Ⅳ　戦略課題 4：大規模生命データ解析
（統括：宮野悟・東京大学医科学研究所）

特定高速電子計算機施設を中核とする HPCI に最適化した最先端・大規模シークエンスデータ解析基盤を整備した上で、生命プログラムの複雑性・多様性や進化をゲノムによって理解する研究と同時に、ゲノムを基軸とした生体分子ネットワーク解析研究を行う。それにより、薬効・副作用予測、毒性の原因の推定、オーダーメイド投薬、予後予測などへの応用に貢献することを目指す。

Ⅳ－１　宮野　悟（東京大学）
大規模データ解析によるがんのシステム異常の網羅的解析とその応用

Ⅳ－１－１　実施計画

がんは、親から受け継いだ遺伝的要因（ゲノム）、腫瘍細胞に蓄積した遺伝子変異（がんゲノム）、環境要因によるゲノムの修飾（エピゲノム）、これらの違いや異常が、正常な細胞の営みを司っている遺伝子ネットワークやシグナル伝達・代謝などのパスウェイに入り込み、システム異常を起こした時空間で進化するヘテロな細胞集団である。そして、血管内皮細胞や免疫炎症細胞などの正常細胞を操り、抗がん剤に対して耐性を獲得していく。ゲノム変異が大きく異なっている複数の原発が進化することも報告されている。こうした複雑さを背景にして、がんは抗がん剤などに対する薬剤感受性や予後の良・不良等、様々な個性を持つ。そして、そのシステム異常の中心で遺伝子の発現を調整しているメカニズムが遺伝子ネットワークであり、がんの個性の一つの捉え方である。

本委託研究は、戦略プログラムの「課題 4　大規模生命データ解析」研究の一環として、多数のがんサンプルデータを用いて大規模・網羅的に遺伝子ネットワークを中心に解析し、そのような多様な個性を生み出すがんのシステム異常の実態をシステムとして暴きだすことを目的とする。これにより、がんの生命プログラム及びその多様性の理解を深め、薬効・副作用予測、毒性の原因の推定、オーダーメイド投薬、予後予測などへの応用に貢献することを目指す。国際がんゲノムコンソーシアムは、50 種のがんについて全部で 25,000 人のサンプルをシークエンスし、がんのゲノム異常のカタログをやがて完成させるが、これだけでがんの多様な個性をシステムとして細やかに捉えることには限界があり、個々人のがんの病態に密接に関係する遺伝子ネットワークを網羅的に解析することが、がんの複雑さと個性を理解するためのチャレンジとなっている。しかし、ゲノムワイドな遺伝子ネットワーク解析だけでも大きな計算コストを要し、その網羅的解析は京コンピュータの稼働以前では夢物語であった。そこで、平成 25 年度は、グランドチャレンジプログラムで開発したアプリケーションを利用・発展させ、京コンピュータの資源を十分に用いることで、以下の研究を実施する。

（１）　700 以上のがん細胞株の遺伝子発現プロファイルデータと 100 以上の薬剤に対するがん細胞株の薬剤感受性（IC50 スコア）データを、複数のゲノムワイドな遺伝子ネットワーク推定法で解析し、薬剤感受性とがん細胞株との関係をシステムの違いとして調べる。

（２）数万の臨床検体の遺伝子発現プロファイルデータ及び関連するオミクスデータと臨床データを用いた大規模・網羅的遺伝子ネットワーク解析を行う。

（３）以上（１）と（２）の大規模データ解析によって有用な知見が得られることが期待されるが、業務協力者との連携により知見またはその発展形の検証を様々な観点から実施する。その検証のために必要となるデータを外注業務により取得する。

（４）以上の研究を遂行する中で必要となる新たな大規模生命データ解析の方式の研究を併せて実施する。

「戦略課題 4：大規模生命データ解析」研究統括では、以下の２つの大学で実施される平成２５年度の研究課題の実施項目について、適宜、関連する研究者とワークショップや研究打合せを行い、また業務協力者に対してはそれぞれの専門の立場から知見とアドバイスを仰ぎ、関係者のとりまとめを行うとともに、理化学研究所と連携して、研究開発の統括を行う。

① 大規模生体分子ネットワーク解析による脂肪細胞組織の刺激応答の網羅的解析とその応用（松田秀雄・大阪大学）

② 次世代シークエンサデータ解析のための情報処理システムの開発(秋山泰・東京工業大学)

Ⅳ－１－２　実施内容（成果）

（１）英国サンガーセンターが公開している Sanger Genomics of Drug Sensitivity of Cancer の遺伝子発現データを解析し、抗がん剤に対するがんの多様な抵抗性をネットワークとして抽出した。このデータセットでは、13,435 遺伝子の RNA 発現プロフィールが 728 のがん細胞株において計測されており、142 種類の抗がん剤や医薬品候補化合物について、これらの細胞株における IC50 の値も合わせて計測されている（図１）。IC50 (half maximal (50%) inhibitory concentration) とは、その薬剤（もしくは候補化合物）が標的としている生物学的プロセスの半数を阻害するのに必要な濃度を表しており、値が小さい細胞ほどその化合物の効果が高いと言え

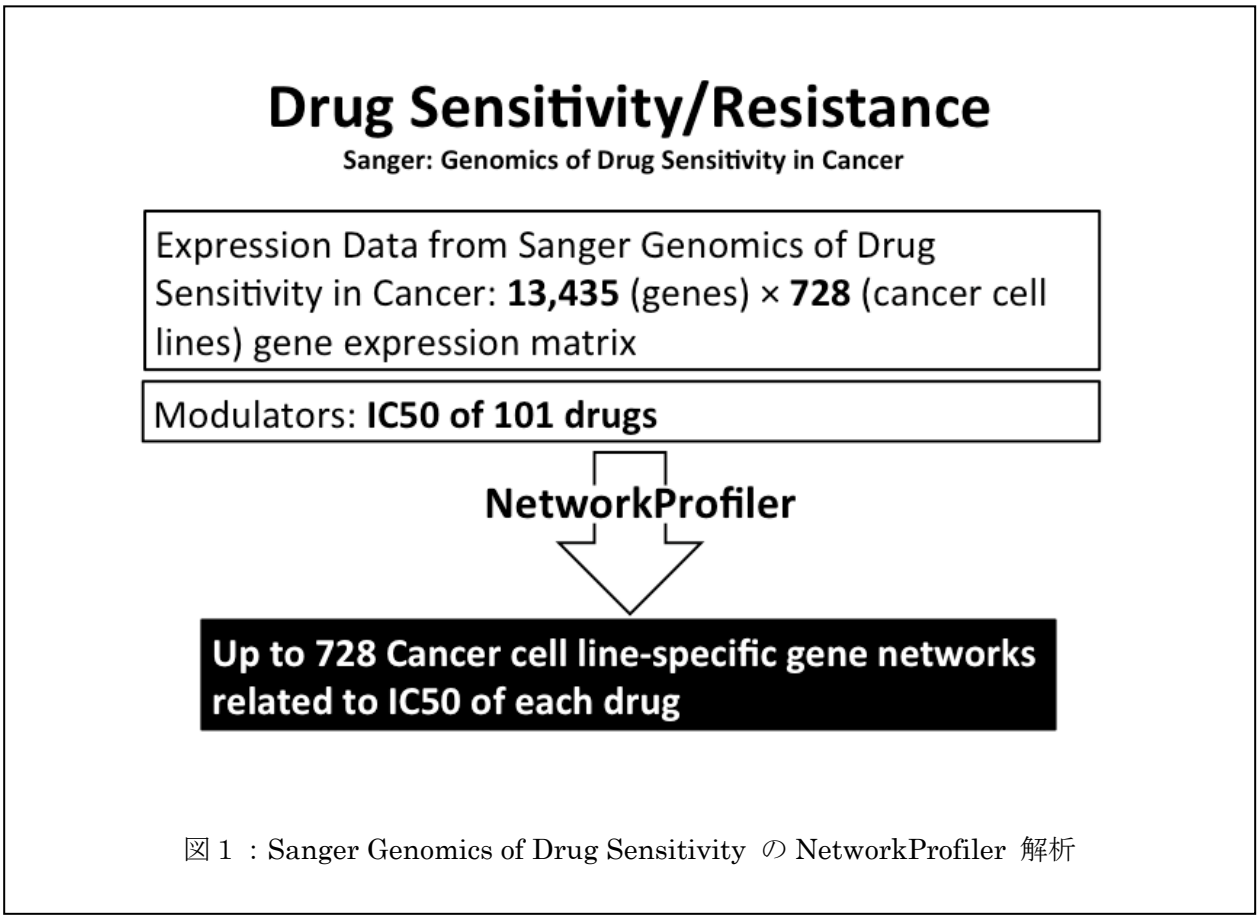


図１：Sanger Genomics of Drug Sensitivity の NetworkProfiler 解析

る。そこで、NetworkProfiler において、それぞれのがん細胞株の IC50 の値を Modulator として使用し、それぞれの薬剤に対する抵抗性を決定している候補サブネットワークの抽出を行った。

142 種類の化合物について 728 細胞株全てで IC50 が計測されているわけではない。ここでは、600 以上のがん細胞株において IC50 の値が計測されている化合物 101 種を解析対象とした（平均サンプル数 650）。従って、Modulator は 101 種類あることになる。これら 101 個の化合物について 1 個あたり平均 650 個のネットワークを推定する必要がある。この計算には K Computer を利用した。8000 コアを利用し約 3 日間の計算により約 7 万個のネットワークの計算に成功した。

一つの化合物について約 650 個のネットワークが推定される。このネットワーク情報を解析し、化合物の効果を規定しているサブネットワークの抽出を試みた。今、NetworkProfiler によって推定された、ある制御因子（転写因子）とそれが制御する遺伝子のペアに着目する。一つのネットワークでは、そのペアについて制御関係が推定されたかどうかを隣接行列によって表現することにより、推定されたネットワークの情報は 2 次元マトリックスとして表現される。しかしながら、一つの化合物について約 650 個のネットワークが推定され、しかも、それらのネットワークは IC50 という順序を有している。この情報を同様にまとめると、それは 3 次元マトリックス、いわゆるテンソルとして表現される。このテンソルにより表現される情報を要約し情報を抽出するため次の解析を行った。

推定された 1 つのネットワークにおいて、この制御因子から被制御遺伝子への影響度を測る統計量 regulatory effect を推定された構造方程式モデルの係数と制御因子の発現量の積により定義した。ある制御因子と被制御遺伝子のペアについて regulatory effect は各ネットワークにおいて計算されるため、そのプロフィールを示したのが図 2 右下の折れ線グラフである。この最大値と最小値の差を regulatory effect change として定義した。すなわち、regulatory effect

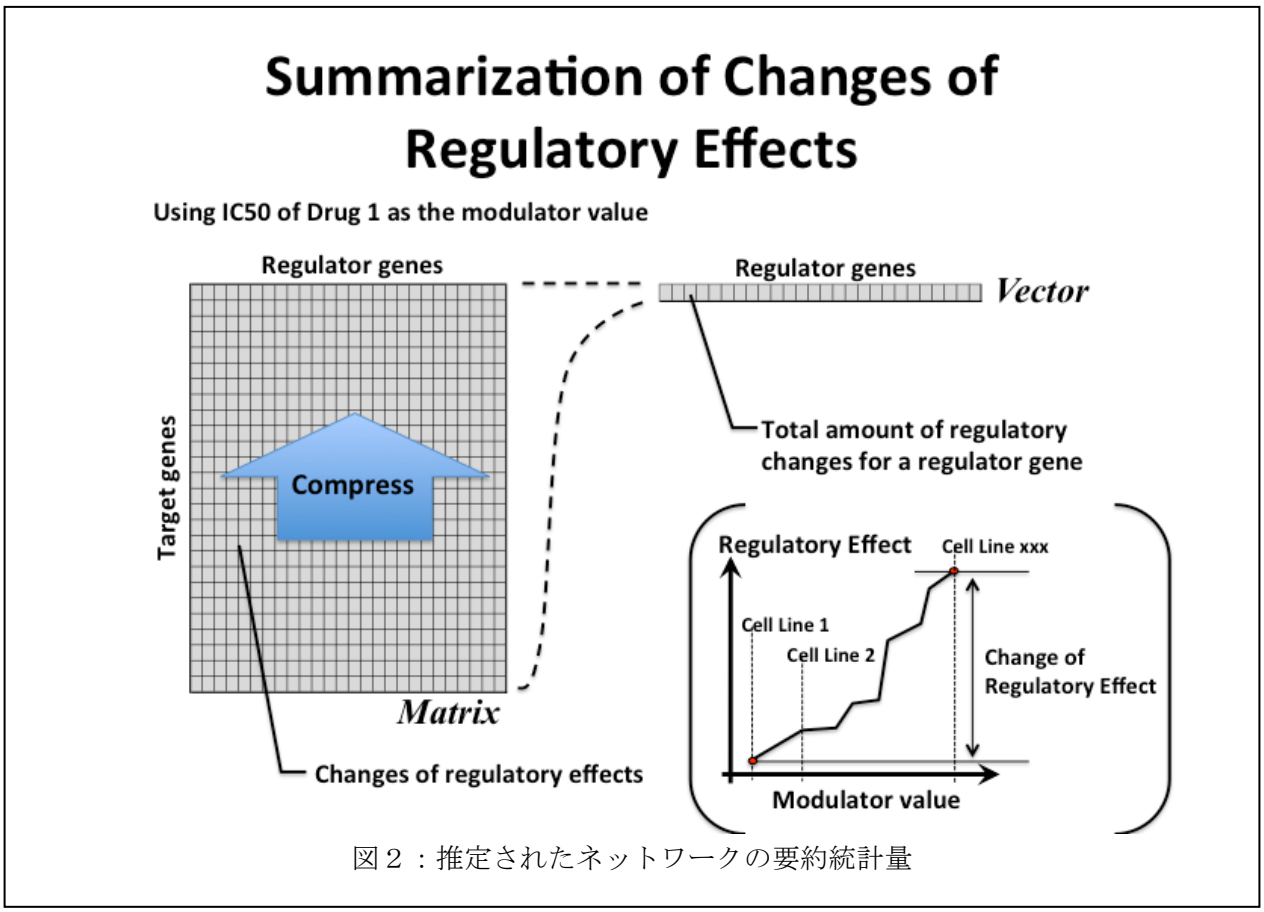


図 2 ：推定されたネットワークの要約統計量

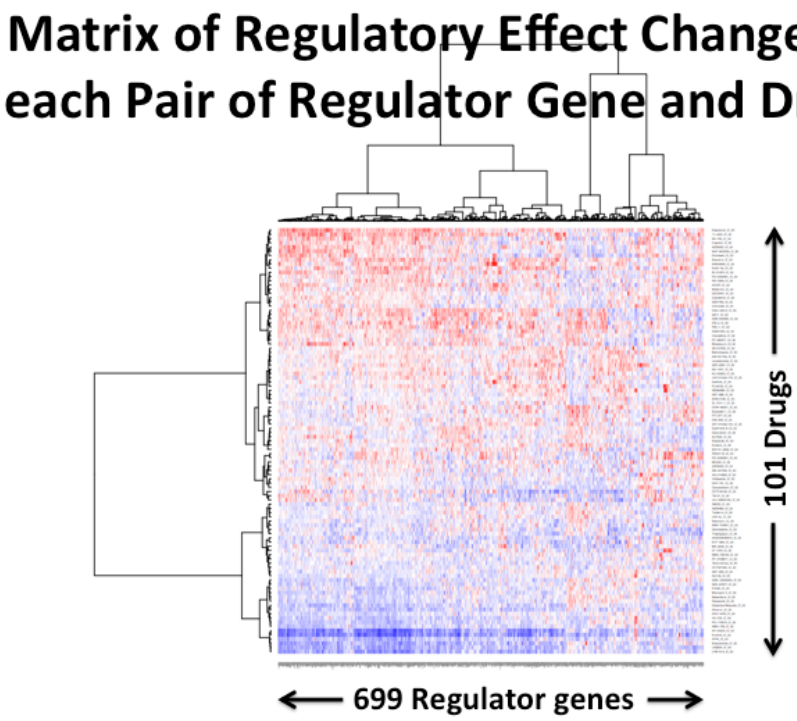


図３：101 化合物に対する regulatory effect change マトリックス

change が大きいペアは、薬剤耐性感受性の違いと制御の活性の違いが相関することになり、このような制御関係の違いが薬剤の耐性感受性を規定している可能性がある。図２左の行列は、列が転写因子、行が被制御遺伝子を表し、その要素は各転写因子–被制御遺伝子のペアに対する regulatory effect change の値である。この regulatory effect change の行列が 101 薬剤について得られることになる。これら 101 個の行列の情報を統合するために、一つの行列において列方向に要素の絶対値の和を取ったベクトルに変換する。すなわち、一つの行列は一つの転写因子数次元のベクトルとなる。この 101 個のベクトルを再度まとめて一つの行列とした。

その行列を階層型クラスタリングにかけヒートマップにより可視化したのが図３である。行が転写因子、列が薬剤を表している。赤は値が大きく、青が小さいことを表す。各要素は転写因子から被制御遺伝子群への regulatory effect change の総和（絶対値）なので、赤ければ赤いほどその転写因子からの制御は薬剤の耐性感受性に強く相関している事になる。逆に、青はその転写因子からの制御は薬剤の耐性感受性にはその強さは相関していない事になる。左に位置する転写因子は、上に配置される薬剤に対しては、その制御の強さが薬剤耐性感受性と相関するが、下の方に配置された薬剤の耐性感受性には、どの転写因子の制御もあまり相関していないことが分かる。この行列において、薬剤をサンプル、転写因子を変数と捉え主成分分析によって２次元に射影したものが図４である。横軸が第一主成分、縦軸が第二主成分である。第一主成分は平均に相当し、多くの薬剤について regulatory effect change が大きい転写因子が大きな値を取る。ここでは、第一主成分の値の大きい３つの薬剤、Elesclomol、17-AAG、BIBW2992 に着目した。

Elesclomol は既に承認されている薬剤である。また、17-AAG は未承認の薬剤候補化合物であり、この二つはともに heat shock protein を標的としている。まず、Elesclomol に対する耐性感受性に相関して regulatory effect change が大きい転写因子の上位 10%（耐性で活性が高い 5%と感受性で活性が高い 5%）を抽出した。

図５は Elesclomol に対する解析結果を示している。横軸はがん細胞株を Elesclomol に対する IC50 の順に並べたもの、上のパネルの縦軸は regulatory effect を表す。例えば AIRE は、

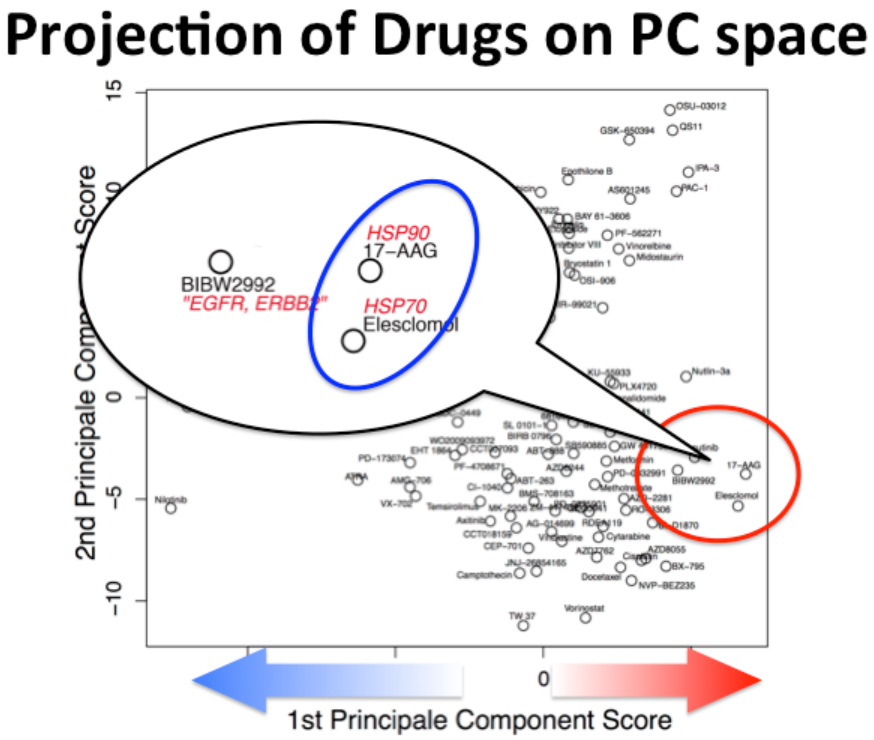


図４：Regulatory effect change マトリックスを主成分分析により２次元空間に射影

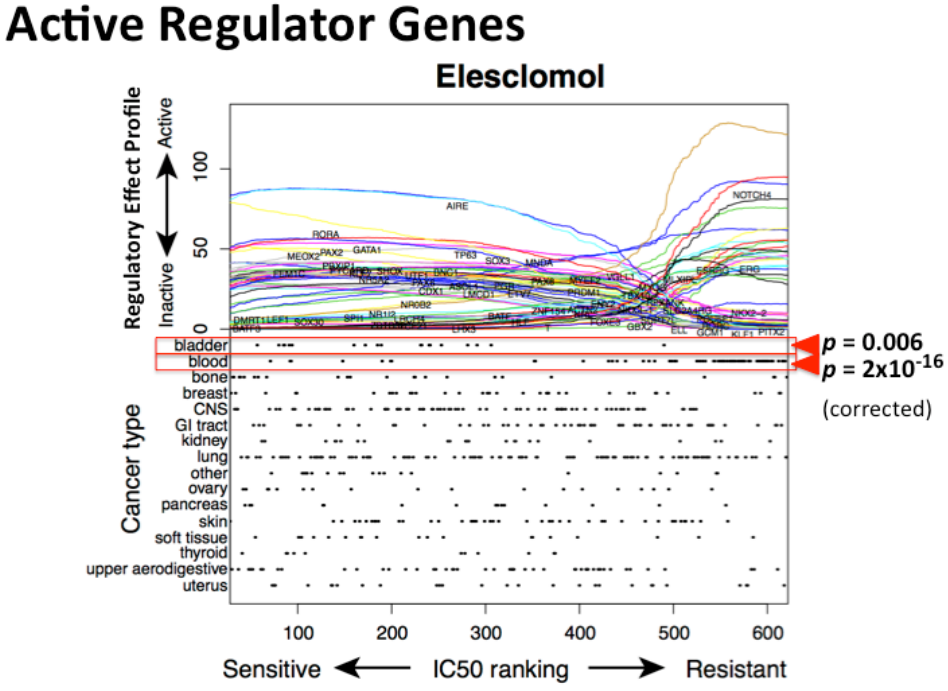


図５：Elesclomol に対する各転写因子の regulatory effect profile

Elesclomol に対して感受性の高いがん細胞株では強く被制御遺伝子をコントロールしているが、耐性がん細胞株においては影響度が小さいということがわかる。下のパネルはがん種による偏りを表している。bladder は若干感受性に偏り、blood は耐性に偏っているが他のがん種では有意な偏りは見られない。すなわち、がん種を超えた Elesclomol 耐性感受性の機構が表れているこ

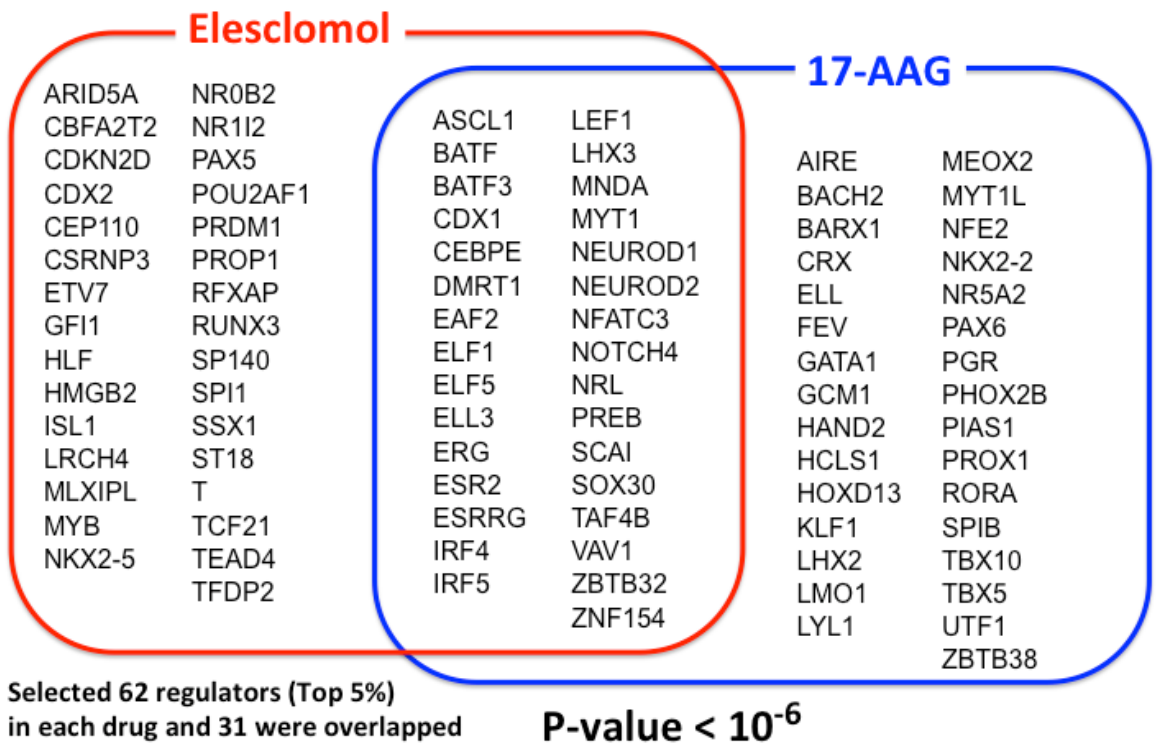


図 6 : Elesclomol と 17-AAG の耐性がん細胞株において活性化している転写因子候補は統計的に有意にオーバーラップする

とが期待される。

Elesclomol と 17-AAG は共に heat shock protein を標的とした薬剤、薬剤候補化合物であった。それらに対して耐性細胞株で活性が高い転写因子上位 5%を抽出しそれらのオーバーラップを示したのが図 6 のベン図である。統計的に有意にオーバーラップしており、このオーバーラップは heat shock protein を標的とした際にその耐性を決定している転写因子群の候補と考えられる。

次に、BIBW2992 について結果を紹介する。この薬剤候補化合物は EGFR、ERBB2 を標的としており、その結果を図 7 に示している。BIBW2992 に対して感受性の高いがん細胞株において活性が高い上位 5%の転写因子に着目し解析を進めた。

NetworkProfiler によって遺伝子ネットワークを推定した際に制御因子としては転写因子に絞っていたため、BIBW2992 の標的遺伝子である EGFR や ERBB2 は制御因子に含まれていない。そこで、これら上位 5%の転写因子のさらに上流にどのようなシグナル伝達経路があるか探索した結果を図 8 にあげる。2 つの遺伝子がこれら転写因子の共通の上流因子として有意であるという結果を得た。その 2 つの遺伝子が HDAC と ERBB2 である。ERBB2 は BIBW2992 の標的遺伝子そのものであり、その下流にある NetworkProfiler によって予測された転写因子群の活性が BIBW2992 に対する感受性を規定している可能性がある。また、HDAC については HDAC inhibitor が 101 薬剤に含まれるため、その薬剤との関連をこの結果から探索していくことで更に情報を抽出できる可能性がある。

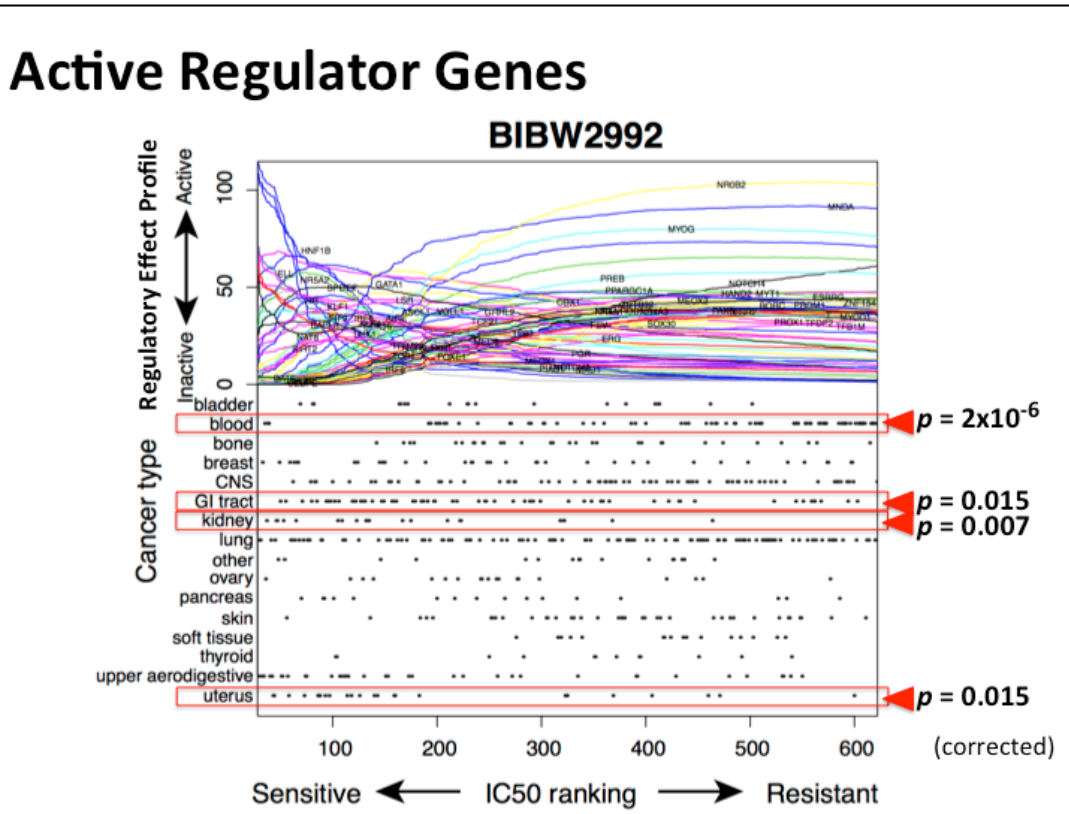


図７：BIBW2992 の耐性感受性を決定している転写因子群の解析

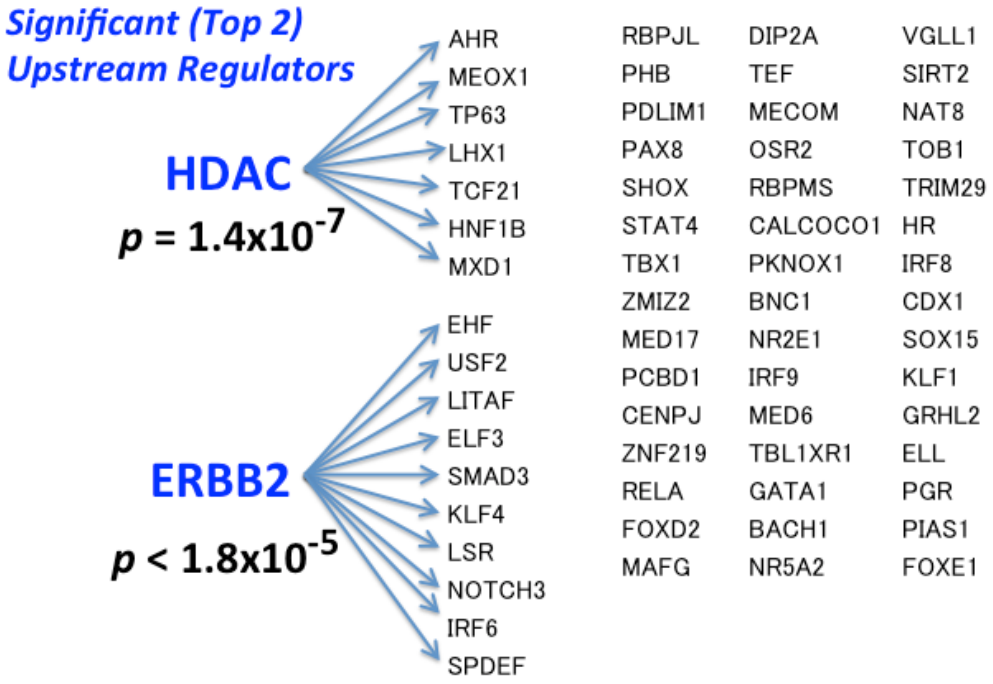


図８：BIBW2992 の薬剤耐性感受性ネットワークをコントロールするシグナル伝達経路の解析

（２）数万の臨床検体の遺伝子発現プロファイルデータ及び関連するオミクスデータと臨床データを用いた大規模・網羅的遺伝子ネットワーク解析に関する成果

１）大規模・網羅的遺伝子ネットワークデータベースの準備

平成 24 年度までに、がん関連公開データ 256 データセット 30,261 サンプルに対してベイジアンネットワークを用いた２種類の遺伝子ネットワーク推定手法 SiGN-BN HC+Bootstrap および SiGN-BN NNSR を適用し 512 個の遺伝子ネットワークの推定を、「京」を用いて行っている。推定した遺伝子ネットワークはデータベース化し、さらにウェブ検索システムを構築し誰もが検索し推定結果を利用できるウェブデータベースを The Cancer Network Galaxy(TCNG)として公開している（http://tcng.hgc.jp)。平成 25 年度は、このデータベースの拡張を目的として新たに EGF 関連 1520 遺伝子の遺伝子セットと前述の 30,261 サンプルを利用して EGF 関連遺伝子ネットワークの構築を行うこととした。遺伝子ネットワークの推定手法として SiGN-BN HC+Bootstrap を用いた。まず前述の 256 データセット 30,261 サンプルのデータと EGF 関連 1520 遺伝子のリストを用いて遺伝子ネットワーク推定となる入力データセットを構築した。すべてのサンプルで 1520 遺伝子のデータが計測できているわけではないので、構築されたデータセットに含まれる遺伝子のリストはデータセット毎に異なる。次に作成した 256 データセットを入力とする遺伝子ネットワーク推定の計算ジョブを「京」に投入した。1 ジョブで計算できる遺伝子ネットワークの数は 1 個である。投入できる計算ジョブ数に制約があるためジョブの実行状況を常時監視し順次 256 個のジョブを投入した。SiGN HC+Bootstrap 法は並列度がブートストラップ回数の制約を受ける。10000 回のブートストラップ回数を採用しているため、最大並列度は 1251 並列となる。これまで SiGN-BN HC+Bootstrap 法を利用したネットワーク推定は 500 遺伝子程度のデータセットに対して行っていたが、今年度の計算対象のデータセットは 1520 遺伝子と、以前と比較するとかなり多い。そのため計算に非常に時間のかかる結果となり、いくつかの計算ジョブは「京」の利用制限である 24 時間以内に終了しなかった。途中、「京」のコンパイラに不具合を発見し、それまでの計算がすべて正しく計算できていないことが判明した。そのやり直しの計算などを含め最終的に平成 25 年度中は、「京」の 2,616,592 ノード時間利用し、250 個のネットワークの推定に成功した。

２）RNA-seq データからのがんの網羅的融合遺伝子の探索

がんの遺伝子変異として重要な融合遺伝子(fusion gene)を網羅的に検出することは、大規模・網羅的に遺伝子ネットワークを理解するために重要である。そのため、東京大学医科学研究所ヒトゲノム解析センターで開発され稼働している Genomon-fusion というデータ解析パイプラインを京コンピュータへ移植した。以下、Genomon Fusion K computer を略して GFK と呼ぶ。

Genomon-fusion は C/C++等で記述されたアラインメントなどを行うフリーソフトウェアおよび perl、シェルスクリプトなどを組み合わせた、Whole Transcriptome データ解析用パイプラインである。本移植の目的は、CCLE(米国 Broad Institute の Cancer Cell Line Encyclopedia http://www.broadinstitute.org/ccle/home)や TCGA(米国 NIH の The Cancer Genome Atlas http://cancergenome.nih.gov/)などのデータベースで提供されている、大規模ヒト検体データを処理するシステムの構築である。CCLE 全検体（780 サンプル）の処理を考えた場合、ヒトゲノム解析センターのスーパーコンピュータ Shirokane2（AMD Opteron 6276, 2.3GHz, 16,128 コア）を 3 か月程度の占有利用が必要な計算量となり、事実上解析不能である。移植にあたり、必要な作業は主に以下の 3 点であった。

① 各ツール等の「京」上でのビルド
② パイプラインの並列化（MPI 化）およびステージング対応
③ 「京」と外部ストレージの協調

上記を踏まえ、本移植に置いては最も計算資源を必要とする本処理（アラインメント部分）とそれ以外（前処理、後処理）にパイプラインを区分し、本処理部分を「京」計算ノードで処理し、それ以外をプレポストノード（ppb）で処理する方針を採用した。前処理、後処理ではファイル分割・結合等の処理が含まれ、計算ノードのローカルストレージに収まらない可能性があること、また後処理の一部に計算ノードに未実装の JAVA が必要なためである。図 9 に GFK の京コンピュータ上での実装の概略を示す。

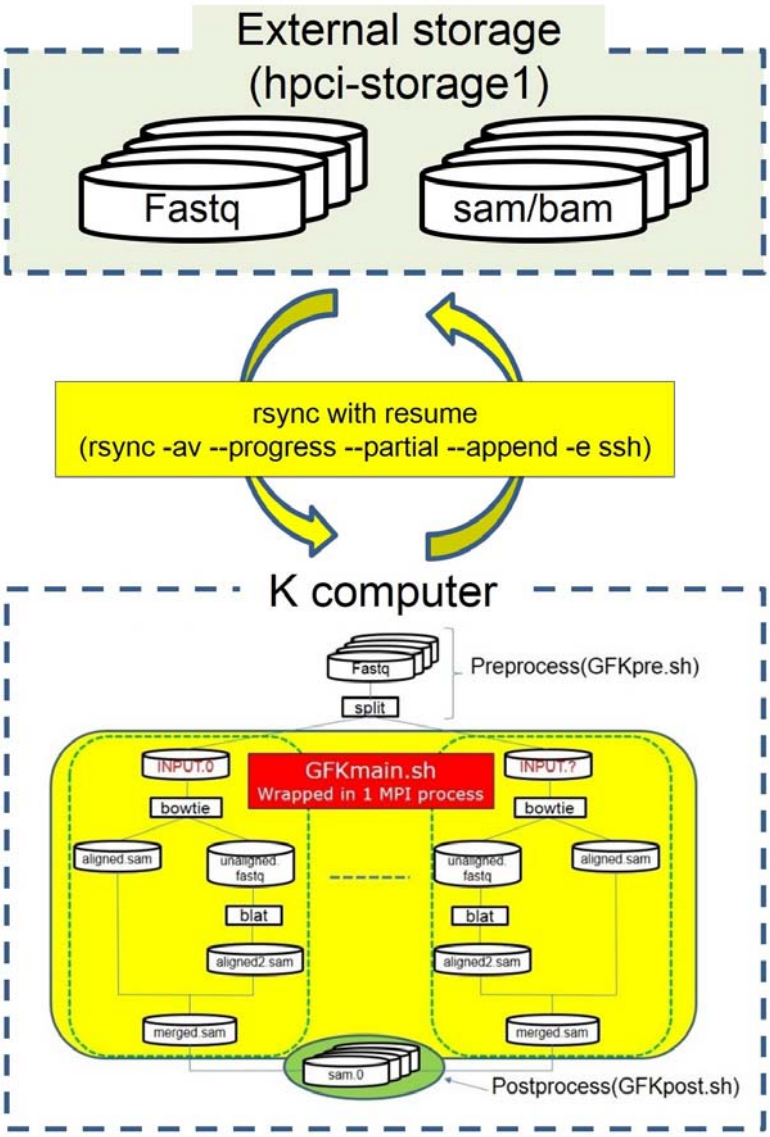


図 9 ：GFK の概略図

ビルドの必要があるソフトウェアは二つのアラインメントツール bowtie および blat である。両ソフトウェアともに C/C++言語で記述されているが、bowtie に関しては富士通コンパイラによるビルドでは正しい結果を返さない（ヘルプデスクによる調査の結果、bowtie 側のバグと判明）ため、GNU C コンパイラを用いたビルドを行った。

MPI 化に関しては当初、図 9 における各処理に対して個別の MPI 化またはラッパープログラ

ムによる MPI 化を実施したが、処理ごとの同期待ちにより全体の処理時間が増大することが判明したため、メイン処理を一つの MPI プロセスの中に収める（図９の黄色部分）方針を採用した。また、bowtie が先述の理由により GNU コンパイラでビルドしていたため、MPI 化した bowtie が京コンピュータ上で実行できなかったことも上記方針の理由の一つである。

ステージング対応については、本パイプラインでは多数検体を split により一括処理し、分割されたファイルの名前を共通プレフィクス＋MPI ランク番号とすることで、MPI ランクとインプット/アウトプットファイルとの連結およびステージングの際のファイル識別を行うようにした。分割された各ファイルの検体識別は、前処理時にテーブルを作成することで後処理時に分離可能になっている。

また、GFK の前処理では処理時間の平滑化に工夫を行った。シークエンスデータは DNA 断片（リードという）の数億～数十億本分の集合体であるが、リードごとの処理時間に大きな差があり、結果として分割したプロセスごとの処理時間のばらつきに著しい差を生じていることが GFK を用いた予備試験で判明した。図１０に、CCLE　の 1 細胞株データを GFK を用いて約 600 プロセスで並列計算した際の各プロセスの計算時間を示す。平均 3 時間のジョブであるが、30 分程度から 8 時間弱と、プロセスごとに非常に計算時間にばらつきがあることがわかる。プロセスごとに時間がばらつくことから、処理時間の短いリードと長いリードにある程度のグループ化が読み取れ、グラフから時間の長時間プロセスと短時間プロセスが全体に散らばっていることから、シークエンサーの特徴（具体的には、試薬による蛍光を取るフローセルの形状と大きさ）と各リードの処理時間の関連と思われる。そのため、分割の際にこの空間依存性を排除する高速かつ簡便な方法としてラウンドロビンを導入した。その結果、同条件での並列計算において、ほぼ均一な処理時間を実現した。（図１０右）

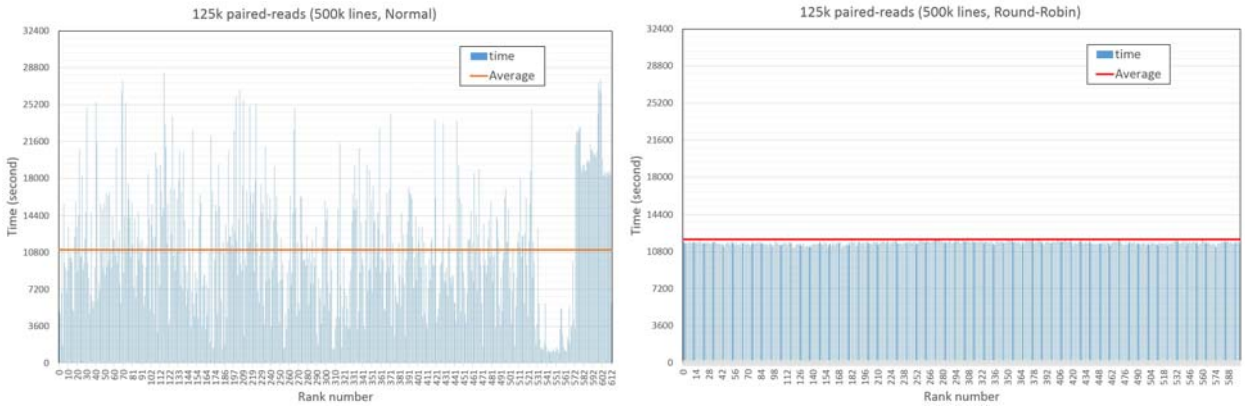


図１０：GFK を用いた 1 サンプルのプロセスごとの処理時間（左：split のみ、右：ラウンドロビン付き split）

このようにして移植したシステムを実際の大規模データで計算する際に問題となったのがストレージ容量である。今回、計算の対象としたのは CCLE（Cancer Cell Line Encyclopedia、http://www.broadinstitute.org/ccle/home）の RNA-seq データである。検体数 780、データ容量はおよそ 20TB である。本データを計算するためには

１．専用ソフト（GeneTorrent）を用いたデータダウンロード、
２．サンプルごとに異なるフォーマットの Fastq フォーマットへの変換と確認、
３．GFK 前処理、
４．GFK 後処理

の 4 つの処理が必要である。GeneTorrent は「京」/プレポストノードの双方で動作しないため、分野 1 の持ち込みストレージである hpcistorage1 へ DL した。DL したデータは今回、約 100 検体ずつの 8 回に分けて計算と結果回収を順次行う形でジョブ実行した。これは、先述の 2、3、4 と処理をするにつれデータ容量が何倍にも増えるためである。今回、/data 以下を最大 100TB 程度消費している。全検体を一度に処理した場合、この数倍～10 倍程度を一時的に消費することになり、課題割当量の 200TB すべてを使っても不足する。今回は最初に 100 検体分を前処理、次の 100 検体分を前処理しつつ本計算、後処理をしつつ次の前処理と本計算という形で hpcistorage1 と「京」でのデータ授受を行いつつ、本計算を協調的に実施することでストレージの問題を回避した。hpcistorage1 とのデータ授受の際にはレジュームオプションを利用した rsync による送受信と、md5sum による通信前後のファイル健全性確認を行っている。

このようにして CCLE780 検体分の計算を完了した。計算時間は 7,027,933,648 ノード秒（976,102 ノード時間）、499,401 コアを用いた。約 12,600 の fusion gene を検出、そのうち 109 個が既知の fusion gene、残り約 12,500 個が未知の fusion gene となっており、現在結果の詳細を分析中である。この予備的研究により現在データとして利用可能な 1 万検体を超えるがんの RNA-seq データ解析の可能性が見え、”Landscape of Cancer Fusion Genes”の解明という国際的にみてどこかの大きな研究グループが必ず挑戦するであろう問題の解決へ技術的には見通しがたった。しかし、同時に、この研究により、本課題に配分されている計算資源では不可能なことも同時に判明した。

３）EEM の「京」への移植

EEM 法(Extraction of Expression Module)は遺伝子セット情報に基づいて mRNA 発現データ中で共発現している遺伝子群、発現モジュールを抽出するソフトウェアであり、大規模な遺伝子ネットワーク解析に有効な方法である(図１１)。この EEM 法の特徴は生物学的に解釈しやすい結果を得られるところにあり、実際 EEM 法を複数のがんの発現データに適用し、がんの転写プログラムを明らかにしている。EEM のソースコードは java で書かれているが、今後、数万の臨床検体の遺伝子発現プロファイルデータ及び関連するオミクスデータと臨床データを用いた大規模・網羅的遺伝子ネットワーク解析を行うために「京」への移植が必要である。そのために、C++への翻訳、並列化(分散データ構造設計および MPI 化)をおこなった。また現在マニュアルを整備して公開準備中である。その応用については（３）－２）で述べる。

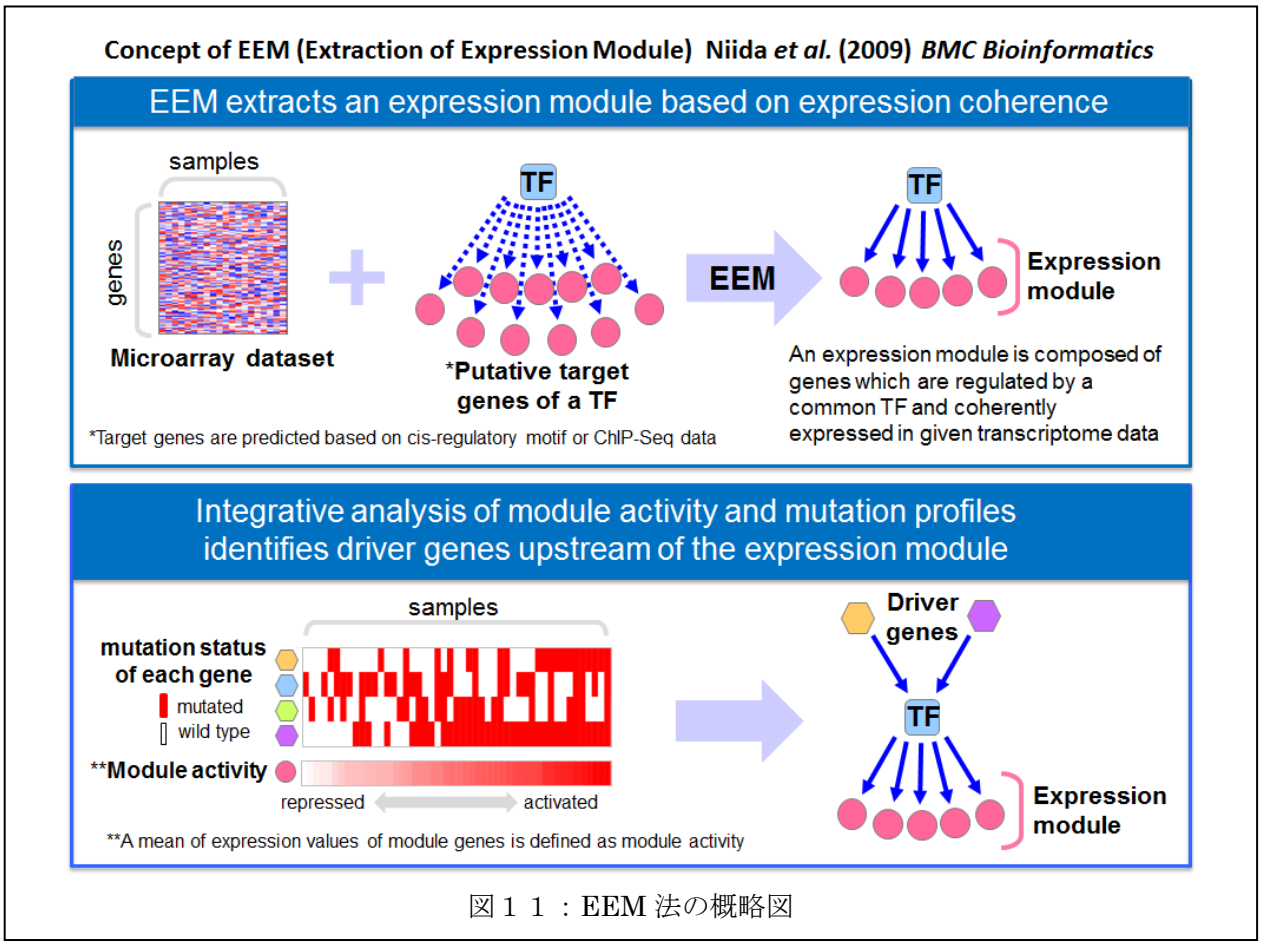


図１１：EEM 法の概略図

４）可視化技術：Multilayer Heatmap の開発

現在のがん研究においては様々なタイプのオミックスデータが大量に算出されおり、解析方法と同時にその結果を解釈するための可視化方法の開発が不可欠である。本研究ではそのような多次元のオミックスデータをクラスタリングして heatmap として表示する multilayer heatmap を開発した（図１２）。この手法を a）TCGA　project のゲノム、mRNA 発現、メチル化の多次元オミックスデータ、b）ENCODE の各種ヒストン修飾プロファイルデータ、及び c）数千の mRNA 発現データセットから予測したネットワークデータに適用したところ既知の生物学的知見の確認、及び新規仮説の生成に成功した。今後、数万の臨床検体の遺伝子発現プロファイルデータ及び関連するオミクスデータと臨床データを用いた大規模・網羅的遺伝子ネットワーク解析結果の可視化に multilayer heatmap は有用である。

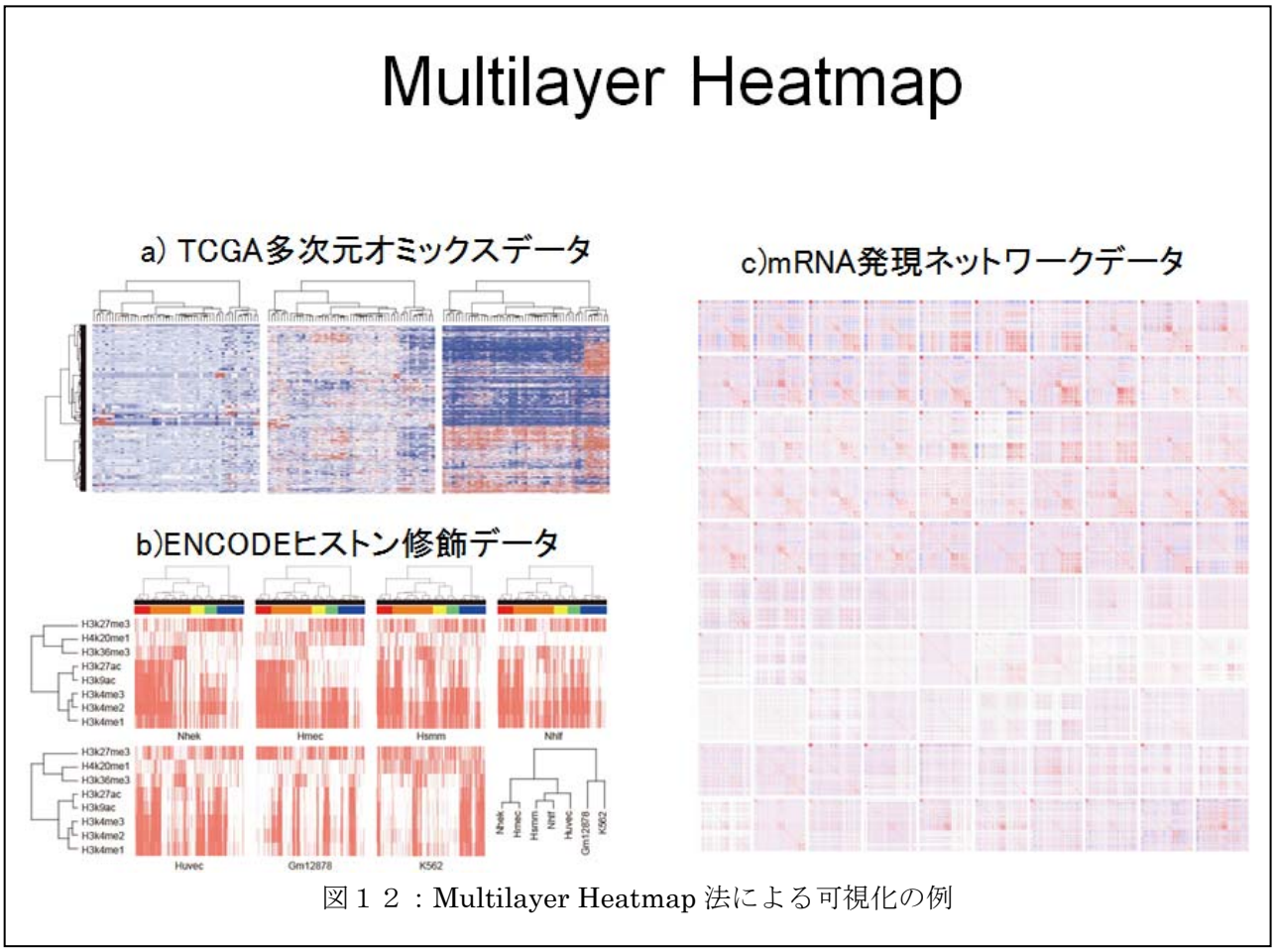


図１２：Multilayer Heatmap 法による可視化の例

（３）業務協力者との連携状況

１）EEM を用いた食道がんの解析

業務協力者である九州大学病院別府病院外科、三森功士教授らのグループは食道扁平上皮癌のエクソーム解析により *NFE2L2* が高頻度に変異していることを見いだした。*NFE2L2*(NRF2)は肺扁平上皮癌等、いくつかの癌腫で変異が見つかっている。

抗がん剤、放射線等は活性酸素の発生によりがん細胞のゲノムを傷つけてがん細胞を殺すが、*NFE2L2*(NRF2)の変異により活性酸素に対抗する還元酵素群が活性化され酸化ダメージに耐性ができると考えられている。また発現マイクロアレイデータに EEM を適用することで *NFE2L2* が制御する Oxidoreductase pathway gene set が有意に共発現し発現モジュールを構成していると予測された。更にその module 活性が *NFE2L2* 変異ありのサンプルで有意に亢進していることを見いだした。以上の事から変異に加え *NFE2L2* の mRNA もモジュール活性に相関しているので変異と mRNA 発現上昇の二つの機構で Oxidoreductase モジュールを活性化していると考えられる(図１３)。今後、数万の臨床検体の遺伝子発現プロファイルデータ及び関連するオミクスデータと臨床データを用いた大規模・網羅的遺伝子ネットワーク解析をこのようなアプローチで「京」を用いて行う準備ができた。

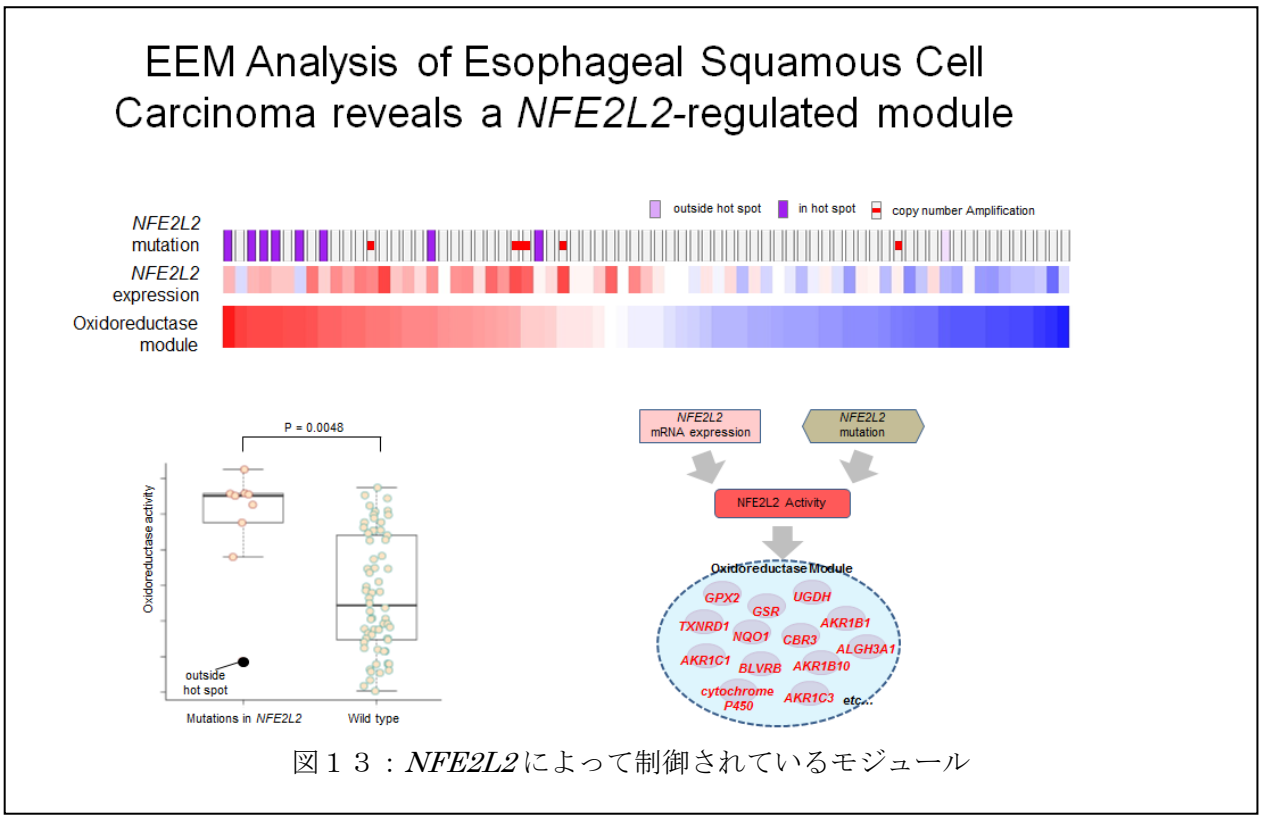


図１３：*NFE2L2* によって制御されているモジュール

２）大腸がんの腫瘍内不均一性のゲノム解析及びシミュレーション

（１）及び（２）の準備的解析の結果より異なる患者の腫瘍間においてゲノムレベル、ネットワークレベルでがんは広汎な多様性を有することが明らかになった。業務協力者である九大別府病院外科、三森功士教授らのグループは一人の患者が有する大腸がん一腫瘍内においての不均一性を解析するために、大腸がん一腫瘍内の複数の領域の細胞集団から DNA を取得し(multiregional sampling) 、エクソーム、コピー数解析をおこなった（図１４）。その結果すべての切片で共通する、がんの進化の早い段階で得られたと考えられるファウンダー変異が存在する一方で、すべての切片には含まれないがんの進化の遅い段階に得られたと考えられるプログレッサー変異が一腫瘍内においても広汎な不均一性をゲノムレベルで産み出していることが明らかになった。更にこのようなゲノム変異パターンから腫瘍のクローン進化の系譜を推定する事に成功した。またゲノムデータと共に mRNA 発現、DNA メチル化も取得し、トランスクリプトーム、メチロームレベルでもがんの進化の過程で不均一性が生み出されていることも確認した。

このことから、がんのシステム異常を網羅的に解析し、理解するためには、以上の実験結果をふまえて、このような広汎な腫瘍内不均一性を生み出す機構の解明をシミュレーションにより試みることが不可欠と考えるに至った。一細胞を一つの agent とする agent based model を用いて細胞を増やしながら腫瘍が成長する様子をシミュレーションにより再現した（図１５）。更に京コンピュータ及びヒトゲノム解析センターのスーパーコンピュータを用いて、シミュレーションモデルを様々なパラメータセットで実行し、実験データを同様の変異パターンを生み出す条件を探索した。その結果、多数のドライバー遺伝子の存在、高い変異率、がん幹細胞の存在が大腸がんの公汎な腫瘍内不均一性を伴う進化に重要であることを見いだした。今後さらにモデルを改良し詳細な条件探索を行うことが必要であることが判明した。

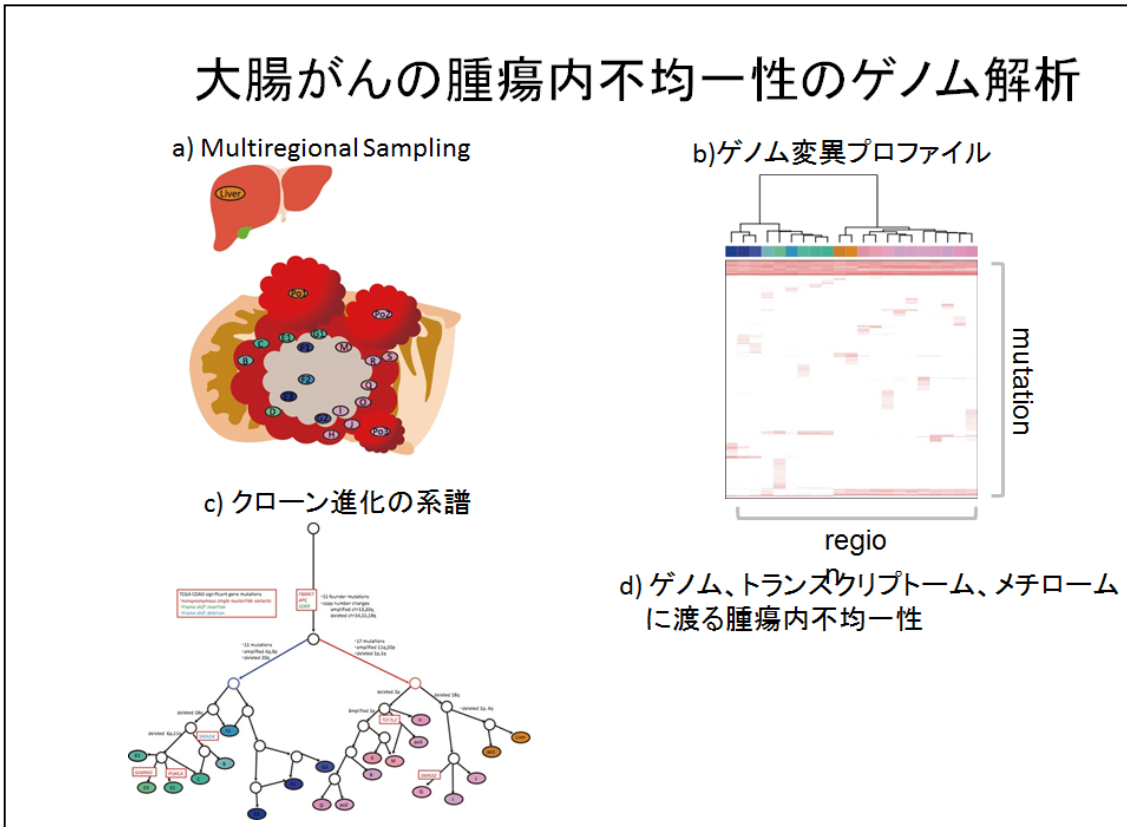


図１４：大腸がん一腫瘍内の複数の領域の細胞集団からその進化と不均一性が判明

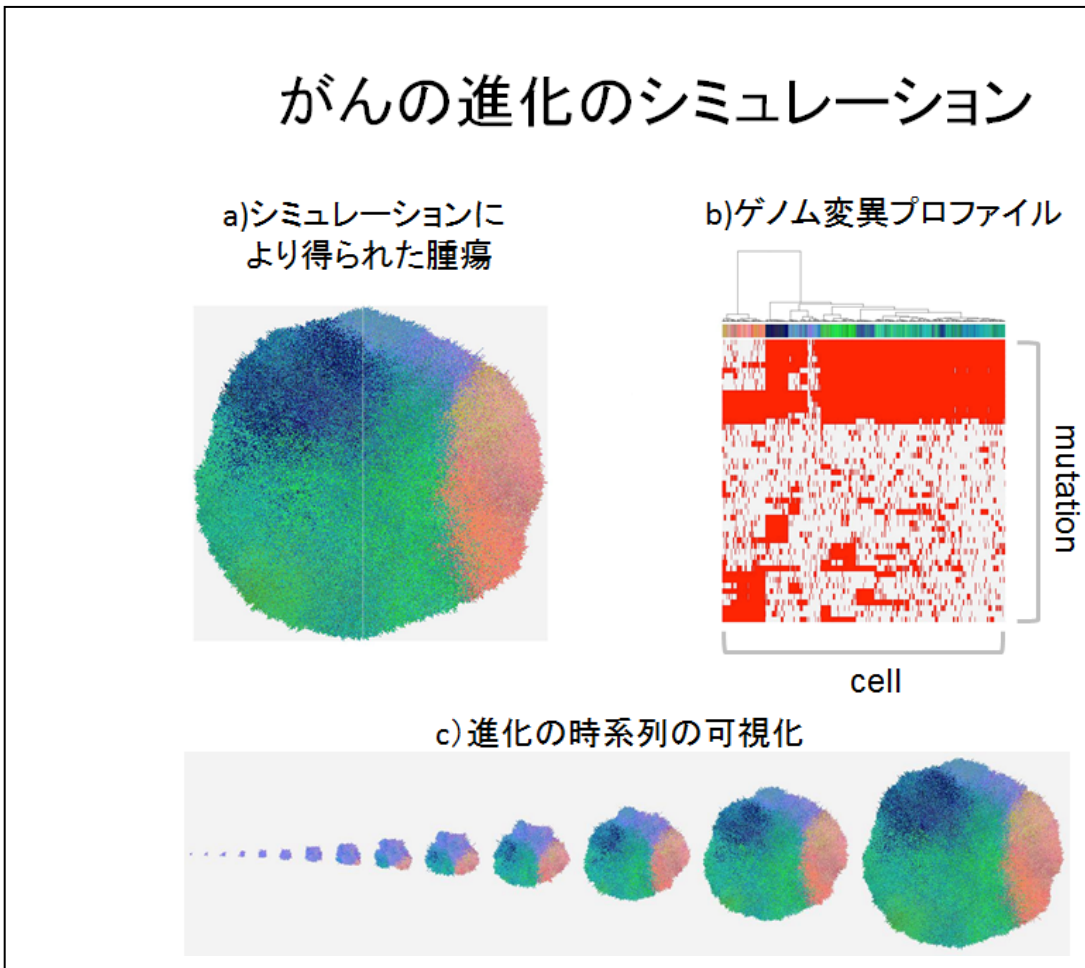


図１５：進化シミュレーションのスナップショット

３）がんの大規模シークエンスによる網羅的変異探索

京都大学院医学研究科の小川誠司教授等のグループは、全国、全世界の研究者と連携して、全部で数千のがん臨床検体の全ゲノム及びエクソーク解析を行い、変異をがん及びゲノム網羅的に探索している。そのデータは膨大であり、そのデータ解析を宮野・小川で開発した Genomon を用いてヒトゲノム解析センタースーパーコンピュータで実施し、多くの発見があった。このデータ解析には京コンピュータは用いなかった。その理由は、「京」の運用がこのデータ解析に適していないこと、「京」では利用できないプログラム言語を用いたソフトウェアが非常に多く用いられていること、臨床検体を扱うための計算科学研究機構側の準備のめどがたっていなかったこと、Genomon の移植には多くの時間と労力がかかるため、「京」で利用できることを待っていては、世界のがん研究から大きく遅れを取ること、など様々な理由がある。こうしたことからゲノムデータ解析に万全の体制が構築されている東京大学医科学研究所ヒトゲノム解析センターのスーパーコンピュータシステムを限界まで利用する必要があった。その一旦は、（２）－２）で述べた、GFK の苦労に象徴されていることからも理解されると考える。小川教授らの要求を十分に満たすには、情報系だけで、１０人員規模の訓練された人員の確保と１０倍の予算が不可欠である。これについてはヒトゲノム解析センター等からの持ち出しでやっている。

４）業務協力者である東京大学医科学研究所松田浩一准教授等のグループと連携して、（１）及び（２）の大規模データ解析から得られた知見に関して、同グループの実験データを解析することによって得た結果を比較することで検討を行った。具体的には p53 欠損マウスへの X 線照射実験により多臓器から取得された RNA-seq データを対象とした。転写因子 p53 はがん抑制遺伝子の一つであり、ヒトのがん細胞において高頻度に変異が認められることが知られている。故にその機能を欠損したマウスに対して、細胞障害を誘発する X 線刺激を与えることで、どの臓器でどのような転写産物が発現するかを明らかにすることは、がんの発生・進展のメカニズムに関しても深い知見を与えることが期待される。しかしながら多臓器（20 臓器、表１）かつ多条件下(野生型＋X 線刺激あり、野生型＋X 線刺激なし、p53 欠損+ X 線刺激あり、p53 欠損+ X 線刺激なし)から同時に得られた RNA-seq データから有用な情報を抽出する方法は自明ではない。そこで我々はまず、各 RNA-seq データをリファレンス配列にマッピングを行い、RNA の発現量を得た。解析には松田准教授の研究の進展を遅らせないため「京」は用いず、東京大学医科学研究所ヒトゲノム解析センターのスーパーコンピュータを用いた。次に臓器ごとに、タンパクをコードしている遺伝子の RNA (mRNA)に対して発現差解析を行い有意に変動している遺伝子群を抽出し、それらの変動遺伝子群を全臓器において比較して興味深い変動パターンを示す遺伝子群を抽出した（図１６）。それらの遺伝子群と、（１）と（２）の解析で推定されたネットワーク上の p53 近傍の遺伝子群との比較検討を行った。

表１：臓器名および ID

Tissue id	Tissue Name	
	JP	EN
1_bl	血液	blood
2_th	胸腺	thymus
3_he	心臓	heart
4_lu	肺	lung
6_ki	腎臓	kidney
7_sp	脾臓	spleen
8_li	肝臓	liver
9_bl	膀胱	bladder
11_es	食道	esophagus
12_st	胃	stomach
13_co	大腸	colon
14_sm	小腸	small intestine
16_te	精巣	testis
17_pa	精巣上体	parorchis
18_se	精嚢	seminal vesicle
20_mu	筋肉	muscle
21_bo	骨髄	bone marrow
23_to	舌	tongue
24_ey	眼球	eye ball
25_ce	大脳	cerebrum

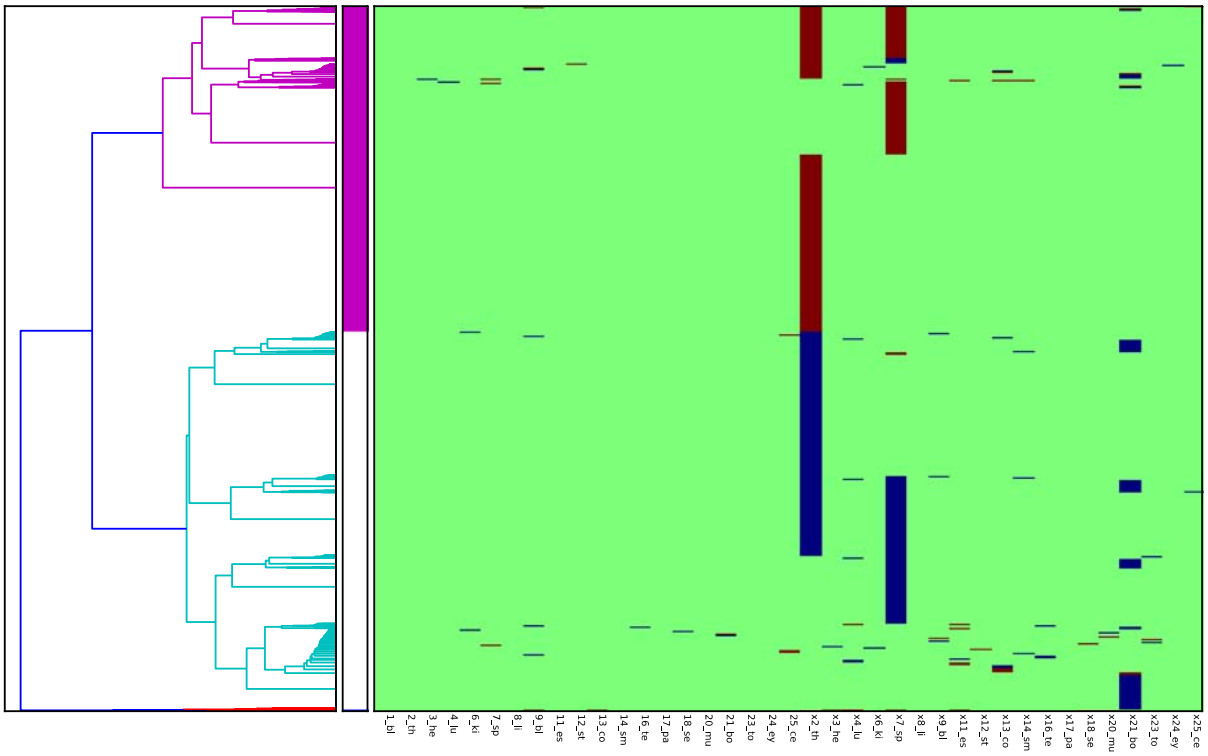


図１６：多臓器変動遺伝子群比較

（４）（１）の研究を遂行する中で新たな大規模生命データ解析の方式が、計算時間と精度の関係から必要となった。多検体で計測された RNA 発現プロフィールの統計解析において、Lasso タイプのスパース学習に基づくモデリングは、標準的な方法として用いられている。しかしながら、特に、応答変数が２値で定義されるクラス分類の問題において、通常の Lasso は、必ずしも分類予測能力は高くなく、変数選択（ここではモデルに取り込む遺伝子の選択）においても偽陽性が多いことが知られている。そこで、我々は、２クラス分類問題において、単変量の Wilcoxon rank sum test の結果を利用し、遺伝子毎に適応型に Lasso パラメータの重みを与えることで高いクラス予測能力と遺伝子選択の正確性を達成する新しい統計解析モデル（Wilcoxson penalized logistic regression model: WPLRM）を開発した。

表２：人工データを用いたシミュレーションによる提案手法 WPLRM の性能評価

n	p	Method	Classification			Variable selection		
			T.N	T.P	Ave	T.N	T.P	Ave
25	100	L	0.838	0.821	0.830	0.996	0.286	0.641
		E	0.879	0.868	0.874	0.843	0.856	0.850
		A	0.847	0.836	0.842	0.996	0.402	0.699
		W	0.885	0.875	**0.880**	0.855	0.864	**0.860**
	1000	L	0.801	0.797	0.799	0.998	0.255	0.626
		E	0.809	0.813	0.811	0.965	0.699	0.832
		A	0.998	0.008	0.503	1.000	0.001	0.501
		W	0.814	0.818	**0.816**	0.965	0.724	**0.845**
50	100	L	0.857	0.858	0.857	1.000	0.346	0.673
		E	0.923	0.929	0.926	0.892	0.941	0.916
		A	0.873	0.878	0.876	0.999	0.505	0.752
		W	0.923	0.934	**0.928**	0.907	0.933	**0.920**
	1000	L	0.863	0.863	0.863	1.000	0.343	0.672
		E	0.904	0.892	0.898	0.968	0.899	0.933
		A	0.933	0.275	0.604	1.000	0.050	0.525
		W	0.906	0.897	**0.901**	0.970	0.906	**0.938**
75	100	L	0.869	0.860	0.865	1.000	0.362	0.681
		E	0.944	0.939	0.942	0.910	0.973	0.941
		A	0.892	0.883	0.888	1.000	0.575	0.787
		W	0.947	0.944	**0.945**	0.924	0.974	**0.949**
	1000	L	0.862	0.867	0.865	1.000	0.380	0.690
		E	0.920	0.926	0.923	0.971	0.966	**0.969**
		A	0.841	0.716	0.778	1.000	0.203	0.601
		W	0.922	0.928	**0.925**	0.972	0.966	**0.969**
100	100	L	0.874	0.877	0.876	0.000	0.392	0.696
		E	0.955	0.954	0.955	0.935	0.982	0.958
		A	0.900	0.902	0.901	1.000	0.626	0.813
		W	0.958	0.956	**0.957**	0.945	0.983	**0.964**
		L	0.876	0.866	0.871	1.000	0.395	0.697

開発した WPLRM は、人工データを用いたシミュレーションによりその性能を評価した。生成したデータは、サンプル数が 100、75、50、25、遺伝子数が 1000、100 の組み合わせのもので、真にクラス決定に寄与している遺伝子数は 10 である。実際の遺伝子発現データを模倣するため、各遺伝子には相関を持たせている。結果を表２に示す。L、E、A は既存手法であり、L は Lasso、E は Elastic net、A は Adaptive Lasso、W が提案手法となる。Classification がクラス予測の正解率、Variable Selection が遺伝子選択の正解率の結果を表している。T.N は True Negative rate、T.P は True Positive rate であり、Ave がその平均を示している。どのサンプル数、遺伝子数の組み合わせでも開発した WPLRM が既存手法を上回る結果を示すことができた。

この開発した WPLRM を用いてサンガーセンターの Cancer Genome Project のデータから５つの化合物（Gemcitabine, Bleomycin, RDEA1190.97, LFM.A13, PF.562271）を選び、その IC50 により耐性、感受性を欠く細胞株で予測する解析を行った。IC50 の四分位点を用い、２５％点以下を感受性、７５％点以上を耐性と定義し、モデルの学習に使う training set、モデルの評価に使う test set にデータを分け、結果を評価した。まず、training set において、主成分分析を用いてサンプル数と同じ次元の空間に射影し説明変数に相当するメタ遺伝子プロファイルを構成し、解析

に用いた。各化合物について耐性感受性を予測するモデルを WPLRM により構成し、耐性感受性について予測能力を有する主成分について寄与の大きい上位５遺伝子をまとめたものが表３である。２つ以上の化合物で表れる遺伝子には色を付けて区別している。

表３：耐性感受性予測モデルに寄与の大きい遺伝子

Drug	PC	1st	2nd	3rd	4th	5th
Gemcitabine	6	VIM	LGALS1	TACSTD2	KRT18	SFN
	9	TMSB4X///TMSL3	B2M	HLA-B	HLA-C	ACTB
	23	COL1A2	IGL@IGLC1I	TMSB10	MT2A	IGL@
Bleomycin	2	S100A6	ANXA2	FTL	FTHP1	LGALS1
	5	TUBA3	RPS3A///LOC439992	HMGB1	K-ALPHA-1	RPL3///LOC653881
	8	TUBA3	NGFRAP1	HLA-B	UBC	LAPTM5
	9	TMSB4X///TMSL3	B2M	HLA-B	HLA-C	ACTB
	16	KRT18	TACSTD2	S100A6	GPNMB	S100A2
	21	AKR1C1	AKR1C2	FTL	AKR1C3	LGALS3///GALIG
	30	AKR1C1	AKR1C2	IGL@IGLC1I	AKR1C3	IGL@
RDEA1190.97	1	PLK1///RPL37A	RPL23A	HUWE1	TPT1	LOC644808
	5	TUBA3	RPS3A///LOC439992	HMGB1	K-ALPHA-1	RPL3///LOC653881
	10	PPIA	GAPDH	MT2A	LGALS3///GALIG	LDHA
	13	FTL	AKR1C1	B2M	AKR1C2	TPT1
	43	MT2A	HLA-H	RPL13///LOC388344	S100A6	CD24
	48	COL1A2	S100A6	FN1	CD24	AKR1C1
LFM.A13	1	PLK1///RPL37A	RPL23A	HUWE1	TPT1	LOC644808
	9	TMSB4X///TMSL3	B2M	HLA-B	HLA-C	ACTB
	13	FTL	AKR1C1	B2M	AKR1C2	TPT1
	42	IGL@IGLC1I	MT2A	IGL@	TUBA3	IGHM
	52	MT2A	FN1	HLA-A	HLA-H	HLA-B
PF.562271	2	S100A6	ANXA2	FTL	FTHP1	LGALS1
	8	TUBA3	NGFRAP1	HLA-B	UBC	LAPTM5
	33	COL1A2	IGL@IGLC1I	TMSB10	MT2A	IGL@
	40	LDHA	CD24	HUWE1	TPT1	ALDOA
	58	FN1	TMSB4X///TMSL3	CAV1	TUBA3	RPL13///LOC388344

次に、この耐性感受性予測モデルを test set に適用し、耐性感受性予測能力を評価した結果を

表４：耐性感受性予測能力の評価

	Gemcitabine	Bleomycin	RDEA1190.97	LFM.A13	PF.562271
L	0.7024	0.6873	0.7708	0.7048	0.7230
	(0.122	(0.123)	(0.144)	(0.146)	(0.143)
E	0.7236	0.7442	0.8295	**0.7548**	0.7458
	(0.110	(0.125)	(0.122	(0.118)	(0.136)
A	0.7054	0.7026	0.7982	0.7192	0.7334
	(0.134	(0.125)	(0.131)	(0.129)	(0.132)
W	**0.7244**	**0.7489**	**0.8334**	0.7509	**0.7465**
	(0.109)	(0.127)	(0.109)	(0.125)	(0.127)

表４に示す。

前出のシミュレーション結果と同様に、開発した WPLRM を既存手法である Lasso（L）、Elastic net（E）、Adaptive Lasso（A）と比較した。４つの化合物において最も高い性能を達成できた。LFM.A13 では Elastic net がトップの予測能力（0.7548）であったが、WPLRM は、ほぼ同等な性能（0.7509）を示しており、この結果から実データにおいても WPLRM は耐性感受性予測において高い性能を有していることを示すことができた。

「戦略課題４：大規模生命データ解析」研究統括では、以下の２つの大学で実施される平成２５年度の研究課題の実施項目について、適宜、関連する研究者とワークショップや研究打合せを行い、また業務協力者に対してはそれぞれの専門の立場から知見とアドバイスを仰ぎ、関係者のとりまとめを行うとともに、大規模生命データ解析ワークショップを開催し、理化学研究所と連携して、研究開発の統括を行った。

① 大規模生体分子ネットワーク解析による脂肪細胞組織の刺激応答の網羅的解析とその応用（松田秀雄・大阪大学）

② 次世代シークエンサデータ解析のための情報処理システムの開発(秋山泰・東京工業大学)

Ⅳ－２　秋山　泰（東京工業大学）

次世代シーケンサデータ解析のための情報処理システムの開発

Ⅳ－２－１　実施計画

「大規模生命データ解析」では、ゲノムを基軸とした大規模生命データ解析により生命プログラムとその多様性を理解することを目標としている。本研究では、これを実現するために最も重要な基盤となる次世代シーケンサから産出される大量のゲノム配列情報の超高速解析を実現するための研究開発を実施する。

また、「次世代シーケンサデータ解析のための情報処理システムの開発」の研究を行う上で、関連する研究者と必要な協議等を行うとともに、本格実施に必要な研究体制の整備を行う。

平成２５年度は、リード配列の相同性解析のための並列ソフトウェア(GHOST-MP)のコード最適化を継続して行い「京」での実行効率をさらに高める。またヒト体内細菌の大規模なメタゲノム解析を中心として、80,000 ノード級の計算をともなう超大規模研究を実施する。既に平成２４年度までに「京」の全系を用いた測定を一度行っているが、その後に計算機能の拡張やコード最適化を実施したので、平成２５年度は 80,000 ノード級での性能測定を複数回実施する。また単なる性能測定ではなく、実際にヒト体内共生細菌（当面は口腔内細菌を対象)に関する大規模なメタゲノム解析を実施し、実用面における実証的な評価を得る。また、I/O 負荷を相対的に軽減することを目的として、少数のノードのみが I/O を担当し、Tofu ネットワークを用いてデータのブロードキャストにより I/O 衝突を減らす新手法について様々な条件下での性能評価を行い、方式として優れていればGHOST-MP の基本機能として組み込みを実施する。さらに、パイプラインへのジョブの投入および実行状態モニタなどを簡便に行うためのインタフェースの第一版の開発を終了し、「京」との遠隔接続の可能性および通信性能などを調査する。

Ⅳ－２－２　実施内容（成果）

平成２５年度は、当初計画していた大規模メタゲノム解析の並列ソフトウェアの開発と、それを利用した大規模なヒト体内細菌メタゲノムの解析に加え、がんゲノム解析パイプラインの「京」への実装を行った。

（１）　メタゲノム解析パイプラインの開発

当該メタゲノム解析パイプラインは、一連の処理を通して次世代シーケンサから得られる各リード配列に対して既知配列データベースとの比較参照を行うことで、サンプルデータ中に含まれる遺伝子の機能の相対存在度に基づいた解析を可能とする。実施計画に基づき、以下のように解析パイプラインに関して改良および調査を行った。

●GHOST-MP の 80,000 ノード級での性能測定

解析パイプラインで中心的な役割を担っている並列相同性解析ソフトウェア GHOST-MP の大規模環境での性能測定を行うとともに実データの解析を行った。41,472 ノードで 6 回、20,736 ノードで 8 回、10,368 ノードで 7 回と大規模並列環境で性能測定と実データの解析を並行して行った。全計算ノードを用いた性能測定は、測定を繰り返すことが困難であったため行えなかったが、大規模並列計算環境での複数回の性能測定の結果を受けて実行時パラメータのチューニングを行うことで、最大 37%の速度向上が達成された。

●大規模メタゲノム解析による実証評価

Human Microbiome Project によるヒト微生物叢のメタゲノムデータのうち、口腔内細菌叢のデータ解析を行った（表 1）。口腔内 9 部位計 418 サンプルのデータに含まれる約 261 億本のリード配列に対し、解析パイプラインを用いて OTU および機能（KEGG Orthology）についてアノテーションを行った。さらに、得られたアノテーションに基づいて各サンプル間の比較を行い、口腔内部位間の微生物叢ゲノムの類似性を明らかにした。

表 1　HMP の口腔内細菌叢メタゲノムデータ

Oral Site	# of samples	read counts ($x10^6$)
Attached Keratinized Gingiva	6	361
Buccal Mucosa	121	7478
Hard Palate	1	54
Palatine Tonsils	6	373
Saliva	5	278
Subgingival Plaque	8	517
Supragingival Plaque	128	7965
Throat	7	393
Tongue Dorsum	121	8708
Total	418	26131

●並列相同性解析ソフトウェア GHOST-MP の I/O 負荷軽減

GHOST-MP の計算時間を削減する試みとして、I/O 負荷軽減のために少数の代表ノードのみがファイル I/O を担当する手法を開発し、性能評価を行った。前項で述べたヒト口腔内細菌叢のメタゲノムデータの相同性解析について、使用ノード数を 12 ノードから 41,472 ノードまで変えながらウィークスケーリングを測定したところ、10,368 ノードまでは良好なスケーリングを示し、解析速度の向上も確認されたが、20,736 ノードからスケーリングが悪化した（図 1）。原因は確定していないが、代表ノードの数を変化させながら、相同性解析やデータベースとクエリ配列の通信時間など各計算要素に要する時間を測定したところ、代表ノードへの通信待ちが発生していると考えられた。通信の合間に行っているファイル出力のために、代表ノードの送受信性能が低下していると考えられる。対策として、通信とファイル I/O を別々のスレッドで実行することや計算の粒度を大きくし、通信頻度を減らすことが挙げられる。ただし、代表ノード数を大きく変化させても計算時間への影響が大きくない結果も得られており、上記の仮説では説明できないため、今後より詳細な調査が必要である。

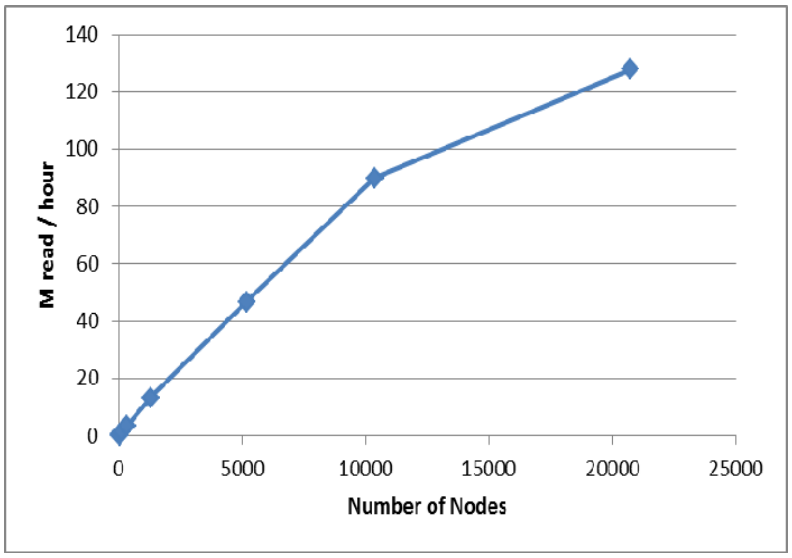


図 1　GHOST-MP のスケーリング。41,472 ノードを使用した場合、計算が終了しなかったためプロットしてない（<82M read / hour）

●ウェブインターフェースを通した遠隔接続の可能性および通信性能の調査

遠隔地にあるメタゲノムデータの解析に対する、本解析パイプラインの利用可能性の調査を行った。次世代シーケンサから得られるメタゲノムの配列データのサイズが大きいため、遠隔地と「京」の間のデータ転送に時間を要することが予想される。本解析パイプラインの解析速度を有効活用するために、解析時間に対して転送時間が十分小さいことが要請される。そこで、実データを用いて、データ転送時間と解析時間の測定を行った。

メタゲノムデータのファイルサイズは様々であるが、Human Microbiome Project が Illumina 社の Genome Analyzer IIx で行った全ゲノムシーケンシングの場合、概ね数百 MB から数十 GB 程度となっている。出力(解析結果)のサイズは、入力サイズに関わらず 100KB から 200KB 程度であるため、出力の転送時間は入力であるメタゲノムデータの転送時間に対して無視できる。東京工業大学から「京」のグローバルディスクへファイルの転送を行ったところ、29MB/s から 47MB/s の転送速度であった。ヒトの頬粘膜のメタゲノムデータ（264MB、350 万リード）を入力として、KEGG GENES データベースを用いて 1,440 ノードを利用して解析を行った場合、データの転送に約 6 秒、計算に 280 分を要した。パイプラインで用いているアルゴリズムは、計算時間が入力サイズに比例するため、より大きなデータに対しても「京」を利用して高速な解析が可能である。繁忙期には「京」に計算を投入した後に、数時間から数日の実行待ちの可能性があるものの、「京」にデータを転送し、計算を行い、結果を得るというワークフローは有効であることが確認できた。

（２）がんゲノム解析パイプラインの「京」への実装

課題内のチーム間の連携を強化すべきであるとの助言に従い、平成 25 年度 6 月より本研究の範囲をがんゲノムデータ解析にも拡張し、「京」へのがんゲノム解析パイプラインの実装を始めた。「京」の大規模計算能力を生かし、「京」全系を利用すれば 1,000 人分のゲノムデータ解析を一日で完了することが可能な並列プログラムを開発することを目指す。大規模ながんゲノム解析基盤を「京」上で実現することにより、今後見込まれる数百万人規模の個別ゲノムデータ解析への道を開くことが期待できる。実装するプログラムは Exome 解析(東京大学医科学研究所 宮野研究室開発の Genomon-exome を移植)、Transcriptome 解析（同 Genomon-fusion を移植)、Whole genome 解析（Ion

Proton データに対応、宮野研究室で開発中)の三つを予定している。

今年度は、東京大学医科学研究所のスーパーコンピュータで動作している Genomon-exome を「京」で稼働させることを目標として開発を進めた。現在までに、複数サンプルの解析を「京」の複数ノードで並列して行う仕組みの実装が完了し、小規模なテスト実行に成功した。今後、よりタスクの並列度を高められるようコードの改修を行い、大規模な実データ応用と計算性能測定を行う。

Genomon-exome は、Exome シーケンス結果の FASTQ 形式データを入力とし、(i)指定したリファレンス配列に対するマッピングとアラインメント、(ii)マッピング結果の統計解析により有意な突然変異の候補を出力する（図 2)。このソフトウェアは様々なオープンソースプログラムを組み合わせたパイプラインである。

Genomon-exome を「京」で動作させるにあたり、解決すべき技術的な課題は主に二点あった。一つは、パイプラインに含まれるオープンソースプログラムのそれぞれを「京」上でコンパイルし、動作を確認する必要があることである。マッピングとアラインメントを行う BWA, SAM/BAM ファイル操作を行う SAMTools については、「京」用にソースコードを改変し、エンディアン変換についてのバグの修正を行った結果、コンパイルと実行に成功した。研究開始時点では「京」では動作していなかった R 言語については、アラインメント結果の統計解析のために必須であるため、富士通社と連携してソースコードに必要な改変を行った上でクロスコンパイラによりコンパイルして利用した。Java 言語によるツール（Picard）による処理については、C 言語で書かれた他のツールの同等機能に置き換えを行った。

もう一つの課題は、効率的な並列計算のための仕組み作りである。「京」上で Genomon-exome の効率的な並列計算を行うためには、MPI を利用してパイプライン中の各タスクを管理する機能が必要になる。この際、各ジョブの依存関係の管理も必要となる（図 3 図 4)。この機能を実現するために、後述する MPIDP を用いた。今年度末までに、この MPIDP を用いて Genomon-exome を実装し、小規模なサンプルデータの処理を一通り「京」上で実行することに成功した。また、サンプルデータの実行結果が Genomon-exome を本来の実行環境（東京大学医科学研究所ヒトゲノムセンターのスーパーコンピュータ shirokane1）で実行して得た結果と一致することを確認した。

今後は、各段階の処理を分割して管理するよう変更し、ステップジョブとの併用を検討するなど、処理の並列性を高めるための改良を試みた上で、大規模な実データ応用での性能計測を実施する予定である。

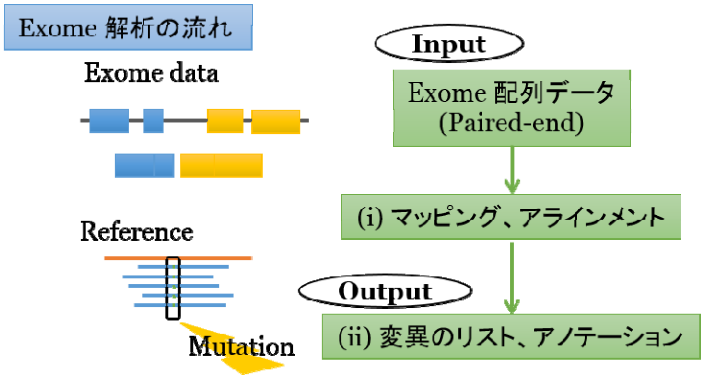


図 2　Genomon-exome による処理の概要

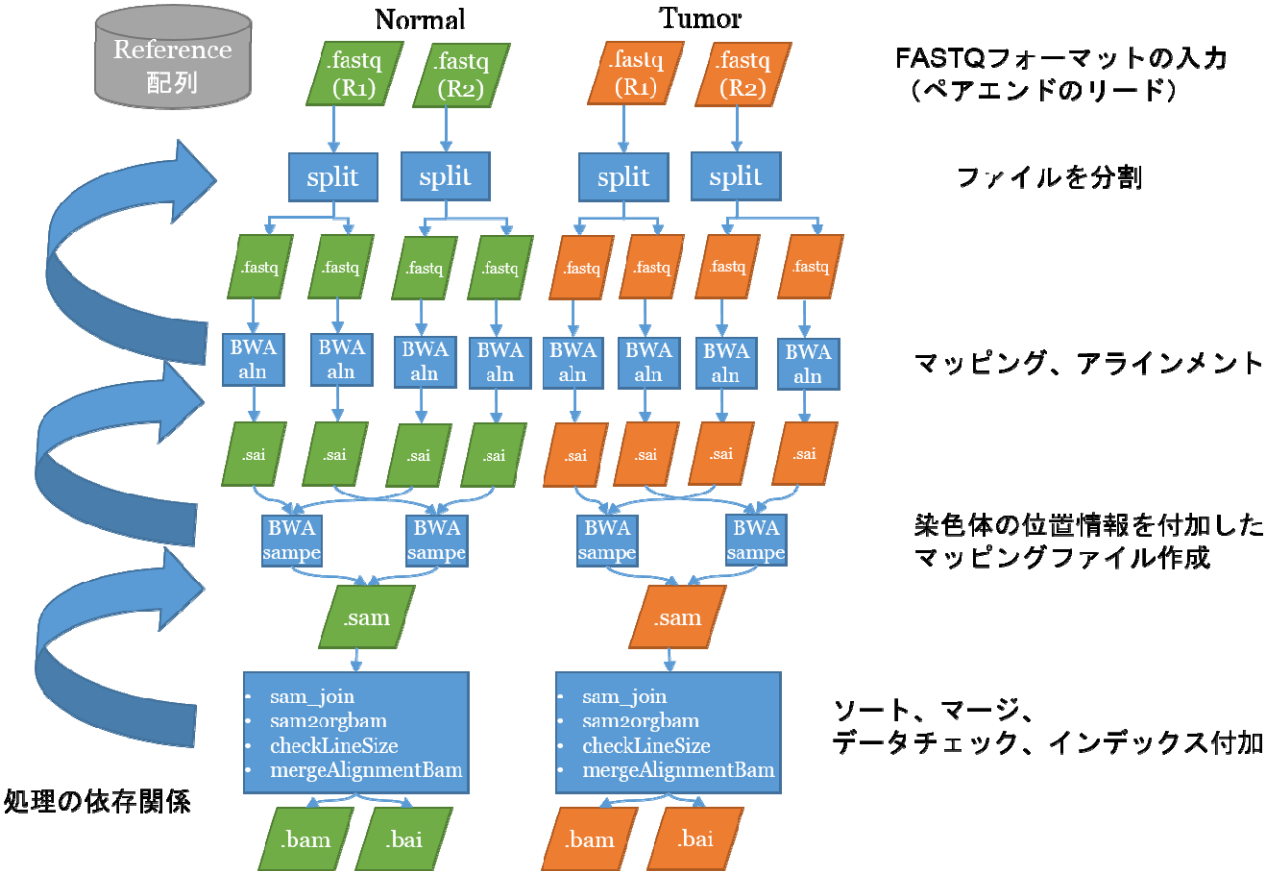


図 3　マッピングとアラインメント処理

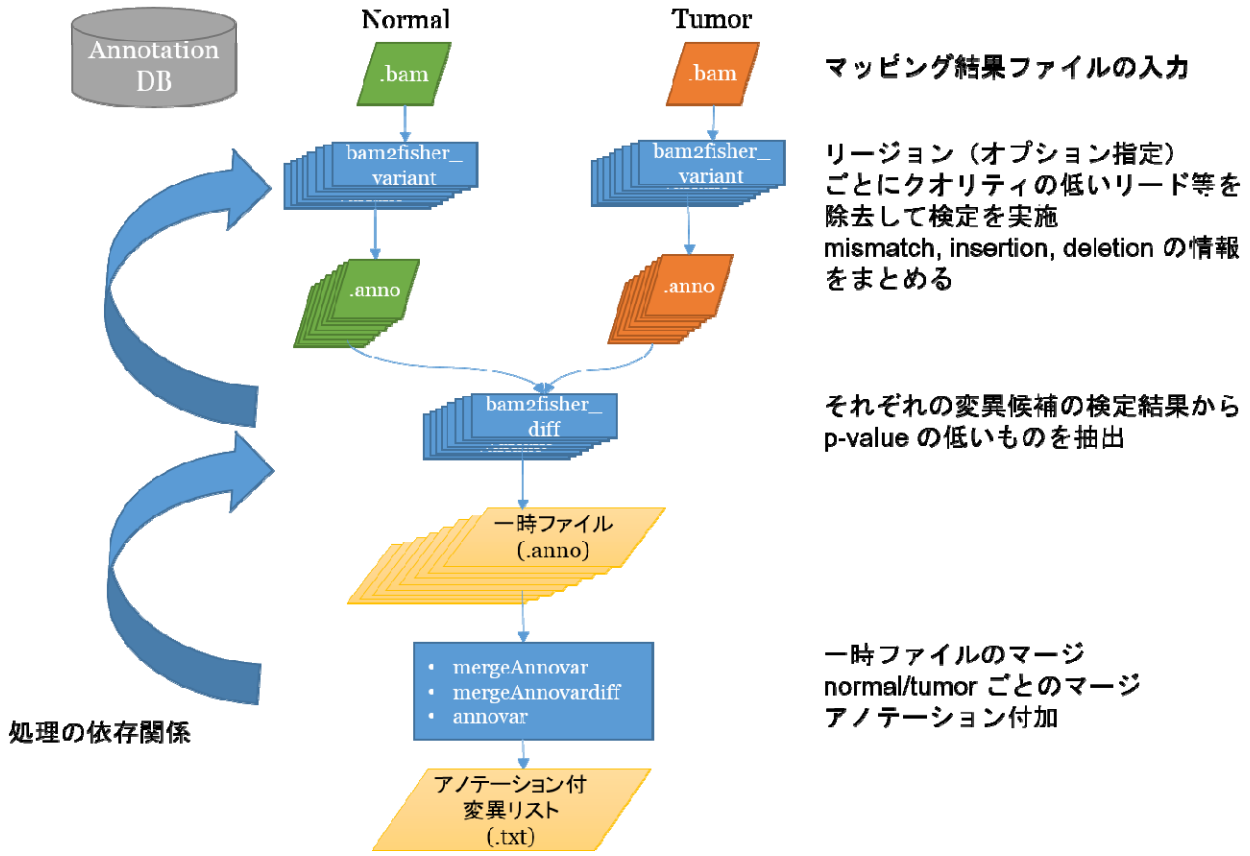


図 4　変異リストの作成

（3）　MPI ライブラリを利用したタスク管理プログラムの改良

本研究で開発する解析パイプラインの実行効率を高めることを目的として、「京」のジョブスケジューラでは困難である細かい粒度のジョブのスケジューリングを可能とするために、タスクの管理および MPI ライブラリを利用したタスクの分配を行う MPIDP を開発した。（1）および（2）のメタゲノム解析パイプラインとがんゲノム解析パイプラインの両パイプラインで、この MPIDP を共通に利用できるように変更を加えたうえ、がんゲノム解析パイプラインで必要とされるタスク間に依存関係のあるジョブの実行を可能とするための機能追加を行った。タスク間の依存関係は非循環有効グラフで表し、Master-Worker モデルの Master が依存するタスクの終了を確認次第、次のタスクを Worker に発行することで、依存関係のあるタスクの処理を可能とした。

Ⅳ－３　松田　秀雄（大阪大学）

大規模生体分子ネットワーク解析による脂肪細胞組織の刺激応答の網羅的解析とその応用

Ⅳ－３－１　実施計画

本研究では、「戦略課題４：大規模生命データ解析」の目標である、特定高速電子計算機施設を中核とする HPCI に最適化した最先端・大規模シークエンスデータ解析基盤を継続して整備し、ゲノムを基軸とした大規模・網羅的な生体分子ネットワーク解析により、生命プログラム及びその多様性を理解するために必要となる、大規模生体分子ネットワーク解析による脂肪細胞組織の刺激応答の網羅的解析とその応用のための研究開発を実施する。

また、「戦略課題４：大規模生命データ解析」の研究を行う上で、関連する研究者と必要な協議等を行うとともに、本格実施に必要な研究体制の整備を行う。

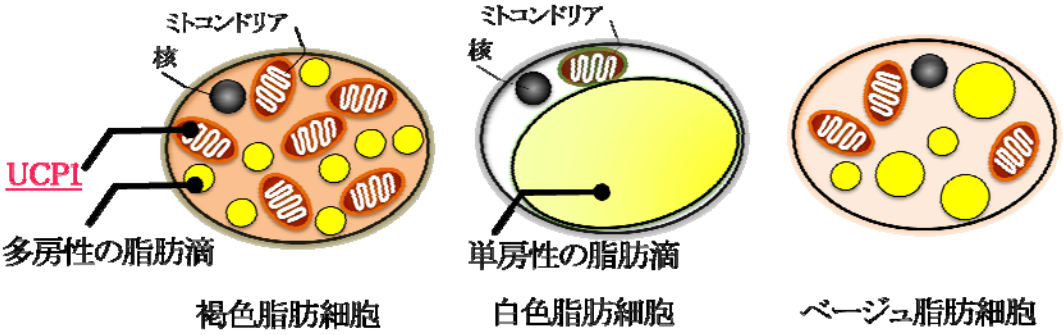
平成 25 年度は、種々の脂肪細胞組織が持つエネルギー貯蔵と高い熱産生能という相反する多面性を、刺激応答に対する脂肪細胞内部の生体分子の経時的変化のデータから、平成 24 年度までに開発した大規模生体分子ネットワーク解析のソフトウェアを用いて、大規模かつ網羅的に生体分子ネットワークを解析することで、脂肪細胞が状態を変化させエネルギー消費に向けて働く機構と生活習慣病の改善との関係を探る。

Ⅳ－３－２　実施内容（成果）

（１）マウス脂肪細胞組織の寒冷刺激下での時系列遺伝子発現プロファイルの追加取得

本研究では、マウスを 4℃の低温環境下で飼育する寒冷刺激を加えたときの、精巣部由来の白色脂肪細胞、鼠蹊部由来の白色脂肪細胞、肩甲骨周辺に由来する褐色脂肪細胞の３種類の脂肪細胞組織の応答の違いに着目している。

褐色脂肪細胞は寒冷刺激の有無によらず脱共役タンパク質 UCP1 が高発現することで高い熱産生を行う（下図左）が、精巣部由来の白色脂肪細胞は脂肪を貯蔵するのみ（下図中）で寒冷刺激に応答しない。これに対して、同じ白色脂肪細胞であっても鼠蹊部由来の白色脂肪細胞は寒冷刺激に応答し、ミトコンドリアが増えることで細胞の褐色化が起こり、褐色脂肪細胞のように高い熱産生を行う細胞（白色と褐色の中間的な性質を持つことからベージュ脂肪細胞と呼ばれる）に変化する（下図右）。

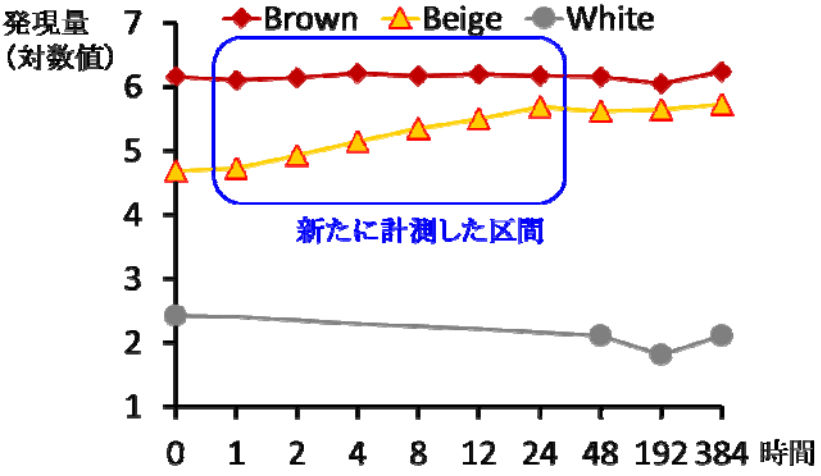


平成 24 年度に、これらの３種類の脂肪細胞組織について、寒冷刺激前、刺激後 48 時間、192 時間、384 時間の４時点での RNA 検体を抽出し、Agilent SurePrint G3 Mouse 8x60K マイクロアレイを用いて時系列発現プロファイル（各時点３レプリケート）を取得していた。平成 25 年度は、刺激直後の遺伝子発現量の変化を調べるため、褐色脂肪細胞とベージュ脂肪細胞の２種類の細胞について、追加で寒冷刺激後１時間、２時間、４時間、８時間、12 時間、24 時間の６時点の時系列遺伝子発現プロファイル（各時点３レプリケート）を同じマイクロアレイを用いて取得した。

これらの時系列発現プロファイルにより、３種類の脂肪細胞の寒冷刺激下での UCP1 の発現

変化を右図上に示す。図では3レプリケートの発現量の平均値を取っている。

図で褐色脂肪細胞（Brown）と白色脂肪細胞（White）については寒冷刺激を加えてもUCP1の発現量はあまり変化しないが、ベージュ脂肪細胞(Beige)では寒冷刺激によりUCP1の発現量が大きく上昇していることがわかる。しかも平成25年度に新たに取得した時系列発現プロファイル（右図の青枠で囲んだ部分）により、UCP1の発現誘導は特定の時点で生じるのではなく、刺激に応じて約24時間かけて徐々に起こっていることがわかる。UCP1の発現誘導は、白色脂肪細胞からベージュ脂肪細胞への褐色化を引き起こしていることから、褐色化とそれに伴う熱産生能の獲得が比較的長い時間をかけて徐々に起こっていることはこれまでに報告されておらず、褐色化の機構を解明する上で有益な手掛かりになると考えられる。



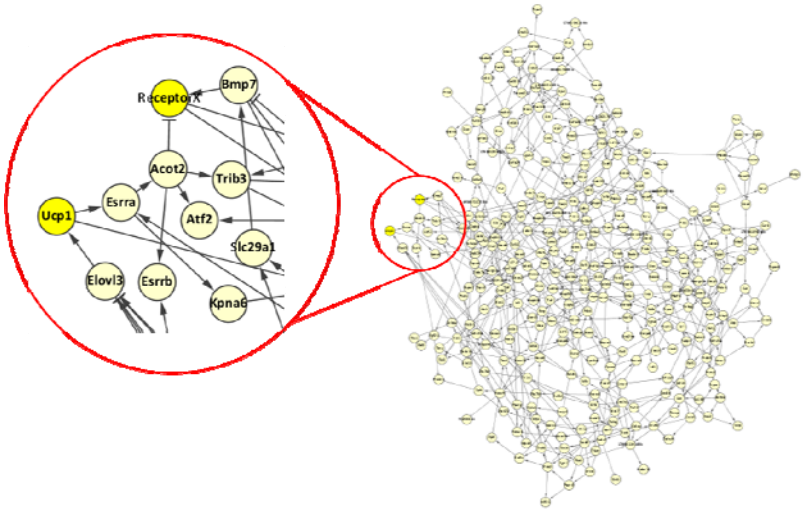
寒冷刺激を加えたときの3種類の脂肪細胞でのUCP1の発現量の経時変化

（2）生体分子ネットワーク解析によるベージュ脂肪細胞への褐色化に関連する新たな機構の発見

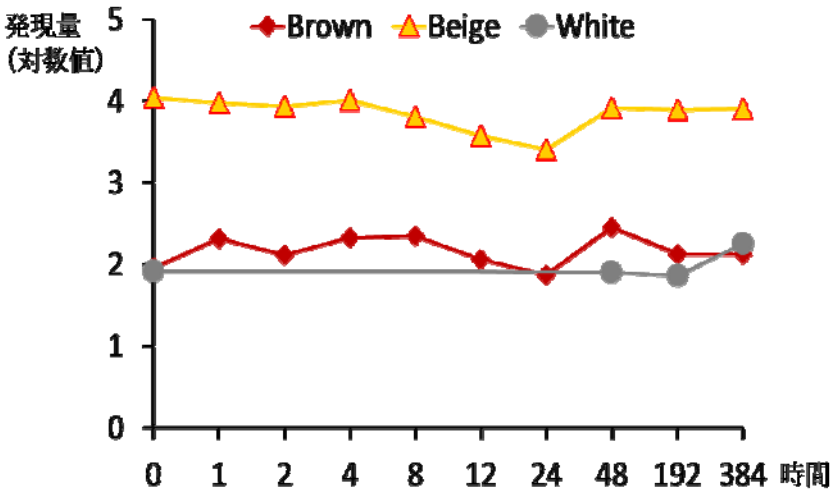
（1）で取得した、寒冷刺激下でのベージュ脂肪細胞への褐色化時の時系列発現プロファイルから、平成24年度までに開発したネットワーク解析ソフトウェアを使って、「京」上で生体分子ネットワーク解析を行った。

得られたネットワークを右図中に示す。赤い丸で囲った拡大図で示されているように、図のネットワークではUCP1の周囲に、これまでベージュ脂肪細胞との関連が報告されていなかった炎症性サイトカインの受容体の遺伝子が現れた（現在、論文準備中のため、右図中では遺伝子名を伏せてReceptor Xと表記している）。

この遺伝子の発現量の経時変化を3種類の脂肪細胞について調べたところ、右図下に示すように、この遺伝子はベージュ脂肪細胞で特異的に



寒冷刺激下での時系列発現プロファイルの時点数を10時点に拡大したときのベージュ脂肪細胞での生体分子ネットワーク

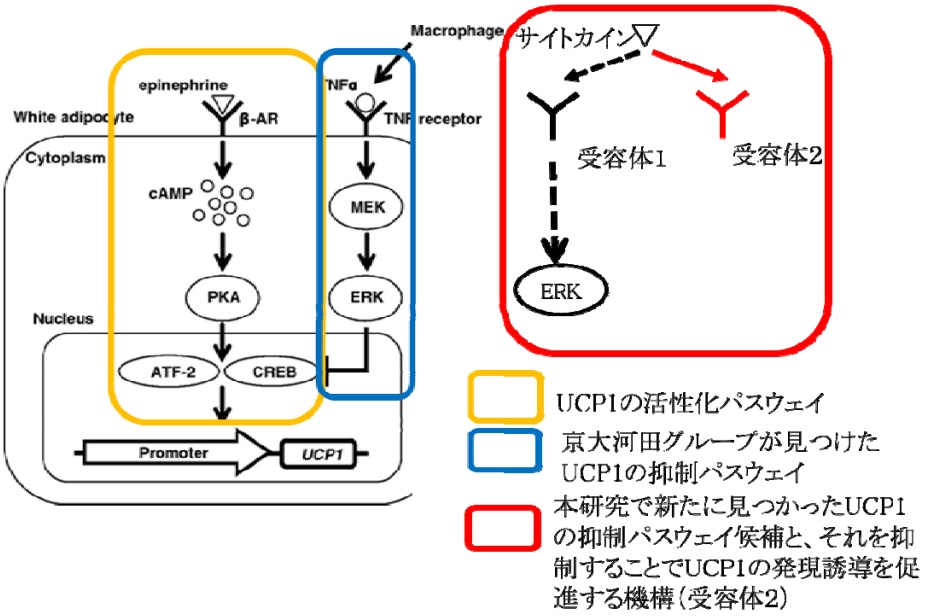


寒冷刺激を加えたときの3種類の脂肪細胞でのサイトカイン受容体遺伝子の発現量の経時変化

発現していることがわかった。さらにこの遺伝子の産物である受容体タンパク質の構造を調べたところ、このタンパク質はリガンドである炎症性サイトカインと結合する受容体ドメインは持っているものの、結合したときにシグナルを伝えるシグナル伝達ドメインを欠いていることがわかった。この遺伝子には、同じサイトカインをリガンドとしてシグナル伝達ドメインを持つ相同遺伝子が存在し、その産物である受容体タンパク質（以下、受容体１と表記する）はリガンドであるサイトカインが結合すると炎症反応を促進することが知られている。一方、前述のベージュ脂肪細胞で特異的に発現する ReceptorX 遺伝子の産物である受容体タンパク質（以下、受容体２と表記する）はシグナル伝達ドメインが欠失しているため、サイトカインと結合しても炎症反応を促進せず、返って受容体１とサイトカインを取り合うことで受容体１がサイトカインと結合することを妨げ、結果的に炎症反応の進行を阻害することが示唆された。

本研究の研究協力者である京大の河田教授のグループは、別の炎症性サイトカインである TNFαの受容体が TNFαと結合することで、UCP1 の発現が抑制されることを発見している（T. Sakamoto et al., Am J Physiol Cell Physiol, 2013）。これは TNFαと受容体が結合することで生じるシグナル伝達が、UCP1 の発現を制御する転写因子の働きを抑制するためであることが河田グループにより報告されている（下図の青線で囲まれた部分）。

今回見つかったベージュ脂肪細胞での受容体２の特異的な発現は、炎症性サイトカインが受容体１と結合することを妨げていることが示唆される。サイトカインが受容体１に結合すると、TNFαの受容体と同様のシグナル伝達が生じて UCP1 の発現が抑えられることが推測される。寒冷刺激を与えても精巣部由来の白色脂肪細胞では UCP1 の発現誘導が見られない原因として、このような抑制パスウェイの関与が示唆される。一方、鼠蹊部由来の白色脂肪細胞で寒冷刺激により UCP1 の発現が誘導され、褐色化してベージュ脂肪細胞に変化することに受容体２が関与していることが検証できれば、これまで未解決だった特定の白色脂肪細胞だけが寒冷刺激により褐色化する機構の解明につながると期待される。現在、河田グループにより検証実験が進められている。



脂肪細胞の褐色化を促進する新たな機構

（３）microRNA の時系列発現プロファイルの取得と microRNA と mRNA のネットワーク解析

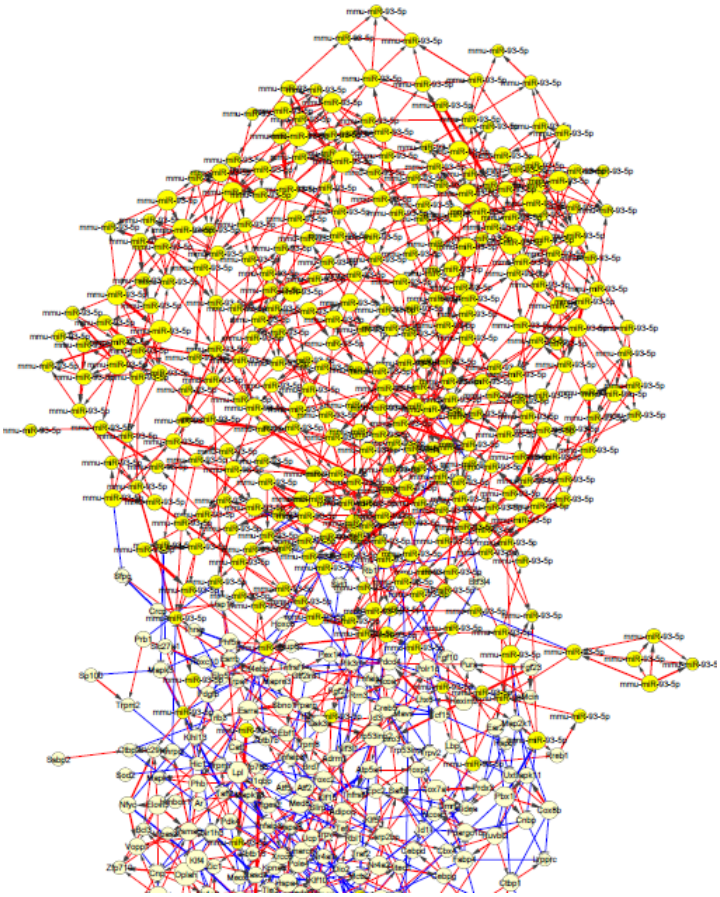
近年、microRNA が脂肪細胞での転写制御に重要な働きをしているという報告が相次いでいる。例えば、最近の総説（McGregor and Choi, Current Molecular Medicine, 2011）では脂肪細胞での転写制御において 22 種類もの microRNA の関与が報告されており、これ以降も多数の新たな microRNA の関与の報告がされている。

（１）で述べたように、本研究ではマウスの褐色脂肪細胞とベージュ脂肪細胞の 2 種類の細

胞組織について、寒冷刺激前と、寒冷刺激後の数時間おきに RNA 検体を取得している。この検体を microRNA のアレイにかけることで、寒冷刺激下でのマウス脂肪組織の時系列発現プロファイルを取得した。

具体的には、Agilent Expression Array Mouse miRNA 8x60k Rel.19.0 を用いて、褐色脂肪細胞とベージュ脂肪細胞の 2 種類の細胞組織について、寒冷刺激前と、寒冷刺激後 1 時間、2 時間、4 時間、12 時間、24 時間の計 6 時点で microRNA の時系列遺伝子発現プロファイル（各時点 3 レプリケート）を取得した。

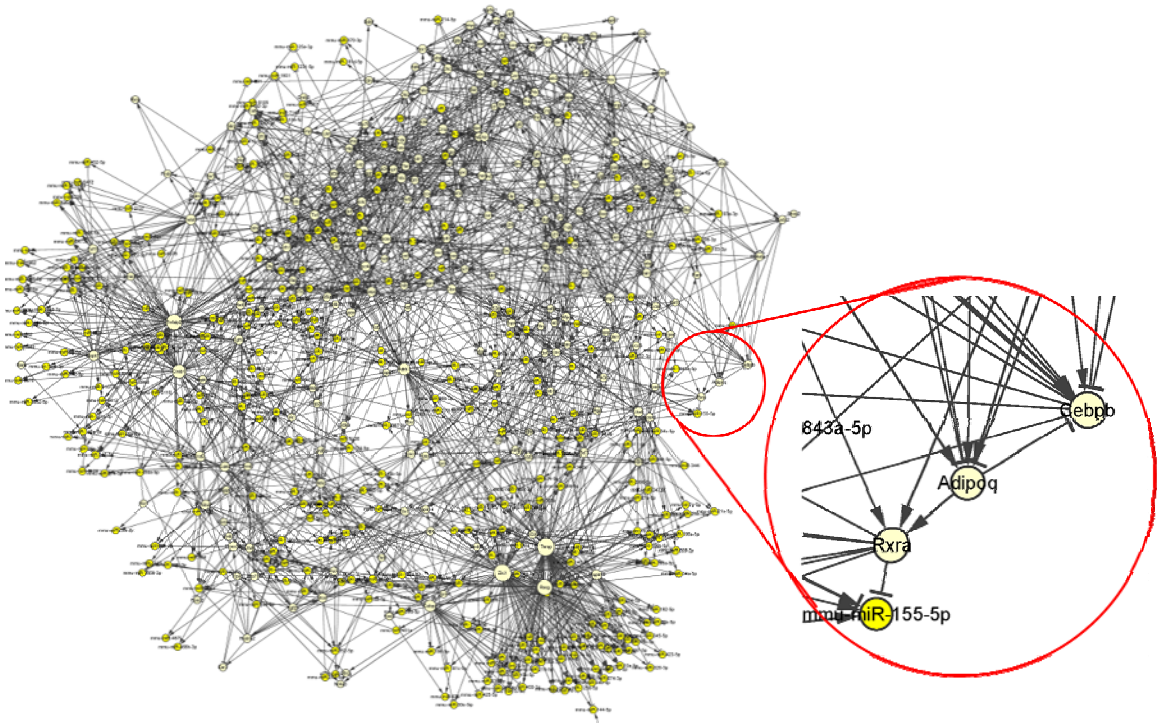
これと、（1）で示した mRNA の時系列発現プロファイルを併合することで、microRNA と mRNA のネットワーク解析を「京」上で行った。しかし、単純に microRNA と mRNA の発現プロファイルを併合してネットワーク解析を行うと、右図のように microRNA のネットワーク（図の上半分の黄色でマークしたノードからなるネットワーク）と、mRNA のネットワーク（それ以外のノードからなるネットワーク）が分離されてしまった。これは、microRNA の発現量が mRNA と比べて微弱であるため、発現レベルで microRNA と mRNA が分かれてしまったためと思われる。本来は、microRNA が mRNA の転写を制御している関係を解析したいのだが、このままだとそのような制御関係を見ることができない。



そこで、東大宮野グループの玉田助教らの協力を得て、ネットワーク解析ソフトウェアで使っているダイナミックベイジアンネットワークモデルの事前確率の設定を調節し、microRNA 間の制御関係の事前確率を 0.01 に抑えるとともに、microRNA と mRNA 間の制御関係の事前確率を 0.5 に設定することで、microRNA と mRNA 間の制御関係を優先してネットワーク解析するようにした。

ベージュ脂肪細胞の 6 時点の microRNA と mRNA の時系列発現プロファイルで、事前確率を設定してネットワーク解析を行った結果を次ページに図で示す。事前確率の設定により、microRNA のノード（黄色でマーク）と mRNA のノード（黄色でマークされていないもの）の間の制御関係が多数出現するようになり、多数の microRNA が mRNA の発現を調節に関与しているようすが見て取れる。例えば、赤い丸で囲んだ部分では、脂肪細胞の分化を制御する転写因子である C/EBPβ（赤丸内の一番右）の周辺に microRNA の一つである miR-155（赤丸内の左下の黄色でマークしたノード）が来ているが、最近、miR-155 が C/EBPβを制御することで褐色脂肪細胞とベージュ脂肪細胞の分化を調節しているという報告（Y. Chen, et al., Nature Communications, 2013）があり、それと整合した結果となっている。

今後は、事前確率の設定の最適化を進め、microRNA が UCP1 など重要なマーカー遺伝子の発現調節に関与している可能性を探るとともに、得られた解析結果について河田グループの協力を得て検証実験を行う予定である。


843a-5p
Cebpb
Adipoq
Rxra
mmu-miR-155-5p