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による遺伝機能システム学

プロジェクトディレクター：　倉田 のり　教授（国立遺伝学研究所）

〔1〕 研究計画・研究内容について

(1) 目的・目標

本プロジェクトでは、超大規模ゲノム配列情報や遺伝子発現情報のデータ解析手法と多元的な生物表現型多様性の統計モデリング手法を開発する。両者を統合することにより複雑な遺伝的相関構造を描出するための方法を開発し、モデル生物に適用してゲノム機能と遺伝的ネットワーク抽出を行う。これにより、多数の遺伝因子の高次連関から形成される生物多様性を、システムとして理解することを目指す。これらの研究の効率的推進のため、本プロジェクトは以下の３つのサブテーマを設定して行う。

（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：情報研／遺伝研）

（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（代表、栗木　哲：統数研）

（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：遺伝研）

サブテーマ１では、最新のゲノムテクノロジーを駆使して年あたりペタバイト級の超大容量ゲノム・遺伝子関連データを系統的に生産し、統計情報解析研究と融合させることによって、生命システム原理についてのデータセントリックな理解を目指す。

サブテーマ２とサブテーマ３においては、まずマウス、イネ、ゼブラフィッシュ、ショウジョウバエなどのモデル生物のゲノム配列・遺伝子型、遺伝子発現表現型、表現型因子、背景因子、時系列情報などのゲノムおよび表現型の多型・多様性データの抽出法を確立する。抽出データの多重組み合わせや相関解析などを行う統計的解析手法の開発と、その手法によるデータ解析を行い、生命・遺伝現象に関するメカニズムを新たな手法で解明することを目的とする。とくにデータ取得技術の進展によって、時系列、e-QTL、遺伝集団構造解析の展開で生み出される新たな形での情報取得と、それによって生ずる課題に対応できる手法の開発を目指す。それらの方法論に立脚して、生物表現型多様性を多元的に様々な角度から説明できる「遺伝機能システム学」の流れを作り、生命システム分野の展開を図る。

(2) 必要性・重要性（緊急性）

ヒトゲノム解読計画以来、生命科学の研究スタイルは、大量のゲノム情報を基盤とするデータ駆動型へとパラダイムシフトが進行しつつある。近年の超並列大規模シーケンサの登場は世界的にもその勢いを加速させているが、我が国の大学研究機関の対応は大きく後れており、特に遺伝学の研究分野における大量のデータ処理と情報解析に不可欠な統計分野、情報分野との研究交流は、ごく一部のグループを除いては殆ど行われていない。このため、今後の情報ゲノム科学の発展に不可欠の研究コミュニティー形成や、社会にこれらの超大量データを用いた科学知の必要性を認知させる仕組みは脆弱であり、早急の基盤形成が必須である。このような基盤形成を行う場として、本プロジェクトの掲げる研究連携を、生命・遺伝研究の一つの中心拠点である遺伝研と、情報処理・解析の拠点である情報研、統計数理の拠点である統数研を有する情報・システム研究機構で担う事は、大きな意味がある。

世界での研究分野の現状を考えるとき、日本でのこのような試みは遅きに失した感もあるが、緊急に進めるべき分野である。これまでの研究期間に成果も出始めており、さらに推進速度を上げ、新たな展

開を図る事が重要である。

(3) 期待される成果等（学問的効果、社会的効果、改善効果等）

本プロジェクトの成果は、生命科学、統計学、情報学の各研究コミュニティーに対して異分野間融合研究の有効性を強くアピールするものとして期待できる。同時に、本融合研究を通して作られる遺伝学と情報学、統計学の共同研究の土壌は，いずれの研究コミュニィーにも学問的に非常に良い影響を与えることが期待される。また、わが国の学術分野で欠落している multi-disciplinary な人材育成の土壌が参加各機関に形成されることが期待される。

多くの局面で統計的データ解析手法は生命遺伝現象データ解析に決定的に有用であるが，測定技術進歩にデータ解析が追いついていないのが現状である．また歴史的には統計学の起源のひとつは，生命・遺伝現象のデータを解析するために生み出されたものであり，現在においても生命遺伝データの解析を目的として開発される手法が普遍化されることにより統計学全体に還元されることは少なくない。それらの方法論は遺伝研究以外においても活用できる汎用的な性質を有するため，統計学への還元も期待される。

これらの新たな研究開発現場の中で学ぶ事により、大学院生、ポスドクなどが新たな学問の形成に寄与しつつ、育って行く事が望まれる。大容量、多様なデータに基づく多次元の視点からの研究が生み出す新たな研究、新たな問題が次の時代の学問の醸成につながる。いずれの視点からも、本課題が遺伝学研究、統計学、情報科学に大きな貢献をもたらすのみならず、これらの分野を超えた新しい融合領域の創成が期待される。

(4) 独創性・新規性等

本プロジェクトで研究対象とするイネ、マウス、ゼブラフィッシュ、ショウジョウバエ等のモデル生物は、国立遺伝学研究所が独自に開発した実験系統と自然変異を豊富に包含した独創的な遺伝資源である。これらの研究資源の大容量ゲノム関連情報の生産は、世界でトップクラスのゲノム解読能力を有する国立遺伝学研究所のシーケンスセンターが担当する。また、大容量情報のデータ解析には、我が国の情報学の拠点である国立情報学研究所と統計数理解析の拠点である統計数理研究所が担当する。このように、独自の遺伝資源を軸に据え、遺伝学、情報学、統計学の統合的解析を行う「遺伝機能システム学」の構築は、世界に類を見ない新規性の高い試みであり、本プロジェクトの推進はこのような融合的研究体制を組織できる情報・システム研究機構以外ではあり得ない。さらに、本プロジェクトは、情報・システム研究機構外の大学・研究機関からも研究者が参加し、相互に密接な関係を保ちながら研究を進める点でも極めて新規性が高く、大学共同利用機関としての役割を担うものでもあり、他の研究機関で実施することは不可能である。

(5) これまでの取り組み内容の概要及び実績

サブテーマ 1 については、１）これまでに新型シーケンサの導入とゲノム解読への利用を進め、大腸菌、線虫などの生命研究に有用な変異体のゲノム変異部位の特定や、野生イネ系統についての比較ゲノム解読、ゲノム機能領域の大規模解析等を行っている（Huang, Kurata, *et al. Nature*, 2012; Shang, *et al. Genome Res.*, 2010 等）。２）また、次世代シーケンシングシステムの開発・改良を進め、2008 年 7 月から 2013 年 3 月までに累積で約 80×10^{12} 塩基のデータを生産するとともに、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いる。３）超大規模データ生産とその直接的情報処理に加え、大規模ゲノムデータの情報・統計処理解析とゲノム工学等の融合を進めている。４）染色体機能領域研究については、特に動原体領域に存在するタンパク質の同定やゲノム配列上の特徴、ゲノム改変技術を開発

してきたが（Hori et al., Cell, 2008; Amano et al., J. Cell Biol., 2009）、さらにゲノム工学を活用して人工的に動原体を構築する実験系を確立した（Nishino et al., EMBO J. 32: 424-436, 2013 ; Hori et al., J. Cell Biol. 200: 45-60, 2013）。５）これに加え、主にシロイヌナズナ及びイネを用いて、ゲノム DNA の修飾と機能との関連解析研究を進めている。

サブテーマ２では、関連する以下の成果を得ている。１）機械学習とベータダイバージェンスの方法による形の表現型の計量化と QTL 解析のロバスト化（Mollah et al. Neural Processing Letters, 2007 など）。２）ロバスト推定に関する基礎的な研究 : 外れ値の割合が大きい場合にも潜在バイアスを小さくすることが可能な方法を提案した（Fujisawa & Eguchi, Journal of Multivariate Analysis, 2008）。３）高い相関構造を持つ多重検定の研究：多重積分の必要なしに簡単に陽に計算ができ、近似精度が高く、保守的でもある検定方法を開発した（Ninomiya & Fujisawa, Biometrics, 2007）。４）連鎖解析のエピスタシスの解析において、逐次解析やチューブ法を含むいろいろな方法で検定の多重性調整 p 値を与える方法を開発した（栗木哲「QTL 解析の統計モデルと検定の多重性調整」、21 世紀の統計科学、東京大学出版会、2008; Kuriki, Harushima, Fujisawa & Kurata, Annals of the Institute of Statistical Mathematics, 2014）。５）遺伝子発現差解析に関連して、既存の外れ値の処理方法を大幅に変更してより良い結果を得ることに成功した（Fujisawa, H., Horiuchi, Y., Harushima, Y., Takada, T., Eguchi, S., Mochizuki, T., Sakaguchi, T., Shiroishi, T. and Kurata, N., BMC Bioinformatics, 2009）。６）いくつかの古典的 QTL 解析において、影響関数を定義し、マウスのデータについて解析を行った（Dou, Kuriki, Maeno, Takada & Shiroishi, Biometrical Journal, 2014）。７）新しい並べ替え技法による P 値の推定法を提案した。一般的に行われてきた並べ替え技法による P 値の推定は必ずしも妥当ではない。P 値推定が妥当になる条件を整理して、そのクラスの中で最適な検定方法を導出した（Fujisawa, H. and Sakaguchi, T., TEST, 2012）。８）不均一な集団を記述するためのベルンシュタインコピュラの方法を開発し、マウスコンソミックデータに適用した（Dou, Kuriki, Lin & Richards, Computational Statistics & Data Analysis, in press）。これらの成果のうち、１）、４）については、サブテーマ３とも関連して、第 1 期新領域融合研究プロジェクト「生物多様性解析」を中心とする成果、６）については第 1 期からの継続中のテーマである。

サブテーマ３については、サブテーマ２と同様に第 1 期新領域融合研究プロジェクト「生物多様性解析」に基づく成果を論文化し、さらに第 II 期の研究においても、以下のような展開が進み、新たな成果が得られている。

マウスにおいては、第 I 期の量的形質情報の基盤整備や遺伝的解析に加えて、ゲノム配列多型（Takada et al., Genome Res.,2013[朝日新聞　2013.6.7　紹介記事掲載]）さらには、この多型情報を取り入れた複合量的形質の解析（Oka et al., PLos Genetics, 2014）に進展があった。また、表現型情報を使用した統計解析手法（Dou X et al., Biom. J.2014）をサブテーマ 2 との連携で報告した。量的形質情報の基盤整備に関しては、Mamm. genome 2012; Exp. anim.2012 に総説を報告している。マイクロ CT 画像の情報学的解析では、情報研との共同研究による成果（Roy S et al. Med. Image Comput. Comput-Assist. Interv. 2013）や遺伝解析（Tamura et al. Hum. Mol. Genet.,2013）についても論文を報告した。現在は、これまでに得られた複合量的形質データおよびゲノム配列多型情報など、大規模データを利用した各種遺伝解析を推進している。

マウスの社会行動を隠れマルコフモデルにより自動解析するフリーウェア（DuoMouse）を完成させ(Arakawa, Takahashi, Shiroishi, Tsuchiya, Koide 他　Journal of Neuroscience Methods 2014)、遺伝研のウェブサイトから公開した（http://www.nig.ac.jp/labs/MGRL/DuoMouse.html)。その成果はプレスリリースされ、新聞報道された（http://www.at-s.com/news/detail/1049459354.html）。さらに、マウス行動の遺伝的基盤解析系の確立のために基礎的研究を進めた（Kato, Kuriki, Koide 他 Heredity,

2014; Takahashi, Koide 他 PLoS ONE 2014; Goto, Shiroishi, Koide 他 Genes Brain Behav, 2013; Umemori, Uno, Koide 他　BMC Genomics, 2013）

１）マウスにおいては、これまでにゲノム配列多型に基づいたマウス複合形質の解析基盤の整備と遺伝的解析についての以下の成果を得た。（Takada et al., Genome Res.,2008, Oka et al., Genetics, 2007, Amano et al.,Dev. Cell, 2009, Oka et al., PLos Genetics, 2013） ２）マウス成体のマイクロ CT 画像の情報学的解析では、各種造影剤の組み合わせによる軟組織、各種臓器のイメージング技術の検討、および情報研との共同研究による表現型の情報学的自動抽出法の開発を進めている。３）マウスの社会行動を隠れマルコフモデルにより自動解析するフリーウェア（DuoMouse）を完成させ、国際誌に論文掲載された。DuoMouse は遺伝研のウェブサイトからフリーで公開し、その成果はプレスリリースされ、新聞報道された。４）上記 2013-2014 年以前の成果として、マウス行動解析の遺伝的基礎研究（Takahashi et al., Behav Genet.,2006, Mamm. Genome,2008, Behav. Genet.,2009, Umemori et al. BMC Genetics, 2009）の成果を発表した。５）イネにおいては、系統間ゲノム多型および発現差検定法 SNEP の開発（Fujisawa et al., BMC Bioinfo.,2009）や生殖隔離因子の相互作用解析（Mizuta et al., PNAS 2010）、野生イネ多数系統の系統進化関係と栽培イネ起源地を特定（Huang, Kurata, Wang, Wei,et al., Nature,2012）した成果を出し、これらの成果をさらに取り込むため、多数系統の野生イネを用いた GWAS（genome-wide association study）解析のための形質調査を行い、GWAS 解析の手法改良をサブチーム２と合同で進めた。６）また、イネ遺伝子発現制御ネットワーク解析のため、系統間の大量発現データの解析および e-QTL 用材料作成を行った。７）ゼブラフィッシュにおいては、細胞・組織・器官特異的に Gal4 を発現するトランスジェニックフィッシュを選別する大規模スクリーニングを行い、数百に及ぶ系統を作製した。それらトランスジェニックフィッシュを用いて、国内共同研究者と初期発生、器官形成、神経科学に関する共同研究を行い多くの成果を得た（成果多数のため 2013 年のみ記載：Umeda et al., Neurosci Res, 2013; Wada et al., PNAS, 2013; Kishimoto et al., Nature Neurosci, 2013; Asakawa et al., Frontiers in Neural Circuits, 2013; Muto et al., Frontiers in Neural Circuits, 2013; Banjoet al. Nature Comm, 2013; Wada et al. Curr Biol, 2013; Satou et al. Development 2013; Kwon et al. Development 2013; Nakayama et al. Mech Dev 2013; Sittaramane et al. Dev Biol, 2013）。８）CRISPR/Cas9 エンドヌクレアーゼを用いたゲノム編集技術をショウジョウバエに応用し、生殖細胞特異的発現プロモータを組み込んだ Cas9 遺伝子導入ハエ系統を用いて平均 50%という高効率でターゲット遺伝子の突然変異を誘発することができた。この手法により、ショウジョウバエのヌル変異体をハイスループットで体系的に取得する技術として発表した（Kondo & Ueda, Genetics, 2013）。

(6) 国内外における関連分野の学術研究の動向

次世代シーケンサの普及により、大規模ゲノムデータにもとづく情報処理、統計処理が個別的な研究テーマにも求められるようになり、大学等の研究室が有するゲノム解析技術レベルとの乖離が深刻化している。また、我が国では諸外国に比して特に遅れていたヒトゲノム研究材料の系統的収集を行うバイオバンクとゲノム情報解読とが一体化した東北メガバンクが稼働し、既に千人規模での日本人ゲノムデータが生産されている。シーケンス技術開発も活発に行われている。高速化、微量化、大規模化、簡便化、1 分子シーケンス等の様々なキーワードの機種開発が進んでいる。また、ナノポア技術を応用した第 4 世代シーケンサの市販が始まった。

遺伝子関連データ解析の観点からは、本プロジェクトで扱うゲノムデータ解析のための並べ替え検定手法の開発、多重性調整への応用、LASSO などの疎性・機械学習アプローチはその実用性が強く期待できる分野である。多重性に関する研究については、国際的に医学統計の観点での研究が多く見られるが、本プロジェクトが主にあつかう実験交配生物に関する多重性に関するものは多くない。

本研究で主に扱うマウス、イネ、ゼブラフィッシュ、ショウジョウバエなどのモデル生物においては、世界中で多様なゲノム情報の抽出や比較解析が行われている。しかし、本プロジェクトで扱う遺伝資源と実験系統は、世界の中でも独自のリソースであり、これらの遺伝的特性、特に表現型や行動パターン、複合形質、発現遺伝子変異、時系列変化など多様な特性の抽出、およびそれらの統合的情報解析を統計学、情報学、ゲノム解析を駆使して解明しようとする試みは、未だ非常に少ない。材料の優位はあるが、方法論の開発は世界中で時を追って進んできており、どのような系統やデータを用いるかの独自性を有効にするため、マッチした方法論との組み合わせの開発は緊急の問題となりつつある。動物行動研究において、社会行動や音声コミュニケーション、さらに多個体を同じケージに入れた状態での行動テストなどにより得られた時系列データの自動解析の重要性が増してきており、本プロジェクトで進めている研究は重要である。それらの行動形質に関わる多因子の解明に向けて、多数系統を交雑したヘテロジニアス集団を用いた遺伝解析法の確立は多くの研究者の注目を集めつつある。

〔2〕 研究計画

(1) 全体計画

全体としては、3 つのサブテーマ間で機動的、融合的にデータ生産、方法論開発、データ解析を繰返しつつ、「生命システム」としての「遺伝機能システム学」を創成し、抽出データ、解析方法論、多重ゲノムデータの体系的な表現型解析、遺伝機能システム解析の成果をコミュニティーに公開・発信して行く。

サブテーマ１

新型 DNA シーケンサの利用技術開発を進め、特にサブテーマ２、３、と連携しながら遺伝学研究所が有する遺伝資源に対して豊かなゲノム情報を付加し、研究資源として高度化する。また、解析対象の微少化等のウエットサイドでの技術開発が、最終的には単一細胞レベルでの解析が可能になるまで進むことが期待できるため、例えば発生過程における精細な遺伝子発現プロファイル時系列データなど、従来研究では実現不可能であった、生物学者の『夢』ともいえる細胞レベルでの定量解析の実現をめざす。このためには、プロジェクト内研究者との連携に加え、「地球生命システムプロジェクト」や「データ同化プロジェクト」との連携、さらには共同利用研究機関であることの特長を生かして国内外の研究コミュニティーと連携しながら融合の視点で研究を遂行する。一方、本プロジェクトで導入した DNA シーケンサは初期型であり、今後のアップグレードには対応できないため、第 4 世代装置の導入も含めて設備更新の検討が必要である。

サブテーマ２

3 つの課題

(a) ゲノムデータ解析のための並べ替え検定手法の開発と多重性調整への応用

(b) LASSO などの疎性、機械学習アプローチの利用

(c) 生物・遺伝データ特有の特徴をとらえる統計データ解析

を軸として、遺伝研メンバーが取得したデータから遺伝的知見を引き出すためのデータ解析を行うことを通して、遺伝学上の発見につながる貢献をするとともに、新たなデータ解析のための方法論を開発する。

サブテーマ３

多様なゲノム情報、表現型情報をもつモデル生物、マウス、イネ、ゼブラフィッシュ、ショウジョウバエを用いて、遺伝機能システム学の基盤を作る。野生系統の多様性、ユニークな変異体集団など独自の系統群の持つ遺伝的変異のパワーを多面的に引出し、サブテーマ１、２と共同して多面的、多重的データの相関解析法の開発と解析を行い、グラフィカルモデル理論や情報の階層的組立てにより、遺伝機

能システム学を展開する。各生物種で取り扱う具体的内容は、以下の通りである。

(a) マウスについては、野生系統を含む多系統交雑によるヘテロジニアス集団や 2 系統間染色体置換系統群等を用いた、複雑形質（生体内構造、骨格形態、行動パターンなど）の定量化、データ抽出と統計解析手法の開発。ゲノムデータの効率的で信頼性の高い解析手法の確立。ゲノム多様性データとの相関解析。

(b) イネについては、野生イネを中心に、ゲノム構造、発現遺伝子の質・量・変異解析。組み換え自殖系統の遺伝型評価、系統毎の表現型抽出、e-QTL 解析。野生イネ集団を用いた集団遺伝構造解析、association 解析。

(c) ゼブラフィッシュについては、トランスポゾンを用いた遺伝学的方法論に基づく大規模遺伝子改変系統の作製とそれらからの多様な表現型の抽出および表出法の確立を行なう。

(d) ショウジョウバエについては、RNAi knock-down 変異系統群を用いた、翅形態の変異の抽出および変異遺伝子群との相関解析を行い、また CRISPR/Cas9 を用いた変異体の作成と遺伝子変異-表現型相関解析の基盤を作る。

(2) 各年度の計画

平成 25 年度

3 つのサブテーマ間で機動的、融合的にデータ生産、方法論開発、データ解析を行い、相互にフィードバックしつつ「生命システム」解析に迫る。データ抽出法、解析方法論の開発を通じて、多重ゲノムデータと表現型の体系的な相互関連解析、様々な系統特性とゲノム関連情報を結びつける遺伝機能システム学として成果を挙げる事を計画している。リソース毎に用いる特性や相関解析の目標および進捗度合いが異なるため、生物種ごとの計画および成果を参照の事。すでに 25 年度については、生物種ごとの計画に沿って進んでおり、成果も得られた。

サブテーマ 1 では、第二世代シーケンサを用いたゲノム配列データの生産と解析研究を継続する。器官形成、個体形成、環境適応など基本的な生命現象に関わる遺伝子発現調節機構について体系的な時系列遺伝子発現プロファイル作成をめざし、試料調製手法や、個別の生命現象に特異的な大規模データを取り扱うための情報解析手法の開発を行う。

サブテーマ 2 では、多重性調整に関しては、提案した疑似相関を利用して検出力を上げる新しい多重性調整法について、その実用性を吟味する。分散の同等性検定に対して外れ値に強い手法を考える。また主成分スコアに基づいたスパース回帰モデルについて検討を始める。グラフィカルモデルについては、昨年度明らかになった問題点を再検討するとともに、グラフィカルモデルの解析が優位性を持つようなデータの取得を、サブテーマ 3 の協力の下に行う。またグラフィカルモデルの新たな可能性として、3 すくみを記述できるようなモデルの開発をはじめる。LASSO を用いた遺伝子探索（QTL 解析）については、引き続きその妥当性を吟味する。ベルンシュタインコピュラを用いたデータ解析については引き続きデータ解析における実用性を吟味する。

サブテーマ 3 では、複雑形質（生体内構造、骨格形態、行動パターンなど）の定量化、データ抽出と統計解析手法の開発（マウス）、大規模遺伝子改変系統の多様な表現型の抽出法の確立を行なう（ゼブラフィッシュ）。また、形質とゲノム構造多型、発現多型間の相関解析を GWAS を中心に進める（イネ）。ゲノム上の遺伝因子群の特定とネットワークとしての因子群の相互連関を視覚化し記述する手法を開発する。

平成 26 年度

サブテーマ 1 では、前年度の計画を進めるとともに、時期的に市販化が予想される第 4 世代 DNA シ

ーケンサ及び新第二世代シーケンサの導入を視野に入れ、それらに対応した情報解析手法についての検討を行う。

サブテーマ２では、多重性調整に関しては、H25 に開発した疑似相関を用いた多重検定手法についての論文を完成させるとともに、さらなる応用を模索する。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に関しては、構築した方法がうまく働くかどうかを様々な数値実験と実データ解析を通して検証する。LASSO を用いた QTL 解析については、より多くの交互作用を含めた回帰モデルを当てはめ、その結果を遺伝学的な観点から議論する。ランキング解析については、H25 に行った 3 すくみ交互作用に基づく階層モデルの方法を理論化する。さらに、評価項目が複数ある場合に結果を統合する方法を開発する。

サブテーマ３では、野生系統を含む多系統交雑によるヘテロジニアス集団や 2 系統間染色体置換系統群等を用いた、複雑形質（生体内構造、骨格形態、行動パターンなど）の定量化、データ抽出と統計解析手法の開発（マウス）、一連の系統の e-QTL 情報の取得、ゲノム上の遺伝因子群の特定とネットワークとしての因子群の相互連関解析手法を構築する（イネ、マウス）。変異体の表現形質変動については、ゲノム機能との連関解析を行い、ゲノム機能ネットワークの抽出を試みる（ゼブラフィッシュ、ショウジョウバエ）。

平成 27 年度

サブテーマ１では、前年度の計画を進める。特に、解析対象の微量化技術の実現が期待できるため、発生過程における精細な遺伝子発現プロファイル時系列データ解析等の当初計画の実現を図る。

サブテーマ２では、多重性調整に関しては、Graphical LASSO の情報量規準を与え、遺伝子間の関連を調べる解析に適用する。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に関しては、線形回帰モデルで作った方法を他のモデルにも拡張する。LASSO を用いた QTL 解析については、時刻を考慮した活動量の時系列モデルを確立し、遺伝学的な考察を行う。また全課題を通して、H22～26 で開発した多重性調整法を一般化・汎用化させた後にパッケージ化する。

サブテーマ３では、ゲノム上の遺伝因子群の特定とネットワークとしての因子群の相互連関を視覚化し記述する手法を確立する（マウス、イネ）。交雑固定化系統および突然変異系統における画像データや時系列による多元的な表現形質変動についても、ゲノム機能との連関解析を行い、ゲノム機能ネットワーク全体の描出を試みる（マウス、イネ、ゼブラフィッシュ、ショウジョウバエ）。

全プロジェクトを通して得られた研究成果をとりまとめ､論文発表と Web を通した公開を行う｡特に､次世代シーケンサで生産したゲノム関連情報については適宜アノテーションをつけてデータベース化して web 公開する。新規に開発したアルゴリズム、ソフトウェアについても Web 発信する。最後に研究成果を公開するための国際シンポジウムを開催する。

平成 28 年度以降

未定

〔3〕 研究推進・実施体制

サブテーマ１では、国立遺伝学研究所シーケンスセンターが直接的なデータ生産に関わるが、従来のゲノム解読では対象になっていなかったエピゲノム修飾、機能的ヘテロクロマチン領域の研究グループや、メタゲノム、個人ゲノム、単一細胞ゲノム等の国立遺伝学研究所の先端的ゲノム研究グループが参加する。また、基礎生物学研究所、理化学研究所、霊長類研究所からゲノム研究グループが参加する予定である。情報解析手法の研究開発は、国立情報学研究所、国立遺伝学研究所、東京工業大学、京都大

学、慶應義塾大学から、ゲノム情報解読やシステム生物学の研究グループが参加する。また、新領域研究「地球生命システムプロジェクト」からの試料解析も実施する。さらにサブテーマ２および３関連のデータ取得では、多様な野生および実験系統からのゲノムおよび発現遺伝子解析等も行うため、これらのデータを各生物のコミュニティーで共有したり、比較研究解析の基盤として利用する事も含め、国内外の研究グループと協力、連携する体制を促進する。すでにマウス、イネなどにおける国際連携体制は整っており、具体的データにより今後の展開を図る。そのためには、シークエンサーのフル稼働とデータ解析部隊の充実は必須である。

サブテーマ２および３の実施に当たっては、データ取得のために遺伝研に研究スタッフを配置する。多次元の遺伝情報、発現変異情報、表現型情報、時系列情報の大量データ取得と解析が鍵になる。しかし、本プロジェクトの目指す遺伝的多様性を軸に据えた同様な研究は未だ国内外で部分的にしか取り組みはなく、今後国内外の状況を見ながら順次連携やコミュニティー形成を推進して行く。統計解析の体制としては、統数研、東京大、九州大などから、統計推測（ロバスト、影響分析、グラフィカルモデル、多重性調整など）と知識を有し、かつデータ解析についても経験と興味を有する研究スタッフをおく。全メンバーが協力してデータ解析の方法論の開発にあたる。

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

・研究代表者

〔国立情報学研究所〕　藤山秋佐夫

・共同研究者

〔国立遺伝学研究所〕　豊田　敦、野口英樹、角谷徹仁、樽谷芳明、深川竜郎、堀哲也、中村保一、神沼英里

〔新領域融合研究センター〕　丸山多恵子、松﨑肖子、辰本将司、程　朝陽、商　維昊、望月孝子

〔東京工業大学〕　黒川　顕

〔九州工業大学〕　矢田哲士

〔慶應義塾大学〕　榊原康文

〔基礎生物学研究所〕　長谷部光泰

〔東北大学〕　黒木陽子

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

・研究代表者

〔統計数理研究所〕　栗木　哲

・共同研究者

〔統計数理研究所〕　藤澤洋徳、間野修平、加藤昇吾

〔国立遺伝学研究所〕　城石俊彦、倉田のり、高田豊行、小出　剛、高橋阿貴

〔新領域融合研究センター〕　Dou Xiaoling、木曽（岡）彩子

〔新潟大学〕　原　尚幸

〔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

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

・研究代表者

〔国立遺伝学研究所〕　倉田のり

・共同研究者

〔国立遺伝学研究所〕　久保貴彦、城石俊彦、高田豊行、川上浩一、武藤　彩、上田　龍、小出　剛、高橋阿貴

〔国立情報学研究所〕　北本朝展、宇野毅明

〔統計数理研究所〕　栗木　哲、藤澤洋徳、加藤昇吾、小山慎介

〔新領域融合研究センター〕　堀内陽子、春島嘉章、木曽(岡)彩子、和田浩則、後藤達彦、Dou Xiaoling、近藤伸二

〔政策研究大学院大学〕　土谷　隆

〔新潟大学〕　中谷明弘、阿部貴志、原　尚幸

〔高知大学〕　清澤秀孔

〔九州大学〕　二宮嘉行

〔京都工芸繊維大学〕　高野敏行

〔首都大学東京〕　相垣敏郎

〔愛知工科大学〕　荒川俊也

〔大阪府立大学〕　川野秀一

〔理化学研究所〕　田村　勝、若菜茂晴

〔東京大学〕　岩田洋佳

〔4〕 研究の進捗状況

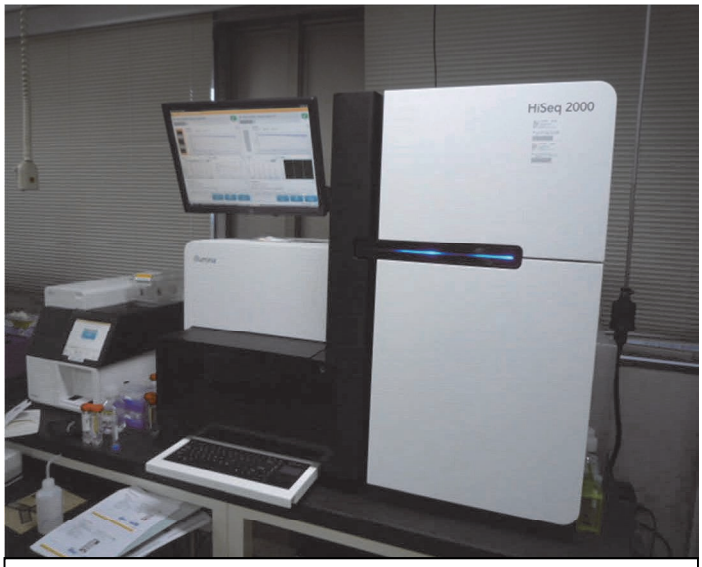
全体としては、3 つのサブテーマ間で機動的、融合的にデータ生産、方法論開発、データ解析が進行しており、体系的な表現型–遺伝型相関解析の基盤が整い、相関解析の成果が生まれつつある。最終年度にむけて「遺伝機能システム学」の新たな展開を図っている。

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」

1−1：新型シークエンサーの稼働状況と大規模データ生産

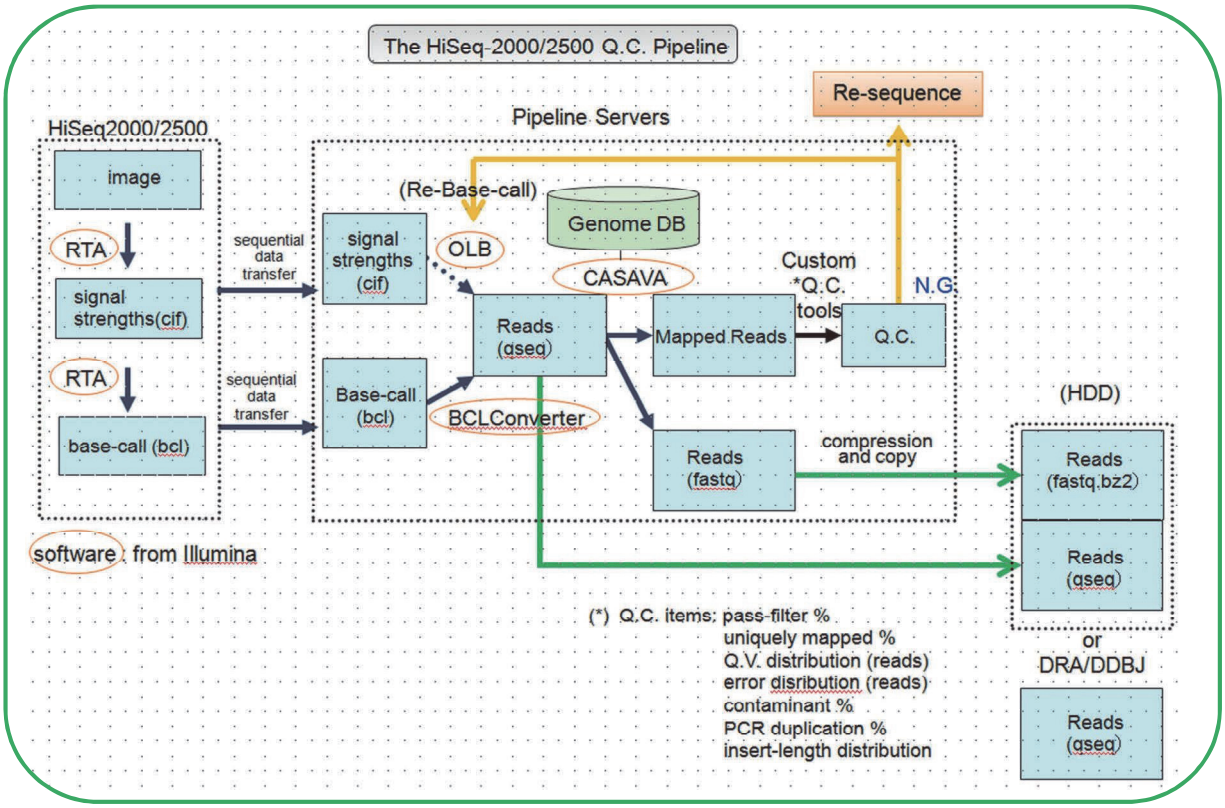
22 年度に、最新の次世代型シーケンサとして Illumina 社製 HiSeq2000 型及び試料調製装置 c-BOT を導入した。同装置は、アジア地域で中国 BGI に次ぐ国内第１号機であり、当初計画よりも数ヶ月遅れて稼働を開始した。本装置の導入当初の性能は、読み取り鎖長が 100 塩基、1 回の運転あたりのデータ生産量が約 200Gb であったが、その後の改良により、読み取り精度の大幅な向上と共に、1 回運転あたりのデータ生産量も 600Gb に増大されている。24 年度に HiSeq2500 タイプにアップグレードし、読



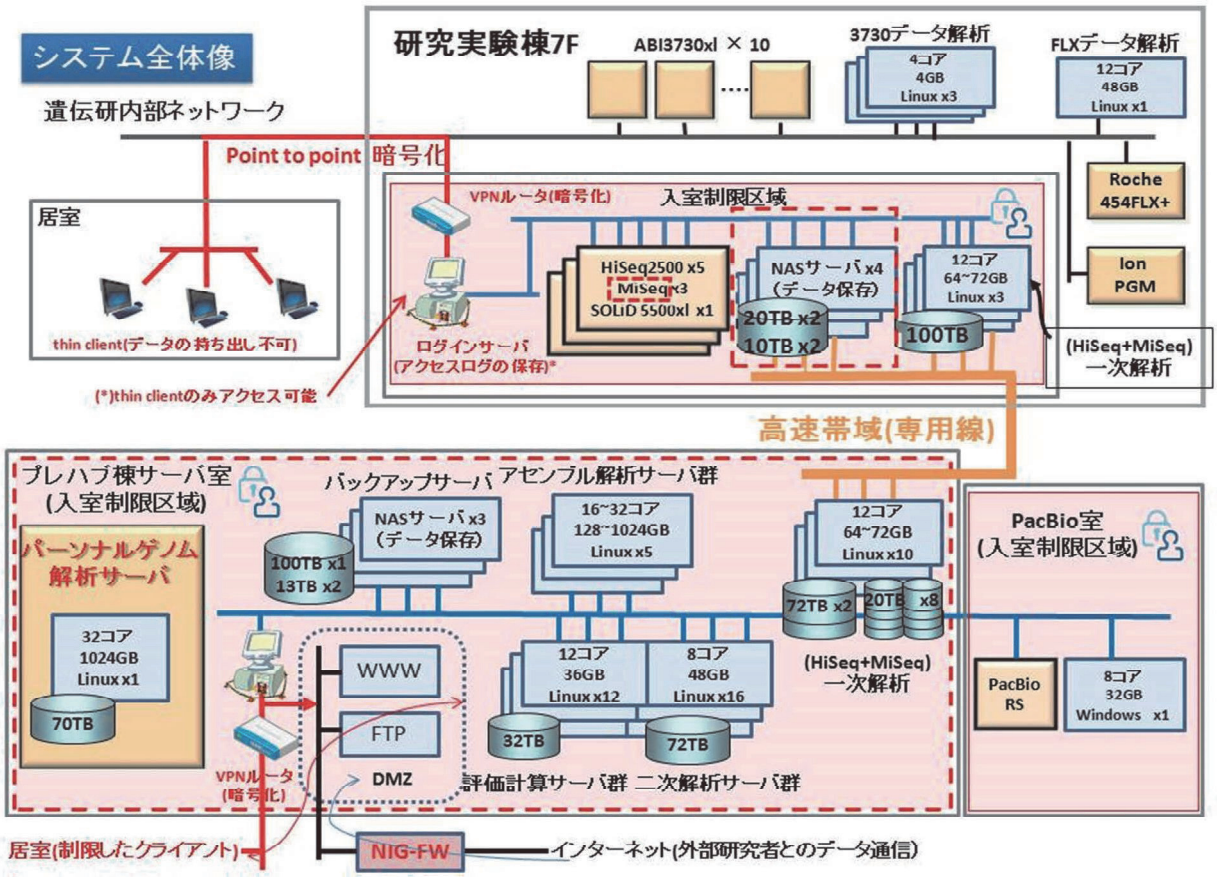
22 年度に導入した試料調製装置 c−BOT(左)と、HiSEQ2500 次世代型シーケンサ(右)。

み取り長 150bpx2、180Gb のデータを 40 時間で産出するモードでの運転が可能となり、時間短縮、マッピングとアセンブリの効率と精度の改善が実現した。しかし、本プロジェクトで導入した DNA シーケンサは初期型であり、今後のアップグレードには対応できないため、第 4 世代装置の導入も含めて設備更新の検討が必要である。シーケンサの運用は、国立遺伝学研究所先端ゲノミクス推進センターのシーケンシング施設で行っており、世界的にも高精度のデータ生産が安定して行われるようになっている。遺伝学研究所先端ゲノミクス推進センターが運用する DNA シーケンス施設では、本プロジェクトで導入したマシン以外に、HiSeq シーケンサ 5 台と MiSeq シーケンサ 3 台、PacBio RSII 1 台、ABI3730 シーケンサ 6 台が主に使われている。

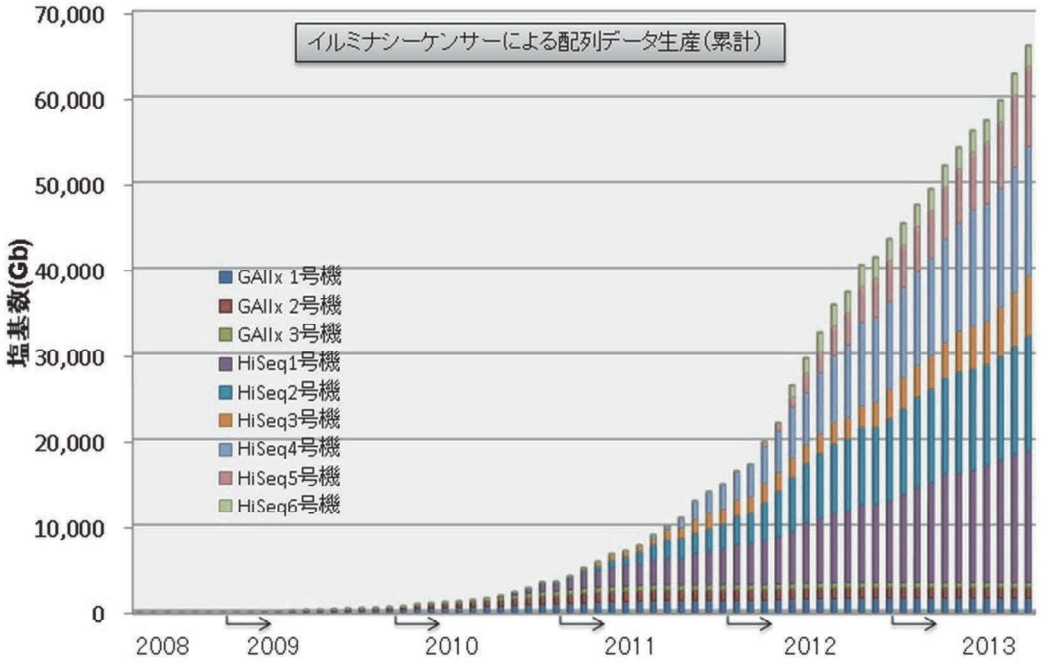
これらのシークエンサーを駆使し、超大規模ゲノム関連情報の生産/解析基盤システムの構築を進めている。現在のデータ生産速度は 180Gb/40hrs、リード長 150bp に達しており、遺伝研スーパーコンピュータを効率よく利用するパイプライン構成と併せて国内トップクラスのゲノム解読能力を実現している。このシステムを活用し、当プロジェクト及び地球環境システム由来の試料について、累計で約 112Tb の塩基配列データを生産してきた。



HiSeq2500 シーケンサのデータ評価/一次解析パイプライン



新型シーケンサによるデータ生産実績



年度	プロジェクト	レーン数	リード数	塩基数 (bp)
2011	融合研究	145	20,463,232,493	2,046,323,249,300
2012	融合研究	194	58,460,266,022	5,888,495,617,400
2013	融合研究	96	29,579,295,428	3,756,579,030,900
	合計	435	108,502,793,943	11,691,397,897,600

＊ 2013年度は11/19までの解析終了分のみ