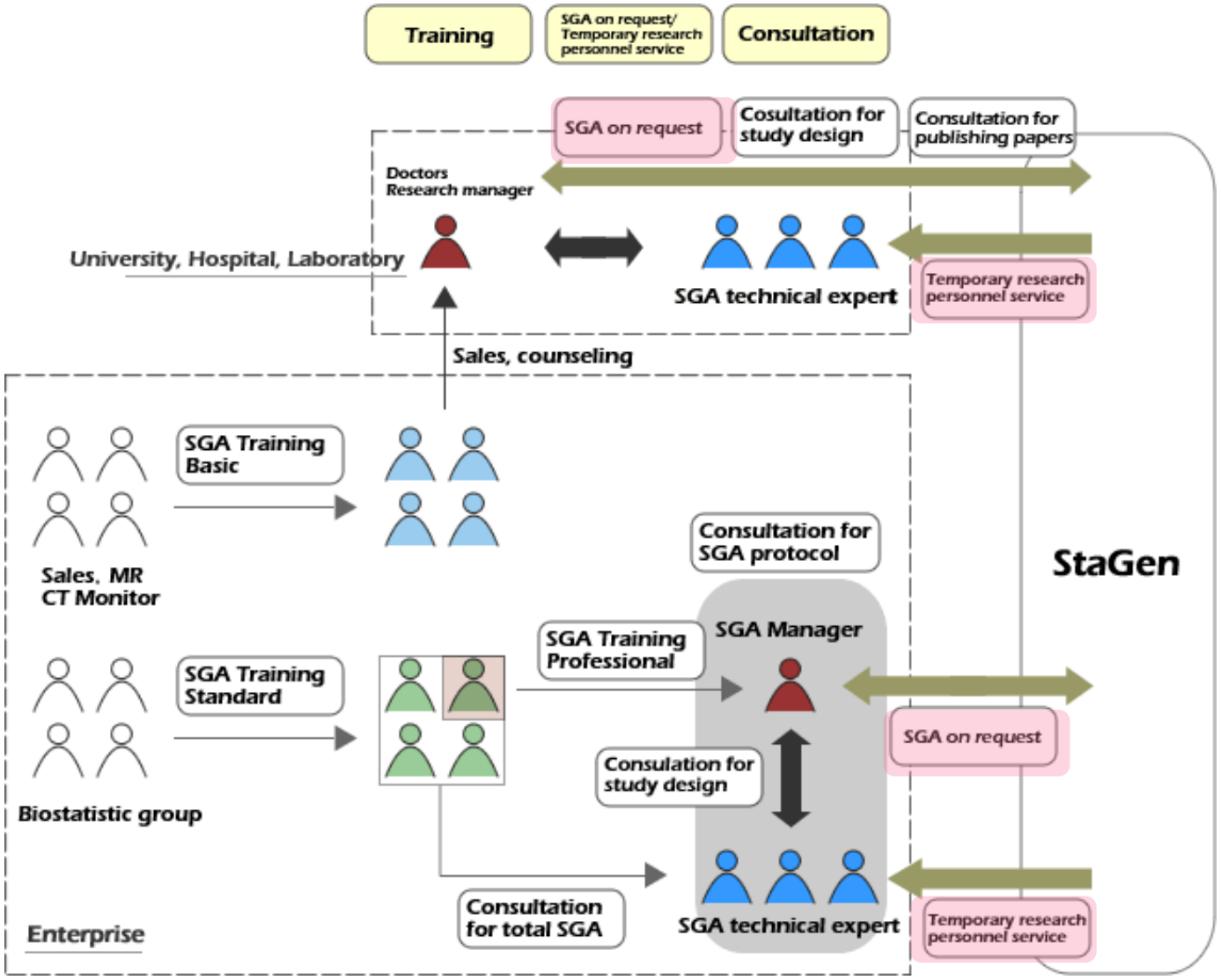


Training for SGA managers as supervisors of PGx studies

SGA Training service



SGA on request and temporary research personnel service



SGA on request and temporary research personnel service

Analysis of your data by SGA specialists in either your laboratory or our laboratory

- High quality SGA based on our protocol
- Analysis based on original algorithms developed by Naoyuki Kamatani, M.D., Ph.D, Tokyo Women's Medical University

ERS (in Japan)

Qualification of temporary personnel service (in Japan)

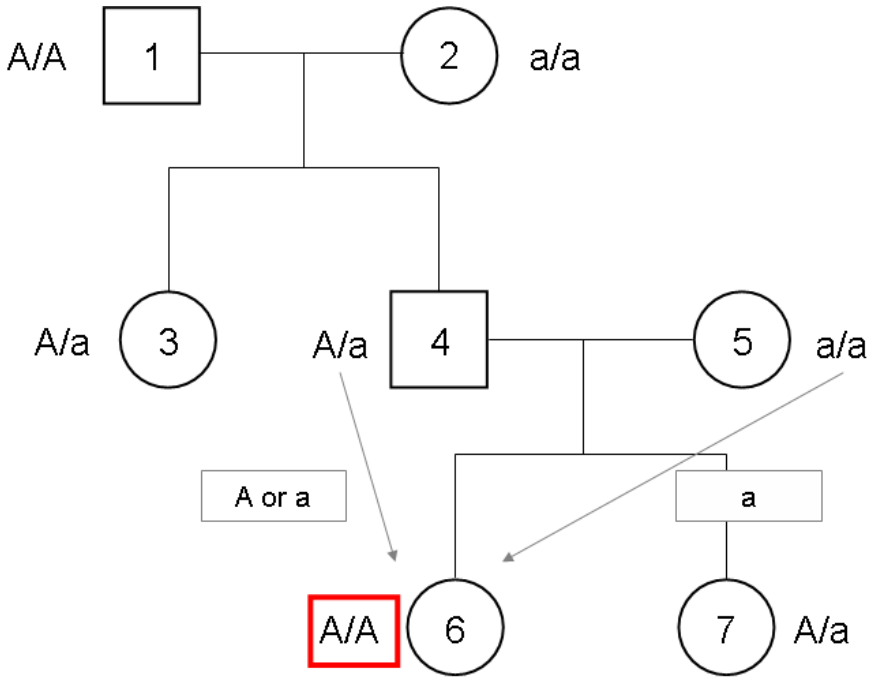
PGx study system under the leadership of SGA manager

Examples of SGA

Category	Contents	
Establishment of framework	Construction of the data format	
	Construction of the framework for SGA	
Preparation for research	Designing of research study	
	Evaluation of previous researches	
Data cleaning	Removal of inappropriate data	Outoff threshold for call rate
		Evaluation by replication test
		Evaluation by MAF
		Comparison of the data between different conditions
		Estimation of typing error rate
		Use of the data for sex chromosomes based on biological laws
	Fitness to the laws of inheritance	Test of goodness of fit to HWE
Association study	Genome-wide SNP research	Statistical test
		Evaluation by log QQ p-value plot or FDR (Q-Value)
	Diplotype configurations (haplotype) association study	Determination of LD block
		Test of association based on diplotype configurations study
		Inference of diploype configurations
		Haplotype inference and test of association based on haplotypes
	Pedigree data analysis; linkage analysis	Parametric linkage analysis
		Nonparametric linkage analysis
		TDT(Transmission Disequilibrium Test)
Others	Examination of population structure	
	Evaluation by meta analysis	

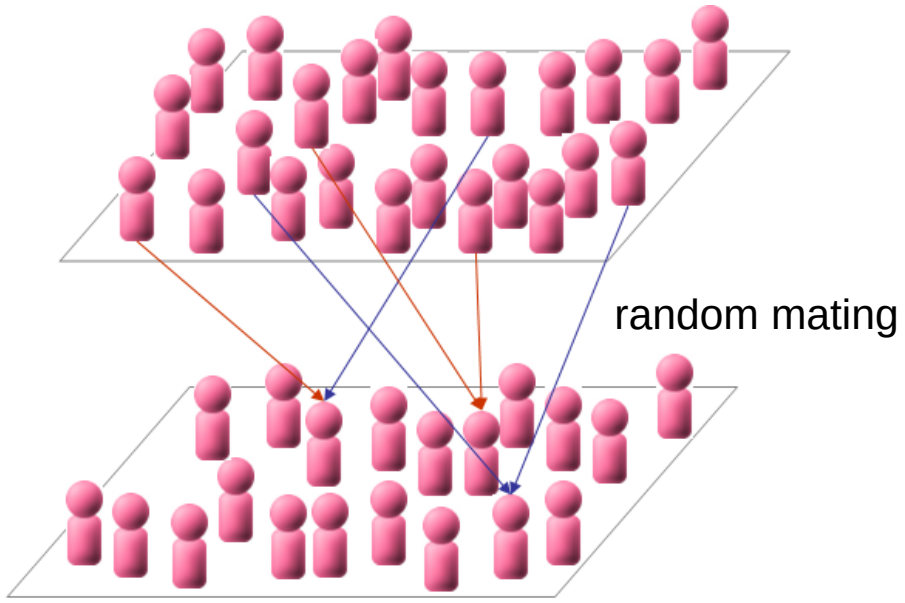
Quality evaluation of observed genotype data

Pedigree data



It is important to eliminate contradictions concerning Mendelian inheritance from genotypic data with pedigree information.

Population data



Hardy-Weinberg equilibrium (HWE)

The genotype frequencies in a population remain constant or are in equilibrium from generation to generation.

Analysis based on laws of inheritance for genomic variations

phenotype	age	BMI	SNP	
case	48	23.2	TT	
case	51	25.4	AT	
control	46	21.3	AA	…
control	57	19.8	AA	
case	28	24.7	AT	
control	40	13	AT	
		⋮		

Logistic model

$$\log\frac{p_i}{1-p_i} = \beta_0 + \beta_1 \mathrm{Age}_i + \beta_2 \mathrm{BMI}_i + (\beta_3, \beta_4)\,\mathrm{SNP}_i + \cdots$$

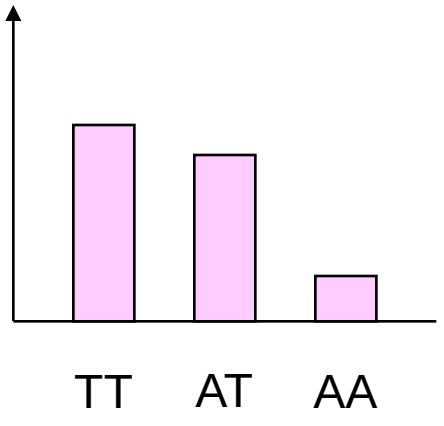
p_i : probability the ith individual is affected

$$(\beta_3 \quad \beta_4)\begin{pmatrix} 0 & 0 \\ 1 & 0 \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} 0 \\ \beta_3 \\ \beta_4 \end{pmatrix} \begin{matrix} \text{AA} \\ \text{AT} \\ \text{TT} \end{matrix}$$

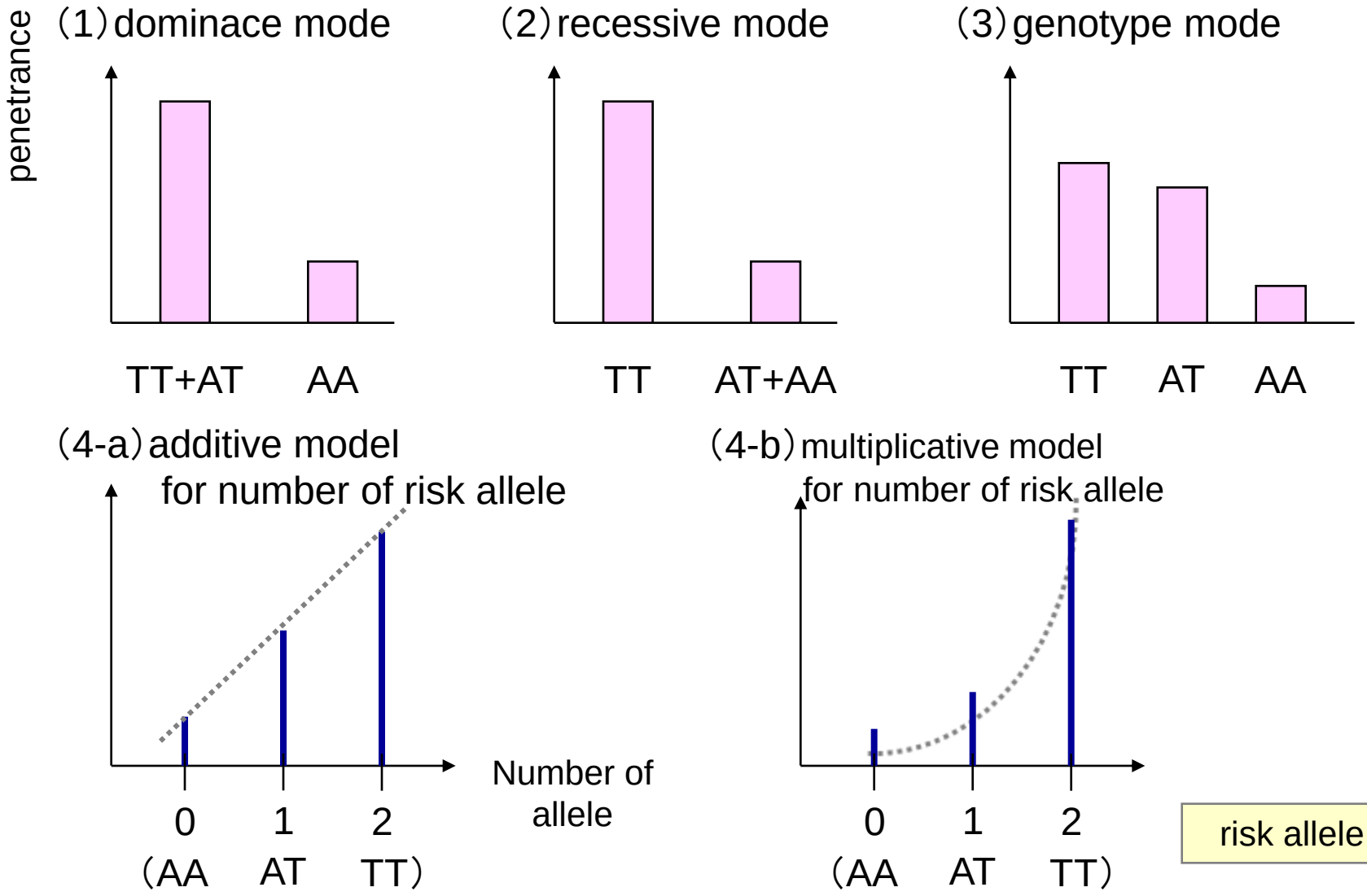
assume

Laws of dominance

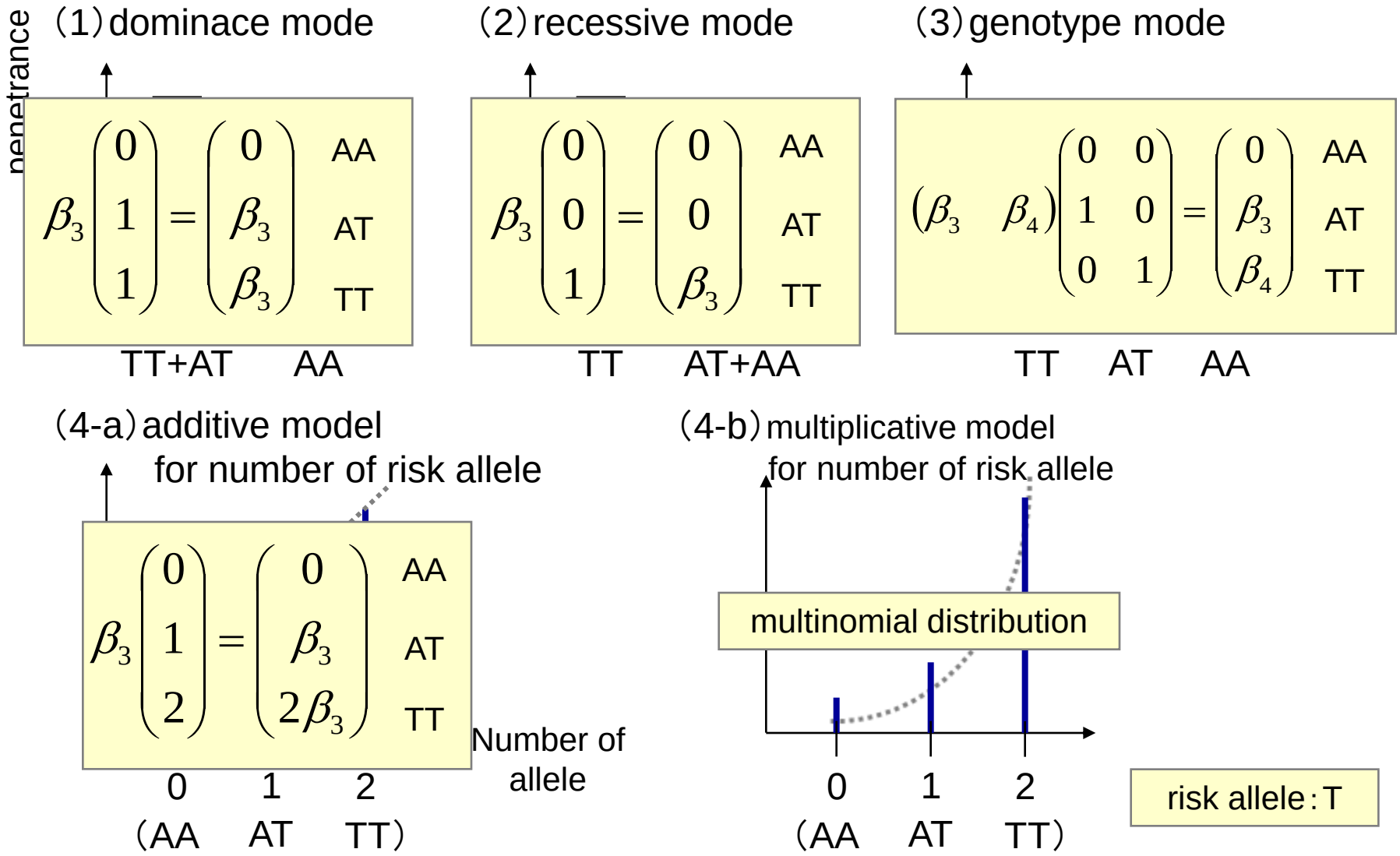
genotype mode



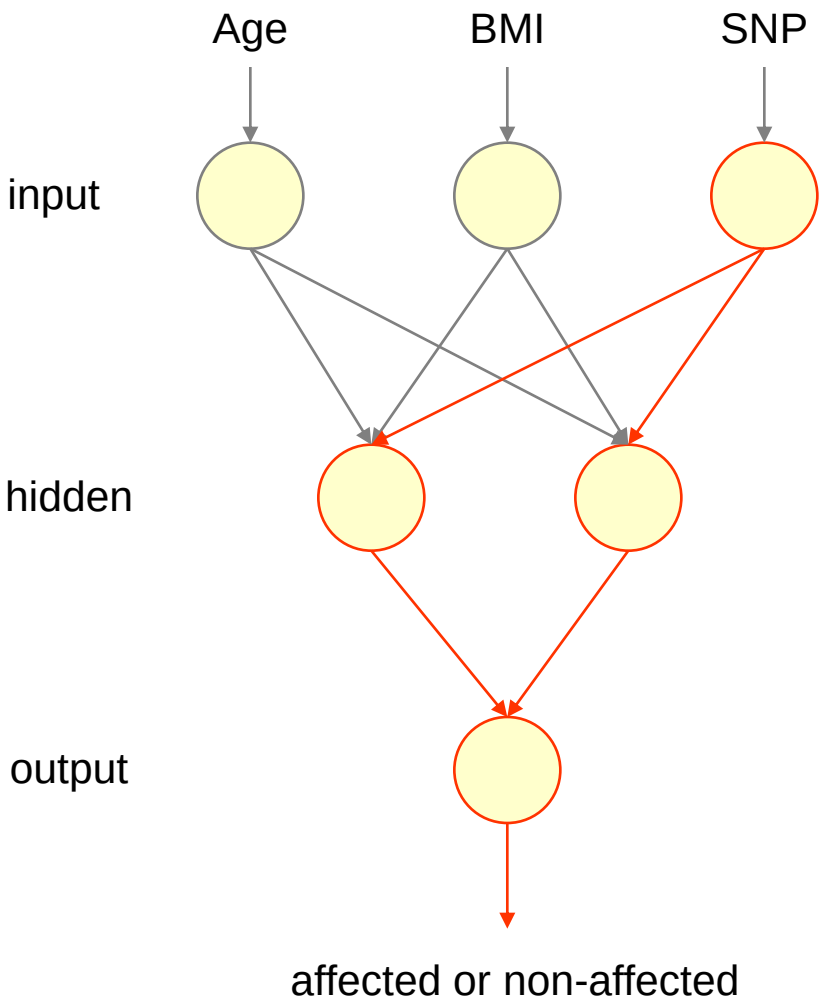
Laws of inheritance (Mendelian inheritance); Laws of dominance



Laws of inheritance (Mendelian inheritance); Laws of dominance



Analysis based on laws of inheritance for genomic variations

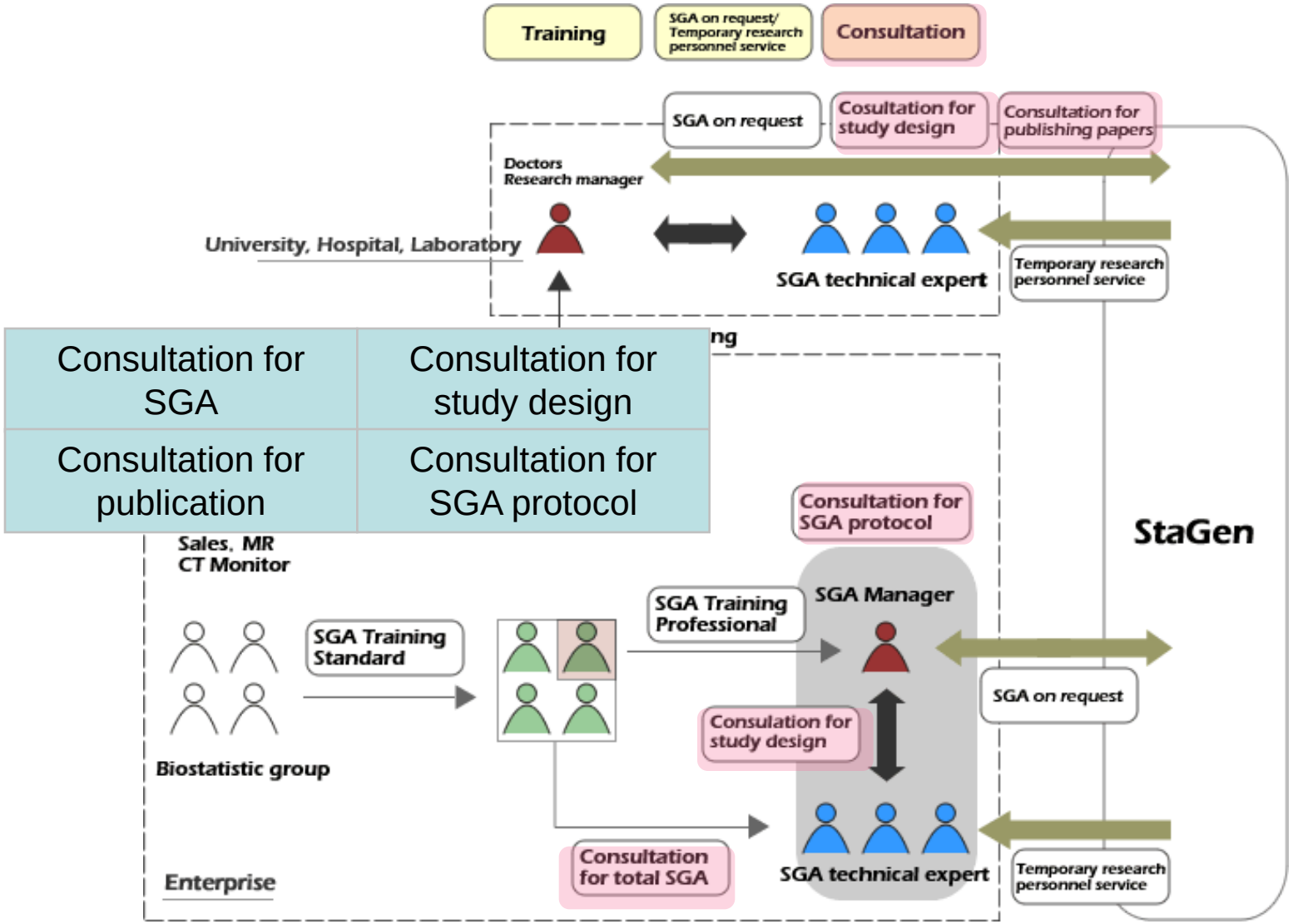


Is this network model taking the laws of inheritance into account for SNP?

The doctors and genome researchers are confusing by results obtained from complicated model.

In pharmacogenomics, FDA or MHLW (Ministry of Health, Labour and Welfare, Japan) may not confirm the results obtained from complicated model leaving genetics out of consideration.

Consultation services



Consultation for study design

Study designs for candidate gene-based and genome-wide association studies

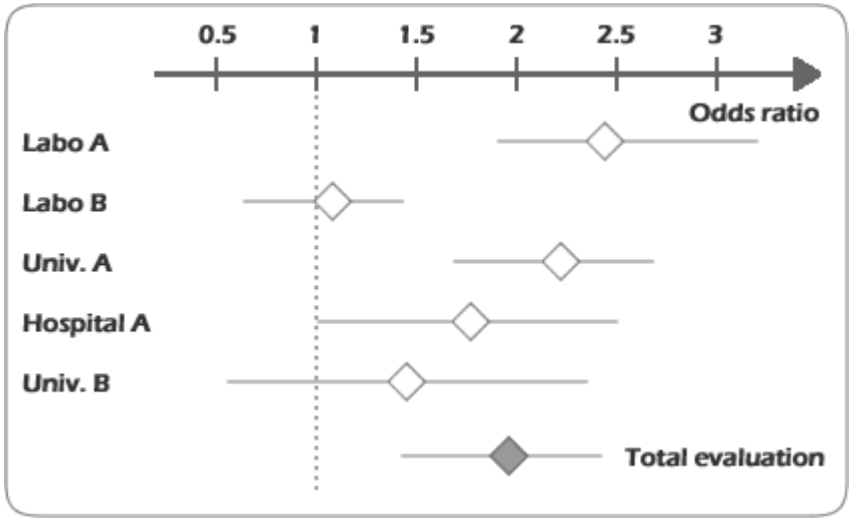
Study designs for candidate gene-based association studies

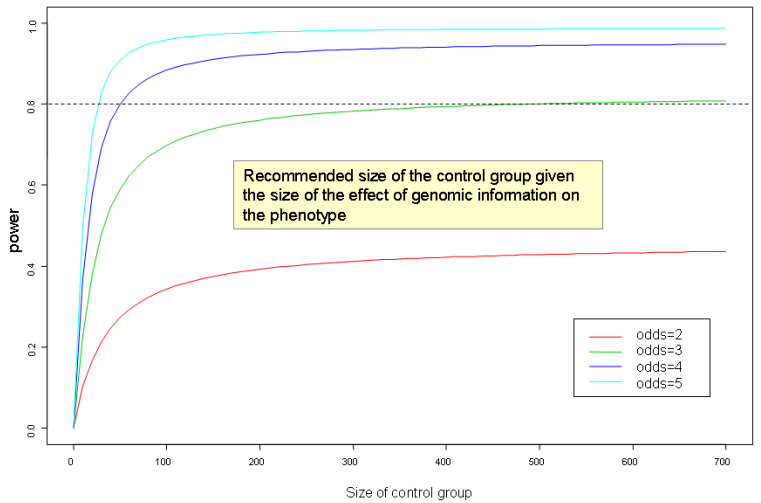
Previous researches are evaluated and

studies are designed based on the evaluation

- Evaluation using FPRP or power
- Integrated evaluation of previous researches using meta analysis
- Simulations for sample size and power under various conditions

There are some reports that a mutation in a gene is associated with adverse events of a medicine. Although we plan to conduct a case-control study based on the evaluation of the previous researches, it is difficult to collect more than 30 people as the case group. How many control people are needed for the significant result?





Other business

Training Staff

- Kobe University・Translational Research Informatics Center (Dr Kamatani, Kamitsuji)
  - Training Unit; Clinical Genome Informatics (CGI)
- Tokyo Medical and Dental University (Kamitsuji)
  - Training Program; Bio-Omics informatics

Media

- Nikkei Business Publications, Inc. monthly PDF magazine "BTJ journal" (in Japanese)
  - 「Welcome to statistical genetics!」 serialized from March 2007

Our requests to educational institution and research institution

Data science is very important in the field of pharmacogenomics

- A need to promote the education and research of Data science
  - We hope to promote sufficient education of Data science in the faculty of medical sciences and pharmaceutical sciences.
- Researchers of statistics join to Genetics field
  - Statisticians should join early stage startup of genome research.
  - If you perform the data analysis of genome data, we hope you perform the analysis without leaving genetics out of consideration.

StaGen Co., Ltd. Statistical Genetics Analysis Division

4-31-10, Kuramae Orashion Building 9F, Kuramae, Taitou-ku,
Tokyo, 111-0051

tel: +81-3-5835-2137 / 2138, fax: +81-3-5835-2139

URL: http://www.stagen.co.jp/, e-mail: info@stagen.co.jp