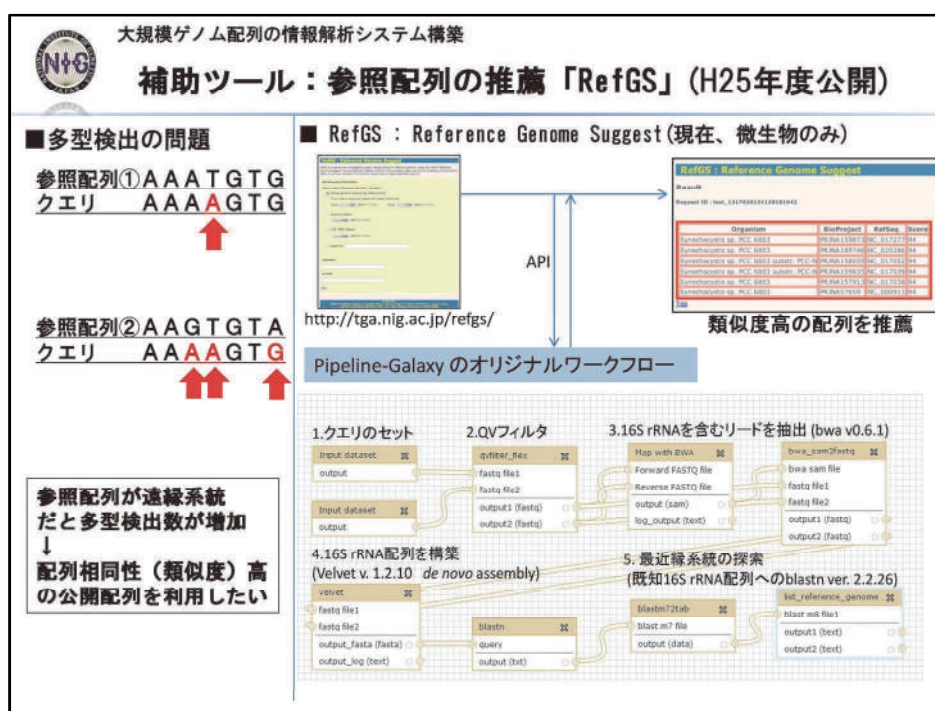
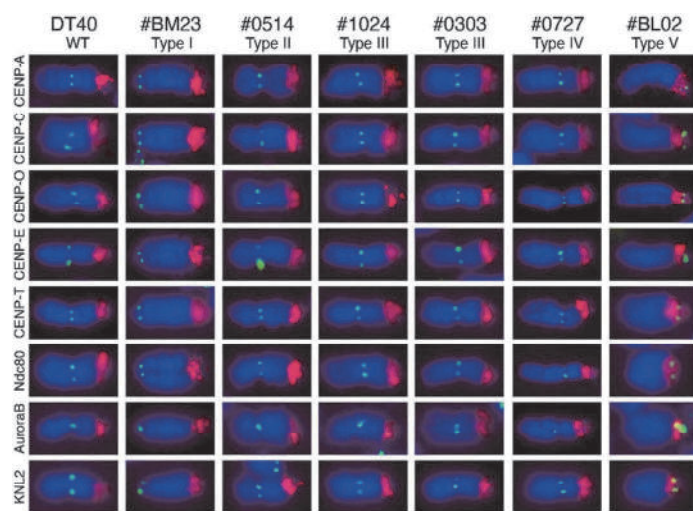


1-2：データ解析・管理システム：データ処理パイプラインの機能拡張

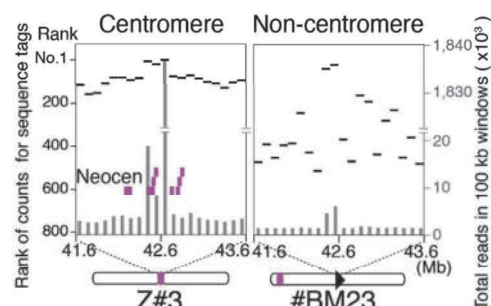
サブテーマ1 中村グループでは、データ処理をクラウド型で行うパイプライン開発を進めてきた。基本的なデータ処理パイプラインのシステム構成は昨年度までと変更はない。平成25年度は、参照配列アライメント利用時に、参照配列を探索する時間がかかることから、配列相同性（類似度）が高いゲノム配列を公開データベースから抽出する参照配列探索補助ツール「RefGS」を構築した。平成23年度に構築したDNA多型同定ワークフローでは、(1)既知の注釈情報を使って、遺伝子領域のSNPsを検出する。(2)遺伝子領域のSNPsが同義置換／非同義置換のどちらを起すかを検出する機能を持つ。SNPsの検出は、参照配列系統が解析対象系統より遠縁である程、同定数が多くなる。例えば柑橘での場合、リファレンスとなる *Citrus clementine* に対して近縁系統地中海マンダリンのSNP数は約26万、遠縁系統晩白柚のSNP数は約67万である。変異SNPの探索の場合などには、より近縁系統のゲノムを参照配列として利用できれば、候補SNP数を減らす事ができる。参照配列探索補助ツール「RefGS」の解析フローを下図に示す。RefGSサイト <http://tga.nig.ac.jp/refgs/> にアクセスして、NGS由来クエリ配列と解析条件を入力する。サブミットボタンを押すと、既存パイプラインの高次部である Galaxy workflow で RefGS 用に構築した解析がバックグラウンドで動作して、配列類似度の小さい参照配列候補をランキングして提示する仕組みである。Galaxy workflow は国立遺伝学研究所スーパーコンピュータの計算機資源を利用する。現在は微生物のみを対象に、類似度計算に16S rRNA公開データを利用している。16S配列は種識別用データベースDNA Barcodeでも利用されており、系統識別レベルでは同位で差が出ない場合がある。今後、識別精度を向上させるために、類似度計算時にITS領域への拡張を行いたい。



1-3: 高次ゲノム機能領域の解析



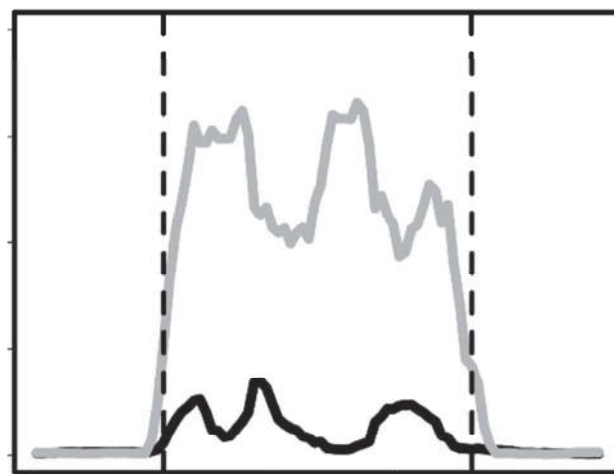
各種ネオセントロメアのセントロメアタンパク質抗体による免疫染色像



セントロメア領域の周辺には、非セントロメアの CENP-A が散在している

サブテーマ 1 のチーム内では、次世代シーケンサを活用した大規模シーケンシングを基盤とする融合的研究が進み、昨年度からの研究を継続発展させている。深川グループによる染色体高次構造に関する研究では、深川らが独自に開発したネオセントロメアを培養細胞で構築できる実験系を活用した。その実験で構築したネオセントロメアのゲノム領域を解析した。一般にセントロメアを構成するゲノム領域は、巨大な反復配列から構成されており、一般的なゲノム解析の対象から除外されているが、ネオセントロメアを構成するゲノム領域に反復配列は存在しないので、ゲノム解析が可能となった。解析には、深川グループの研究に加え、ゲノム解析を専門とする藤山のグループと、情報学専門とする榊原のグループが融合することで行なわれた。今年度の解析では、セントロメアのサイズは、どのセントロメアにおいても～40kb 程度の長さで保たれること、セントロメアタンパク質 CENP-A は、セントロメア領域意外にも存在して、本来のセントロメアが失われた際に新規セントロメア形成の種になることが明らかになった。これらの成果に関しては、Dev Cell, 24: 635-648, 2013; Dev Cell, 27: 367-168, 2013; Nucl Acids. Res. 42: 1644-1655, 2014 へ論文報告した。

サブテーマ 1 の角谷グループは、ゲノム機能調節の重要なメカニズムの一つであるメチル化とトランスポゾン転移について、シロイヌナズナやイネをモデルとした研究を進めている。DNA 低メチル化状態にすると転移するシロイヌナズナのトランスポゾンの一つ VANDAL21 のコードするタンパク質の機能について調べた結果、そのうちのの一つが、トランスポゾン DNA の低メチル化と転写脱抑制を引き起こすことがわかった。DNA 低メチル化効果の特異性を知るためにゲノムワイドで DNA メチル化を調べた。その結果、配列の似た一群のトランスポゾンで特異的にメチル化が低下し、また標的となるトランスポゾンの全長にわたって低メチル化がおこっていた (図)。大規模ゲノムシーケンシングによってこの解析が可能になった (Fu et al EMBO J)。



抗サイレンシング因子の発現によるトランスポゾンの脱メチル化(黒色)。灰色はメチル化された対照系統

サブテーマ 2

「大量ゲノム関連データと多角的な生物表現型多様性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」

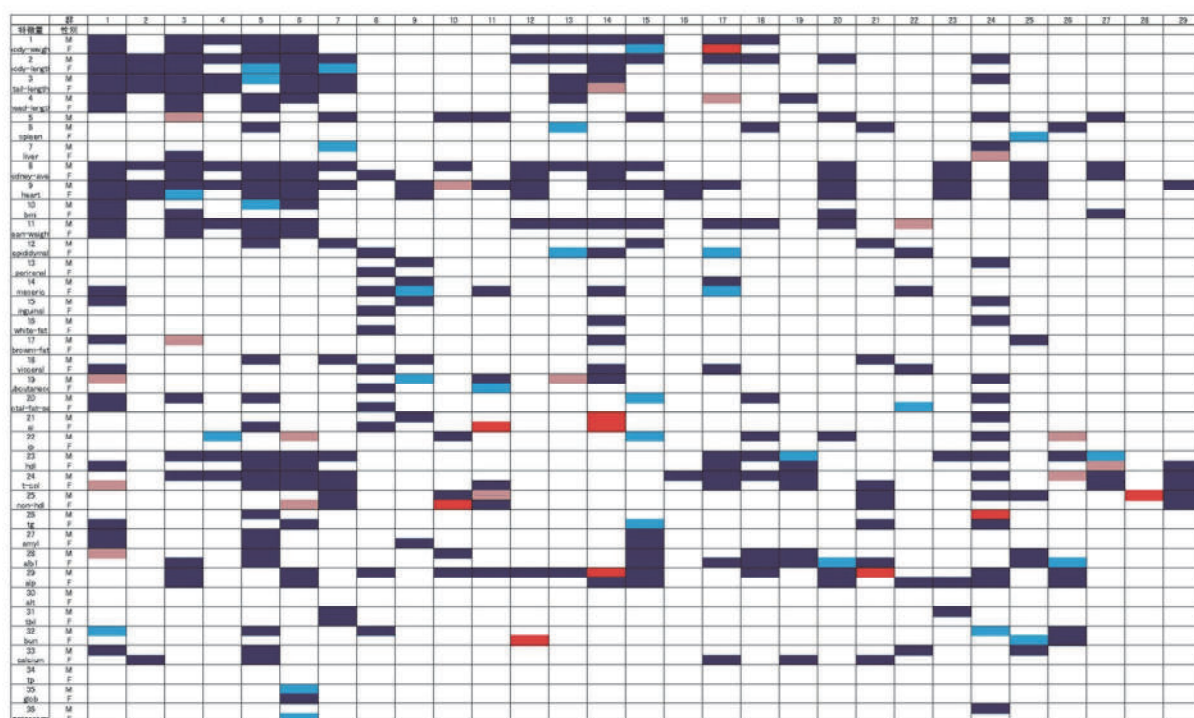
2-1: 疑似相関を利用した新しい多重性調整手法の開発 [二宮、栗木、高田、城石]

前年度、未知の相関をもつ複数の t 検定からなる多重検定問題に対し、ある種の疑似相関を未知の相関の代わりに用いる新しい多重性調整法の原型となるものを開発した。本年度の研究は、その手法を完成させ、モチベーションとなった融合研究における実際問題にそれを適用したことである。

実際問題とは、純系のマウス MSM とそのコンソミックマウスとの特徴量を比較する多重検定問題である。特徴量は、大きさに関する 11 種、脂肪に関する 10 種、血液化合物に関する 15 種の計 36 種となり、相関の存在が想定される。このデータに対し、MSM とコンソミックの特徴量（の期待値）は等しいという帰無仮説と等しくないという対立仮説とを比較する t 検定 36 個からなる多重検定を構成した。コンソミックマウスは 29 種あり、また各々のマウスにオスとメスがあるため、58 種の多重検定を行い、有意となったものに色を付けたのが下図である。

最もシンプルな多重性調整法は、相関を考慮せずにボンフェロニ法を用いることである。本データには明らかに有意となるデータが多かったため、これだけで 316 個の検定が棄却される（濃青で色付けられた箇所）。相関を考慮するものの中での自然な多重性調整法は、未知の相関に対して通常の標本相関を代入する方法である（例えば Dudoit & van der Laan, 2008, § 2.6 と § 5.2）。これをシングルステップで用いるとさらに 29 個、ステップダウンで用いるとさらに 17 個の検定が棄却された（それぞれ淡青、淡赤で色付けられた箇所）。そして新しい方法の適用結果であるが、このデータには有意か否かが微妙なものが少なかったため、期待されるほどの効果は出なかったものの、それでも 10 個の検定を新たに棄却した（濃赤で色付けられた箇所）。

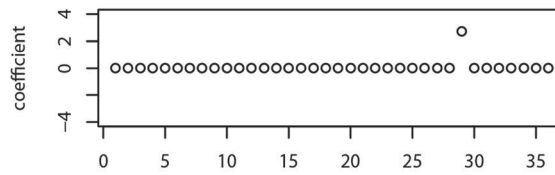
元々の解析ではボンフェロニ法を用いており (Takada et al., 2008)、それに比べると 56 個の仮説をさらに棄却したことになるため、データに関する新たな解釈がなされるはずである。また、この解析結果により、相関を考慮することには無視できない影響があることが確認されたため、今後様々な多重検定問題に対して同様の解析を進めることを検討中である。



2-2：高次元説明変数の次元圧縮を予測目的に合わせて行う方法 [川野、藤澤、高田、城石、片山]

サブテーマ 3（城石プロジェクト）のコンソミック系統および両親系統、計 33 種類について、36 形質の測定データがあった。そこで、コンソミック系統を形質によって特徴づけられないか、というテーマが考えられた。これは統計学では、形質を説明変数として系統を目的変数とした判別分析に対応する。

遺伝研側では、まずはデータを主成分分析で解析してみた。しかし、特徴は特に何も得られていなかった。主成分分析はある種の分散を大きくしているだけであり、判別に役に立つかどうかは考慮されていない。そこで、判別モデルとして標準的なロジスティック回帰モデルを使って、判別に役に立つ形質を選択することを考えた。今回は形質数が大きいので、そういう場合に有効であるスパース推定の考え方も使った。結果として、系統に特徴的な形質がいくつか現れた。



その後に、遺伝研の考えであった主成分分析を取り入れるため、まずは主成分分析を行い、固有値が大きい幾つかの主成分スコアを新しい説明変数と見直して回帰を行う手法を考えていた。これはよく利用されている 2 段階法である。しかし主成分分析で新しい説明変数をいくつか選んだとき、予測の意味で良いものを選んでいとは限らない。これは、目的が違う手法を組み合わせる 2 段階で行っているからである。

そこで我々は、主成分分析と回帰モデルを同時に 1 段階で行うことを考えた。予測の意味で有望な主成分スコアを選ぶためである。しかし普通はそれだけではうまく行かない。なぜなら明らかにパラメータ過多になり、推定不能になるからである。本研究では、その問題の克服を目指している。

まずは判別分析よりも扱いやすい、標準的な線形回帰モデルで考えている。最初に通常の予測二乗誤差を用意する：

$$L_{\text{reg}}(B, \gamma) = \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n \left(y_i - \sum_{j=1}^q \gamma_j \sum_{k=1}^p \beta_{jk} x_{ik} \right)^2$$

普通はこれの最小化でパラメータを推定するのだが、既述したように、パラメータ過多の問題があって、そのままではパラメータ推定がうまく行かない。まずは、主成分分析の考えを取り入れるために、予測二乗誤差に以下の主成分誤差を罰則として導入することを考えた：

$$L_{\text{PC}}(B, A) = \sum_{i=1}^n \|x_i - AB^T x_i\|^2$$

これだけではまだ、パラメータ推定値に回転の自由性が残り、パラメータがうまく推定されない。そこで、この自由性を取り除くためとパラメータ推定値にスパース性を導入するために、スパース罰則も取り込むことにした。より一般的な手法にするためにリッジ罰則も導入することにした。結果的に構築された目的関数は以下である：

$$L(B, \gamma, A) = L_{\text{reg}}(B, \gamma) + wL_{\text{PC}}(B, A) + \lambda_1 \sum_{j=1}^q \|\beta_j\|_1 + \lambda_2 \sum_{j=1}^q \|\beta_j\|^2 + \lambda_\gamma \|\gamma\|_1$$

この目的関数の最小化によってパラメータ推定を行うこととした。現在は、簡単な数値実験でその方法の良さを確認したところである。

2-3: 回帰分析を用いたマウス活動量データの遺伝解析 [加藤、栗木、小出]

今年度は、昨年度までに得られた結果をまとめた論文を執筆・投稿した。論文は一度の改訂の後、学術誌「Heredity」へ採択された。今年度に執筆および改訂した論文の概要は以下の通りである。

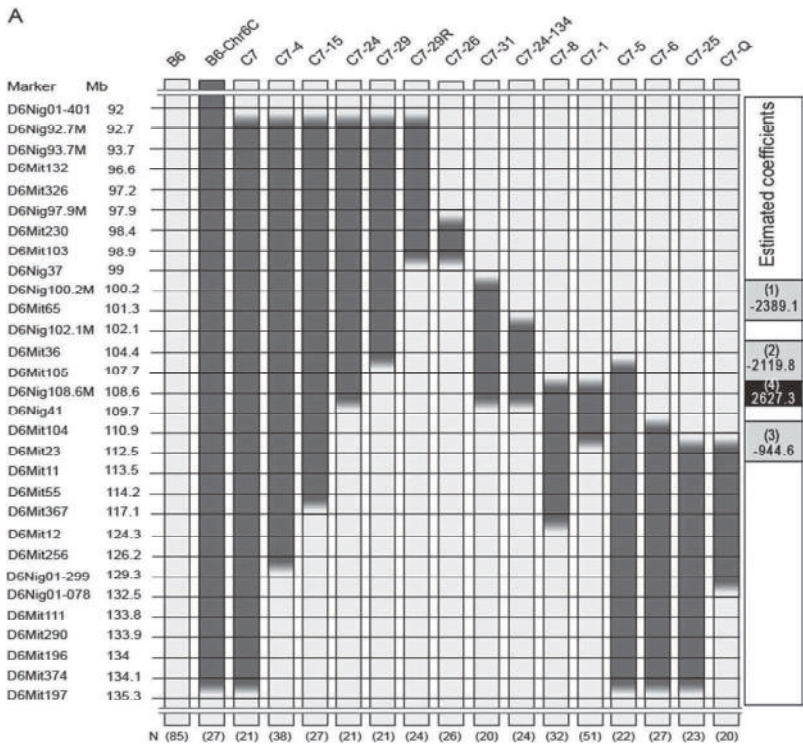
Nishi *et al.* (2010) は、野生由来マウス系統 (MSM) と実験用マウス系統 (B6) の行動形質を行い、2 系統のマウスの活動量 (home-cage activity) に有意な差があることを明らかにした。さらに、第 6 番染色体の遺伝子型の違いが、活動量に大きく影響していることも示した。

そこで論文では、第 6 番染色体のどの領域がマウスの活動量に影響するのかを調べるため、第 6 番染色体の一部を MSM 由来のもので置換した B6 を親系統とするコンジェニック系統を 15 系統樹立し、それらの系統とその親系統の活動量データから回帰分析の手法を用いて活動量に影響する領域の特定を試みた。

回帰分析においては、マウス系統の活動量を被説明変数、microsatellite marker における遺伝子型を説明変数とした。しかし今回のデータに対して、通常最小二乗法に基づく回帰モデルを当てはめることには大きな問題点があった。それは、microsatellite marker (説明変数) の数がマウスの系統数 (標本数) よりも大きくなるという問題、つまり $p \gg n$ 問題、である。このようなデータに対しては、通常最小二乗誤差に基づく回帰モデルでは解が一意に求まらないことが統計学において知られている。

この問題を克服するため、論文では以下のように解析を行った。まず初めに、lasso (Tibshirani, 1996) を用いて変数選択を行い、説明変数が標本数よりも小さくなるまで変数を削減した。それから、通常最小二乗法による回帰モデルを当てはめ、all subsets を用いてさらなる変数選択を行った。

その結果、活動量に有意に影響を与える染色体領域を推定することに成功した。また、下図に示されているように、回帰係数の推定値から、推定された染色体領域において活動量がどの程度増加または減少するのかもあわせて推測することができた。今回の解析により活動量を大きくすると推定された領域は、追加して行った実験の結果とも一致した。

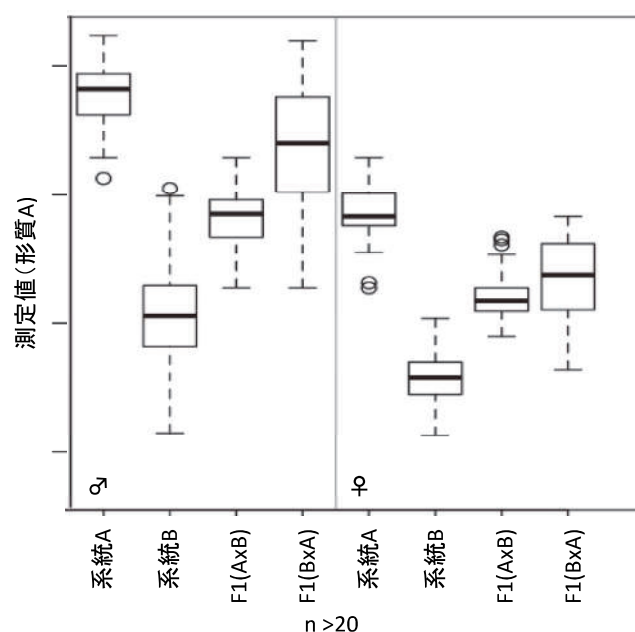


2-4: 分散の同等性検定 [藤澤、坂口、高田、岡、城石]

平均の同等性検定はしばしば行われているが、分散の同等性検定はあまり行われていない。本研究では、サブテーマ3（城石プロジェクト）のマウス形質の測定データにおいて、平均の差というよりも、分散の差が特徴的であるというデータが存在していたため、それに対処するための方法を扱うこととなった。図を参照して欲しい。二つの系統をかけ合わせる時、どちらがオスであるかによって、測定値（形質A）の平均にはあまり差は見られないが、分散には明らかに差が見られた。

標準的な方法は、正規分布を仮定した等分散性の検定である。しかしながら、データに明らかに外れ値が存在していたので、標準的な方法はバイアスを生じる。そこで、外れ値に妥当に対処するための方法が必要であった。まずは、データの正規性を確認した。それは大丈夫そうだったので、正規分布を仮定した下で、外れ値に強い等分散性の検定を構築することにした。外れ値に強い標準的な手法として、アンサリ・ブラッドレイ検定があるが、正規性の仮定を使う方が強力な手法が作れるため、その想定を使うことにした。そして、ガンマ・ダイバージェンスを使ってパラメータ推定を行い、その推定量に基づいて検定を行うことを考えた。P 値の計算は漸近近似を使うことにした。そのプログラムを作って遺伝研側に使ってもらった。幾つかの場所で有意差がきちんと確認できた。

その後、別のマウス（コンソミック系統など）を使用したアレイデータの解析において、やはり、等分散性の検定を行う必要が生じた。このデータに対しては、正規性の仮定が疑われたため、また、ばらつきが何であるかということ自体もはっきりしなかったため、外れ値に強く、また、オムニバスの手法であるアンサリ・ブラッドレイ検定を行った。



2-5: 攻撃・従属データに基づく上下関係の決定 [栗木、小出、高橋]

行動遺伝学において、動物個体のあるグループ内における強弱の上下関係をデータより推定し、その推定結果を表現型データとして QTL 解析等を行うことがしばしば行われている。

表 1（左、右）はその目的のためのデータ（攻撃・従属行列）の例である。5000 番台の数字はサル の 個体番号である。およそ 3 ヶ月にわたって、4 匹のサルを囲いにいれ、1 回 45 分のセッションの 21 回 について、サル 間の攻撃行動をカウントしたものである。このようなデータから、4 匹のサルにどの ようなランク付け、あるいはスコア付けを行うかがここでの問題である。

表1: サルの攻撃行列 (Morgan, et al. [1])

	5834	5837	5397	5390		5391	5396	5382	5381
5834	—	7	17	11	5391	—	5	62	30
5837	1	—	24	18	5396	0	—	0	9
5397	4	5	—	35	5382	0	0	—	39
5390	0	1	0	—	5381	1	3	6	—

統計学では、この種のデータは一对比較とよばれ、そのためのモデルとしてブラッドリーテリー (Bradley-Terry) モデルが知られている。さらにデータの同モデルからの乖離は「3 すくみ」として解釈できることが知られている([2])。本稿では、3 すくみを存在を考慮した一般の設定で、一对比較データに対する階層モデルの族を提案し、表1のデータに適用する。

一对比較データを $(n_{ij})_{p \times p}$ ($p = 4$) とおき、確率モデル

$$n_{ij} \sim \text{Poisson}(\lambda_{ij}), \quad \log \lambda_{ij} = \beta_{ij} + \gamma_{ij}$$

を想定する。ここで (β_{ij}) は対称行列 ($\beta_{ij} = \beta_{ji}$)、 (γ_{ij}) は歪対称行列 ($\gamma_{ij} = -\gamma_{ji}$) である。順位の情報は γ_{ij} が担っている。

新たなパラメータ

$$\eta_{ijk} = \gamma_{ij} + \gamma_{jk} + \gamma_{ki}$$

を定義する。 $\eta_{ijk} = 0$ のとき、別のパラメータ (α_i) を用いて、 $\gamma_{ij} = \alpha_i - \alpha_j$, $\gamma_{jk} = \alpha_j - \alpha_k$, $\gamma_{ki} = \alpha_k - \alpha_i$ とおくことができる。ここで α_i は個体 i の強さと解釈できる。 $\eta_{ijk} = 0$ のとき、個体 i, j, k は一次元的な尺度 (α_i) で順位付けできるので、パラメータ η_{ijk} は個体 i, j, k 間の 3 すくみを意味するパラメータと解釈できる。また $\eta_{ijk} = 0$ のとき、さらに $\gamma_{ij} = \alpha_i - \alpha_j = 0$ であるならば、それは個体間 i, j に強さの差は無いと解釈することができる。

$p = 4$ の場合、 γ_{ij} ($1 < i < j < 4$) のパラメータ数は 6 つ、3 すくみのパラメータ η_{ijk} ($1 < i < j < k < 4$) は 4 つ (自由度は 3) ある。これらのパラメータのいくつかを 0 と制約することによって階層モデルの族が得られる。ただし、制約 $\gamma_{ij} = 0$ は $\eta_{ijk} = 0$ ($\forall k$) の場合にのみ考えるものとする。

表1のデータについて、ここで提案した階層モデルを当てはめ、AIC によってモデル選択を行った。選択されたモデルとパラメータの推定値は以下の通りである。

表1 (左): 選択されたモデルは γ_{124} , γ_{134} を 0 とするモデル。非ゼロパラメータの推定値は

$$\hat{\gamma}_{ij} = \begin{cases} \hat{\alpha}_i - \hat{\alpha}_j & (i, j) \neq (2, 3), \\ \hat{\alpha}_i - \hat{\alpha}_j + 1.36270 & (i, j) = (2, 3), \end{cases}$$

$$\hat{\alpha}_1 = 2.95167, \quad \hat{\alpha}_2 = 1.70374, \quad \hat{\alpha}_3 = 2.28213, \quad \hat{\alpha}_4 = 0.$$

強さの尺度 $\hat{\alpha}_i$ を与えることはできるが、それは個体 2 と 3 の間の順位を反映していないことが分かる。

表1 (右): 選択されたモデルは、全ての η_{ijk} ($1 < i < j < k < 4$) および γ_{23} を 0 とするモデル。非ゼロパラメータの推定値は $\hat{\gamma}_{ij} = \hat{\alpha}_i - \hat{\alpha}_j$ とおくと、

$$\hat{\alpha}_1 = 2.9217, \quad \hat{\alpha}_2 = \hat{\alpha}_3 = 0.7792, \quad \hat{\alpha}_4 = 0.$$

この場合は、強さの尺度 $\hat{\alpha}_i$ は 4 個体の順位を忠実に反映することが分かる。

今回数値計算には、統計パッケージ R の関数 `glm` を用いた。このプログラムは重み付き最小 2 乗法を用いるため、データ n_{ij} の数が小さいとき、特に 0 を含む場合は精度良い解を与えていない恐れがある。その場合は、正確な最尤法を用いる必要がある。

また、動物個体の強弱関係は、複数の攻撃・従属行列データに基づいて想定する場合が多い。表ごとに異なった結論が得られた場合に、どのように結論を統合するかは今後の課題である。

引用文献

1. D. Morgan, K. A. Grant, O. A. Prioleau, S. H. Nader, J. R. Kaplan, and M. A. Nader (2000). Predictors of social status in cynomolgus monkeys (*Macaca fascicularis*) after group formation, *American Journal of Primatology*, 52, 115-131.
2. 栗木哲 (1991). 一対比較モデルに関する同時信頼領域の構成と多重比較法, *応用統計学*, 20, 127-13.

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」

3-1：次世代シーケンサーによる対立遺伝子発現バイアスの遺伝的制御機構の解明

[豊田、藤山、阿部、中谷、近藤、清澤、藤澤、栗木、二宮、原、岡、高田、城石]

本テーマの目的は、ゲノムのリファレンス配列が整備されている 2 種類のマウス系統間のゲノム多型情報と、両系統を使用した計画的交配により得られた系統および個体群について、各種組織の遺伝子発現情報を解析して、対立遺伝子間の発現量差を制御する遺伝機構の解明に繋がる研究を行うことである。我々が使用するマウスは、最初にゲノム配列が明らかにされた汎用実験用系統である西ヨーロッパ産由来の C57BL/6J (以降 B6 と表記)、および国立遺伝学研究所が樹立し、我々がその塩基配列を明らかにした日本産野生由来の MSM 系統である。両系統間には、莫大な SNP 情報があるため、次世代シーケンサーを用いた RNA-Seq 法で両親由来の対立遺伝子発現バイアスのプロファイリングが可能になる。昨年度までに、B6 および MSM の SNP 情報を取り入れ、両アレルを区別した遺伝子発現解析プラットフォームを構築できたので、本年度は本解析のためのデータ取得をサブテーマ 1 との連携で行った。解析に使用したのは、MSM、B6、F1 (B6/MSM)、F1 (MSM/B6) の 10 週齢の雌雄、各 6 個体である (下表)。遺伝子発現解析はまず、肝臓および精巣について行った。

系統	性別	組織	解析数 (RNA-Seq)
B6	Male	Testis	6
MSM	Male	Testis	6
F1 (B6 ♀ x MSM ♂)	Male	Testis	6
F1 (MSM ♀ x B6 ♂)	Male	Testis	6
B6	Male	Liver	6
MSM	Male	Liver	6
F1 (B6 ♀ x MSM ♂)	Male	Liver	6
F1 (MSM ♀ x B6 ♂)	Male	Liver	6
B6	Female	Liver	6
MSM	Female	Liver	6
F1 (B6 ♀ x MSM ♂)	Female	Liver	6
F1 (MSM ♀ x B6 ♂)	Female	Liver	6

RNA 保護試薬から取り出した組織サンプルから RNA を精製し、アジレント社製、バイオアナライザーによる品質検査をパスしたものをイルミナ社製、TruSeq Stranded mRNA Sample Prep Kit より解析試料を作製した。塩基配列解読用のサンプルは、バイオアナライザーにより再度品質検査を行い、サブテーマ 1 によるデータ解析に供した。イルミナ社製、HiSeq2000 あるいは HiSeq2500 により生産された 1 次データは、サブテーマ 1 のサーバーよりデータを受け取り、国立遺伝学研究所 DDBJ スーパーコンピュータの「tricmouse」アカウントで 2 次・3 次解析を行った。結果のマッピングには、マウスの基

準配列である B6 系統の mm9、mm10、我々がゲノムを公開している MSM の多型情報を使用した。なお、2 次・3 次解析を行う過程で、データの閲覧および解析手法の検討を「tricmouse」アカウントを利用してメンバー間で行った。図 1 にこれまでに得られた結果の一例（肝臓、オス）を示す。緑円は MSM アレルの発現量が B6 アレルよりも高い遺伝子群、青円は B6 アレルの発現量が MSM アレルよりも高い遺伝子群、赤円は、BM および MB どちらの組み合わせでも母方のアレルの発現量が高い遺伝子群、黒円は同様に父方のアレルの発現量が高い遺伝子群を示す。このように、亜種間ゲノム多型を利用した、我々のプラットフォームを使用することで、アレルを区別した遺伝子発現解析を全ゲノムに渡って効果的に行うことが可能である。

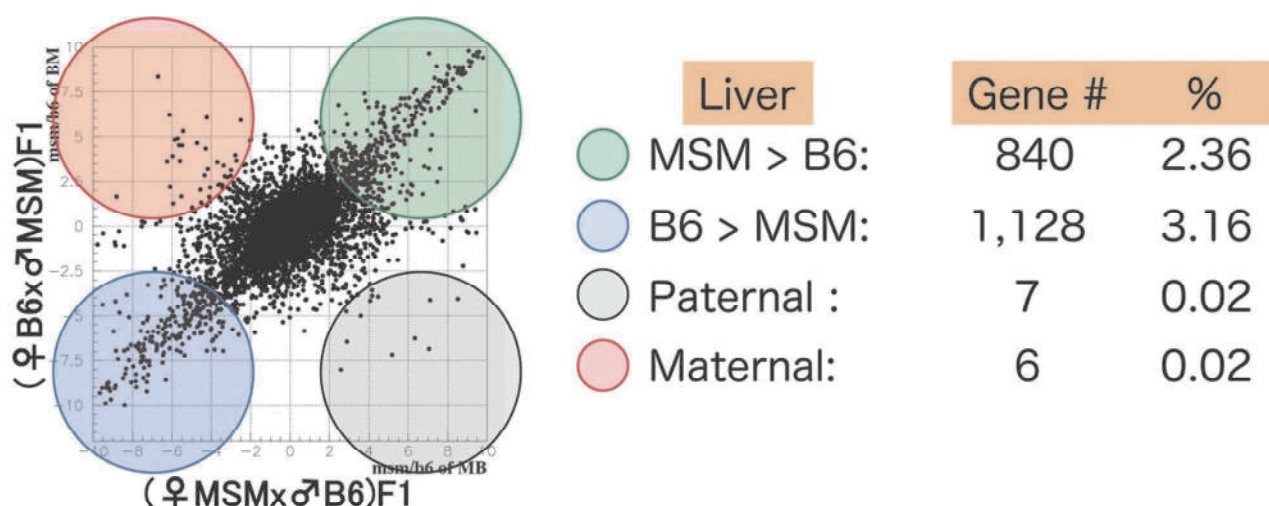


図1. 一組の F1 個体間の肝臓での遺伝子発現の比較結果。non-coding 遺伝子を含む転写産物 35,662 種 (ensemble version 71)を対象にしたプロット図を示す。

今後は、発現遺伝子の検出に関わるパラメータの設定や、検出した遺伝子の他の解析系による検証を行う。さらに他の組織や遺伝子発現制御領域の解析についても順次行う。また、B6 および MSM 系統間の多型情報を利用した遺伝子発現バイアス解析を本格的に開始することができたので、これまでに収集した表現型データと遺伝子発現データとの相関解析をサブテーマ 2 と連携して進める。

3-2: 大規模実験交配集団を対象とした遺伝的構造の統計解析 [栗木、岡、高田、城石]

マウス汎用近交系統および亜種由来系統の大規模交配により F2 個体を作製し、複数の形質データを観測するとともに、シーケノム社製マスアレイを利用した高速 SNP タイピングによる遺伝子型情報を組み合わせ、形質データに影響をおよぼす遺伝的構造を統計的に検出することを目的としている。昨年度までに、B6 および MSM を基盤とした F2 個体 3000 頭超の形質データと全ゲノムに渡る 200 カ所程度の SNP によるジェノタイピングを行った。本年度は、形質データの取りまとめ、および初回の SNP タイピングの際にデータの得られなかった座位について、再度マスアレイによる解析を行い、データを補完する作業を行った (図 2)。

今年度に行った一連の作業により、個体ごとの解析結果の歩留まりと遺伝子型情報量が増加した。今回得られたデータをもって本テーマのデータ生産を終了とし、次年度以降サブテーマ 2 と連携した統計解析を本格的に進める。

CHROMOSOME	1	1	1	1	1	1	1	1	1	1	1	1	1	1	2	2	2	2	2	2	2	2	2	2
	ASSAY_ID	A_0001_m	A_0014_m	A_0027_m	A_0038_m	A_0049_m	A_0061_m	A_0067_m	A_0069_m	A_0070_m	A_0079_m	A_0093_m	A_0104_m	A_0110_m	A_0112_m	A_0124_m	A_0135_m	A_0148_m	A_0155_m	A_0160_m	A_0169_m	A_0183_m	A_0193_m	A_0208_m
	POSITION	3186252	23082046	43228091	63118108	83032013	103319439	113926082	123216845	126361123	143058805	163052397	183043901	193228927	3222853	23035126	43042909	63288507	73889803	80505208	101452036	123548198	143009707	165512149
B6MS_3010		M	-	-	-	-	H	-	H	-	-	B	-	B	B	B	-	-	-	-	-	H	-	B
B6MS_3011		B	-	B	-	-	H	-	H	-	-	B	-	B	M	H	-	-	B	B	B	B	-	B
B6MS_3012		B	B	B	-	-	H	-	H	-	-	H	H	H	-	B	-	-	B	B	B	B	-	B
B6MS_3013		B	H	-	-	-	H	-	B	B	-	B	B	B	-	H	-	-	-	-	H	H	-	B
B6MS_3014		-	-	M	M	-	H	-	H	-	-	B	B	B	-	H	-	M	M	M	-	H	-	B
B6MS_3015		M	-	M	M	-	H	-	H	-	-	H	H	H	-	H	-	-	-	-	-	-	-	B
B6MS_3016		M	-	M	M	M	-	M	M	M	M	H	H	H	-	M	M	-	M	M	-	H	-	B
B6MS_3017		B	B	B	B	-	B	B	B	B	-	B	H	H	-	H	-	-	B	B	B	B	-	B

	Chr	Target_SNP_ID	1	A_0001_A_A1	1	A_0014_A_A1	1	A_0027_A_A1	1	A_0038_A_A1	1	A_0049_A_A1	1	A_0061_A_A3	1	A_0067_C_C3	1	A_0069_A_A1	1	A_0070_A_A2	1	A_0079_A_A1	1	A_0093_A_A1	1	A_0104_A_A1	1	A_0110_C_C1	2	A_0112_A_A1	2	A_0124_A_A1	2	A_0135_A_A1	2	A_0148_A_A1	2	A_0155_C_C3	2	A_0160_J_D2	2	A_0169_D_D2	2	A_0183_A_A1	2	A_0193_A_A1	2	A_0208_J_A2
	Position	3186252	23134610	43340390	63230406	83149564	103388469	113995112	123285876	126430154	143138971	163145851	183137357	193352191	3222853	23068637	43076420	63325725	73927021	80544526	101491354	123682493	143144002	165646444																								
B6MS_3010		M	M	H	H	H	H	H	H	B	B	B	B	B	B	B	B	-	H	H	M	H	H	B																								
B6MS_3011		B	B	B	B	H	H	H	H	H	H	B	B	B	B	M	H	B	-	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	
B6MS_3012		B	B	B	H	H	H	H	H	H	H	H	H	H	H	H	B	B	-	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	
B6MS_3013		B	H	H	H	H	H	H	H	B	B	B	B	B	B	H	H	H	-	H	H	H	H	H	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B
B6MS_3014		H	M	M	M	H	H	H	H	H	B	B	B	B	B	H	H	H	M	M	M	M	H	H	B																							
B6MS_3015		M	M	M	M	H	H	-	H	-	H	H	H	H	H	H	H	-	-	H	-	-	H	B																								
B6MS_3016		M	M	M	M	M	M	M	M	M	M	H	H	H	H	M	M	-	M	M	M	H	H	B																								
B6MS_3017		B	B	B	B	B	B	B	B	B	B	B	H	H	H	H	H	-	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	

図2. MassARRAY により得られた結果の一例を示す。データ補完前のデータ(上)、データ補完により得られた最終データ(下)。各列に個体ID、各行に遺伝子座の染色体番号と座標(セントロメア側から、bp)を示す。B: B6 ホモ型、M: MSM ホモ型、H: B6 と MSM のヘテロ型。- はデータが得られなかった座位を示す。

3-3 : マイクロ CT 画像の情報学的解析による高速・高精細マウス表現型の自動抽出法の開発

[田村、前野、城石、北本]

本研究は、ゲノム・遺伝子機能を解明する為とその礎となる表現型解析を、高速、かつ精細に行う技術の開発を行う事にある。その先ず第一歩として、1) micro computed tomography (micro-CT) を用いて、各種造影剤を組み合わせることにより軟組織、各種臓器のイメージングを行う技術の開発、2) それらイメージデータを用いた表現型の情報学的自動抽出法の開発を行った。1) は、遺伝研・城石グループが、2) は、情報研・北本グループが担当した。これら micro-CT を用いたイメージング技術、及び表現型自動抽出技術は世界的に見ても黎明期にあり、今後の発展が大いに期待される分野である。遺伝研の城石グループは、サンプルにマウス胎児を用い、造影剤としリンタングステン、及びヨード系溶液を、固定液として 4%パラホルムアルデヒド、10%ホルマリン、ブアン溶液を用いて画像データ取得のための条件検討を行った。これらの組合せを比較検討した結果、最もコントラストよく軟組織を画像化出来たのは、4%パラホルムアルデヒド固定+ヨード系造影剤の組合せであった(図1)。次に、その解像度

を検討した結果、胎生 14.5 日胚において、個々の臓器のみならず、肺や脊椎に入る毛細血管、更には個々の筋肉繊維までもが識別可能であり、実際の使用に十分耐え得ると考えられた。そこで実際の突然変異マウス（染色体異常をもつ）胎児を用いて、その同腹野生型 3 個体、ヘテロ変異体 7 系統、ホモ変異体 3 系統の胎生 14.5 日胚の撮影を行い、画像データを取得した。

一方、情報研の北本グループでは、表現型を自動抽出する画像処理手法に関する研究を進めた。本研究は表現型自動抽出技術を、①画像レジストレーション、②表現型候補領域の抽出、③表現型候補領域の評価、④表現型の分類と解析、⑤データアノテーションツールの 5 テーマに分け、特に②、③、⑤のテーマについて研究を進めた。

まず②のテーマでは、平均画像に合わせにくい領域が各個体の表現型に対応するとの仮説に基づき、①のレジストレーションで得られる各種の特徴量（レジストレーション変形特徴量）を活用して候補領域を抽出する方法について、前年度から引き続き研究を進めた。前年度の研究では、実験で抽出した結果に既知の表現型（心室中隔欠損など）が含まれていることを確認していたが、今年度は表現型のアノテーションツール Annotation ROI 3D を開発し、これによって⑤のステップを効率化することができた。また上記ツールは ImageJ のプラグインとして開発し、オープンソースとして公開した。さらに、③のテーマについては、レジストレーション変形特徴量に基づく手法では表現型だけではなく既知の変異（腸などの不定形臓器や性差）なども検出してしまうため、野生型とヘテロ・ホモ型との比較を通して、野生型にも存在する変異は無視するというアイデアに基づく手法を試した。

今後の方針としては、1) レジストレーションと変形特徴量の改善による候補領域抽出の高精度化と高速化、2) 候補領域の確からしさの評価に基づくランキング手法、の 2 点について研究を進める計画である。尚、本研究は、NII Internship Program Student である Sharmili Roy, Parminder Reel, Samuel Kerjose, William Grimes, Zhongliu Xie 氏の研究協力を得て遂行された。

図1. マウス胚の micro-CT 画像。
(左)14.5 日胚 3D 立体再構築像、(右)同矢状断面像

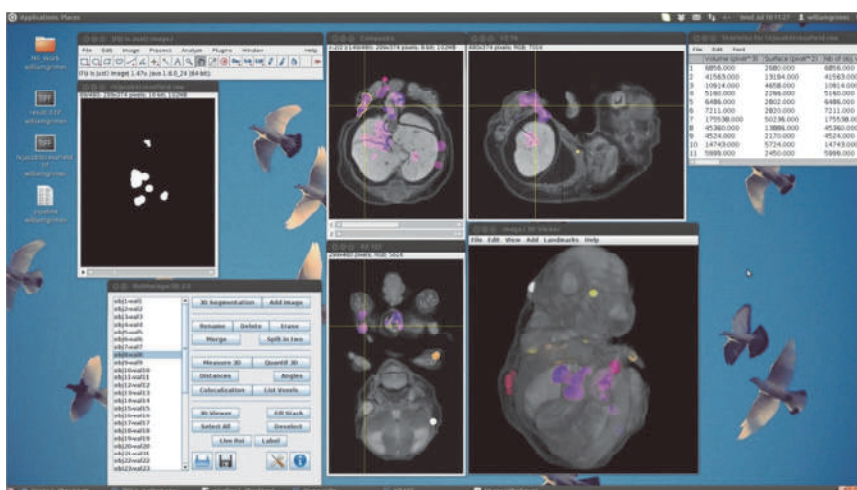


図2. データアノテーションツール Annotation ROI 3D のスクリーンショット。

3-4：野生由来ヘテロジニアスストックによる高精度 QTL 遺伝子マッピング法の確立

〔後藤、高橋、栗木、加藤、二宮、高野、小出〕

行動や体質などの多因子形質の遺伝的基盤を理解することは、遺伝学におけるもっとも大きな課題の一つとなっている。そのためには、全ゲノムを対象とした関連解析（GWAS）などが行われているが、実際には明確な遺伝子同定が困難な状況が続き、目立った成果が見られない。その原因として、小さな効果を持った多数の遺伝子座の存在や遺伝子間相互作用などが考えられているが、その実態はまだ明らかになっていない。このような多因子形質の遺伝解析における問題を打開するために、近年多数のマウス系統を遺伝的に交雑した集団を使う例が多く見られるようになってきた。ヘテロジニアス集団は、複数の遺伝的に異なった近交系統を相互に交配し、その後近交化を避けるためにランダム交配を続けることで作出されるものである。現在我々は、Domesticus 亜種由来の 2 系統（PGN2, BFM/2）、Castaneus 由来の 1 系統（HMI）、それから Musculus 亜種由来の 5 系統（BLG2, NJL, MSM, CHD, KJR）の合計 8 種類の野生由来系統をもとに 16 ペアで一つの集団を維持することで野生由来ヘテロジニアス集団の作製を進めている（図 1）。これら野生由来の 8 系統は、系統間で遺伝的な差が大きく、その遺伝的マーカーおよび形質の多様性を持つことが大きな特徴である。そのため、これらの野生系統から作出されるヘテロジニアス集団内では膨大な多様性が期待できる。私たちは、この集団に対して従順性にかかわる形質の選択交配を行うことで、この形質に関して顕著な違いを示す集団を作製し、選択集団と非選択集団を比較することによる精度の高い遺伝解析を行うことを試みている。また、この野生由来ヘテロジニアス集団は、ほかにもさまざまな多因子形質の遺伝解析に利用できる可能性を持っており、F2 集団などを用いた QTL 解析では十分な精度の遺伝子座情報が得られない場合に、このヘテロジニアス集団を利用した解析を行うことで精度の高い遺伝解析が可能になると期待できる。現在、11 世代目を交配中であり、その作製は順調に進んでいる。

昨年度は、16 ペア（総アレル数 64）で作製しているヘテロジニアス集団中のアレルがランダムドリフトによりどのような頻度で変動するかシミュレーションにより推定した。その結果、16 ペアで集団を維持した際に、アレル頻度は 10 世代目においても総アレル数の半分である 32 を超えることはほとんどないことが分かった。このことから、選択交配により特定の形質を選択した集団について、総アレル数がシミュレーションの値を超えるものが出来れば、それが選択された形質に関わると予想される。今後はこのシミュレーションの方法を用いて、遺伝解析を行う手法の確立を目指す。

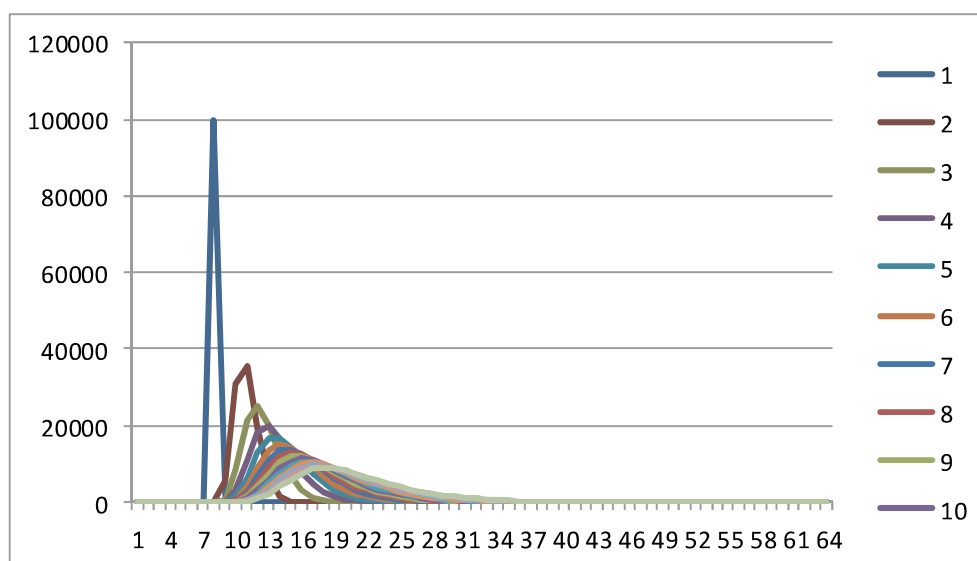


図1. 16 ペア（総アレル数 64）のヘテロジニアス集団におけるアレル頻度の世代ごとのシミュレーション結果
横軸：最大アレル数、縦軸：10 万回のシミュレーションにおける該当する回数、1～10：世代数

3-5：マウス並走状態除去による自動推定アルゴリズム一致率向上の試み

〔土谷、荒川、柿原、高橋、杉本、田邊、小出〕

本プロジェクトではこれまで、社会行動解析に関して隠れマルコフモデルによるコンピュータ学習により自動推定を可能とし、社会行動テスト記録画像から有用な行動データを自動抽出する方法の確立を進めてきた。平成 23 年度はマウス社会行動の自動解析を可能にするソフトウェアの開発を進め、平成 24 年度はこのソフトウェアを用いて、実際のマウス系統の社会行動を自動解析し、実験者の実験結果と比較することで、実験者の観察とほぼ同レベルの解析結果を出した。しかし、マウスペアが並走する場合（図 1）においても追従行動とみなす誤判定が除去できておらず、結果的に実験者の観察結果との一致率向上やソフトウェアの判定精度向上の妨げになっている。

そこで平成 25 年度は、従来のマウス社会行動を自動解析するアルゴリズムを改良し、マウスペアの並走状態を除去する手法を開発した。人間観察により動画画像のフレーム単位で並走状態と判定したデータと、その際のマウス行動のデータを教師データとし、サポートベクタマシン（SVM）を用いて並走状態を学習させた。この学習結果を任意のマウスデータに適用した所、正しく状態推定ができていた箇所はそのままの状態、並走状態を追従行動と誤判定している箇所は除去することができた（図 2）。現在、推定結果を実験者の実験結果と比較し、定量的な評価に基づいて一致率が向上していることを解析中である。

また、平成 23 年度、平成 24 年度に開発したソフトウェアを、フリーウェアの解析ソフト DuoMouse として国立遺伝学研究所にて公開を開始した（図 3）。また、この件は論文を発表と同時にプレスリリースし、新聞にも掲載された。



図1. マウスペアの並走状態

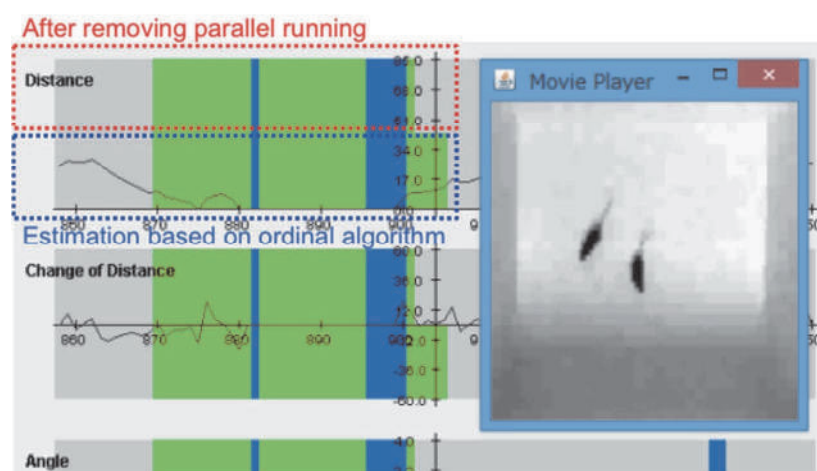


図2. 従来アルゴリズムによる判定と SVM による並走除去後の判定の比較。緑が並走状態を表す。

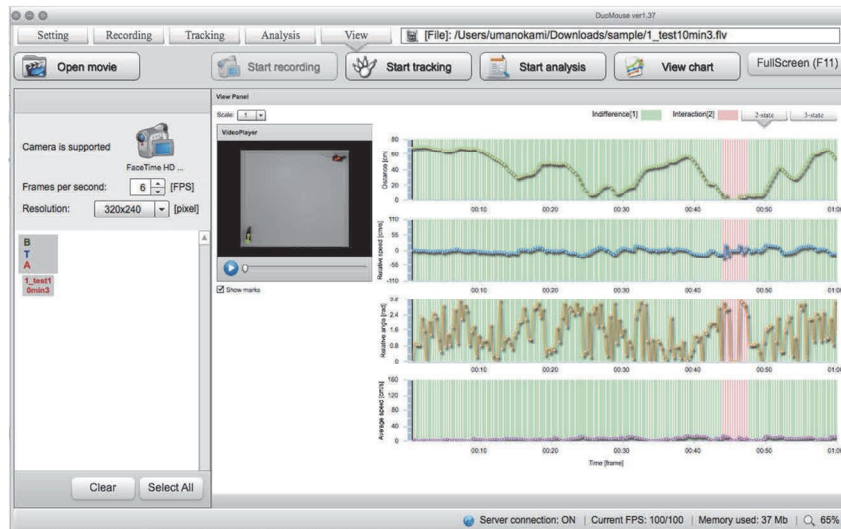


図3. フリーウェア「DuoMouse」

3-6: マウス超音波発声の関数データ解析 [Dou、栗木、杉本、高橋、小出]

マウス超音波発声データには多くのノイズが含まれていて、発声パターンの種類も多いため、データを自動的に解析することは難しい。本研究ではマウスの超音波発声（USV）データのパターン認識を行うための効率的なデータ解析手法を提案し、フリーソフトウェアで解析を行う事を目指している。なお、本研究は大阪大学白旗慎吾名誉教授の協力により進めている。

提案する手法は主にノイズの除去、USV の特定、USV の関数化、関数データのクラスタリングの 4 つのステップで構成されている。これまでの研究で、これらの大きなステップの構成が完成し、連続な発声パターンはもちろん、周波数が突然シフトする発声パターン（ジャンプ）の解析、分類もできるようになった。これらの方法の有用性が BALB/cAnN と C57BL/6 のマウスの USV データの分類により確認されている。しかし、以下に示すようにいくつか細かな課題がまだ残されている。

- 1) ノイズ除去のステップにおいて、ノイズであることを判断するための規準は今まで経験的に決めてきたが、客観的な基準が必要
- 2) USV と USV の間の区間を決定する方法の確立が求められる（図 1）
- 3) 関数化のステップにおいて、USV 曲線にジャンプがある場合（図 1）はジャンプを定義するための変化点を決定することが難しい

これらの課題のうち、USV 間の区間の長さを決めるためには、USV 波形総数が隣接する二つ USV 間の区間の長さに対する変化によって決める方法を考えている。区間を短く設定すると多くの USV 波形が得られ、区間を長く設定すると少ない USV 波形が得られる。そこで、USV 波形の数が安定する区間の長さを適切な区間の長さとする。たとえば、図 2 では、区間の長さが 50（単位時間）程度のときに USV 波形の総数が安定しているため、50（単位時間）が適切な区間の長さと考えられる。また、短い USV 波形をノイズとして考えるときに、短い USV 波形の長さや数を考える必要もある。データ解析を自動化するために、これらのパラメータの値をデータから自動的に決める方法を探る必要がある。

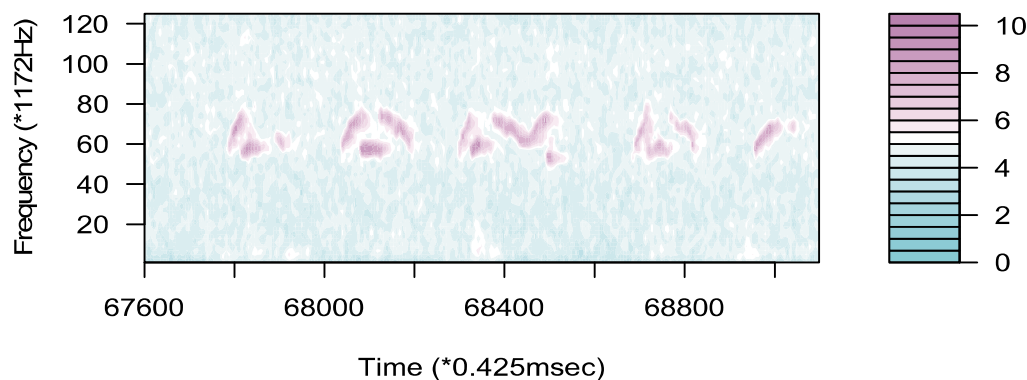


図1. C57BL/6 超音波発声データの一部

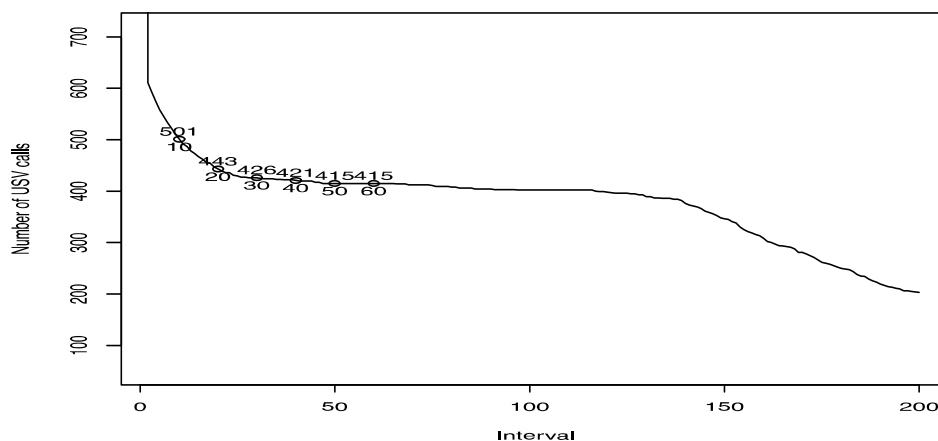


図2. 発声パターンを決める

3-7: eQTL解析による表現形質と遺伝子発現のネットワークの解明 [堀内、春島、藤澤、久保、倉田]

栽培イネ *Oryza sativa* の祖先種である野生イネ *O. rufipogon* は、栽培イネより多様な遺伝的形質を示し、生態、形質調査から数種類の集団構造が存在することが示唆されている。第 I 期融合研究プロジェクトで推進した。*O. rufipogon* の量的形質調査からも、柱頭長、種子稔性、穂の形（着粒数と相関あり）、発根・再生力など多く形質が多年生系統群、一年生系統群で大きく 2 つに分かれることが示唆されている。

このように野生イネ *O. rufipogon* は集団構造を持つため、様々な品種を用いたアソシエーション解析による表現型及び遺伝子発現量の遺伝解析は困難を伴う。そこで本研究は表現型の離れた一年生、多年生グループから代表的な 2 系統 W0106 と W1921 を選抜し、この交雑後代の実験集団を育成する。実験集団を用いた量的な遺伝形質の遺伝解析 (QTL 解析) と遺伝子発現量の遺伝解析 (eQTL 解析) を行い、表現形質を形成するゲノム上の遺伝因子群を特定し、その遺伝子発現のネットワークの解明を目指している。本研究ではより解像度の高い解析集団として組換え固定化集団を確立するために多数の F₂ 種子を取得し、H22、H23 年度に異なる F₂ 個体どうしを交配した F₃ 他殖系統群 333 個体を獲得後、H24、H25 年度それらの自殖 F₄ 系統から薬培養による倍加半数体系統

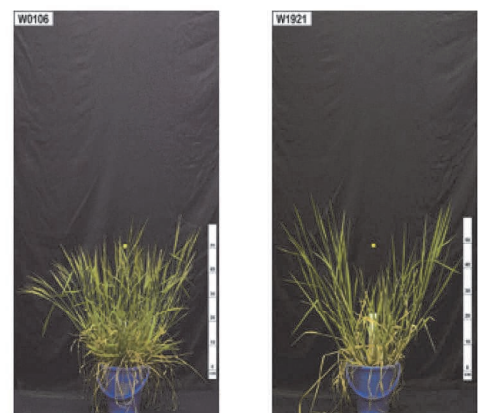


図1. *Oryza sativa* の祖先種 *O. rufipogon* (左)一年生型の W0106、(右)多年生型の W1921。