

第1部 これまでの10年
融合研究の苦労と成果

「遺伝学と統計学の再融合」

近代統計学と遺伝学の再融合

Statistics

Genetics

1936 Fisher R Aの
メンデル論文批判

Fisherの論文の
約半数は遺伝学
(集団遺伝学) 関係

1900 メンデル遺伝則の再発見

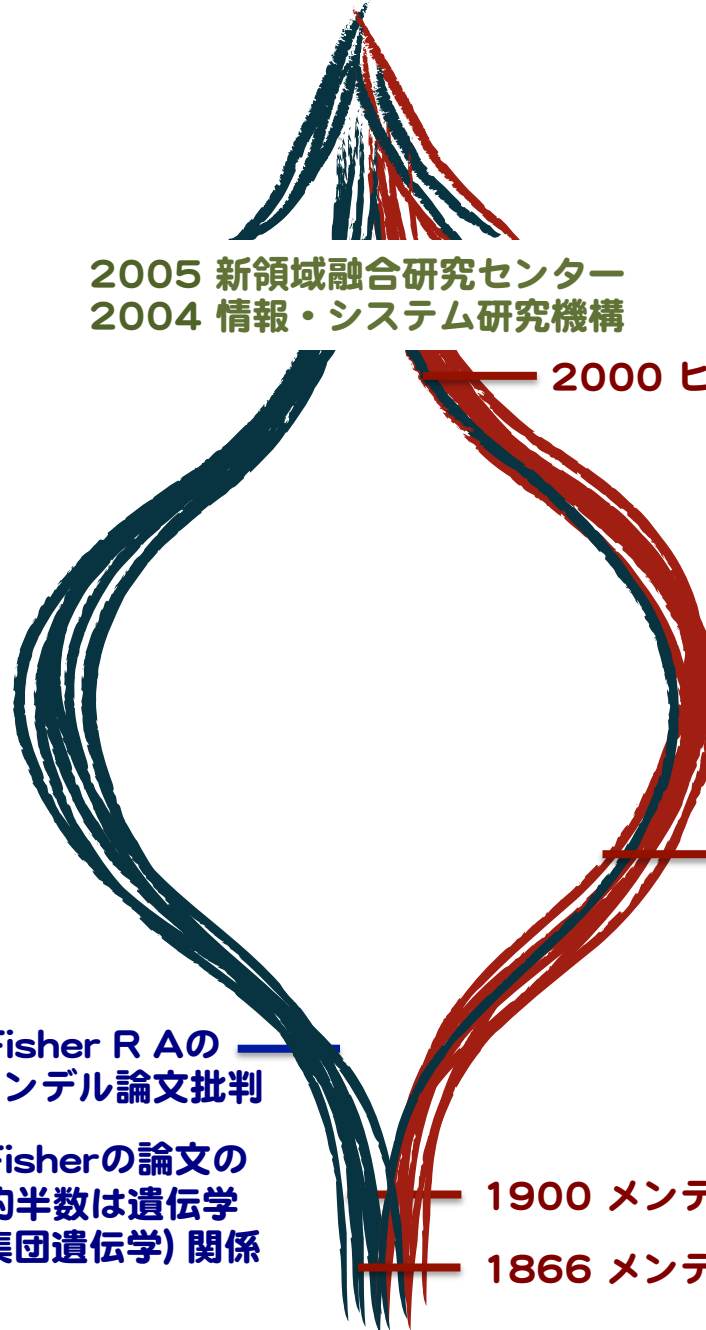
1866 メンデル遺伝則の論文発表

1953 DNA二重鎖構造
の発見

分子遺伝学
の隆盛

2000 ヒトゲノム解読

2005 新領域融合研究センター
2004 情報・システム研究機構



近代統計学と遺伝学

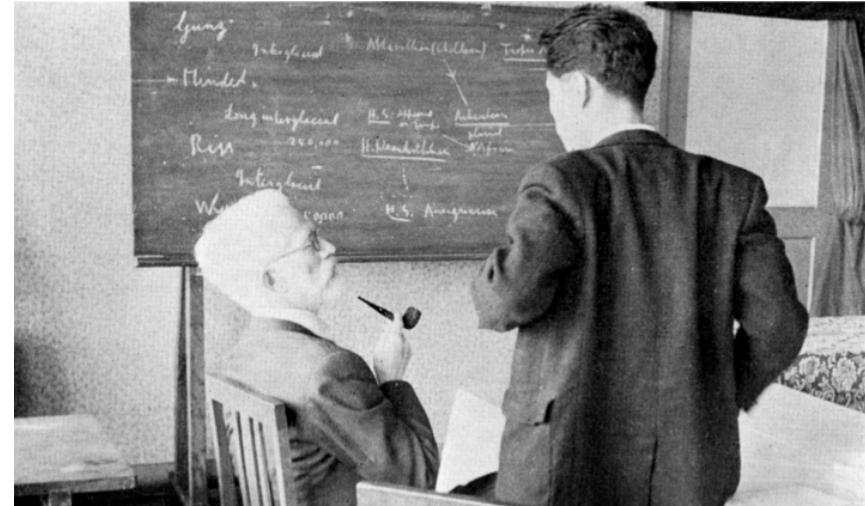
1936年 フィッシャーのメンデル論文 への批判

Fisher, R. A.

“Has Mendel’s work been rediscovered?”

Annals of Science, 1, 115–137.

「カイ二乗検定で、期待値と観察値が
あまりに良く一致し過ぎている」



1960年 Fisher博士の遺伝研来訪。右は木村資生博士。
東京での国際統計協会総会に出席のため来日。
このとき、九大(北川敏男博士)を訪問。

1900年 「メンデルの遺伝則」の再発見



Hugo de Vries

*Comptes Rendus
de l'Académie des
Sciences*

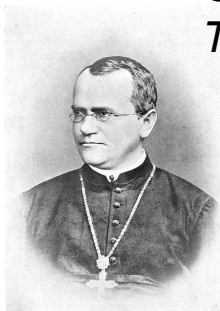


Carl Erich Correns

*Berichte der deutschen
botanischen Gesellschaft*
18: 158–168

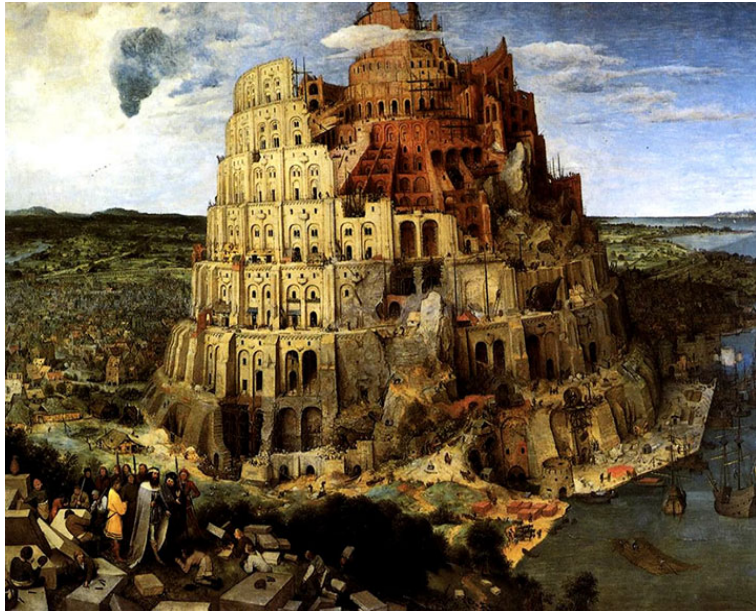
1866年 “メンデルの遺伝則”論文発表

Gregory Mendel: “Experiments on Plant Hybridization.”
The Proceeding of the Natural History Socceity of Brunn.



Seed		Flower	Pod		Stem	
Form	Cotyledons	Color	Form	Color	Place	Size
						
Grey & Round	Yellow	White	Full	Yellow	Axial pods, Flowers along	Long (6-7ft)
						
White & Wrinkled	Green	Violet	Constricted	Green	Terminal pods, Flowers top	Short (<1ft)
1	2	3	4	5	6	7

言語の分化



バベルの塔（ピーター・ブリューゲル）

旧約聖書創世記11章

「神は降臨してこの塔を見、「人間は言葉が同じなため、このようなことを始めた。人々の言語を乱し、通じない違う言葉を話させるようにしよう」と言った。このため、人間たちは混乱し、塔の建設をやめ、世界各地へ散らばっていった。」

統計数理の世界

順問題:

原因から結果を推定

(数理モデルの解析・数理モデルからの予測)



逆問題:

結果から原因を推定

(データから数理モデルを推定: “データ中心科学”)

遺伝学の世界

順遺伝学(フォワードジェネティクス):

結果(表現型)から原因(遺伝子)を同定



逆遺伝学(リバーズジェネティクス):

原因(遺伝子)の破壊から結果(表現型)を評価

統計学と遺伝学の融合研究

平成17-21年

プロジェクト名:「統計・情報技術を駆使したゲノム多型と表現型多様性の連関解析システムの開発」

PD:城石俊彦

サブテーマ1:「表現型数値計測システムの開発と実験データの収集」 (代表:城石俊彦)

サブテーマ2:「生物多様性データの統計的モデリングと解析システムの開発」 (代表:栗木 哲)

【国立遺伝学研究所】 倉田のり 伊藤幸博 春島嘉章 久保貴彦 永久貢 堀内陽子 田村勝 高田豊行 岡(木曾)彩子
細谷正樹 城石俊彦 小出 剛 梅森十三 杉本大樹

【統計数理研究所】 栗木哲 江口真透 奥田将己 川崎能典 坂口隆之 田中英希 種村正美 田村義保 土谷 隆
藤澤洋徳 Dou Xiaoling Nurul Haque Mollah 池田思朗 坂口隆之

平成22-27年

プロジェクト名:「超大容量ゲノム・多元軸表現型データの統計情報解析」 PD:倉田のり

サブテーマ1:「次世代シーケンサによるゲノム関連情報の大規模生産とその情報解析手法の開発」

(代表:藤山秋佐夫)

サブテーマ2:「大量ゲノム関連データと多元的な生物表現型データの統合による遺伝的相関構造描出の
ための統計手法の開発と最適化」 (代表:栗木 哲)

サブテーマ3:「大量で多元的なデータの情報・統計手法を適用したゲノム機能と遺伝的ネットワーク抽出」
(代表:倉田のり)

【国立遺伝学研究所】 倉田のり 春島嘉章 久保貴彦 永久貢 堀内陽子 田村勝 高田豊行 岡(木曾)彩子 城石俊彦
小出 剛 高橋阿貴 後藤達彦 藤山秋佐夫 豊田敦 野口英樹

【統計数理研究所】 栗木 哲 藤澤洋徳 Dou Xiaoling 間野修平 加藤省吾 小山慎介

【国立極地研究所】 近藤伸二

【国立情報学研究所】 宇野毅明 井ノ上克巳

【政策研究大学院大学】 土谷隆 【新潟大学】 原尚幸 阿部貴志

【山形大学】 坂口隆之 【大阪府立大学】 川野秀一 【九州大学】 二宮嘉行 【大阪大学】 片山翔太

【高知大学】 清澤秀孔

業績リスト

融合研究の存在無しでは成立しなかったもの。著者名の下線は融合研究センター所属が記載されているもの、茶色の著者名は統数研、緑字は遺伝研、青色は情報研、銀色は機構外からの著者

2007年

1. Oka A, Aoto T, Totsuka Y, Takahashi R, Ueda M, Mita A, Sakurai-Yamatani N, Yamamoto H, Kuriki S, Takagi N, Moriwaki K, Shiroishi T*. Disruption of genetic interaction between two autosomal regions and the X chromosome causes reproductive isolation between mouse strains derived from different subspecies. Genetics. 175: 185-197 (2007).

2008年

2. Takada T, Mita A, Maeno A, Sakai T, Shitara H, Kikkawa Y, Moriwaki K, Yonekawa H, Shiroishi T*. Mouse inter subspecific consomic strains for genetic dissection of quantitative complex traits. Genome Res. 18: 500-5008 (2008).

2009年

3. Fujisawa H*, Horiuchi Y, Harushima Y, Takada T, Eguchi S, Mochizuki T, Sakaguchi T, Shiroishi T and Kurata N. SNEP: Simultaneous detection of nucleotide and expression polymorphisms using Affymetrix GeneChip. BMC Bioinformatics. (1):131. (2009).

2010年

4. Mizuta Y, Harushima Y, Kurata N*. Rice pollen hybrid incompatibility caused by reciprocal gene loss of duplicated genes. Proc. Natl. Acad. Sci. 47: 20417-20422. (2010).
5. Fujita M, Horiuchi Y, Ueda Y, Mizuta Y, Kubo T, Yano K, Yamaki S, Tsuda K, Niihama M, Kato H, Kikuchi S, Hamada K, Mochizuki T, Tsutsumi N, Kurata N*. Rice expression atlas in reproductive development. Plant Cell Physiology. 51: 2060-2081. (2010).

2011年

6. Sugimoto H., Okabe S., Kato M., Koshida N., Shiroishi T., Mogi K., Kikusui T., Koide T*. A role for strain differences in waveforms of ultrasonic vocalizations during male–female interaction. PLoS ONE (2011).

2012年

7. Huang X, Kurata N, Wei X, Wang Z-X, Wang A, Zhao Q, Zhao Y, Liu K, Lu H, Li W, Guo Y, Lu Y, Zhou C, Fan D, Weng Q, Zhu C, Huang T, Zhang L, Wang Y, Feng L, Furuumi H, Kubo T, Miyabayashi T, Yuan X, Xu Q, Dong G, Zhan Q, Li C, Fujiyama A, Toyoda A, Lu T, Feng Q, Qian Q, Li J, Han B*. A map of rice genome variation reveals the origin of cultivated rice. Nature. 490: 497-501. (2012).

過去2年の業績(*は責任著者)

2013年

8. [Umemori J.](#), [Mori A.](#), [Ichiyanagi K.](#), [Uno T.](#), [Koide T*](#). Identification of both copy number variation-type and constant-type core elements in a large segmental duplication region of the mouse genome. BMC Genomics (2013).
9. [Goto T.](#), [Tanave A.](#), [Moriwaki K.](#), [Shiroishi T.](#), [Koide T*](#). Selection for reluctance to avoid humans during the domestication of mice. Genes, Brain, and Behavior (2013).
10. [Takada T.](#), [Ebata T.](#), [Noguchi H.](#), [Keane TM.](#), [Adams DJ.](#), [Narita T.](#), [Shin-I T.](#), [Fujisawa H.](#), [Toyoda A.](#), [Abe K.](#), [Obata Y.](#), [Sakaki Y.](#), [Moriwaki K.](#), [Fujiyama A.](#), [Kohara Y.](#), [Shiroishi T*](#). The ancestor of extant Japanese fancy mice contributed to the mosaic genomes of classical inbred strains. Genome Res. 23:1329-1338 (2013).

2014年

11. [Kuriki S*](#), [Harushima Y.](#), [Fujisawa H.](#), [Kurata N.](#) Approximate tail probabilities of the maximum of a chi-square field on multi-dimensional lattice points and their applications to detection of loci interactions. Annals of the Institute of Statistical Mathematics. 66: 725-757. (2014).
12. [Dou X*](#), [Kuriki S.](#), [Maeno A.](#), [Takada T.](#), [Shiroishi T.](#) Influence analysis in quantitative trait loci detection. Biom J. 56: 697-719 (2014).
13. [Oka A.](#), [Takada T.](#), [Fujisawa H.](#) and [Shiroishi T*](#). Evolutionarily diverged regulation of X-chromosomal genes as a primal event in mouse reproductive isolation. PLoS Genet. 10, e1004301, (2014).
14. [Arakawa T.](#), [Tanave A.](#), [Ikeuchi S.](#), [Takahashi A.](#), [Kakahara S.](#), [Kimura S.](#), [Sugimoto H.](#), [Asada N.](#), [Shiroishi T.](#), [Tomihara K.](#), [Tsuchiya T.](#), [Koide T*](#). A male-specific QTL for social interaction behavior in mice mapped with automated pattern detection by a hidden Markov model incorporated into newly developed freeware. Journal of Neuroscience Methods (2014).
15. [Kato S.](#), [Ishii A.](#), [Nishi A.](#), [Kuriki S.](#), [Koide T*](#). Segregation of a QTL cluster for home-cage activity using a new mapping method based on regression analysis of congenic mouse strains. Heredity (2014).

本日発表する主な成果

1. ゲノム研究における多重検定

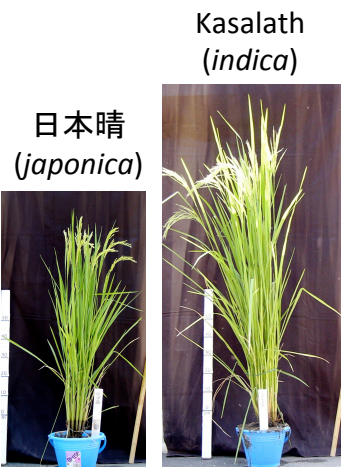
多くの変数を観測するほど、ランダム誤差による見せかけの統計的「発見」がなされる。その多くは、再現実験で確認されず、偽陽性 (false positive) である。ゲノム解析では多数の検定を繰り返すことが多く、偽陽性と真の陽性を区別する必要がある。融合研究では、この多重性を克服する方法を検討した。

2. 対立遺伝子発現量の比較解析研究基盤の構築

体質や疾患の原因の多くは、遺伝子発現を制御するシス配列上の多型(変異)であることがわかってきた。制御配列多型を効率良く探索するための研究基盤として、遺伝子発現量を比較計測する解析プラットフォームを構築した。

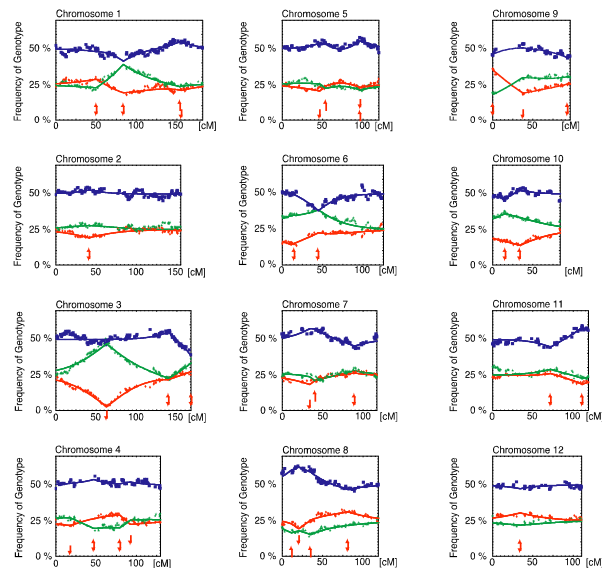
1-1. 生殖隔離障壁に関わるエピスタシスの統計学的検出と多重性調整

栗木哲・藤澤洋徳(統数研) 春島嘉章・倉田のり(遺伝研)

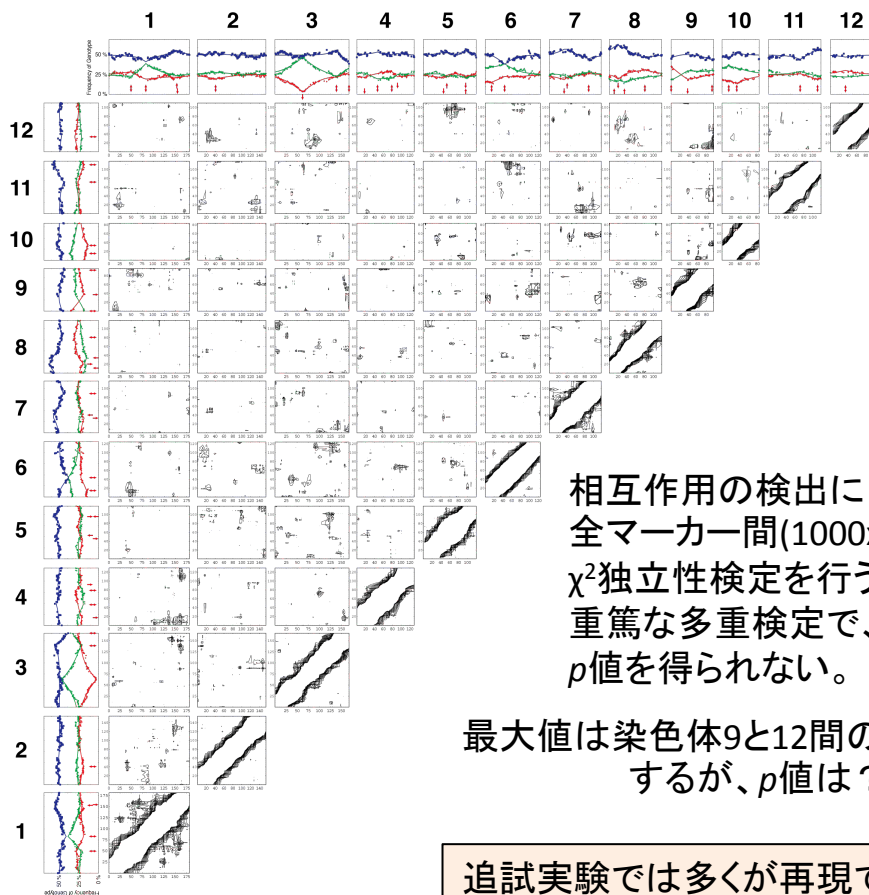


交配可能な亜種間でも種を隔てる遺伝的機構により、
雑種後代で分離比がメンデル則からずれる所がある。

生殖的隔離は相互作用によって起きる



F₂集団の連鎖地図上のマーカー分離比



相互作用の検出に
全マーカー間(1000x1000)の
 χ^2 独立性検定を行うが、これは
重篤な多重検定で、 χ^2 値から
 p 値を得られない。

最大値は染色体9と12間の33.6を観測
するが、 p 値は？

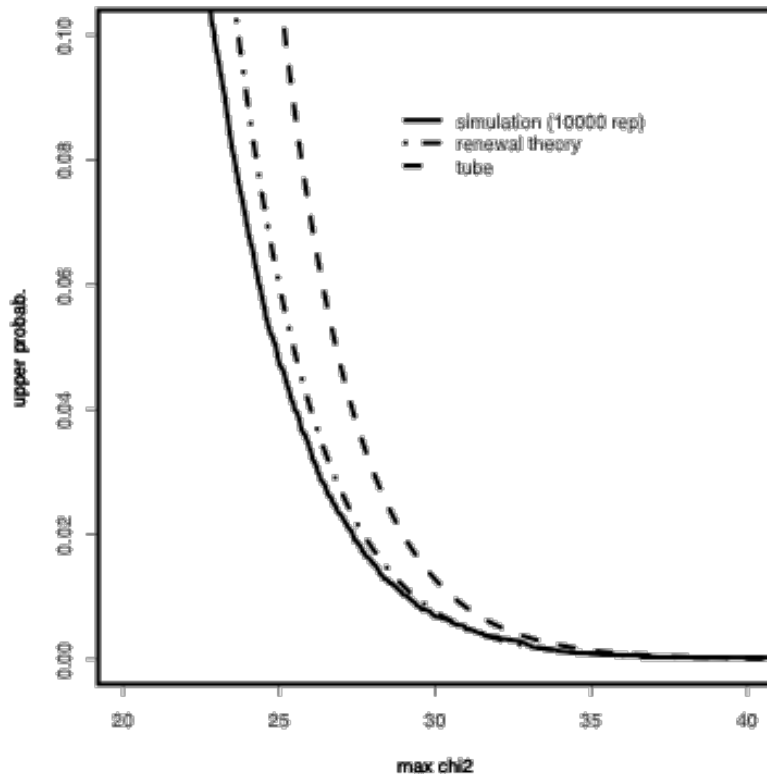
追試実験では多くが再現できない

χ^2 独立性検定の多重性調整

遺伝マーカー間には連鎖しており独立ではない
連鎖の確率構造を考慮し、3種類の方法で調整 p 値を計算

1. Monte Carlo, 2. Nonlinear renewal theory, 3. Tube method

3種類の方法で調整した p 値は一致



高密度連鎖地図における相互作用検定
時の、多重検定の調整 p 値を決定した

Kuriki S, Harushima Y, Fujisawa H, Kurata N.

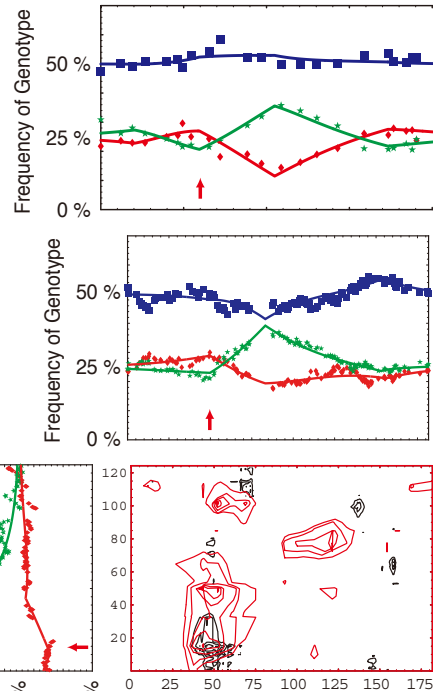
Ann Inst Stat Math 66:725–755. (2014)

F_2 集団で連鎖しないマーカー間で観測された
全ての χ^2 ピークは有意(<0.05)では無かった。

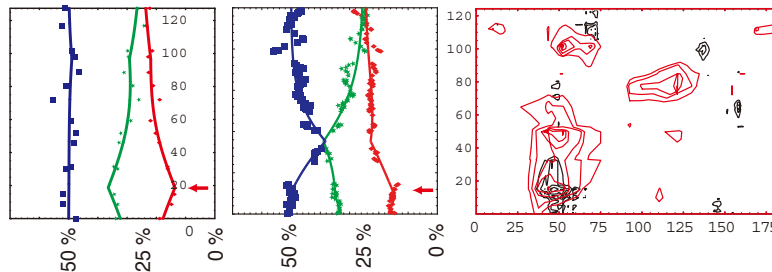
統計学的に有意でないと判定されたピークにも 実験で再現された真の相互作用も有る

Chr 1

染色体1と6の相互作用は、統計学的には有意な値を示さないが、別の集団でも独立性検定で高い χ^2 値を示し(赤い等高線)、二つの集団で同様にマーカーの分離比が歪む(再現性があった)。



Chr6



この相互作用は、重複遺伝子で *japonica*, *indica* で別々に破壊され花粉管発芽に關与する二つの遺伝子によるものであった。

Mizuta Y, Harushima Y, Kurata N. (2010)
Proc Natl Acad Sci USA **107**: 20417–20422.



真の相互作用の判定方法の開発: (論文 in preparation)

遺伝的相互作用モデルを建て、独立性検定値、マーカー分離比、 F_2 生存率を考慮し、複合判定することにより真の相互作用の検出が可能である。

1-2. 疑似相関を利用した新しい多重性調整手法の開発

二宮嘉行(九州大) 栗木哲・藤澤洋徳(統数研) 高田豊行・城石俊彦(遺伝研)

マウス表現型解析データ

系統		表現形質1		表現形質36
マウス C57BL/6 (B6) 系統: 対照系統	サンプル1	1.47	...	2.84
	サンプル2	2.78	...	2.96
	⋮	⋮		⋮
染色体 置換系統1	サンプル1	2.87	...	4.34
	サンプル2	2.56	...	3.81
	⋮	⋮		⋮
⋮	⋮	⋮	⋮	⋮
染色体 置換系統29	サンプル1	1.98	...	2.57
	サンプル2	2.32	...	3.15
	⋮	⋮		⋮

対象とする多重検定問題

C57BL/6J (B6) 系統と染色体置換系統の表現型の差を検定する。表現形質(特徴量)は36種類、系統数は29(雌雄で58)あり、合計で2,000以上の検定数となる。

各多重検定におけるp値:

第j番目の表現形質を比較するt統計量を T_j 、その実現値を t_j としたとき、相関を考慮した検定の多重性調整済p値は次で与えられる。

① $\stackrel{\text{def}}{=} P$ 完全帰無仮説 ($\max_j |T_j| > |t_j|$)

この確率は表現形質間の未知の相関に依存するため、厳密に評価できない。

② 相関を想定しないBonferroni法の上界を使う (Takada et al. Genome Res. 2008)

相関を考慮した多重性調整

表現形質間の相関が大きいとき、②は大きな値を与えずに検出力を下げる

① $\leq$ ② $\stackrel{\text{def}}{=} \sum_j P$ 完全帰無仮説 ($|T_j| > |t_j|$)

Bonferroni法はconservativeな方法で、本当は系統差があっても検出できない場合がある

疑似相関を利用した多重性調整手法の開発

定理 (Dudoit & Van der Laan, 2008)

未知の相関の代わりに通常の標本相関を用いて求めた①(MaxT法)は、漸近的に多重性調整された p 値となる。ステップダウン法を組み込むこともできる。

より検出力を上げるため、通常の標本相関より大きくなる傾向がある疑似相関を用いることを検討する

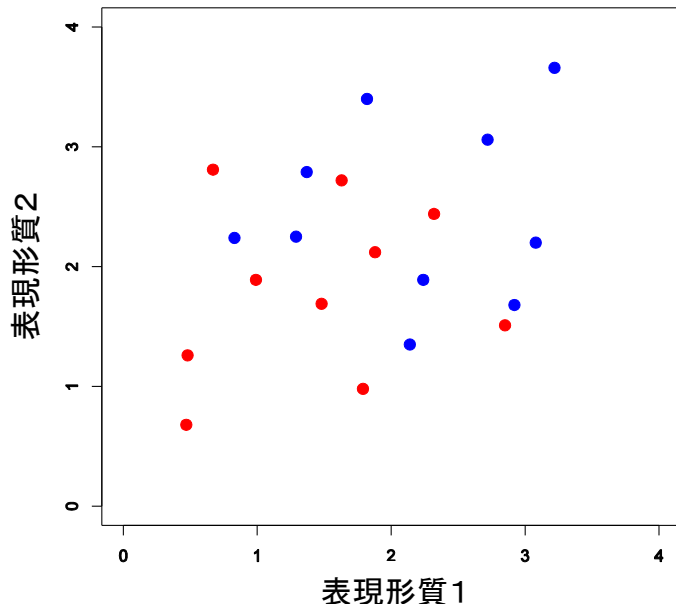


図 疑似相関を用いた多重検定:

C57BL/6系統(対照系統)と比較する染色体置換系統を合わせたデータの標本相関は“疑似相関”である。疑似相関は、帰無仮説が正しいときは良い推定値を与え、帰無仮説が正しくないときは意味の無い値となり、全体として多重検定のエラーに影響を与えない。

図で疑似相関は0.30, 通常の標本相関は0.12(青)と0.18(赤)。赤群と青群の比較で、通常の標本相関を使うと2特徴量とも受容(系統差なし), 疑似相関を使うと2特徴量とも棄却(系統差有)される。

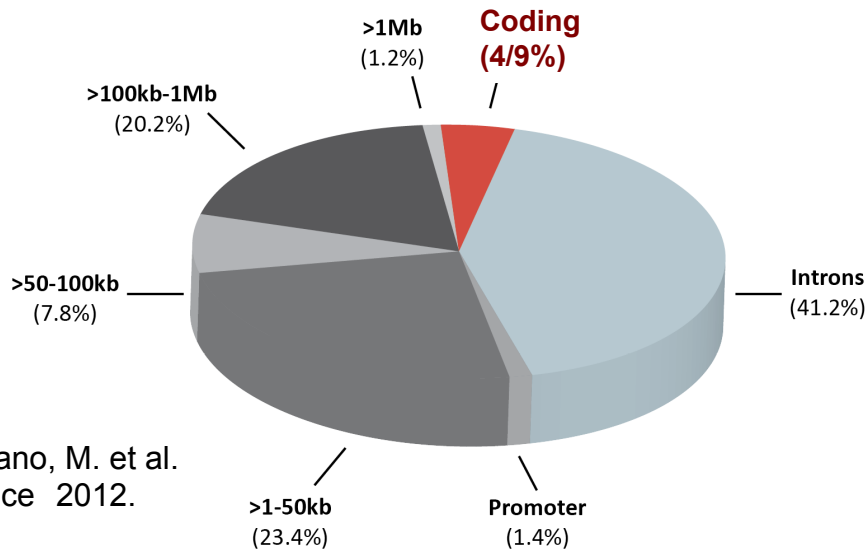
定理 (Ninomiya et al., 論文 in preparation)

ある(本問題では満たされるような)条件のもと、未知の相関の代わりに疑似相関を用いて求めた①は、漸近的に多重性調整された p 値となる

2. 対立遺伝子発現量バイアスの解析プラットフォームの構築

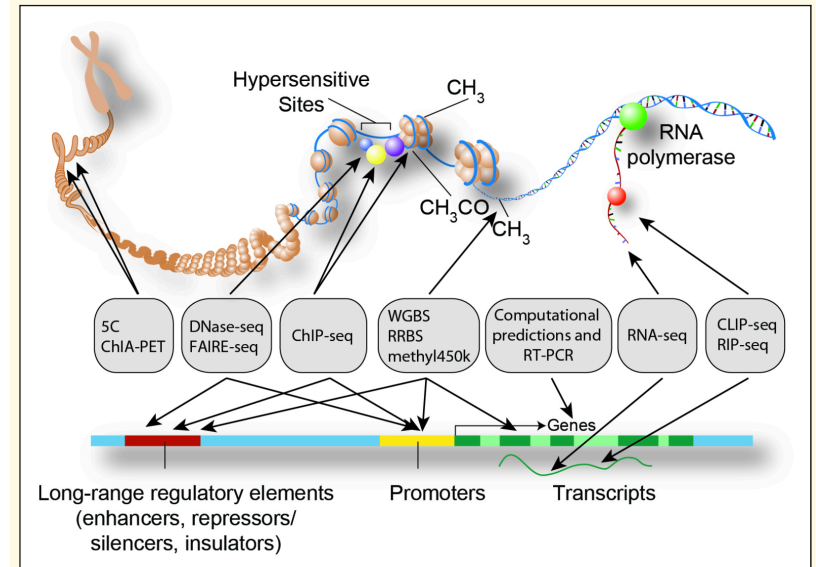
高田・岡・藤山・豊田・城石(遺伝研) 近藤(極地研) 清澤(高知大)
二宮・藤澤・栗木(統数研)

Distribution of GWAS SNPs vs. RefSeq



Maurano, M. et al.
Science 2012.

ヒト集団で体質・疾患の原因ゲノム多型の位置。原因SNPsが翻訳配列(タンパク質をコード)領域にみつかるとは、わずか4.9%。原因の多くは、遺伝子発現を制御するシス配列多型(変異)にある。



Credits: Darryl Leja (NHGRI), Ian Dunham (EBI), Michael Pazin (NHGRI)

翻訳配列はゲノムの僅か2%。残りの98%にある発現制御に働くエンハンサー上のゲノム多型を探すことが重要。発現制御に関わる染色体クロマチンの化学修飾も重要である。

個体差(表現型多様性)は、遺伝子発現量の違いで生じる！

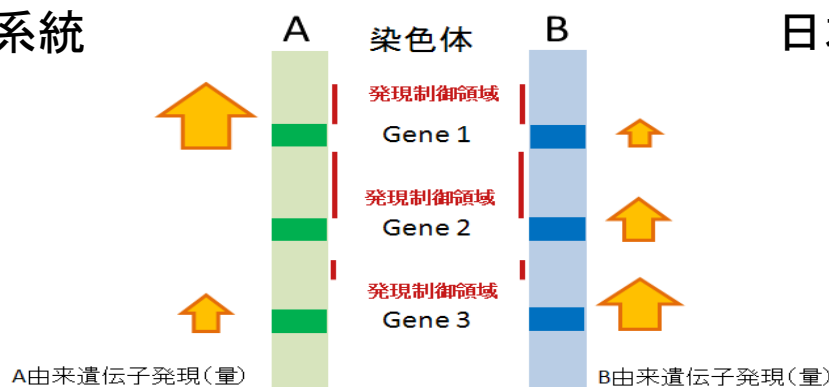
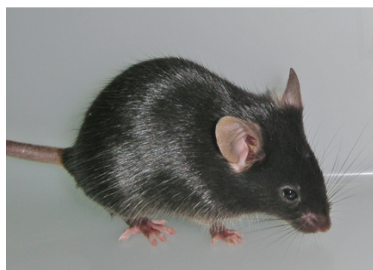
遺伝子発現量の違いを網羅的に検出する研究基盤の必要性

大型、高発がん、
糖尿病傾向、低活動
老齢難聴、etc.

大きな表現型多様性

小型、低発がん、
低糖尿病傾向、高活動
老齢でも健聴 etc.

西欧産 C57BL/6 (B6) 系統



日本産 MSM 系統(遺伝研)



一塩基多型などの
ゲノム多型

一千万以上の一塩基置換

MSM	B6
ATGCGGTCAGGATTACGATTTAGAC	ATGCGGTCAGGTTTACGATTTAGAC

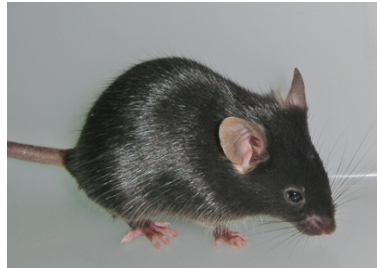
(C57BL/6 x MSM)F1交雑個体
から調整したmRNAを次世代シーケンサで大規模解析

翻訳配列上のゲノム多型 (cSNP) で2系統からの遺伝子
の転写産物 (mRNA) を区別して発現量を正確に計測する

全遺伝子の80%以上が区別される

遺伝子発現量の系統比較計測のための解析プラットフォーム構築

C57BL/6 (汎用実験用系統)



リファレンスゲノム配列

ゲノム配列情報：
UCSC mm9

ゲノム配列情報：
NIG MSM v1&2

MSM/Ms (日本産野生由来系統)



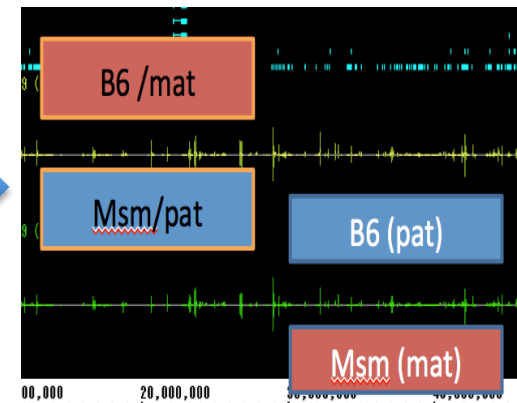
倉田PJ サブテーマ1
次世代シーケンサ
(HiSeq2000) による
RNA-sequencing
藤山秋佐夫・豊田敦 (遺伝研)



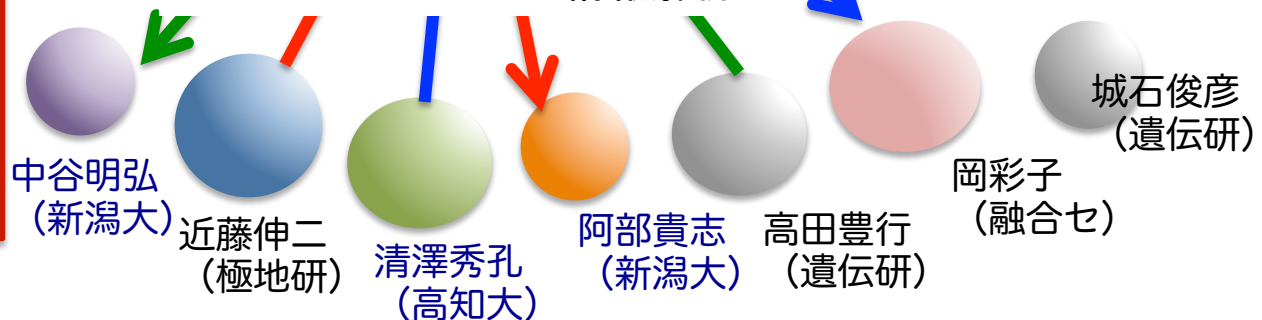
RNA調整
(♀B6 x ♂MSM)F₁
(♀MSM x ♂B6)F₁
肝臓・精巣・骨・脳
高田豊行 (遺伝研)
岡彩子 (融合セ)



遺伝子発現量比の計測



ゲノム・ビッグデータ情報解析



情報・システム研究機構の融合研究で構築した 遺伝子発現量比較解析プラットフォーム：

個体差が遺伝子発現量の違いによることが明らかになり、
発現量の個体差を正確に計測する必要が出てきた。

今後、研究基盤として、
世界のゴールデンスタンダードになる可能性がある。

結語

20世紀初頭、遺伝学と近代統計学は一体となってスタートした。分子生物学が起こると、遺伝学は遺伝子機能の詳細な解析に関心に移り、集団遺伝学を除くと独自の路線は歩むようになった。

ゲノム科学の時代が到来して、遺伝学で扱うべき変数が爆発的に増大すると、遺伝学は統計学との再融合が必要となっている。

情報・システム研究機構における融合研究は、そのような試みの貴重な一歩である。