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、複数のがん種にわたって、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、並びにがん組織のゲノム異常を中心に解析し、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化させる。そして、その理解に基づき、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。平成 25 年度に行った研究により、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の網羅的な解析が京コンピュータの利用により現実のものとなった。この成果をさらに発展させ、平成 26 年度は、承認された「京」の計算及びストレージ資源の範囲内で、以下の研究を実施する。また、「京」で対応できない部分は、東京大学医科学研究所ヒトゲノム解析センタースーパーコンピュータシステム等の計算資源で補う。

（１）複数のがん種にわたり、最大で数万のがん臨床検体の遺伝子発現プロファイルデータにより構築した計 2 万以上の遺伝子を含む 500 以上の遺伝子ネットワークのデータベースを構築していく。これにより、世界最大規模で、遺伝子ネットワークとして、がんのシステム異常を網羅的に明らかにしていく。このデータベースは、様々ながん種についての個々のがん研究とがんの臨床検体を遺伝子ネットワークでつなぐ機能を果たす。

（２）がん細胞株の遺伝子発現プロファイルデータと薬剤感受性、コピー数異常や DNA などの変異から、がん細胞株の薬剤感受性の違いを遺伝子ネットワークの活性の違いとして抽出し、その活性を制御している変異を明らかにしていくことにより、がんがどのようなシステム異常により薬剤耐性を獲得しているかを大規模データから明らかにしていく。

（３）mRNA に加え機能性 RNA（miRNA、lincRNA）を含む 5 万ノード以上の大規模遺伝子ネットワーク解析を実施する。lincRNA は非常に多彩であるため miRNA, mRNA と合わせると 5 万種

類以上の因子の発現制御の推定が必要となっている。これにより、網羅的に機能性 RNA を含めたがんのシステム異常の本態を解明していく。

（４）ゲノムシークエンス、RNA シークエンス、網羅的なメチル化データの統合により、一塩基置換や short indel、Internal Tandem Duplication、fusion gene などの変異に加え、splicing 異常や over-expression など転写異常を引き起こす後天的変異を調べる。これにより、より複雑ながんのシステム異常の原因と実態を様々ながん種について明らかにしていく。また、甲状腺がんについてはバイオマーカーが見つかっていないが、悪性度の高い未分化甲状腺がんについては、極めて貴重な検体（全部で１７検体）のうち１０検体について全ゲノムシークエンスデータ等が得られており、このデータの大規模解析により未分化甲状腺がんを特徴づける異常を追求していく。

（５）がんは細胞のゲノムに変異が蓄積し増殖能力が高いものが進化的に選択された結果生じる。この進化の過程で様々なクローンが生み出され一つの腫瘍内においてゲノムレベルのヘテロ性を生み出していると考えられている。このヘテロ性ががんの薬剤耐性など悪性度に深く係っていると考えられている。この腫瘍内ゲノムヘテロ性を明らかにするために、一腫瘍内の複数の領域から DNA をサンプリングし，次世代シークエンサーを用いてシークエンスする multiregional sequencing が試みられている。このような研究によって、がんの進化の早い段階で得られたと考えられるすべての領域に共通して観察されるファウンダー変異が存在する一方で、すべての領域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高いヘテロ性を生み出していることが明らかになった。しかしながらこのような腫瘍内ゲノムヘテロ性を生み出す機構の詳細はいまだ明らかになっていない。そこで、この腫瘍内ゲノムヘテロ性を生み出す機構を解明するために、1 細胞を一つの agent とする agent based model を構築し、細胞を増やしながら腫瘍が成長する様子を大規模シミュレーションにより解析し、腫瘍内ヘテロ性を生み出す原理を解明していく。数十年の時間の中でヘテロ性を生み出しているがんの進化を「京」で数時間でシミュレーションすることが可能になれば、個々の患者さんにおいて前がん状態から、今後どのようにがんへと進展していくかについての知見が得られることなる。これは「がん組織」のシステム異常の解明というさらに複雑な課題への挑戦となるが、得られる知見は臨床応用へ反映されることが強く期待できる。

（６） 以上（１）から（５）の大規模データ解析を平成 25 年度に取得した甲状腺がんのゲノムデータ、業務協力者の有するゲノムデータ、遺伝子発現プロファイルデータ、Sanger Institute や、TCGA、CCLE など公共データベースにて公開されているゲノムデータ（数万検体規模）により実施する。

（７）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合わせて実施する。

「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２６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）複数のがん種にわたり、最大で数万のがん臨床検体の遺伝子発現プロファイルデータにより構築した計2万以上の遺伝子を含む500以上の遺伝子ネットワークのデータベースを構築していく計画であった。平成２６年度はこれを達成し、大規模・網羅的遺伝子ネットワークデータベースを整備した。具体的には、平成25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に加え、EGF関連1,520遺伝子リストを用いた256データセットを新たに構築し SiGN-BN HC+Bootstrap を用いて250データセット分の計算を「京」を用いて完了し、合計762個の遺伝子ネットワークの推定を完了した。256データセットのうち6個分に関しては「京」の利用制限である24時間以内に計算が終了しなかった分である。SiGN-BN HC+Bootstrap は１つのデータセットからリサンプルされたデータセットを10,000セット作成し, そこから10,000個のネットワークを推定することにより高信頼な遺伝子ネットワークを推定するものである．この並列化はこれまでこの 10,000 回の計算を並列で行う方法であるため、10,000並列（コア）分の計算リソースがある場合、ブートストラップ回数を減らしても、１回のネットワークの推定に必要な時間は変わらないことから、全体でも必要な計算時間は変わらない。10,000回の計算のそれぞれの計算時間はかなりばらつきが大きいことがこれまでの研究で判明しているが、「京」での実装上、そのうち1個でも24時間を超えるものがあれば全体の計算は失敗してしまう。24 時間以内に終わった計算結果を集めると、短時間で計算が終わった結果のみを集めることになり、結果に偏りが生じる。したがってこのままでは256データセットすべての計算を完了できない状態であった。この「京」の運用上の困難を乗り越えることが最初の重要な事項となった。そこで、平成26年度は SiGN-BN のネットワーク構造探索アルゴリズムの高速化のための新たなアイディアを出し、それに成功し、この運用上の困難を乗り越えることができた。その高速化の方法はスレッド並列化および内部設計の緻密な改良である。これまで１つのコアで１つのネットワークを推定していたが、これを１つの CPU（「京」の場合最大８コア）で１つのネットワークを推定する様に変更した。SiGN-BN の構造探索アルゴリズムは逐次性の高いアルゴリズムであるため効率の良いスレッド並列化は一般には困難である。そこで、逐次アルゴリズム内にあるわずかな並列性のある部分を解析して抽出し、それをひとつずつ並列化するという膨大な作業を行った。また計算したスコアを再利用する部分のアルゴリズムをより計算時間の高速なものに改良した。これらの改良によりテストデータで51分59秒かかっていた計算が、2スレッドでは45分17秒に、4スレッドでは30分15秒に、8スレッドでは18分13秒まで高速化することに成功した。このようにして24時間以内に終わらなかった6つのネットワークの計算をすべて完了させることに成功した。これにより目標数以上の768個すべてのネットワーク推定を完了した。

これにより、世界最大規模で、遺伝子ネットワークとしてのがんのシステム異常を網羅的に明らかにできるデータベースが整備され、様々ながん種についての個々のがん研究とがんの臨床検体を遺伝子ネットワークでつなぐ機能を果たすことが可能となった。

（２）がん細胞株の遺伝子発現プロファイルデータと薬剤感受性、コピー数異常や DNA などの変異から、がん細胞株の薬剤感受性の違いを遺伝子ネットワークの活性の違いとして抽出し、その活性を制御している変異を明らかにしていくことにより、がんがどのようなシステム異常により薬剤耐性を獲得しているかを大規模データから明らかにしていくことがこの項目の目標である。平成26年度は以下の成果を得た。

１）がん細胞株の遺伝子発現プロファイルデータと薬剤感受性・耐性を、個々の細胞株の遺伝子ネットワークとして捉えることを目標とした研究を実施した。一般に、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が、高次元データに対する有用性からさまざまな目的において標準的な方法として用いられている。しかし、遺伝子発現プロファイルデータのように数万次元という超高次元データのモデリングにおいては、通常の Lasso は変数選択の性能低下が著しく、それに伴い予測能力も低下し、多くの場合有効に働かないことが判明している。そこで、予測能力向上と変数選択の正確性を同時に高めることを目的に、Lasso タイプの正則化推定法の問題点を見出し、新たな統計解析手法を開発し、がん細胞株の薬剤応答性の予測モデルを構築した。これを「京」に実装し、世界最大規模の網羅的がんの薬剤感受性・耐性遺伝子ネットワーク解析を実施し、「個々人に対する」薬剤耐性・感受性バイオマーカーの推定と耐性・感受性を予測する世界最高精度の方法を開発した。この方法は、Garnett MJ, et al. Systematic identification of genomic markers of drug sensitivity in cancer cells. Nature. 2012 Mar 28; 483 (7391) :570-5 に発表された方法を凌駕することを確認した。さらに、がん種横断的遺伝子ネットワーク解析手法を構築しその有効性を証明した。

i) Statistical method for robust sample specific analysis

Shimamura et al. (2011) PLoS One にて提案された NetworkProfiler をロバスト化し、薬剤感受性の予測においても高い予測能力を達成した (Park et al., 2014, PLoS One)。Kernel-based Lasso タイプの正則化推定化法は、Lasso タイプの正則化推定化法にカネール関数を組み合わせることで局所推定となり、サンプルそれぞれに対する遺伝子ネットワークのモデリングが可能となる。しかしながら、通常の Lasso タイプの正則化推定化法は残差平方和に基づくため、結果が外れ値に大きく影響を受ける。したがって、データに外れ値の混入が避けられないようなハイスループットな計測、例えば、遺伝子発現アレイや RNA-seq によって計測された数万遺伝子に対する遺伝子発現プロフィールの解析において、予測能力と変数選択の正確性は大きく低下することが容易に想定される。しかしながら、データ解析を行う前に外れ値を同定し除去することは、データが高次元であること、およびサンプル数が次元数に比べて決して充足してはいないことから考えて容易ではない。そのため、がんの遺伝子発現プロファイルデータの解析では、解析結果が混入している外れ値から影響を受けにくい、いわゆるロバストなデータ解析手法の開発が必須となった。

そこで、主成分空間で計算されたマハラノビス距離を利用するという着想を得て、高次元ゲノムデータに混入している外れ値をコントロールし、外れ値に対してロバストな新しい Kernel-based Lasso タイプの正則化推定化法 (Robust kernel-based L1-type regularized regression: RKLRR) を開発することに成功した。

外れ値を意図的に混入したシミュレーションデータによりその性能を評価し、同時に Sanger Institute が公開している Cancer Genome Project のデータを解析した。既存手法である Elastic net (ELA)、NetworkProfiler (NP)、および本開発の成果である Robust kernel-based L1-type regularized regression: RKLRR (R) の性能を比較した。T.N は True Negative rate、T.P は True Positive rate であり、 P.E が予測の正解率の結果を表している。表１からわかるように、RKLRR が既存手法を上回る結果を示した。

まず、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、遺伝子数が 200 (1000 個の遺伝子中で分散が大きい 200 個の遺伝子) のもので、5, 10, 15, 20%のサンプルに$N(5,1)$と$N(5,5)$の外れ値を挿入した。実際の遺伝子発現データを模倣

するため、各遺伝子間には最大で相関係数が 0.5 となるような相関構造を仮定している。結果を表１と２に示す。

表１：人工データを用いたシミュレーションによる提案手法 RKLRR の性能評価

Results of simulation 1 with Outlier for $N(5,1)$

		Type 1			Type 2		
		T.P	T.N	P.E	T.P	T.N	P.E
5%	ELA	-	-	0.338	-	-	0.324
	NP	0.71	1.00	0.290	0.70	1.00	0.276
	R	0.71	1.00	**0.285**	0.70	1.00	**0.271**
10%	ELA	-	-	0.325	-	-	0.329
	NP	0.69	1.00	0.290	0.70	1.00	0.310
	R	0.69	1.00	**0.284**	0.70	1.00	**0.303**
15%	ELA	-	-	0.289	-	-	0.294
	NP	0.71	1.00	0.288	0.70	1.00	0.264
	R	0.71	1.00	**0.287**	0.70	1.00	**0.258**
20%	ELA	-	-	0.285	-	-	0.259
	NP	0.71	1.00	0.254	0.69	1.00	0.258
	R	0.71	1.00	**0.244**	0.69	1.00	**0.255**

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 RKLRR の性能評価

Results of simulation 2 with Outlier for $N(5,5)$

		Type 1			Type 2		
		T.P	T.N	P.E	T.P	T.N	P.E
5%	ELA	-	-	0.321	-	-	0.314
	NP	0.69	1.00	0.280	0.70	1.00	0.277
	R	0.69	1.00	**0.278**	0.70	1.00	**0.271**
10%	ELA	-	-	0.298	-	-	0.280
	NP	0.70	1.00	0.266	0.70	1.00	0.251
	R	0.70	1.00	**0.262**	0.70	1.00	**0.249**
15%	ELA	-	-0	0.261	-	-0	0.255
	NP	0.71	1.00	0.227	0.69	1.00	0.240
	R	0.71	1.00	**0.225**	0.69	1.00	**0.231**
20%	ELA	-	-	0.290	-	-	0.229
	NP	0.71	1.00	0.251	0.70	1.00	0.214
	R	0.71	1.00	**0.249**	0.70	1.00	**0.211**

開発した RKLRR を用いてサンガーセンターが公開している Cancer Genome Project のデータを解析した。承認薬、未承認の化合物を含む１０の化合物（FTI.277, DMOG, NSC.87877, AKT.inhibitor.VII, Midostaurin, BMS.754807, Thapsigargin, Bleomycin, Doxorubicin, Epothilone.B）を選び、その IC50 感受性の予測精度を評価した。予測精度の評価は、モデルを学習するためのトレーニングデータと予測するためのテストデータにデータを分割する標準的な方法を用いた。結果を表 3 に示す。

表３：サンガーセンターの Cancer Genome Project のデータによる提案手法 RKLRR の性能評価

Comparison of prediction accuracy of drug sensitivity

	FTI.277	DMOG	NSC.87877	AKT.inhibitor.VIII	Midostaurin
R	0.293	**0.220**	**0.162**	**0.177**	**0.120**
NP	0.291	0.239	0.211	0.232	0.134
Elastic net	**0.269**	0.561	0.323	0.447	0.477
	BMS.754807	Thapsigargin	Bleomycin	Doxorubicin	Epothilone.B
R	**0.099**	**0.120**	**0.044**	**0.153**	**0.621**
NP	0.124	0.131	0.049	0.182	0.725
Elastic net	0.720	0.274	0.279	0.367	0.954

前出のシミュレーション結果と同様に、開発したRKLRRを既存手法と比較した。１０の化合物中９つの化合物において最も高い性能を達成できた。一つの化合物に対しては、RKLRR法が最良とはならなかったが、どの方法も一様に高い予測精度を示しており、この化合物については外れ値がそれほど致命的な影響を与えなかったと推察される。他の９つの化合物に対する結果においては、本手法が大幅に予測精度を改善できていることが分かる。特に、Garnett et al. (2012) Nature の用いた Elastic Net は、外れ値の影響を大きく受けており、本手法は大幅に精度を改善できていることがわかる。開発したRKLRRによって、薬剤感受性の予測性能が大幅に向上したといえる。この研究成果は、抗がん剤の使い分けに加え薬剤感受性に関連する遺伝子ネットワークも同時に推定しているため、新たな薬剤標的分子の同定にも繋がる可能性がある。

ii) Statistical method for identifying gene network via multi-omics data analysis

遺伝子ネットワークの解析は、がんのシステム異常を理解するために重要である。その目的のため、多様なオミクスデータを統合的に解析することが求められている。しかしながら、単純に複数のオミクスデータを重ね合わせただけでは、データの有する構造を有効に活用しがんのシステム異常を捉えることは難しい。例えば、パスウェイの情報を用いるとき、ある遺伝子は複数のパスウェイに属していることがあり、そのような遺伝子をデータ解析において理論的に整合性のある統計学的理論により取り扱わなければ有効な情報抽出は出来ない。そこで、このマルチオミクスデータにおける遺伝子ネットワーク推定問題において、ネットワークの構造学習とネットワークにおける遺伝子選択が同時に可能な新しい統計解析モデル（Sparse Overlapping Group Lasso: SOGL）を開発した（Park et al. (2014) J Comp Biol）。タンパク質間相互作用(Protein-Protein Interaction; PPI) の情報を用いて遺伝子ネットワークを構成する場合には、それぞれの部分PPIネットワーク間に遺伝子の重複が生じる。そこで、我々は、拡張されたデータ空間を定義し潜在変数（latent variable）に基づいた遺伝子ネットワーク推定法を考案し、遺伝子ネットワーク推定と遺伝子選択を同時に実現する統計科学的データ解析手法 SOGL を開発した。

開発した SOGL 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、既存手法である Latent Group Lasso と比較した結果、変数選択の正確性において既存手法よりも高い性能を有していた。

このSOGLを用いてThe Cancer Genome Atlas（TCGA）において公開されているマルチオミクスデータから５つのがん（lung squamous cell carcinoma (LUSC), kidney renal papillary cell carcinoma (KIRP), ovarian serous cystadenocarcinoma (OV), brain lower grade glioma (LGG), head and neck squamous cell carcinoma (HNSC)）の解析を行った。まず、各がんにおいて活性を持っている発現モジュールを同定するために、新井田等の開発したExtract Expression Module法(EEM:「京」に実装済)を用い５つのがんの発現データから有意に共発現している発現モジュールを抽出し、mutation status, TUSON list gene set, copy number variation, expression levels, protein-protein interaction の５種類のオミクスデータに基づいて遺伝子をフィルタリングし（表４）、100組のブートストラップデータに対してSOGLによる回帰モデリングを行うことで、遺伝子ネットワーク推定と遺伝子選択を同時に行った。

表４：Multi-Omics data に基づいた遺伝子フィルタリングとネットワーク構成

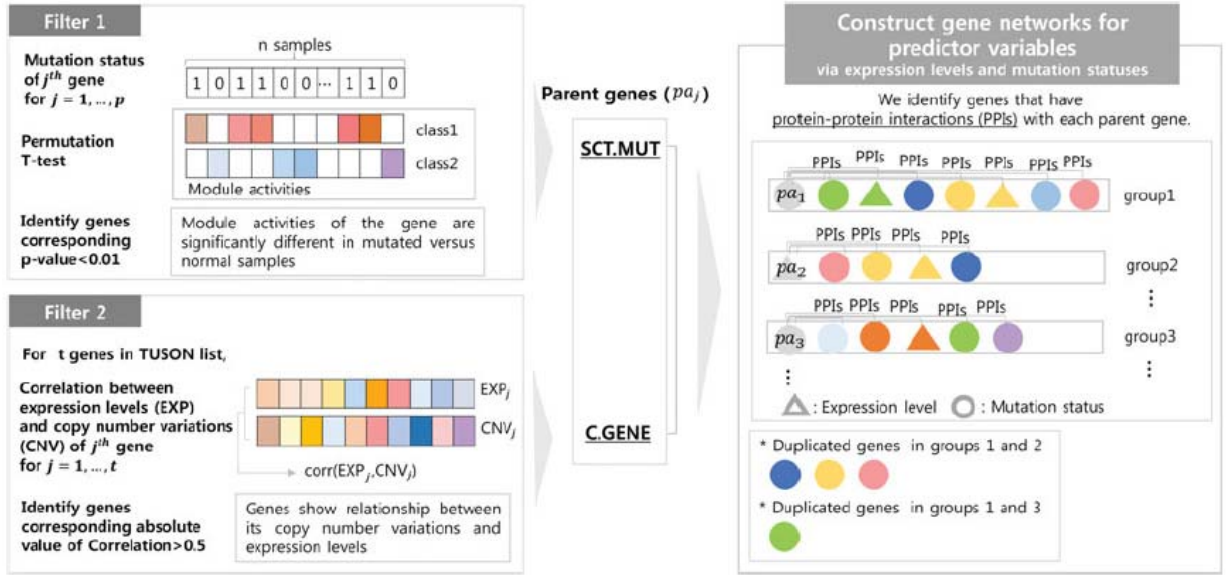

表４で示されたプロセスによって構成された遺伝子ネットワークに基づき、発現モジュールの活性を予測した際の性能を評価した結果、および Gene Set Enrichment Analysis (GSEA) (http://www.broadinstitute.org/gsea/index.jsp)を用いた発現モジュールの解析結果を表５に示している。開発した SOGL は、各がんにおいて活性を持っていると予測される発現モジュールの活性予測において、既存手法よりも高い性能を有していることが示された。

SOGL により構成された５つのがんにおいて活性を持っていると予測される発現モジュールと、発現モジュールに対してその活性を予測する能力を有していると SOGL により同定された遺伝子や変異からなるネットワークが図１である。図１で表示された遺伝子は 100 回のブートストラップに基づく回帰モデリングにおいて 50 回以上選択されたものである。LUSC がんのネットワークにおいて選択された NFE2L2（細胞の生存に重要な防御メカニズムを持ち、がんの予防と治療に重要な役割を担っている）と KEAP1（肺がんでの hypermethylated gene）は、タンパク質間相互作用があり、さまざまな文献で重要ながんドライバー遺伝子として知られている。このような従来の重要な知見との合致も見られ、今後、このネットワークの更なるがん生物学的解析によって新しい発見に繋がることが期待される。

表５：TCGA データにる提案手法 SOGL の性能評価と発現モジュールの Annotations

Module	LUSC		KIRP		OV		LGG		HNSC	
	SOG.La	LG.La	SOG.La	LG.La	SOG.La	LG.La	SOG.La	LG.La	SOG.La	LG.La
1	# 241		# 207		# 253		# 108		# 236	
	Immune System		Immune System		Immune System		Up-regulated in neurons		Immune System	
	0.165	0.177	**0.257**	0.258	**0.189**	0.197	**0.180**	0.181	**0.229**	0.233
2	# 74		# 79		# 81		# 98		# 122	
	Cell Cycle, Mitotic		Cell Cycle, Mitotic		Cell Cycle, Mitotic		Immune System		Cell Cycle, Mitotic	
	0.073	**0.073**	**0.181**	0.183	**0.083**	0.085	**0.142**	0.143	**0.114**	0.119
3	# 50		# 62		# 41		# 97		# 79	
	Beta1 integrin cell surface interactions		Genes annotated by the GO term GO:0005739. A semiautonomous, self replicating organelle that occurs in varying numbers, shapes, and sizes in the cytoplasm of virtually all eukaryotic cells. It is notably the site of tissue respiration.		Cytokine Signaling in Immune system		Cell Cycle, Mitotic		Extracellular matrix	
	0.183	0.188	**0.125**	0.126	**0.131**	0.135	**0.150**	0.151	**0.195**	0.201
4	# 19		# 44		# 27		# 38		# 13	
	Peptide chain elongation		Genes up-regulated in DAOY cells (medulloblastoma) upon knockdown of BMI1 [GeneID=648] gene by RNAi.		Beta3 integrin cell surface interactions		Interferon gamma signaling		Peptide chain elongation	
	0.197	0.201	**0.314**	0.319	**0.348**	0.375	**0.220**	0.225	**0.331**	0.353
5	# 13		# 38		# 21		# 20		# 12	
	RNA Polymerase I Promoter Opening		Respiratory electron transport, ATP synthesis by chemiosmotic coupling, and heat production by uncoupling proteins.		Respiratory electron transport, ATP synthesis by chemiosmotic coupling, and heat production by uncoupling proteins.		Translation		RNA Polymerase I Promoter Opening	
	0.452	0.464	**0.111**	**0.111**	**0.081**	0.083	**0.258**	0.263	**0.571**	0.598
6	# 8		# 17		# 18		# 8		# 11	
	Genes annotated by the GO term GO:0007398. The process whose specific outcome is the progression of the ectoderm over time, from its formation to the mature structure. In animal embryos, the ectoderm is the outer germ layer of the embryo, formed during gastrulation.		Genes down-regulated in epithelial lung cancer cell lines upon over-expression of an oncogenic form of KRAS [GeneID=3845] gene and knockdown of TBK1 [GeneID=29110] gene by RNAi.		Peptide chain elongation		Interferon Signaling		Genes annotated by the GO term GO:0007398. The process whose specific outcome is the progression of the ectoderm over time, from its formation to the mature structure. In animal embryos, the ectoderm is the outer germ layer of the embryo, formed during gastrulation.	
	0.165	0.179	**0.136**	0.137	**0.196**	0.205	**0.307**	0.311	**0.188**	0.191
7	# 7		# 12		# 13		# 7		# 10	
	Genes up-regulated in MEF cells (embryonic fibroblasts) with knockout of NFE2L2 [GeneID=4780] gene.		RNA Polymerase I Promoter Opening		RNA Polymerase I Promoter Opening		RNA Polymerase I Promoter Opening		Interferon alpha/beta signaling	
	0.139	0.141	**0.379**	0.389	**0.472**	0.491	**0.694**	0.744	**0.361**	0.383
8			# 6		# 4		# 5		# 8	
	-		ER-Phagosome pathway		Genes down-regulated in MCF-7 cells (breast cancer) positive for ESR1 [GeneID=2099] MCF-7 cells (breast cancer) and long-term adapted for estrogen-independent growth.		Cell surface interactions at the vascular wall		Metabolism of xenobiotics by cytochrome P450	
			0.245	0.251	**0.354**	0.382	**0.290**	0.294	**0.268**	0.271
9			# 5						# 6	
	-		Generic Transcription Pathway		-		-		Genes up-regulated in DAOY cells (medulloblastoma) upon knockdown of BMI1 and PCGF2 [GeneID=648] [GeneID=7703] genes by RNAi.	
			0.088	0.089					**0.148**	0.150

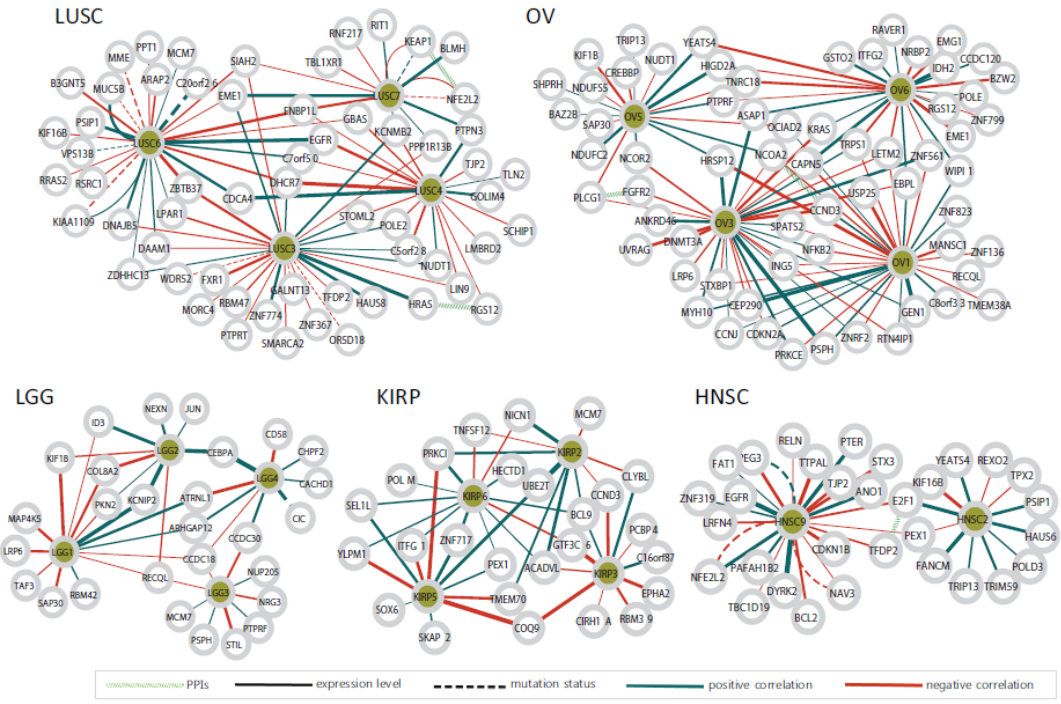

図１：SOGL により構成された遺伝子ネットワークと選択された遺伝子

（３）microRNA に比べて非常に多彩である lincRNA を加えた大規模遺伝子ネットワーク解析を行った。業務協力者である名古屋大学大学院医学系研究科の島村徹平特任准教授、高橋隆教授と梶野泰祐特任助教との共同で、がん遺伝子として知られる MYC 遺伝子の転写活性を制御するモジュレータ因子の網羅的探索に関する研究を開始した。現在、MYC などの転写因子、ターゲット遺伝子、モジュレータ因子の３項関係を数理モデル化し、観測された遺伝子発現データからモジュレータの制御調節の変化を推定する GIMLET (Genome-wide Identification of Modulators with Local Energy statistical Test) と呼ばれる統計的手法を開発中であり、「京」に実装している。フィジビリティスタディとして、公開データベースの 40 研究から 7,988 サンプルの肺腺がん患者の遺伝子発現プロファイルデータを収集し、「京」上で GIMLET を実行し、MYC のモジュレータの網羅的探索の一部が完了した。これまでに、MYC の転写活性を制御する新規ノンコーディング RNA（MYMLR）を同定し、生物学的実験においても興味深い結果が得られつつある（図２）。「京」の計算時間を使い切ったため、その後の解析は平成 27 年度に行う予定である。

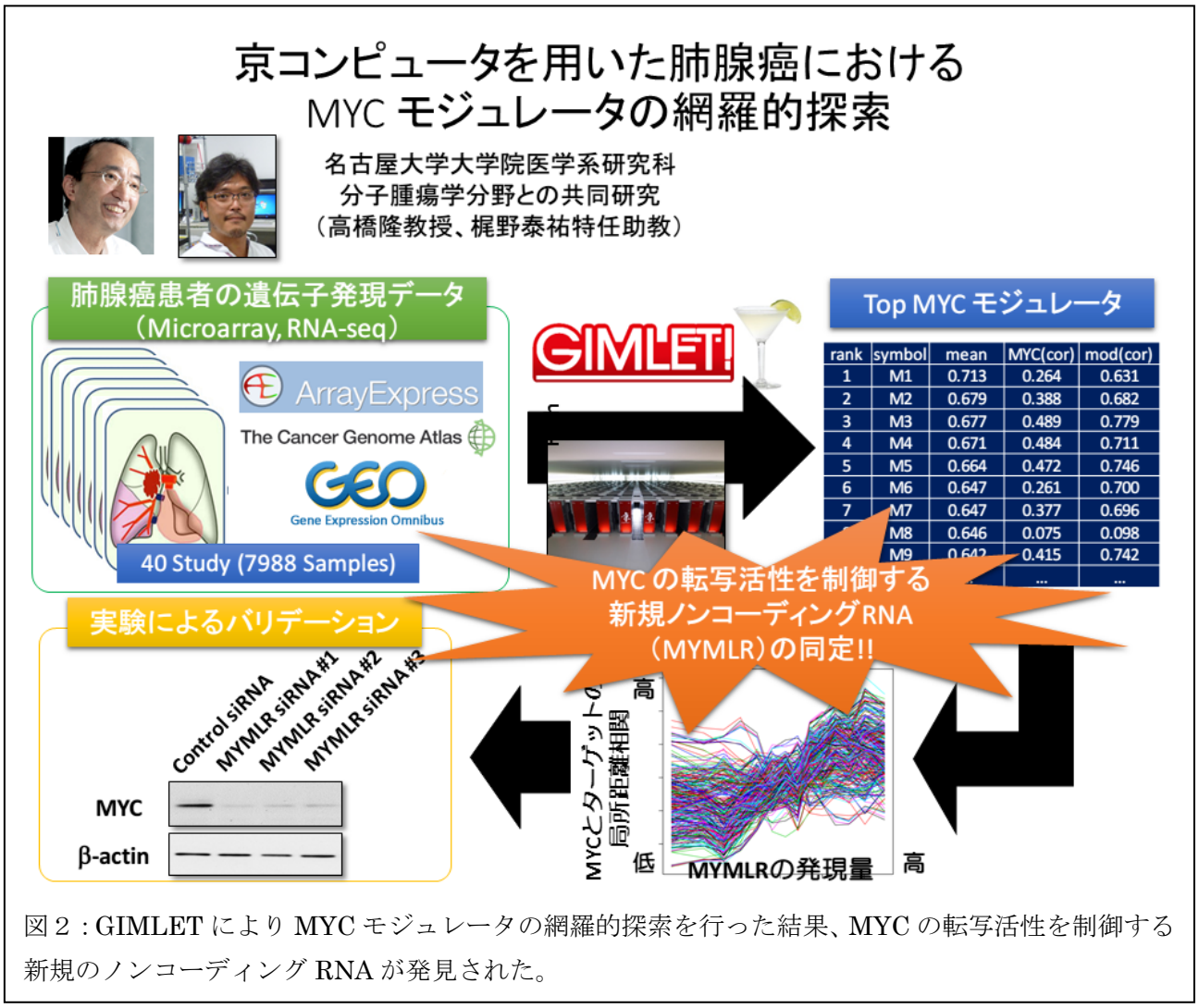

図２：GIMLET により MYC モジュレータの網羅的探索を行った結果、MYC の転写活性を制御する新規のノンコーディング RNA が発見された。

（４）ゲノムシークエンス、RNA シークエンス、網羅的なメチル化データの統合により、一塩基置換や short indel、Internal Tandem Duplication、fusion gene などの変異に加え、splicing 異常や over-expression など転写異常を引き起こす後天的変異を調べるためのツールの開発とデータ解析を実施した。

1）Genomon-fusion という RNA シークエンス解析（融合遺伝子探索パイプライン）の「京」への実装を進めた。平成 25 年度は、Genomon-fusion パイプラインの中で、計算時間の大部分を占めるアラインメント部分について「京」への移植（Genomon-Fusion for K (GFK)）を行った. 米国 Broad Institute から公開されている Cancer Cell Line Encyclopedia (CCLE) の 780 検体の解析を実施し、「京」による大規模解析の有効性の確認ができたが、本年度は新たに判明した問題点に対応するべく、主に以下の 2 点の開発を行った.

1．Blat のスレッド並列化による高速化

CCLE780 検体のアラインメントに要した CPU 時間はおよそ 100 万ノード時間である。一方、データベースに登録される検体数は増加の一途をたどっている。例えば、TCGA（The Cancer Genome Atlas, http://cancergenome.nih.gov/）には RNA シークエンスだけで 2 万検体以上のデータが登録されている。これらの解析を実施し、「融合遺伝子のランドスケープ」を描き出すには、2,500 万ノード時間以上の「京」のリソースが必要であり、事実上その全部の解析は不可能である。この問題を解決するため、アラインメント部分の高速化を行った。

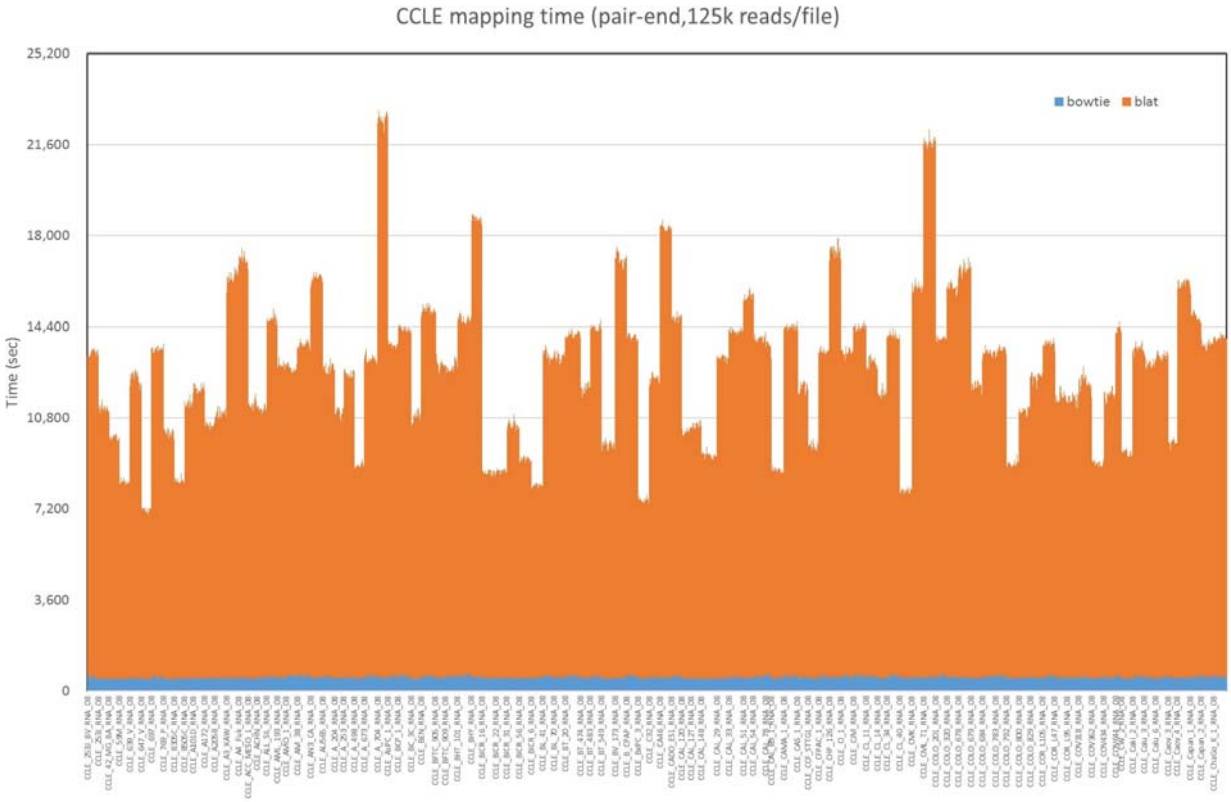

図３：CCLE100 検体の京コンピュータ上でのアラインメント時間

図 3 に CCLE100 検体の「京」における計算時間を示している。図 3 からわかるように、計算時間の 90%以上（橙色部分）を blat が占めている。そのため計算時間の短縮には blat の高速化が必要である。高速化には様々なアプローチがあるが、バイオインフォマティクス分野で注意しなければならないのは、ソフトウェアの更新頻度が非常に速い点である。速いものでは一か月程度で新しいバージョンが登場する本分野において、半年～1 年単位の時間をかけることも珍しくないレジスタ・キャッシュ等に合わせたマシンごとの最適化（チューニング）は不適切である。

ソフトウェアに対する修正作業を最小化しつつ十分な高速化を達成する方法として、今回は OpenMP によるスレッド並列化を採用した。OpenMP はスレッド並列を実現する一般的な規格で

あり、現在普及しているコンパイラの多くが対応している。もちろんであるが、「京」に搭載された富士通コンパイラも OpenMP に対応しており、利用を推奨している。また、OpenMP はホットスポット（演算負荷の高い部位）のループ計算に対してコンパイラ・ディレクティブを挿入するだけでループを自動的にスレッドに分割する。Blat は 1 プロセスあたりのメモリ使用量が非常に多い（6GB 程度）。そのため、「京」では 1 ノードあたり 2 プロセスしか割り当てができず、8 コアのうち 6 コアが未使用の状態であった。スレッド並列により、この遊んでいたコアに計算させることでリソースの有効活用にもなる。

ホットスポット特定のため、blat の実行時間測定を行った。その結果、処理時間の 90%以上を占める主要演算部分（図４）を特定することができた。

```
static struct dnaSeq seq;
struct lineFile *lf = lineFileOpen(fileName, TRUE);
while (faMixedSpeedReadNext(lf, &seq.dna, &seq.size, &seq.name))
    {
      searchOneMaskTrim(&seq, isProt, gf, outFile,
                        maskHash, &totalSize, &count);
    }
lineFileClose(&lf);
```

図４：Blat の主要処理部分のソースコード

処理内容を見ると、ファイルからリードデータを読み込み（faMixedSpeedReadNext）、それをマッピングしファイルに書き込む（searchOneMaskTrim）という処理を繰り返している。スレッド並列化は

- 読み込み部分の分離とメモリ展開化
- 出力ファイルをスレッドごとに独立ファイル化

することにより図５のように書き換え可能である。

```
#pragma omp parallel private(i, ii, thread_num)
{
   thread_num = omp_get_thread_num();

   #pragma omp for  // Mail loop
   for( ii = 0; ii < lcount; ii++ ) {

   searchOneMaskTrim(&seq[ii], isProt, gf, outFile[thread_num],
                     maskHash, &totalSize, &count, thread_num);
   } // End of for
} // End of parallel region
```

図５：OpenMP 化した blat 主要部分のソースコード

メモリ展開化に関して補足すると、Genomon-fusion では blat の前に別のアラインメントソフトウェアを掛けることで、既知のリード情報をスクリーニングしてデータ量を削減している。本研究で開発したロードバランシング機構の効果と合わせると、インプットデータの容量は 9Mbyte 程度に抑えられており、スレッドごとにデータを分割してメモリ展開が可能なデータ量になっている。スレッド並列化した blat の実行時間測定結果を図 6 に示す。計測に用いたインプットデータは CCLE の RNA シークエンスデータ（CCLE_253J_BV_RNA_08）を 613 分割したうちの一つである。データサイズは約 84Mbyte、26,111 リードの fasta 形式データとなっている。計測は「京」および Shirokane1（ヒトゲノム解析センターのスーパーコンピュータ）の 1 ソケット 1～4 コアを用いた。また、各マシンのスペックを表 6 に示す。4 スレッドでの計算で、「京」では約 2.4 倍、Shirokane1 では 2.8 倍程度の高速化に成功した。本高速化作業に要した時間はおよそ一か月であり、非常に効果的に高速化を実現できたといえる。

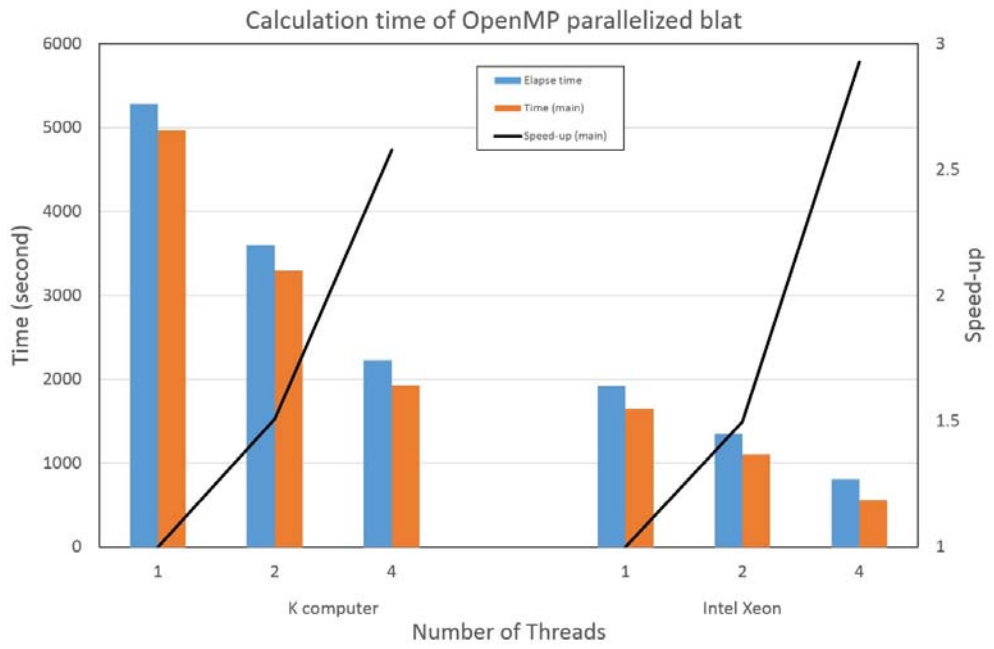

図 6 ：Blat の計算時間

表 6 ：テスト環境の仕様

	京コンピュータ	Shirokane1
CPU	SPARC64 VIIfx 2.0 GHz 8 core/socket, 1 socket/node 128Gflops/socket	Intel Xeon E5450 3.0 GHz 4 core/socket, 2 socket/node 48Gflops/socket
RAM bandwidth	64 GB/s/socket	10.66GB/s/socket
ファイル I/O (IOR) *	600 MB/sec/socket	750 MB/sec/socket
ファイル I/O (original) **	130 MB/sec	240 MB/sec
運用開始	2012 年	2009 年

* Corresponding value from actual measured value of IOR benchmark

** Actual measured value by “time dd if=/dev/zero of=tempfile bs=1M count=10000”

２．パイプライン全体の移植

二つ目の問題はデータ移動である。CCLE 等のデータベースからデータをダウンロードするには専用ソフトを使う必要がある。その専用ソフトウェアを「京」上で利用できなかったため、当初はヒトゲノム解析センターのマシン上に一旦ダウンロードしてから「京」への転送を行っていた。しかし､数十～数百 TB にもなるデータを中継ダウンロードしていたのでは時間がかかりすぎるため，「京」上で専用ソフトを利用できるように環境整備（ソフトウェアビルド環境や P2P ソフトウェア利用の運用機関からの許諾取得など）を行った。

残る問題は解析結果の転送である。融合遺伝子検出にはアラインメントの結果ファイルに対して PCR 重複除去を実施したうえで検出パイプラインにかける必要がある。この処理には外部ソフトウェアを複数利用しており、「京」上でコンパイルできない、もしくは、ソースが公開されていないなどの理由で実行できない部分が多かった。そのため、インプットファイルより大きな結果ファイル（SAM ファイル）をヒトゲノム解析センターのスパコンへ転送して検出を行っていた。

平成 26 年度は未移植部分を含めたパイプライン全体の移植を行い、最後まで「京」上で解析できるようにした。移植に際し問題になったのは以下の 3 点である。

- 計算ノードで JAVA が使えない。
  Samtools で必要な処理を代替した。
- Samtools が動かない。
  ソフトウェアの詳細な解析の結果、データ型定義部分のバグであることを突き止め、正しい型に修正し、富士通コンパイラで動くことを確認した。
- CAP3 が動かない。
  De novo assembly を行う本ソフトウェアはソースコードが非公開であり、「京」に対応するバイナリが公開されていない。そのため、同様の機能を持つオープンソースソフトウェア「SOAP denovo-Trans」で代替し、周辺プログラムを調整した。

これらの対応により、パイプライン全体を「京」で動かすことが実現した。完成したパイプラインのフローチャートを図 7 に示す。Genomon-fusion 全体を 3 つのパートに再構成し、それぞれを GFKalign, GFKdedup，GFKdetect としている。平成 25 年度に移植した部分は GFKalign に相当する。GFKdedup および GFKdetect のインプットファイルは、それぞれ GFKalign，GFKdedup の結果をインプットとするように設計されており、順番にジョブを投入することにより最終結果である fusion.txt ファイルを得られるようになっている。fusion.txt は数十 Kbyte 程度であり、転送は一瞬で終わるサイズである。各ステップはすべて MPI による並列化により、多数検体の同時処理が可能な実装となっており、各検体ごとに fusion.txt ファイルが得られる仕様である。パッケージ化して公開する予定である。

２）未分化甲状腺がんのゲノム解析

甲状腺がんのほとんどはヒトのがんのなかで最も予後のよい分化型甲状腺がんであるが、まれにヒトのがんのなかで最も予後の悪い未分化型甲状腺がんに転化する。分化型がんにさらなる遺伝子異常か蓄積 された結果、未分化転化が引き起こされていると考えられているがその詳細は明らかになっていない。連携研究先である野口病院（大分県別府市）より甲状腺未分化がん組織 10 例、正常甲状腺組織 6 例提供を受け、全ゲノムシークエンスを行った。「京」を用いて全ゲノムデータ解析が進行中である（図８）。今後、更にサンプル数を追加し、また分化型がんゲノムの公開データ（TCGA）と比較することで、未分化転化の責任遺伝子異常の探索を行う予定である。

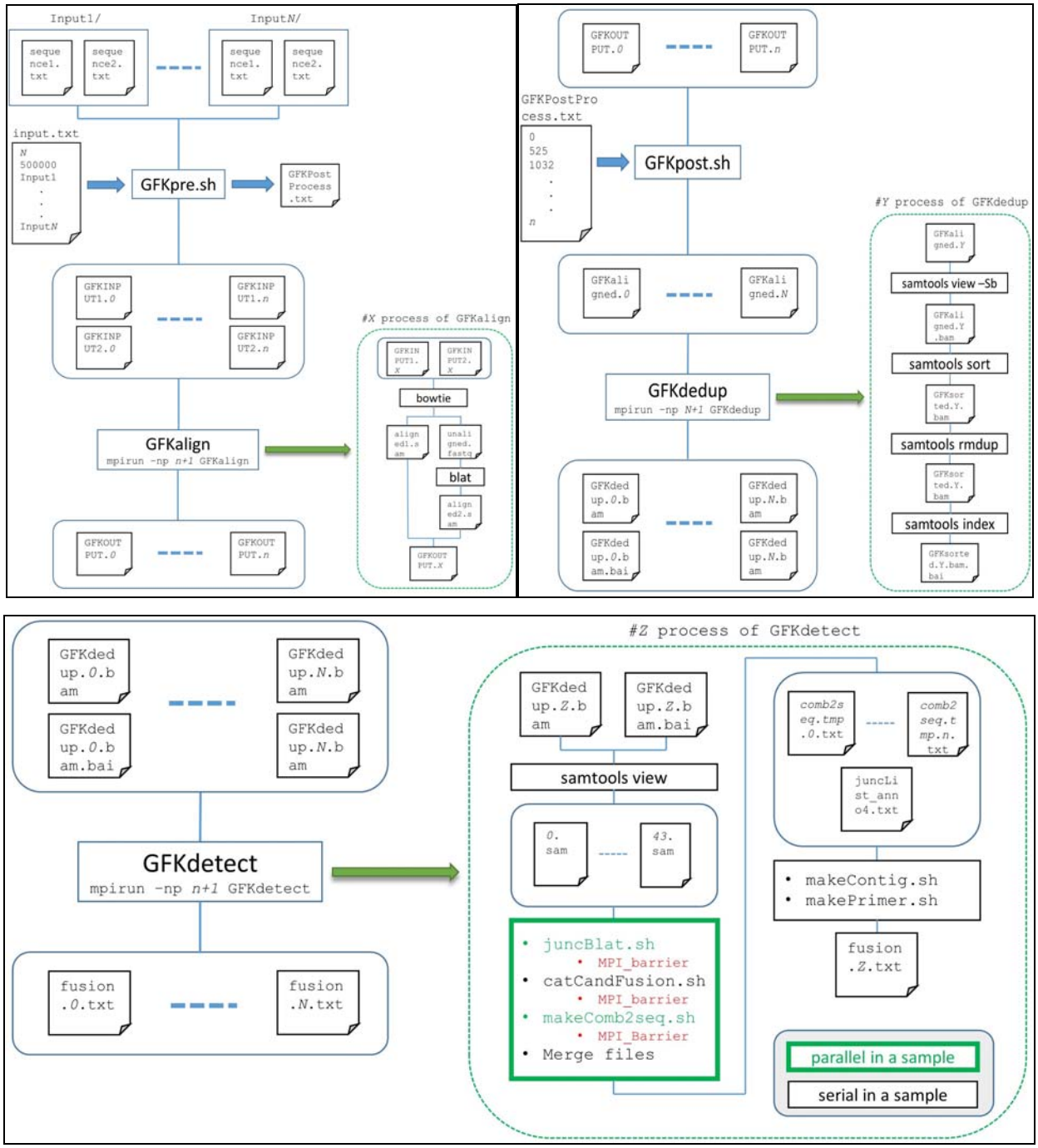

図７：GFKalign, GFKdedup, GFKdetect のフローチャート

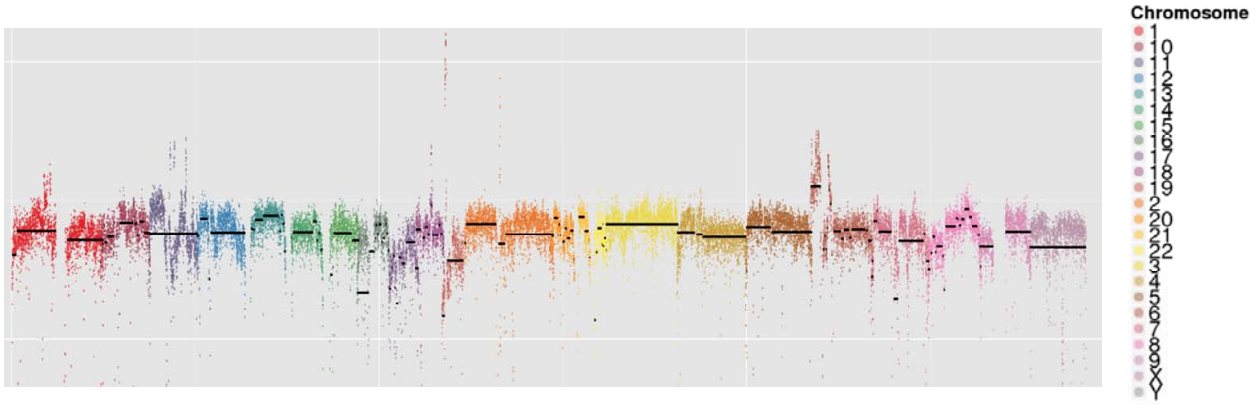

図８：「京」により明らかになった未分化甲状腺がんコピー数プロファイル

２）EEM を用いた MYC co-regulator の同定

平成 25 年度に発現モジュール同定手法 EEM 法の「京」への移植を行ったが、平成 26 年度は、この EEM を用いた大規模データ解析を行い、様々ながん種で遺伝子増幅しているがん遺伝子かつ転写因子、MYC の co-regulator の同定に成功した。はじめに業務協力者である三森功士教授が病院長を務める九州大学病院別府病院で取得した大腸がん 130 検体の遺伝子発現プロファイルデータに EEM を適用し MYC の転写標的遺伝子群、MYC 発現モジュールを同定した。MYC 発現モジュールは MYC の発現と有意に相関していたが、その相関は強くなく、他の co-regulator の存在が示唆された。またそれと並行して大腸がんゲノムコピー数データの解析を行い MYC 遺伝子座と 20 番染色体短腕（20q）の相関を見出した。これらの観察に基づき、20q に新規 MYC co-regulator が存在するという仮説をたて、20q の各遺伝子について MYC 発現レベルと相加的に MYC 発現モジュールにその発現が相関する遺伝子を多重回帰により探索した。その結果、発現パターンが強く相関し、細胞分裂を制御するタンパク複合体をコードする二つの遺伝子、AURKA 及び TPX2 を同定した。更に EEM を公開データベース NCBI GEO の様々ながん種を含む 257 発現データセット（計 15,360 臨床検体）に適用し、同様の相関関係が幅広い癌腫に存在することを確認した（図９）。最終的にこの大規模データ解析から得られたモデル（図１０）に基づいて九大別府病院において分子生物学的実験を行い、大腸がん細胞において AURKA 及び TPX2 が MYC と協働的に細胞増殖を制御することを確認しモデルの正しさを証明した。MYC の阻害剤は開発が難しく、MYC を標的とした分子標的医薬は存在しないが、本研究から AURKA/TPX２複合体の働きを阻害することで、MYC パスウェイを阻害できる可能性が示唆される。実際、MYC 高発現がん細胞株で AURKA 阻害剤が効率的に細胞増速を阻害することを実験により確認している。この成果は、Takahashi, Y, Sheridan, P, Niida, A, Sawada, G, Uchi, R, Mizuno, H, Kurashige, J, Sugimachi, K, Sasaki, S, Shimada, Y, Hase, K, Kusunoki, M, Kudo, S, Watanabe, M, Yamada, K, Sugihara, K, Yamamoto, H, Suzuki, A, Doki, Y, Miyano, S, Mori, M and Mimori, K, The AURKA/TPX2 axis drives colon tumorigenesis cooperatively with MYC, Ann Oncol, 2015 として発表した。

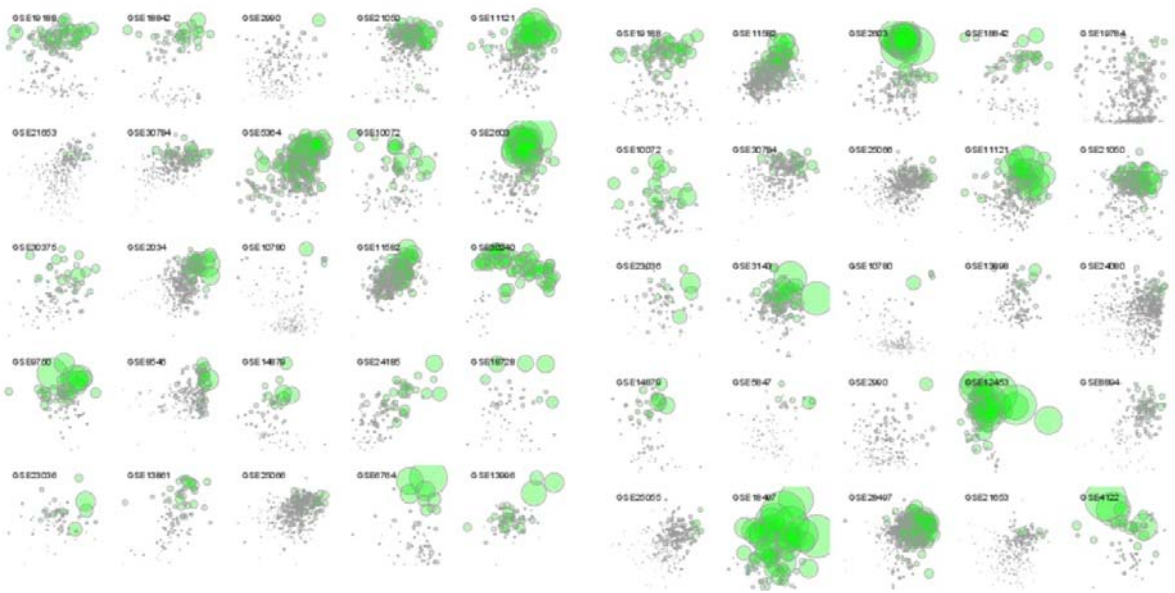

図９：様々ながん種で確認された MYC 遺伝子発現（縦軸）、TPX/AURAK 遺伝子発現（横軸）、MYC 発現モジュール活性（円のサイズ）の相関関係

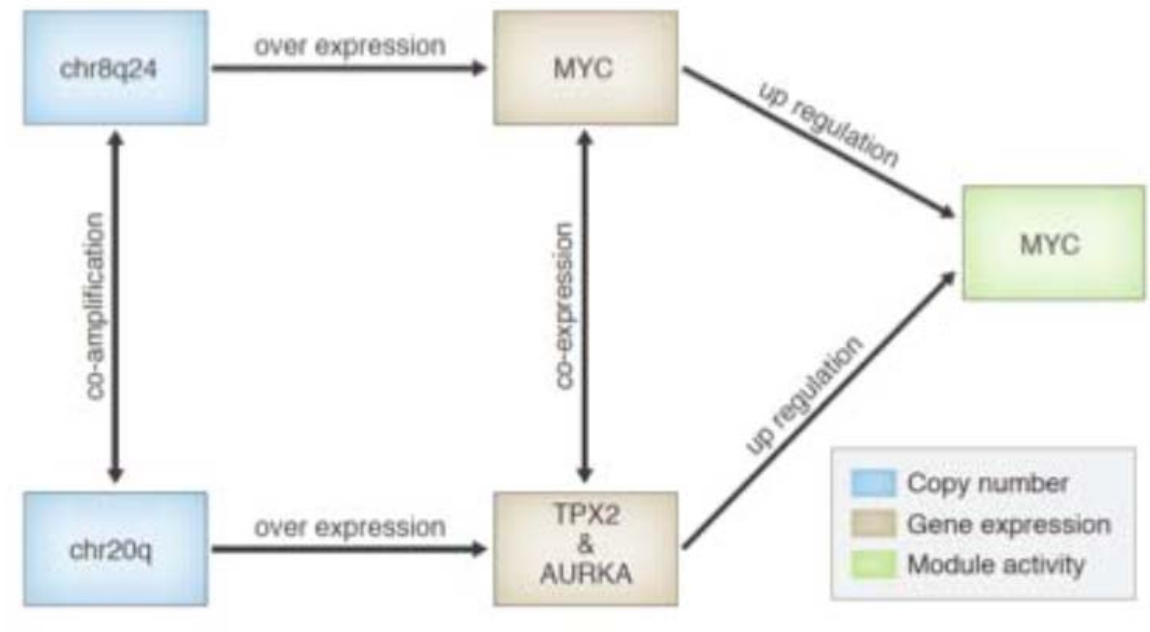

図１０：大規模データ解析により構築した MYC 及び TPX/AURAKA の co-regulation モデル

３）EEM を用いた胃がん腹膜播種制御モジュールの同定

胃がんは腹膜に転移（腹膜播種）することで最終的に人を死に至らしめるが、その分子機構の詳細はいまだわかっていない。業務協力者である三森功士教授（九大別府病院）のグループと共同で、胃がんから樹立した高腹膜播種能を獲得した細胞株と親株のマイクロアレイによる発現プロファイルの比較を行い、発現変化を示す腹膜播種制御因子群の探索を行った。さらに臨床的に意味のある遺伝子群にしぼりこむために EEM を用いて細胞株で発現変化を示す遺伝子群とシンガポール大学で取得した胃がん 200 検体の発現データセットを統合解析することにより腹膜播種制御モ

ジュールの同定に成功した。200 検体中、腹膜播種制御モジュールの活性が高い患者群は有意に腹膜播種が多く、生存期間も短かった（図１１）。更に九大別府病院で腹膜播種制御モジュール中の遺伝子 X について分子生物学的実験をし、腹膜播種を制御していることを確認している。成果は論文として投稿中である。

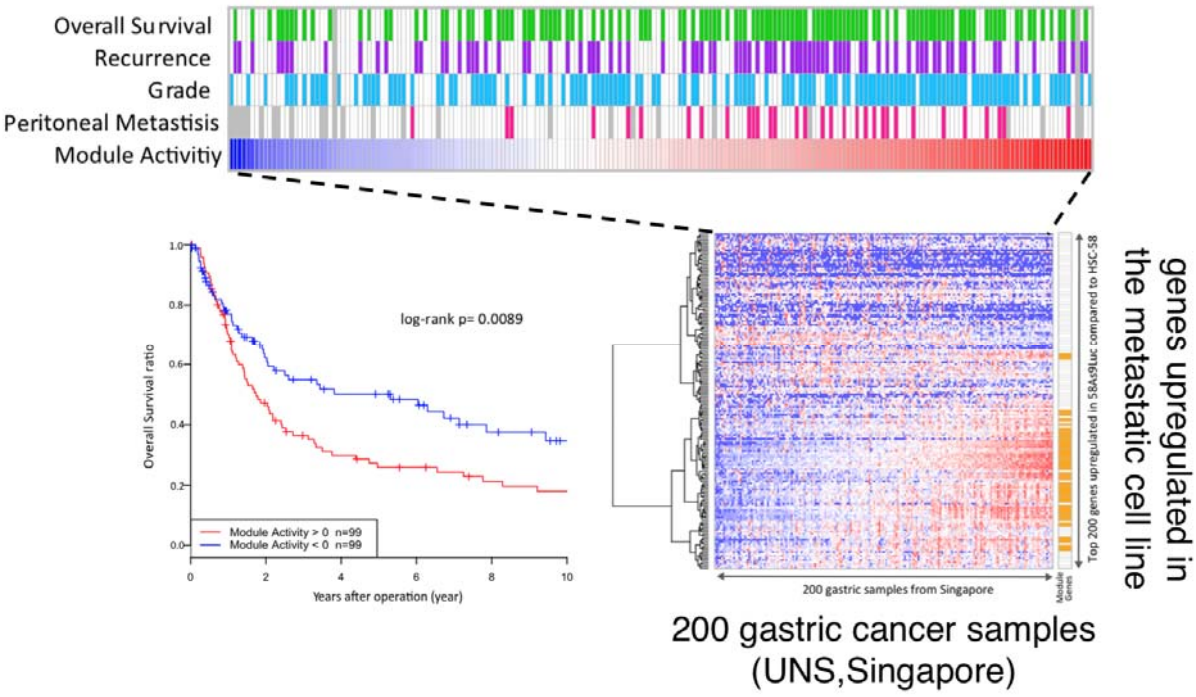

図１１：胃がん腹膜播種モジュールの活性と腹膜播種（Peritoneal Metastasis）及び予後の相関関係

４）p53 遺伝子に関する RNA シークエンスデータ解析

平成 26 年度は、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、同グループの有する遺伝子発現データを対象とし、大規模データ解析を遂行する上で必要となる、i）遺伝子発現シークエンスデータ（RNA シークエンスデータ)の新規解析手法の開発を行い、また ii）実データの解析および結果データの解釈法の開発検討を行った。

i) RNA-seq データからの遺伝子発現量精密推定法開発

具体的には、まず RNA シークエンスデータから遺伝子発現量を精密に推定するために、RNA シークエンスリードのクォリティ値などの様々な特徴量を反映することができる柔軟な統計モデルフレームワークを新たに構築した（図１２左)。さらに、そのフレームワーク上で、RNA シークエンスデータに系統的に含まれるバイアスを補正するために機械学習の一種である深層学習(deep learning)に基づく補正法を開発した（図１２右)。

まず上記のモデリングフレームワークにおいて、遺伝子発現量を推定するために、重み付き EM アルゴリズムを新たに提案し実装した。これにより様々なリード中の特徴量を柔軟に重みとして推定量に反映することができ、推定の精度が向上することが期待される。図１３に、あるシミュレーションデータに対して提案手法および既存の二つの手法（Cufflinks, TIGAR2）を適用して得た、各手法における遺伝子発現予測値と真値の散布図および相関係数を示す。我々の手法（図１３左上）が最も高い相関係数を得ていることがわかる。発現値の推定に影響を与えるリードの特徴量として、

リードのクォリティ値を例にとると、通常、シークエンサーから得られるリード群には、クォリティの高いものと低いものが混在している。このようなデータの解析時に、しばしば取られる方策として、クォリティ値に閾値(例: フレッドクォリティ値 30 以上)を設け、前処理として閾値以下の値を持つリード群を一律に解析から除外するというものがある。しかしながら、個々のデータセットは、様々なクォリティ値の分布を持ち、単一の閾値を設定し適用することが最善とは限らない。そこで、我々の提案したフレームワークでは、重み付き EM アルゴリズムにおいて正規化したクォリティ値を重みとして表現することにより、外的基準によって設定された単一の閾値に依存しない、それぞれのデータセットに適応的な情報抽出が可能となっている。その結果、遺伝子発現値の推定により多くのリードの情報を取り込むことができ、推定精度が向上すると期待される。

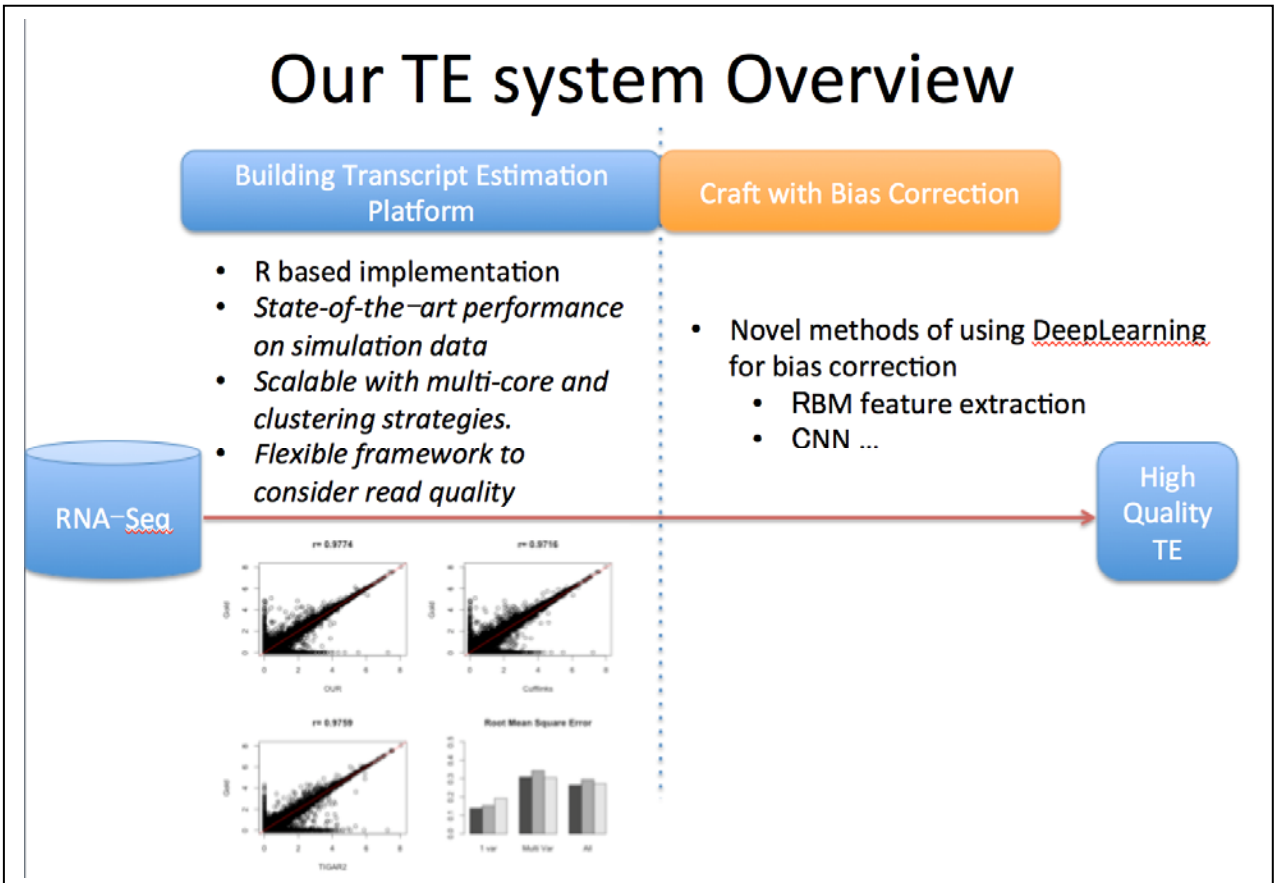

図１２：開発した重み付き EM アルゴリズムに基づく遺伝子発現量推定フレームワーク（左）および深層学習に基づくバイアス補正法（右）

次に、上記で述べた深層学習に基づくバイアス補正法について説明する。RNA シークエンス実験過程においては、最終的に発現推定値に影響を与えうる種々のバイアスが生じうる。通常、個々のバイアスを直接的に表現して補正することは困難であるが、我々の提案したフレームワークでは、バイアスに応じてリードの重みを調整することで、バイアスの補正を図ることができる。今期は、その一例として、まずサンプルライブラリ作成時に生じる、あるバイアスの補正法を考案した。そのバイアスとは、本来転写産物由来のシークエンスを得たい場合、リードの開始点は転写産物配列上の 5'端から 3'端まで一様に分布していることが期待されるが、実際の実験の過程では 6 塩基(hexamer)からなるランダムプライマーによるプライミングを行う際に、転写産物上のリード開始点の分布に不均一性が見られることに由来し、このバイアスは Hexamer Priming Bias（以後 HP バイアスと記述）と呼ばれイルミナ社のシークエンサーから得られた RNA シークエンスデータに

含まれていることが過去の文献で指摘されている。この事は、リファレンス配列上にアライメントされたリードの量に応じて発現量を推定する際に、リードの開始点を一様であると想定した推定方式では、推定値にバイアスを生じることを意味する。図１４に、実データから得られたリード上の位置ごとの塩基種の比率を示す。

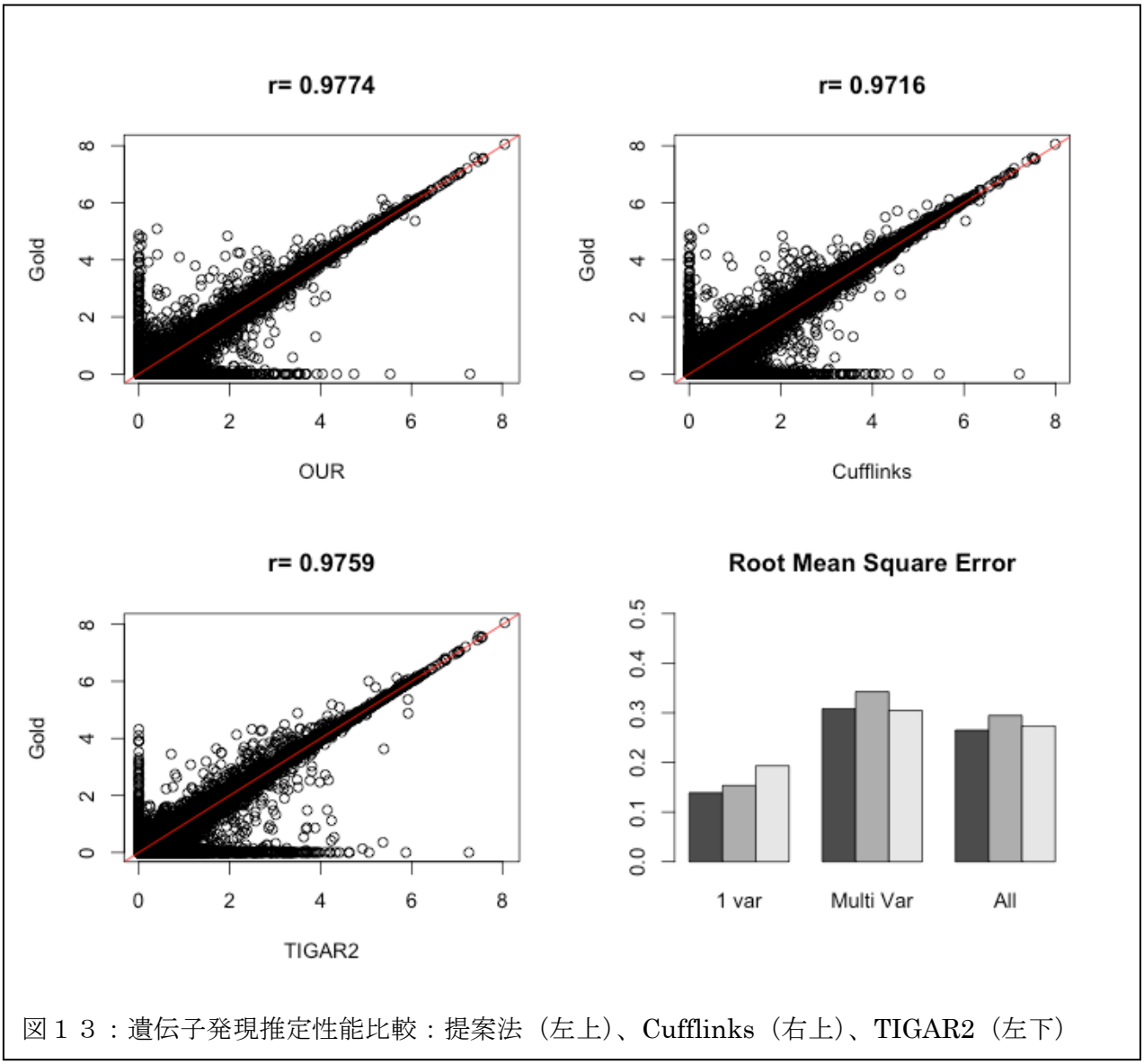

図１３：遺伝子発現推定性能比較：提案法（左上）、Cufflinks（右上）、TIGAR2（左下）

リード開始位置に近い左橋の約 14 塩基において比塩基種比率の非一様性が観測され HP バイアスがあることが確認される。この HP バイアスをモデル化するために、過去の文献では、リード開始位置前後 5 塩基ずつの幅をもったウィンドウ内の配列の分布を、リード配列中の先頭および中間点の配列における各 hexamer の頻度で特徴づける方式が取られていた。しかしこの方式では、ウィンドウ中の配列のパターン数（4^{10}=1,048,576 通り）に応じたパラメータの次元に比して、実際に観測されるパターン数が乏しいという問題が生じる。Cufflinks ではこの問題を、単純化された 3 次 Markov 連鎖を用いて、ウィンドウ中の配列分布を特徴づけるのに必要とされるパラメータ数を削減することで解決を図っている。一方、我々はウィンドウ中の配列分布を特徴づけるために必要な代表配列群を、Restricted Baltzmann Machine (RBM)を用いた深層学習機によるモデル化を通じて自動的に抽出する方式を提案した。図１５に、提案方式の概要を示す。RBM により、元々の配列を縮約表現し､縮約配列に応じた各リードの重みを計算することで HP バイアスの補正を行う。

既存の方式に比べて、より表現力の高い隠れ変数を持つ RBM を用いることにより、モデル化に必要なパラメータ数を縮減しつつ、各データに内在する代表配列群を自動抽出することができ、より精度の高いバイアス補正ができると期待される。現在すでに R による実装が済んでおり、「京」上での最適化を図るため、コアとなるアルゴリズムの最適化および C での実装を進めた。

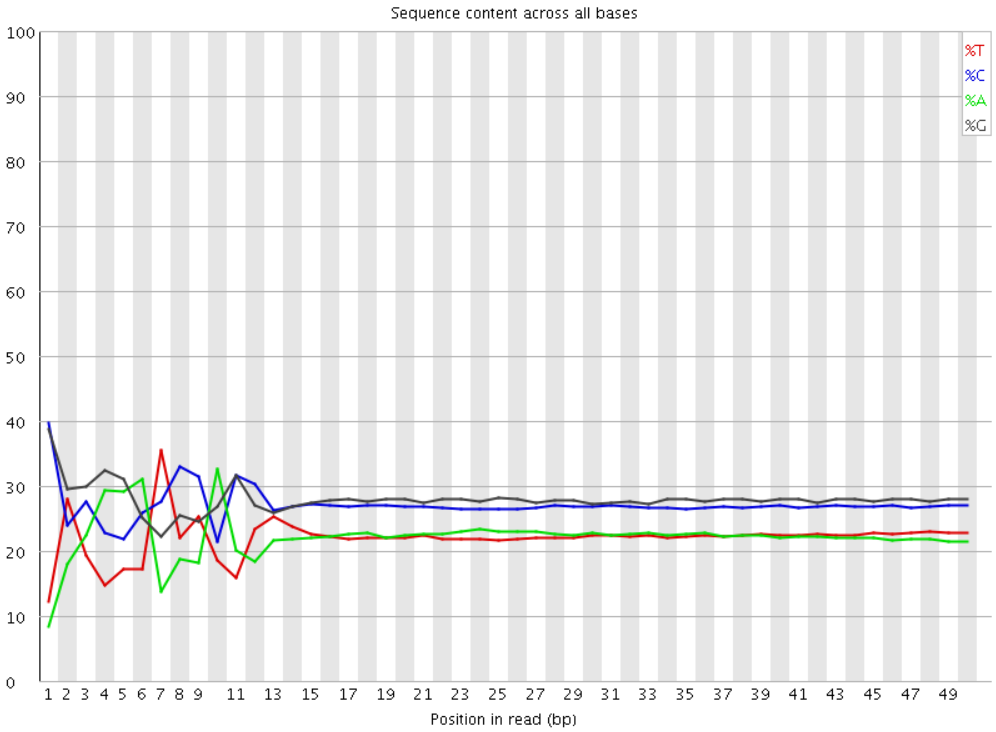

図１４：RNA シークエンス実データから得られたリードの塩基種分布に見られる HP バイアス。横軸：リード中の塩基位置座標。縦軸：比率（%）

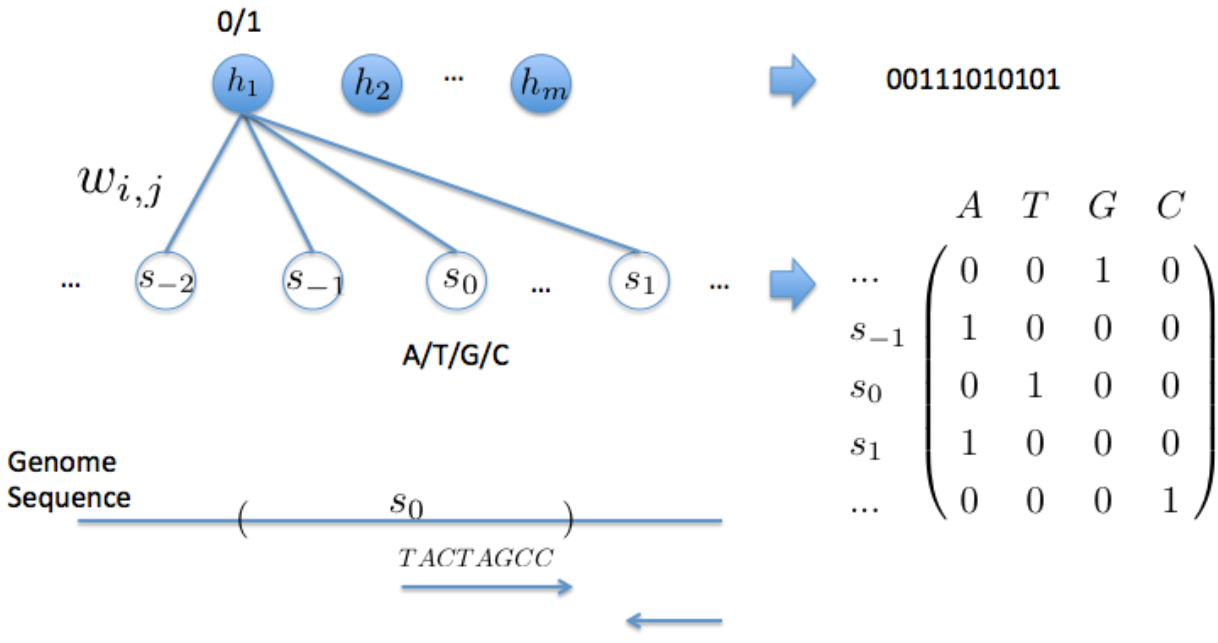

図１５：Restricted Boltzmann Machine (RBM)による、代表的塩基配列の自動抽出法

ii) RNA シークエンスデータからの p53 制御遺伝子群特徴抽出および結果解釈法の開発

平成 25 年度に続き、松田准教授グループの有する 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 シークエンスデータを対象として上記の開発手法の性能評価を進め、大規模データ解析から得られた知見に関して結果を比較することで検討を行った。また解析結果データからの解釈法の検討も行った。

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。そのため、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（24 臓器）かつ多条件下（WX: 野生型＋X 線刺激あり、W: 野生型＋X 線刺激なし、KX: p53 欠損+ X 線刺激あり、K: p53 欠損+ X 線刺激なし）から同時に得られた RNA シークエンスデータから有用な情報を抽出する方法は自明ではない。平成 26 年度は、臓器特異的に発現増加もしくは減少の見られている遺伝子群に着目し、情報抽出を行った。

まず RNA シークエンスデータから遺伝子発現量の推定を行い、下記の条件で臓器ごとに発現増加遺伝子群および発現減少遺伝子群を抽出した（図１６、１７）。

$$\frac{Avg(WX)+1}{Max(Avg(W),Avg(K),Avg(WX))+1} \geq 2 \text{ , for } \underline{up\text{-}regulated} \text{ genes.}$$

$$\frac{Avg(WX)+1}{Min(Avg(W),Avg(K),Avg(WX))+1} \leq 1/2 \text{, for } \underline{down\text{-}regulated} \text{ genes}$$

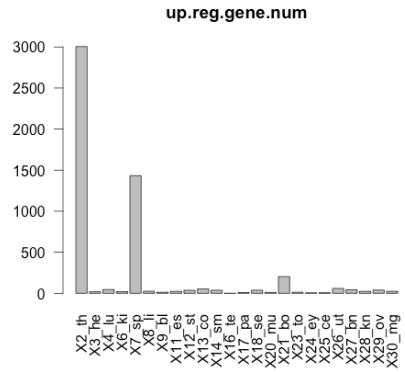

図１６：臓器別発現上昇遺伝子群

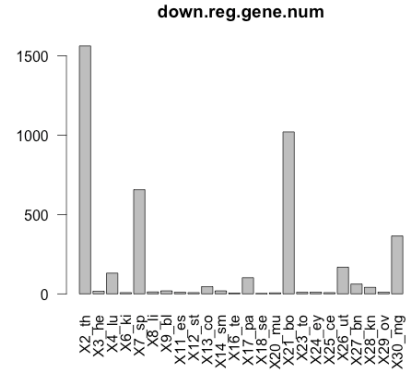

図１７：臓器別発現減少遺伝子群

更に、各臓器間で共通もしくは特異的に発現している遺伝子群による多臓器間の関係性を把握する手法の検討を行い、図１８、１９に示すように、臓器（灰色）、共通遺伝子群(紺色)および臓器特異的遺伝子群（淡青）をつないだ、ネットワーク図を作成した。発現増加遺伝子群および発現減少遺伝子群双方において、脾臓、胸腺、骨髄において、多くの遺伝子群が共有されていることが確認された。更に、各臓器特異的な遺伝子群の機能による特徴づけ手法の検討を行い、フィッシャー正確確率検定に基づく、Gene Ontology で定義される遺伝子セットを対象とした Gene Enrichment 解析を行い、それらを図２０、２１のようにヒートマップの形式で表現し臓器間関係性を探り、大規模データ解析から得られた知見との比較検討を行った。

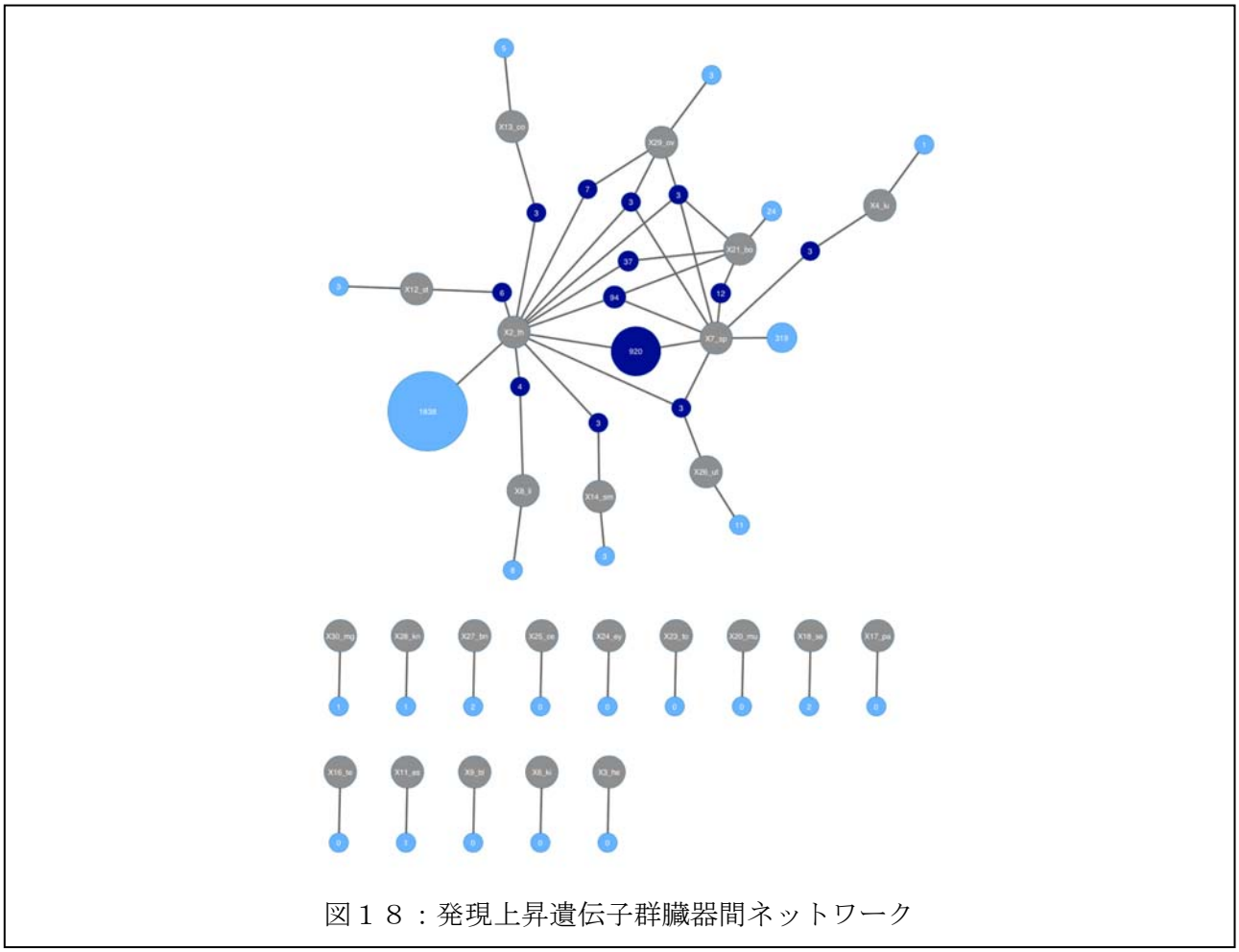

図１８：発現上昇遺伝子群臓器間ネットワーク

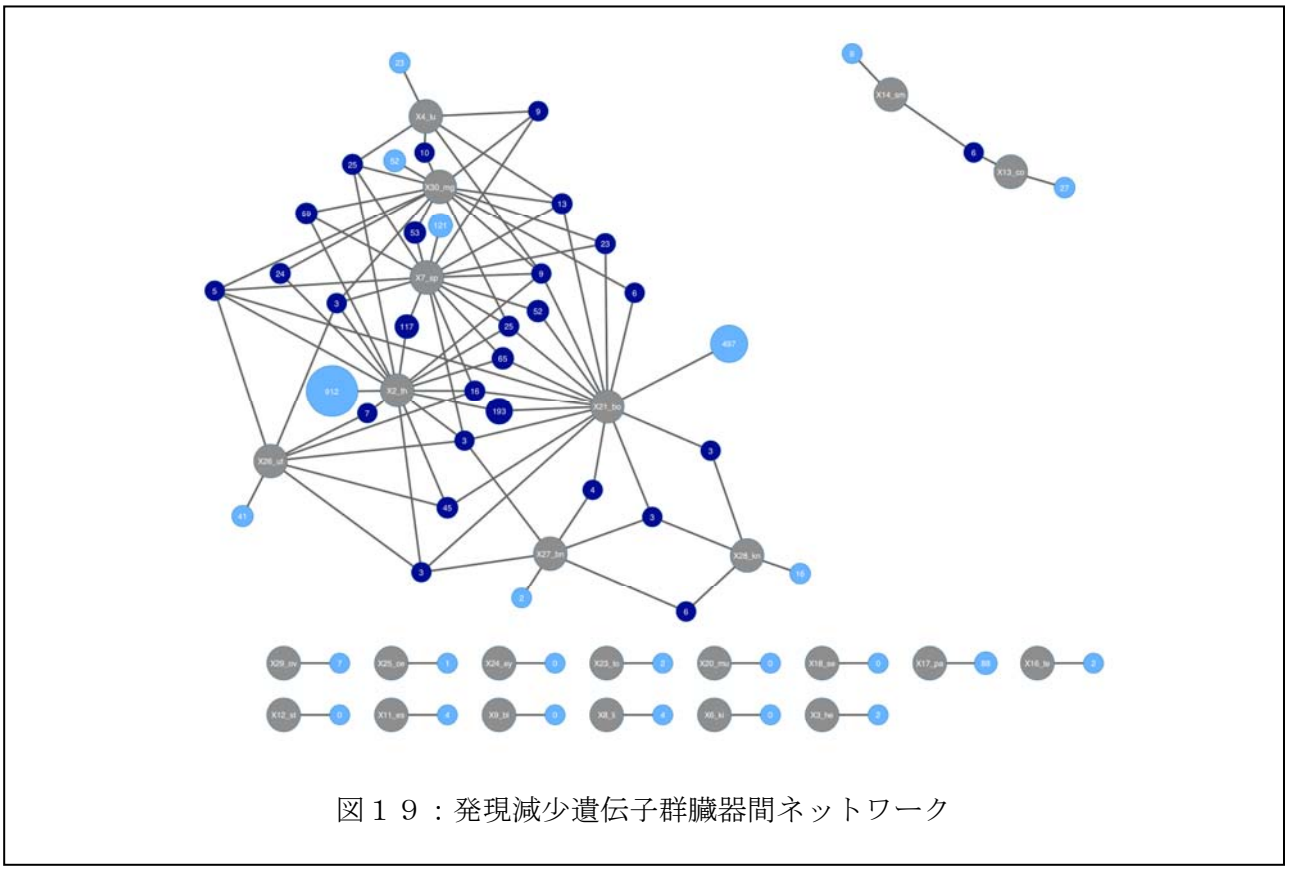

図１９：発現減少遺伝子群臓器間ネットワーク

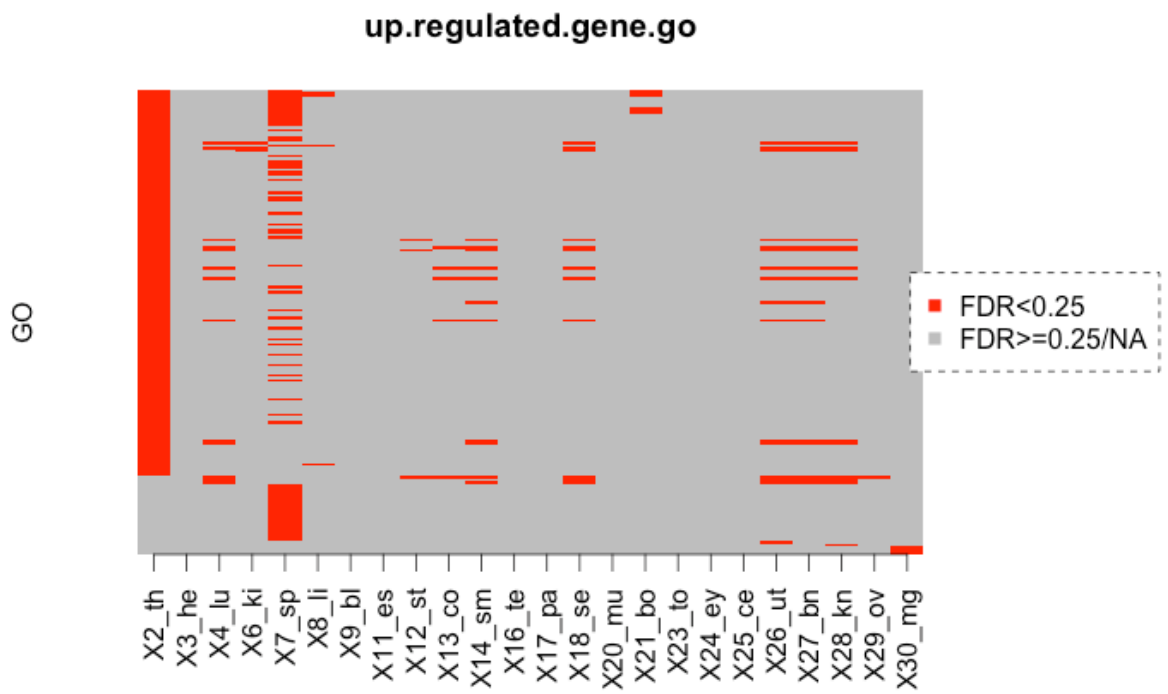

図２０：発現上昇遺伝子群に対して得られた臓器別濃縮 Gene Ontology 群

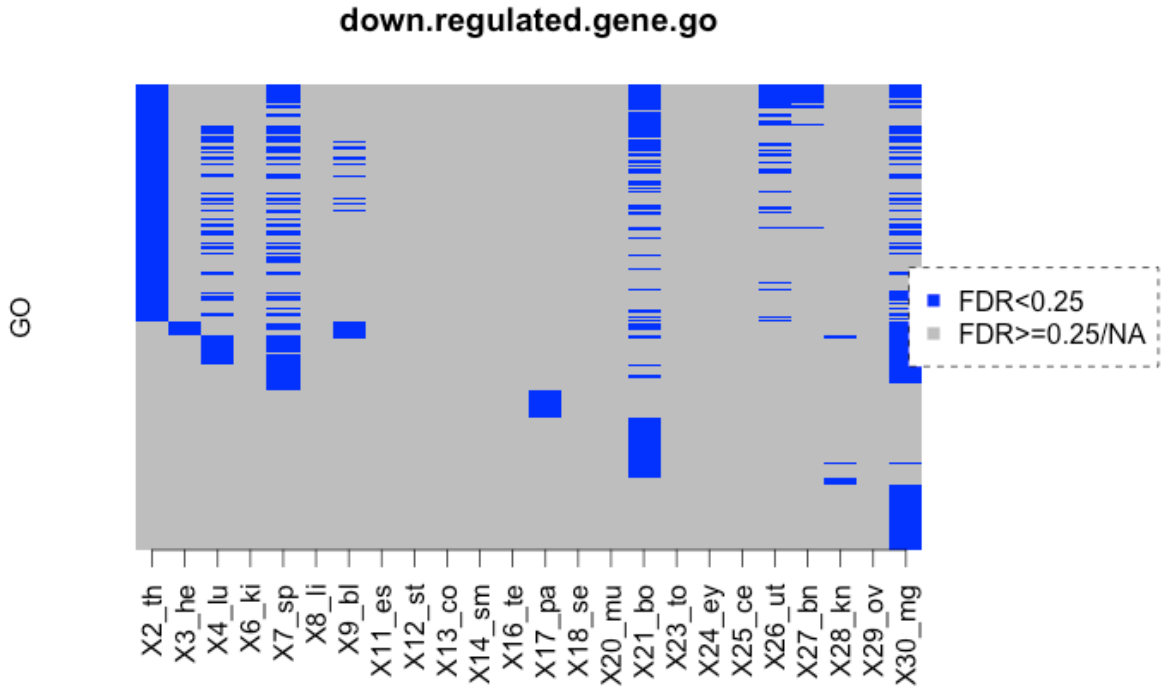

図２１：発現減少遺伝子群に対して得られた臓器別濃縮 Gene Ontology 群

５）大規模がんゲノムの網羅的解析

業務協力者の小川誠司教授（京大大学院医学研究科）と連携して大規模ながんゲノム解析を実施した。RIST の運用対応、利用可能なプログラミング言語の制約、ストレージ量、I/O の脆弱さ、プレポストマシンへの負荷、ソフトウェアの書き直しにかかる時間、国際連携、倫理審査のやり直しにかかる時間などから判断して、過激な世界的競争に勝つために「京」を利用せず、最適な運用を行っ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を用いた。以下は、2013 年からの成果論文リストの一部である。下線は業務参加者及び協力者に記載の者。これらの研究に用いられたゲノム解析パイプラインは、平成 27 年度中に「京」に実装する計画である。

- Sakata-Yanagimoto M, Enami T, Yoshida K, **Shiraishi Y**, Ishii R, Miyake Y, Muto H, Tsuyama N, Sato-Otsubo A, Okuno Y, Sakata S, Kamada Y, Nakamoto-Matsubara R, Tran NB, Izutsu K, Sato Y, Ohta Y, Furuta J, Shimizu S, Komeno T, Sato Y, Ito T, Noguchi M, Noguchi E, Sanada M, **Chiba K**, **Tanaka H**, Suzukawa K, Nanmoku T, Hasegawa Y, Nureki O, **Miyano S**, Nakamura N, Takeuchi K, **Ogawa S**, Chiba S. Somatic RHOA mutation in angioimmunoblastic T cell lymphoma. ***Nat Genet***. 46(2):171-175, 2014.
- Sato Y, Maekawa S, Ishii R, Sanada M, Morikawa T, **Shiraishi Y**, Yoshida K, Nagata Y, Sato-Otsubo A, Yoshizato T, Suzuki H, Shiozawa Y, Kataoka K, Kon A, Aoki K, **Chiba K**, **Tanaka H**, Kume H, **Miyano S**, Fukayama M, Nureki O, Homma Y, **Ogawa S**. Recurrent somatic mutations underlie corticotropin-independent Cushing's syndrome. ***Science***. 344(6186):917-920, 2014.
- Kon A, Shih LY, Minamino M, Sanada M, **Shiraishi Y**, Nagata Y, Yoshida K, Okuno Y, Bando M, Nakato R, Ishikawa S, Sato-Otsubo A, Nagae G, Nishimoto A, Haferlach C, Nowak D, Sato Y, Alpermann T, Nagasaki M, **Shimamura T**, **Tanaka H**, **Chiba K**, Yamamoto R, Yamaguchi T, Otsu M, Obara N, Sakata-Yanagimoto M, Nakamaki T, Ishiyama K, Nolte F, Hofmann WK, Miyawaki S, Chiba S, Mori H, Nakauchi H, Koeffler HP, Aburatani H, Haferlach T, Shirahige K, **Miyano S**, **Ogawa S**. Recurrent mutations in multiple components of the cohesin complex in myeloid neoplasms. ***Nature Genet***. 45(10):1232-1237, 2013
- Makishima H, Yoshida K, Nguyen N, Przychodzen B, Sanada M, Okuno Y, Ng KP, Gudmundsson KO, Vishwakarma BA, Jerez A, Gomez-Segui I, Takahashi M, **Shiraishi Y**, Nagata Y, Guinta K, Mori H, Sekeres MA, **Chiba K**, **Tanaka H**, Muramatsu H, Sakaguchi H, Paquette RL, McDevitt MA, Kojima S, Sauntharajah Y, **Miyano S**, Shih LY, Du Y, **Ogawa S**, Maciejewski JP. Somatic SETBP1 mutations in myeloid malignancies. ***Nature Genet***. 45(8):942-946, 2013.
- Sakaguchi H, Okuno Y, Muramatsu H, Yoshida K, **Shiraishi Y**, Takahashi M, Kon A, Sanada M, **Chiba K**, **Tanaka H**, Makishima H, Wang X, Xu Y, Doisaki S, Hama A, Nakanishi K, Takahashi Y, Yoshida N, Maciejewski JP, **Miyano S**, **Ogawa S**, Kojima S. Exome sequencing identifies secondary mutations of SETBP1 and JAK3 in juvenile myelomonocytic leukemia. ***Nature Genet***. 45(8):937-941, 2013.
- Sato Y, Yoshizato T, **Shiraishi Y**, Maekawa S, Okuno Y, Kamura T, **Shimamura T**, Sato-Otsubo A, Nagae G, Suzuki H, Nagata Y, Yoshida K, Kon A, Suzuki Y, **Chiba K**, **Tanaka H**, **Niida A**, Fujimoto A, Tsunoda T, Morikawa T, Maeda D, Kume H, Sugano S, Fukayama M, Aburatani H, Sanada M, **Miyano S**, Homma Y, **Ogawa S**. Integrated molecular analysis of clear-cell renal cell carcinoma. ***Nature Genet***. 45(8):860-867, 2013.

- Yoshida K, Toki T, Okuno Y, Kanezaki R, **Shiraishi Y**, Sato-Otsubo A, Sanada M, Park MJ, Terui K, Suzuki H, Kon A, Nagata Y, Sato Y, Wang R, Shiba N, **Chiba K**, **Tanaka H**, Hama A, Muramatsu H, Hasegawa D, Nakamura K, Kanegane H, Tsukamoto K, Adachi S, Kawakami K, Kato K, Nishimura R, Izraeli S, Hayashi Y, **Miyano S**, Kojima S, Ito E, **Ogawa S**. The landscape of somatic mutations in Down syndrome-related myeloid disorders. ***Nature Genet***. 45:1293–1299, 2013. Erratum in: ***Nature Genet***. 45(12):1516, 2013.

（５）腫瘍内不均性解明のための大腸がん統合解析及び進化シミュレーション

腫瘍内不均一性は治療抵抗性の一因であると考えられ、その理解は臨床的にも重要な問題である。平成 25 年度に引き続き業務協力者である三森功士教授（九大別府病院）と大腸がん一腫瘍の多数部位から取得したサンプルを解析する multiregional analysis を行った。平成 26 年度は前年度から対象症例を増やし、計 9 検体の 5~20 箇所についてエクソームシーケンスデータ、DNA コピー数、DNA メチ化、mRNA 発現データを取得した。これらのデータを解析した結果、全ての部位に共通して観察される、すなわち不均一性が生み出される進化の前半に蓄積されたと考えられるゲノム、エピゲノムの変化と年齢との相関を見出した。このことは大腸がんの主な原因は加齢によるゲノム、エピゲノムの異常の蓄積であることを示唆している。またこれらのデータを統合的に解析することによりゲノム、エピゲノム、トランスクリプトームが協働して広汎な腫瘍内不均一性を構成していることを見出した（図２２）。

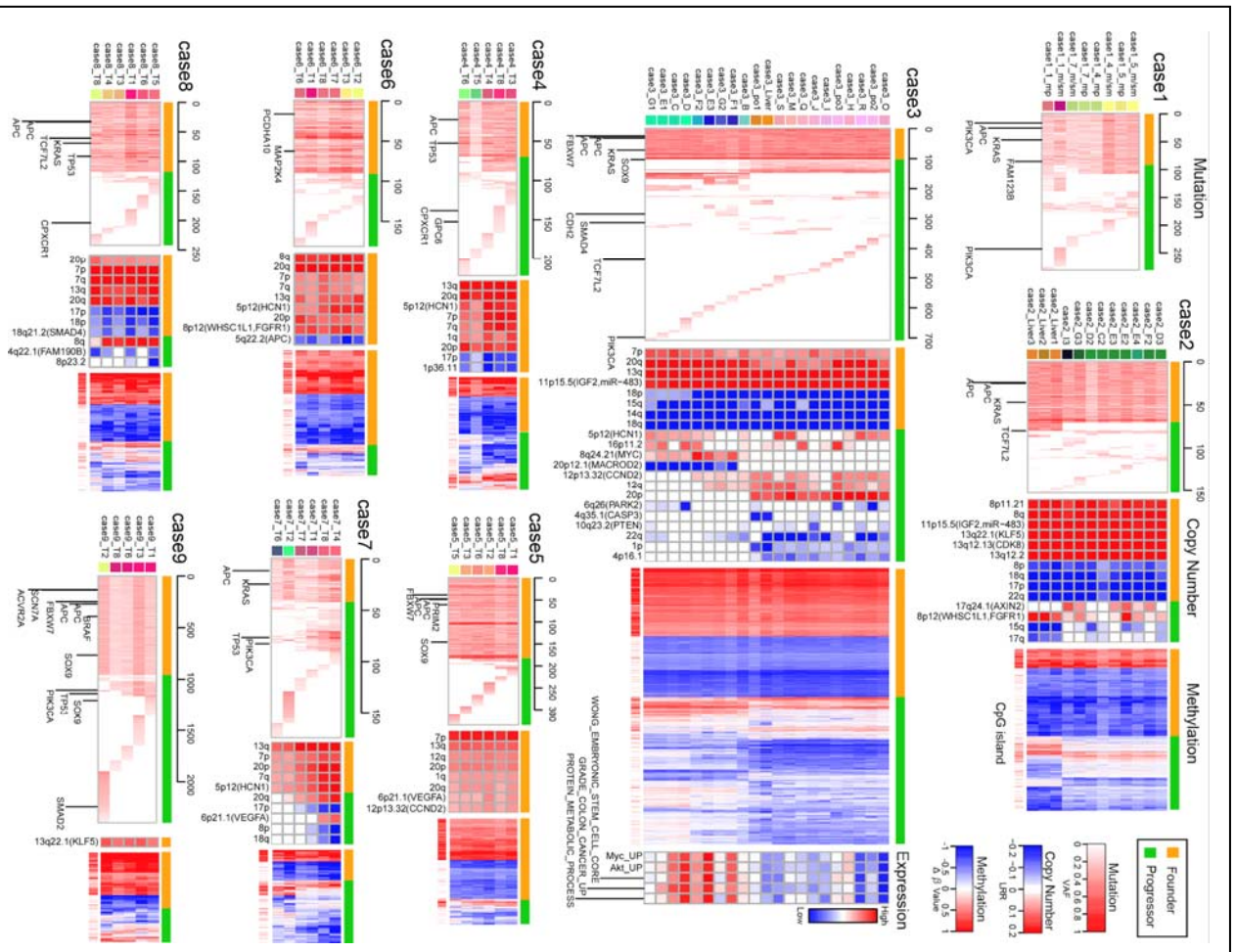

図２２：大腸がん multiregional analysis から得られた統合的 heat map

近年他のいくつかのがん種においても multiregional analysis により腫瘍内不均一性の存在が明らかにされている。しかしながら。腫瘍内不均一性を生み出す原理の探求についての試みはほとんどなされていない。この目的のために腫瘍内不均一性を再現する、がんの進化シミュレーションモデル、Branching Evolutionary Process (BEP) Model を構築した。細胞の中の複数個の遺伝子にランダムに変異を入れながら増殖させていくと、増殖速度を増加させるドライバー変異を蓄積する細胞が選択され増殖能力の高いクローンに進化する。このような進化過程で条件によっては異なる変異を有するクローンに分岐し不均一性を獲得する。この研究では「京」を利用して膨大な組み合わせのパラメーターセットで BEP model によるがんの進化シミュレーションを行うことにより、高い不均一性が生み出される条件の網羅的探索を行った。その結果、高い遺伝子変異率、がん幹細胞の存在を仮定すると、上記の大腸がんの multiregional exome データに観察されるような、高い不均一性を有する遺伝子変異パターンが再現できることを見出した。さらにシミュレーション結果からドライバー遺伝子は全てのがん細胞に共有されている一方で、不均一性を生み出している変異の大部分は細胞の増殖速度に影響を与えない中立変異であることが示唆された。このことは multiregional exome データにおいて全ての部位に共有されている変異に関しては既知のドライバー遺伝子変異が認められる一方で、腫瘍内不均一性を生み出している、全ての部位で共有されてはいない変異についてはほとんど既知のドライバー遺伝子変異が認められないという結果とも一致している。以上、この研究により腫瘍内不均一性を生み出している進化原理の一端が大腸がんゲノム解析と「京」を用いたがんの進化シミュレーションにより明らかになった（図２３）。今後、これまでの結果に基づいてがんの治療抵抗性を克服するための治療戦略の確立を試みる予定である。成果は論文として投稿中である。

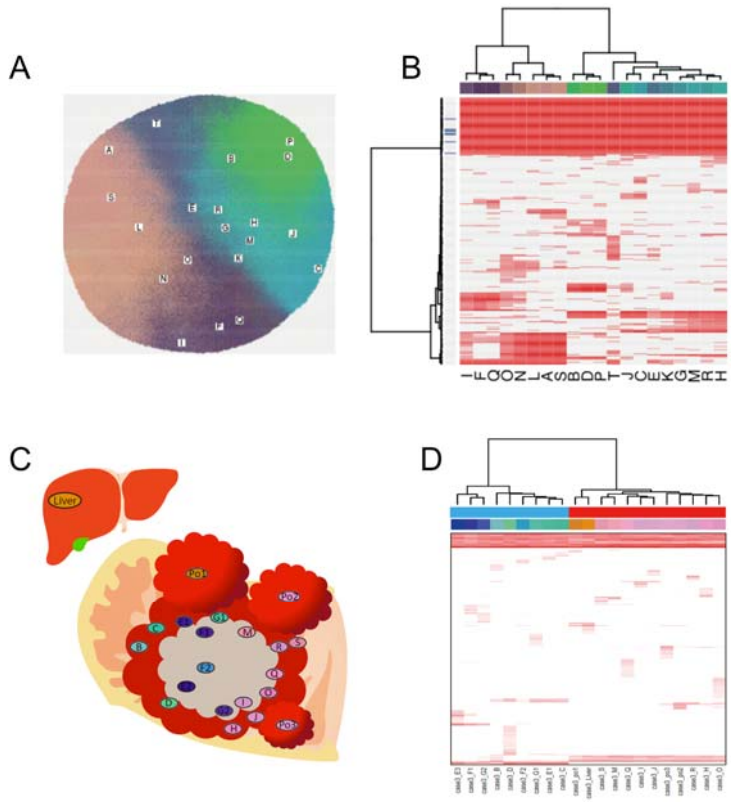

図２３：がんの進化ミュレーション（A,B）をおよび大腸 multiregional exome 解析（C,D）で得られた不均一性の高い遺伝子変異パターン

（6） 以上（1）から（5）の大規模データ解析を平成 25 年度に取得した甲状腺がんのゲノムデータ、業務協力者の有するゲノムデータ、遺伝子発現プロファイルデータ、Sanger Institute や、TCGA、CCLE など公共データベースにて公開されているゲノムデータ（数万検体規模）により実施した。

（7）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合わせて実施した。

「戦略課題 4：大規模生命データ解析」研究統括業務では、以下の２つの大学で実施される平成 26 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）：microRNA を含む遺伝子ネットワーク推定について助言した。

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)：メタゲノム研究者との連携を構築した。

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。

平成２６年度は、前年度に引き続きメタゲノム解析およびがんゲノム解析パイプラインの開発とそれらを利用した大規模ゲノム解析を行い、これらパイプラインで効率的に並列計算を行うための汎用的なプログラムの開発も行う。また、戦略プログラム分野１の課題内で連携し、がん関連遺伝子に対する解析を行う。

平成２５年度より、東京大学医科学研究所　宮野研究室と連携し、「京」上で稼働するがんゲノム解析パイプライン（Genomon-exome）の開発に着手した。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。これまでに開発してきたメタゲノム解析パイプラインのタスク管理の機能が部分的に利用できるため、この機能を共通で利用できるプログラムとして改修し、そのうえで比較的緩い依存関係を有したタスクの並列計算を容易にする機能強化を行う。このがんゲノム解析パイプラインについては、平成２５年度中に MPI ライブラリにより、複数サンプルの解析を同時に実行できるシステムを実装完了する見込みである。平成２６年度は、上記タスク管理プログラムを用いてパイプラインに含まれるタスク間の依存関係に対応する改良を行い、各サンプルの解析を複数のノードに分散させることのできるシステムを開発する。また、このシステムを実データに適用し、性能測定を行う。

メタゲノム解析では、平成２５年度までに、開発したメタゲノム解析パイプラインを用いてヒト口腔内細菌叢を解析した結果、口腔内部位によってオーソロググループの相対存在度が異なることを発見した。平成２６年度は、個々のオーソロググループと口腔内部位との関連を調査するとともに、口腔内細菌叢以外のサンプルでも同様な解析が有効であるか調査する。

Ⅳ－２－２　実施内容（成果）

平成２６年度は、実施計画に基づきメタゲノム解析およびがんゲノム解析パイプラインを開発し、それらを利用した大規模ゲノム解析を行った。両パイプラインで同様な機能については、汎用的なプログラムとして開発を進めた。また、戦略プログラム分野１の課題内で連携し、がん関連遺伝子の大規模な実データに対し解析パイプラインを適用した。

・メタゲノム解析パイプラインの開発

昨年度に引き続き、環境中に存在する細菌叢の機能解析を行うメタゲノム解析パイプラインの開発を行った。メタゲノム解析パイプラインでは相同配列解析部分が、パイプラインの実行時間の大半を占める。平成２５年度までに、この相同配列解析部分として、GHOSTX と名付けた検索アルゴリズムを考案し、これを「京」の大規模並列環境に適応させた GHOST-MP プログラムの開発と改良を継続してきた。GHOSTX アルゴリズムは、クエリ配列とデータベース配列を接尾辞配列というデータ構造としてインデックス化することで、アラインメント候補を効率的に探索し、高速な配列相同性検索を実現している。平成２６年度は、GHOSTX に代わるものとして、GHOSTZ アル

ゴリズム（Suzuki et al. 2015, *Bioinformatics* 31 (8): 1183-1190）を開発した。

GHOSTZ アルゴリズムは、アラインメント候補部位を探索する際に、あらかじめ構築しておいたデータベース配列中の部分文字列のクラスタ代表を利用する。GHOSTZ や GHOSTX だけでなく、BLAST をはじめとした多くの配列相同性検索では、まずはじめに比較的検出が容易な類似度が特に高い部分をアラインメント候補部位として列挙する。それら候補の周辺を探索領域として、アラインメントの作成とアラインメントスコアの計算を行い、最終的に相同配列候補を出力する。アラインメント候補を列挙するために、BLAST は有限オートマトンを用いて、データベース配列中から固定長の文字列一致を探索するが、その検索の高速化を目的としてあらかじめデータベース配列のインデックスを作成する様々な手法が提案されている。一方、GHOSTZ では部分文字列のハッシュテーブルを用いてアラインメント候補部位の探索を行うが、このハッシュテーブルに全部分文字列を登録する代わりに、部分文字列のクラスタ代表を登録し、他のクラスタメンバとの比較を減らすことで高速化を実現した。この目的でデータベースの全部分文字列をあらかじめクラスタリングする。クラスタリングにおける類似度の評価には、高速に計算可能であることと、次に説明する三角不等式が利用できることから編集距離を用いる。クエリの部分文字列とクラスタメンバの類似性の判定は、クエリの部分文字列とクラスタ代表の距離だけを新たに計算すれば、あらかじめ計算されたクラスタ代表とクラスタメンバとの距離から、三角不等式を利用して直接の文字列比較を避けて比較を行える（図 1)。これにより、異なる部分文字列もまとめて扱うことができ、従来のハッシュテーブルを用いた BLAT や RAPSearch などの手法と比較し、検索精度を落とさずに 2.7～5.4 倍の検索速度を実現した（図 2、表 1)。また、相同性検索として標準的に用いられている BLAST と比較して、類似性の高い（E-value $<1.0\times10^{-12}$) 領域ではほぼ同等の精度で 261 倍の検索速度を実現した（図 2、表 1)。現在、GHOSTZ アルゴリズムを GHOST-MP プログラムの検索アルゴリズムとして採用する方向で検討している。

メタゲノム解析パイプラインの開発に関連して、平成２５年度までに解析パイプラインを利用してヒト口腔内細菌叢の解析を行ってきた。米国立衛生研究所の Human Microbiome Project が公開する、ヒト口腔内細菌叢の口腔内 8 部位 381 サンプルのデータを解析してきた（平成２５年度は 9 部位 418 サンプルの実施を報告したが、その後に品質評価の低いサンプルを解析対象から除いた)。解析パイプラインを通して得られる機能アノテーション（KEGG Orthology の相対存在度）を対象に主成分分析の結果を比較することで、口腔内部位によって機能の割合（オーソロググループの相対存在度）が異なることを発見した。本年度は、個々のオーソロググループと口腔内部位との関連を調査し、狭義の口腔、口腔前庭、歯垢の３部位で比較すると、特定のパスウェイ（KEGG PATHWAY: ko00540 (Lipopolysaccharide biosynthesis), ko02030 (Bacterial chemotaxis), ko02040 (Flagellar assembly)）の相対存在度が口腔内部位によって有意に異なることを明らかにした。リポ多糖生合成（ko00540）に関わるオーソロググループは、狭義の口腔で多く存在し、細菌の移動に関する細菌走化性（ko02030）やべん毛（ko02040）に関わるオーソロググループは、口腔前庭で少ないことがわかった（図 3)。

また、メタゲノム解析パイプラインの他の応用として、東京大学医科学研究所国際粘膜ワクチン開発研究センターの植松智特任教授の研究グループと免疫研究を目的としたマウスの腸内細菌叢サンプルの解析について検討に着手した。

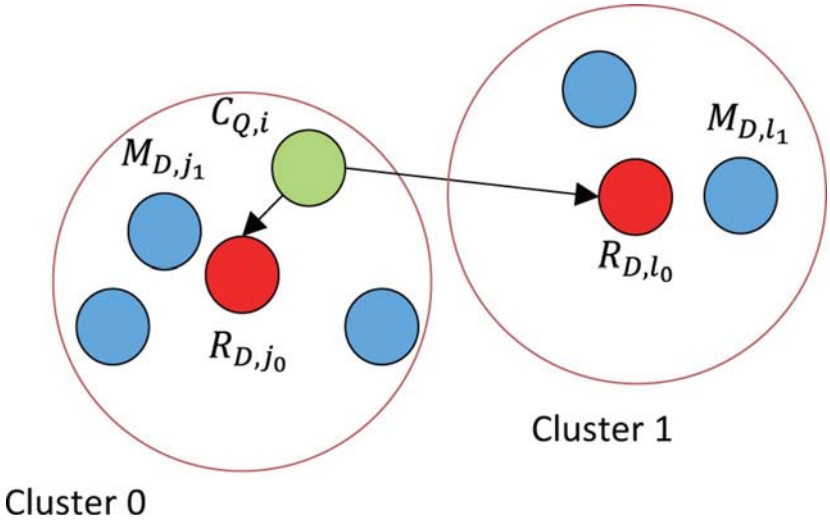

図 1　GHOSTZ のアラインメント候補探索。それぞれ、緑：クエリの部分文字列、赤：クラスタ代表である部分文字列、青：クラスタメンバである部分文字列を表す。クラスタ代表から一定距離内にある部分文字列がクラスタメンバとしてクラスタリングされる。三角不等式を利用することで、クラスタメンバと直接比較することなく類似部分文字列の判定が可能になる（偽陽性を含む）。図では、クエリと２つのクラスタ代表、Cluster 0 と Cluster 1 の部分文字列を直接比較することで、Cluster 0 のメンバをアラインメント候補部位として列挙できる。(図の出典、Suzuki et al. Bioinformatics (2015) 31 (8): 1183-1190 Fig. 2.)

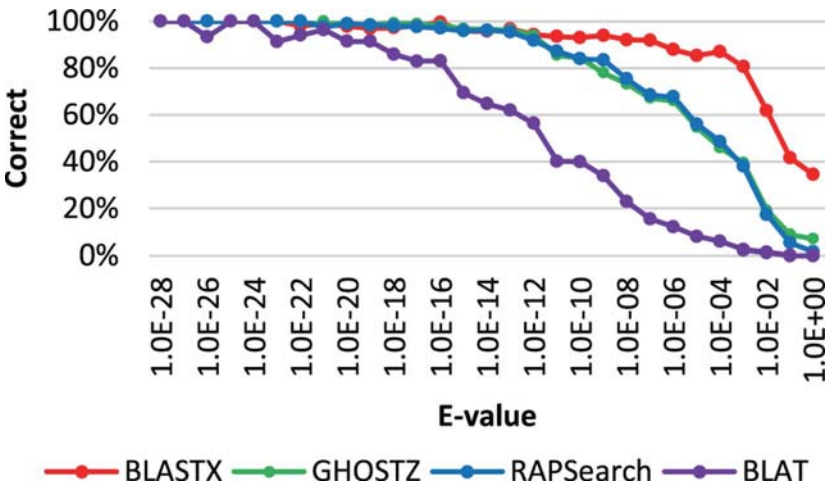

図 2　GHOSTZ の検索精度。SRA Accession Number: SRR407548 から無作為に抽出した 10,000 配列をクエリに、KEGG GENES のアミノ酸配列をデータベース配列として検索を行った結果。Smith-Waterman アルゴリズムの実装である SSEARCH による精密な検索で最もスコアの高かった結果を正解として一致度を縦軸に示した。(図の出典、Suzuki et al. *Bioinformatics* (2015) 31 (8): 1183-1190 Fig. 9.)

表 1　GHOSTZ の検索速度。SRA Accession Number: SRR407548 から無作為に抽出した 10,000 配列をクエリに、KEGG GENES のアミノ酸配列をデータベース配列として検索を行った結果。Acceleration ratio は BLASTX の結果を基準とした。(表の出典、Suzuki et al. *Bioinformatics* (2015) 31 (8): 1183-1190 Table 2.)

	Computation time (sec.)	Acceleration ratio
GHOSTZ	460.8	261.3
RAPSearch	1285.5	93.7
BLAT	2514.9	47.9
BLASTX	120395.2	1.0

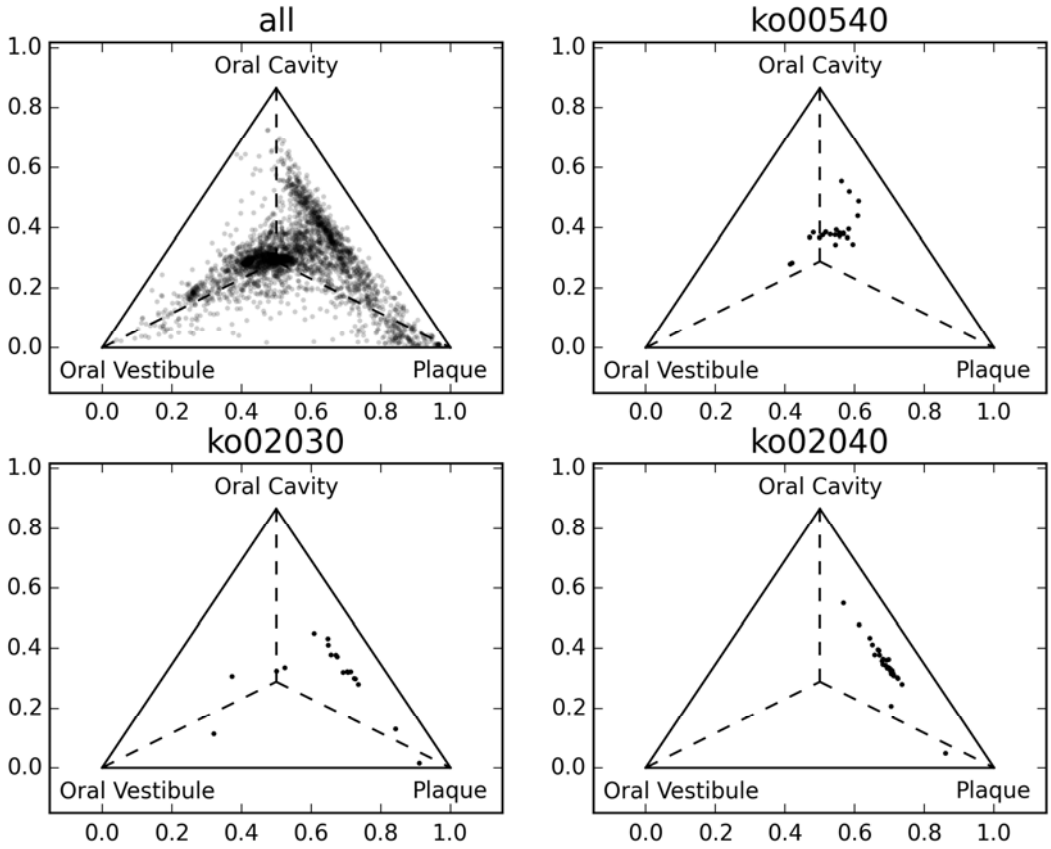

図 3　各機能に関わるオーソロググループの相対存在度の口腔部位間比較。狭義の口腔（Oral Cavity）、口腔前庭（Oral Vestibule）、歯垢（Plaque）の 3 部位のオーソロググループの相対存在度を Ternary Plot で示す。左上の Ternary Plot には、全てのオーソロググループを表示し、他は特定のパスウェイに関わるオーソロググループのみを表示した。

・がんゲノム解析パイプラインの開発

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。このがんゲノム解析基盤の一つとして、東京大学医科学研究所 宮野研究室開発の Genomon-exome に基きエクソーム解析を行うパイプラインの開発を行っている。

平成２５年度に、がんゲノム解析パイプラインとメタゲノム解析パイプラインが共にタスク管理機能を必要としており、その管理方法も似ていたため、メタゲノム解析パイプラインで使用していたタスク管理機能を改修し、プログラムの一部共通化を行った。この際の共通化されたプログラムでは、依存関係を有したタスクの処理順序は、依存関係さえ解決されれば入力順に処理を行っていたが、依存関係グラフから処理時間が短くなるように処理順序を変更するように今年度に改修を加えた。

また、処理の並列数を挙げるため、処理の分割方法と中間出力の出力形式を変更することで解析パイプライン全体の高速化を行った。解析パイプラインの実行時間の多くを占めるマッピングとアラインメント処理では、BWA プログラムでマッピングとアラインメントを行った後、Samtools プログラムで並列処理を行った結果のマージを行う（図 4)。処理の分割方法の工夫として、並列に処理できる部分が残っているにもかかわらず、BWA の処理直後に 1 サンプル毎に 1 つの.sam ファイルへマージを行っていたことを改め、マージ後でないと行えない重複除去（データチェック）とインデックス付加を除く処理を並列に計算する変更を検討した。また、重複除去とインデックス付加の処理は、参照配列（マッピング先）の染色体ごとに完全に独立して行えるため、染色体ごとに中間結果を出力し、重複除去とインデックス付加の処理も並列に計算する方法の検討も行った。中間出力の出力形式の工夫としては、ソート、マージ、データチェック、インデックス付加の各中間出力すべてで行われていたデータ圧縮と展開を行うことを廃止した。中間出力は BAM 形式のファイルとして出力を行っていたが、この BAM 形式は内部で BGZF（Blocked GNU Zip Format）という圧縮形式でデータの圧縮を行っている。BAM 形式の出力と入力のたび、GZIP の圧縮・展開が行われ、実質的な処理内容のソート、マージ、データチェック、インデックスに対して大きな計算時間を占めていた。そこで、中間出力では BGZF の圧縮を廃止し、最終出力のみ圧縮を行うことで高速化を試みた。

肺腺がんのエクソームデータを用いて解析パイプラインの計算時間を評価した結果、マージの遅延、染色体ごとの並列計算、データ圧縮・展開の削減はいずれも高速化に寄与することが判明した。結果の詳細は論文にて発表予定で、現在投稿準備中である。

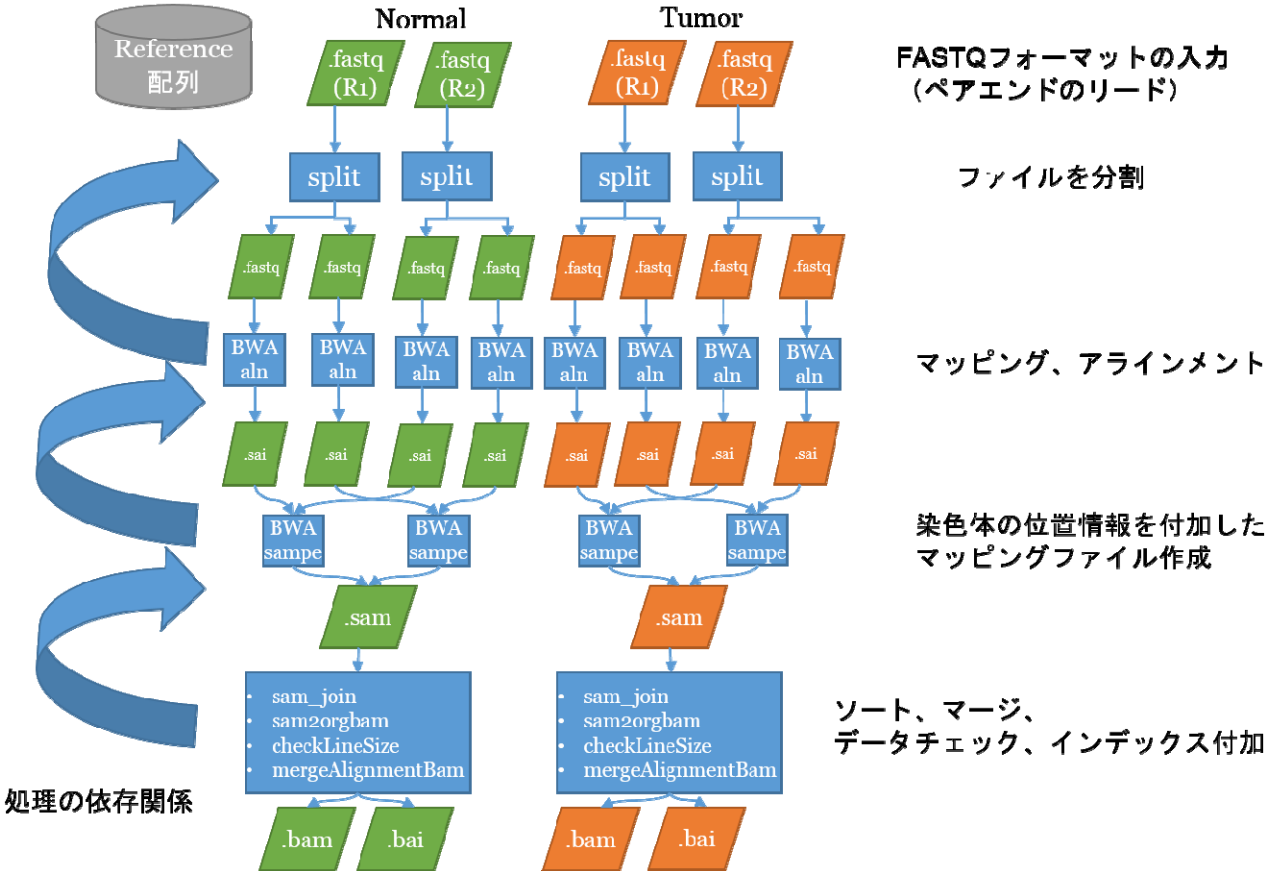

図 4　がんゲノム解析パイプラインのマッピングとアラインメント処理。マッピングとアラインメントは BWA、ソートやマージなどの処理は Samtools を利用して行う。

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 26 年度は、刺激応答に対する種々の脂肪細胞組織中の生体分子の経時的変化のデータから、平成 25 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を解明し、肥満是正のための知見を得る。

Ⅳ－３－２　実施内容（成果）

（１） 生体分子ネットワーク解析による刺激に対するエネルギー消費に向けての脂肪細胞の状態変化を抑制する新たな機構を発見

ヒトを含む哺乳類には大別して 2 種類の脂肪細胞があり、それぞれ白色脂肪細胞、褐色脂肪細胞と呼ばれている。白色脂肪細胞はエネルギーの貯蔵を行う細胞であり、全身のエネルギー要求に応じて蓄積した中性脂肪を分解し、脂肪酸の形で細胞外に放出する。つまり、白色脂肪細胞はエネルギーの貯蔵と放出を行う役割を果たしている。

一方で、褐色脂肪細胞は生体内にごく少量しか存在しないが、ミトコンドリアを豊富に含むことから褐色を呈し、脂肪酸を酸化して体温の恒常性維持のために熱としてエネルギーを放出する。褐色脂肪細胞の熱産生能力(物質を代謝して熱を放出する能力)は骨格筋細胞と比較すると約 100 倍高く、この高い熱産生能力はミトコンドリアに存在する uncoupling protein 1 (UCP1)に起因することが明らかとなっている。

継続的な寒冷刺激を受けると、刺激を脳の視床下部が感知して、交感神経を介してノルアドレナリンが放出される。ある部位（主に皮下脂肪と呼ばれる部位）の白色脂肪細胞は、ノルアドレナリンがβアドレナリン受容体に結合するとシグナルが伝達され、UCP1 の発現が誘導されてミトコンドリアが増えて褐色脂肪細胞に似た色となり熱産生を行うようになる（図１）。この転換過程は「褐色化」と呼ばれ、褐色化した白色脂肪細胞はベージュ脂肪細胞と呼ばれている。白色脂肪細胞の褐色化は蓄積された脂肪の分解とエネルギー消費を伴うことから、新しい視点からの肥満是正の方策として注目を集めているが、同じ刺激を与えても一部の白色脂肪細胞のみしか褐色化しない原因は不明であった。

そこで、本研究では、マイクロアレイ(Agilent SurePrint G3 Mouse GE 8×60K)により、平成 24 年度に計測した褐色脂肪細胞、ベージュ脂肪細胞に加えて、白色脂肪細胞（寒冷刺激を与えても褐色化しない、内臓脂肪にある脂肪細胞）についても寒冷刺激を与えた時の遺伝子の時系列発現プロファイルを取得した。それを解析した結果、白色脂肪細胞では褐色脂肪細胞、ベージュ脂肪細胞と比べて、ある種のサイトカインが高発現していることが示された（図２）。このサイトカインは炎症反応に関与する生理活性物質であり、脂肪組織において炎症反応を起こすことは知られていたが、褐色脂肪細胞や、白色脂肪細胞の褐色化との関係は不明であった。

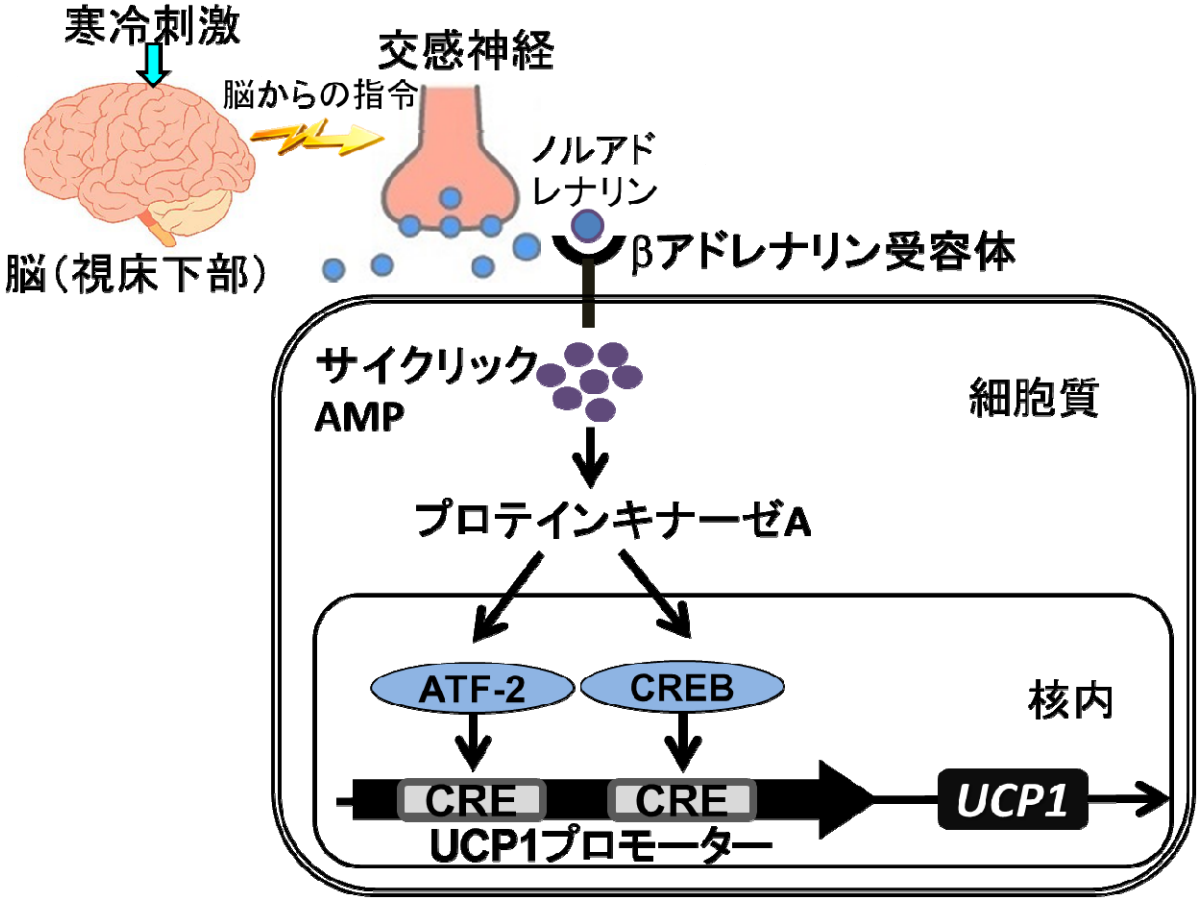

図１　寒冷刺激により UCP1 の発現が誘導されるメカニズム

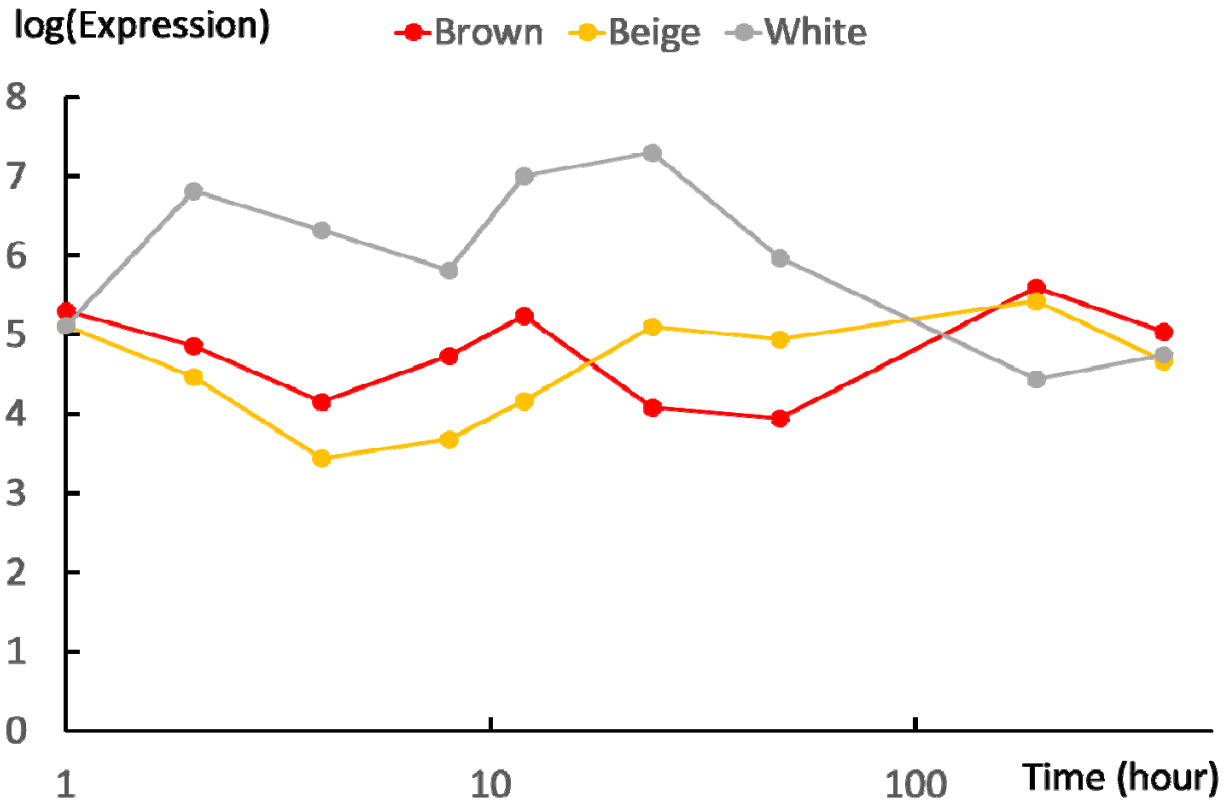

図２　褐色・ベージュ・白色の各脂肪細胞でのサイトカインの遺伝子発現量（対数値）

さらに、マウス個体に４℃の寒冷刺激を与えた時のベージュ脂肪細胞の時系列遺伝子発現プロファイルから、ダイナミックベイジアンネットワークモデルにより遺伝子ネットワークを作成した（図３）。図３のネットワークの一部（青い丸で囲った部分）を拡大すると、UCP1 の近くに上記のサイトカインの機能を抑制する遺伝子（赤い丸で囲ったもの）が位置していることがわかった。前述のように UCP1 は脂肪細胞の褐色化による熱産生能力の獲得に最も重要な働きをする遺伝子であることから、ベージュ脂肪細胞では、白色脂肪細胞で高発現するサイトカインを抑制していることが示唆された。すなわち、寒冷刺激を与えても褐色化しない白色脂肪細胞ではサイトカインの働きで UCP1 の発現が抑制されるが、ベージュ脂肪細胞では寒冷刺激を受けるとサイトカインを抑制する遺伝子が働くためサイトカインの影響を受けずに褐色化するのではないかという仮説が得られた。これを検証するため、研究協力者である河田照雄教授（京大農学研究科）の協力を得て、マウスにサイトカインを投与したときの UCP1 の発現量を計測した（図４）。

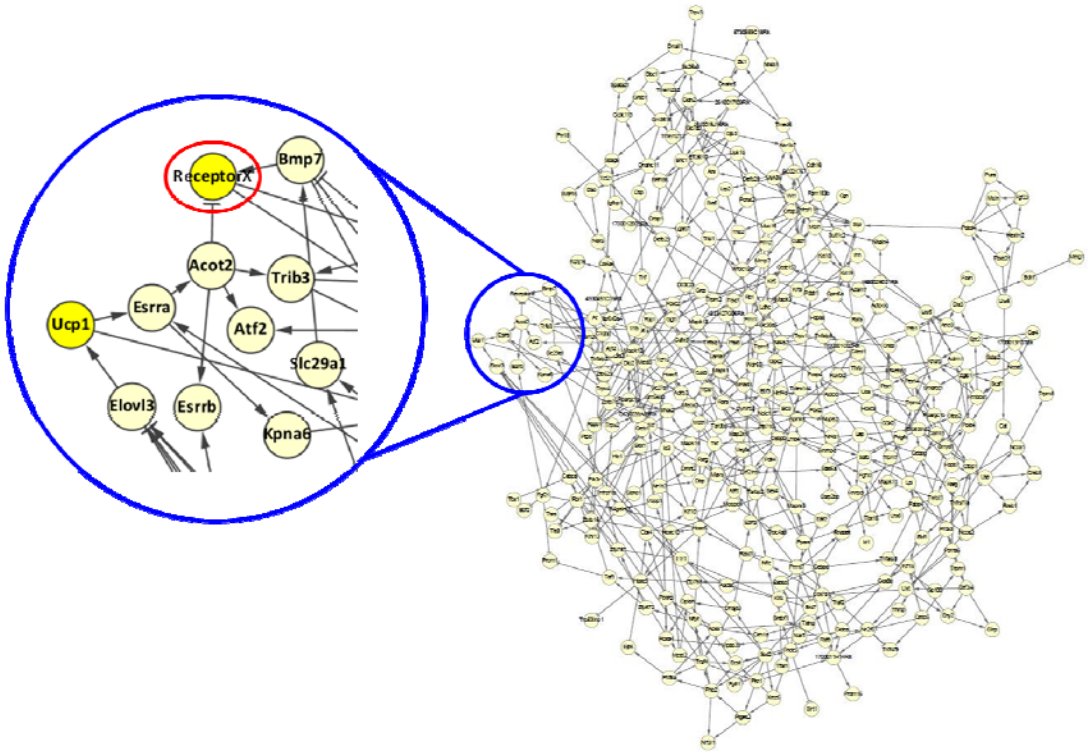

図３　マウス個体に寒冷刺激を与えた時のベージュ脂肪細胞での遺伝子ネットワーク

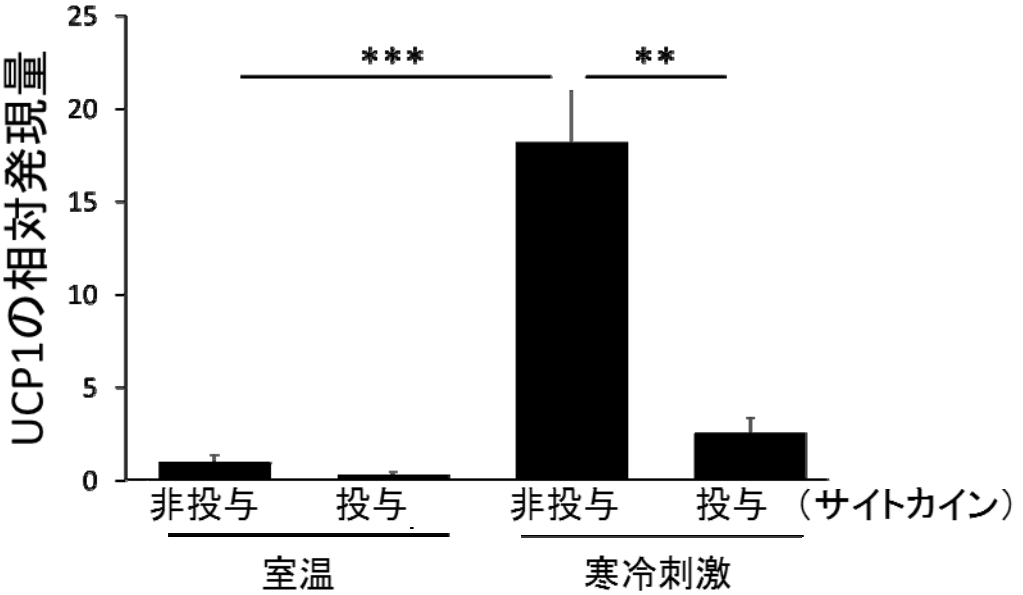

図４　マウス個体にサイトカインを投与したときの UCP1 の遺伝子発現量
（室温でサイトカイン非投与のときの発現量を１としたときの相対値で表示）

図４から、マウスに上記のサイトカインを投与するとUCP1の発現が有意に減少することが示され、このサイトカインが白色脂肪細胞の褐色化を抑制する機能を持っていることがわかった。

一般に、脂肪組織には、脂肪細胞以外に微量ながらマクロファージが存在しており、このサイトカインはマクロファージにより分泌されることが知られている。このため、寒冷刺激を受けても、内臓脂肪組織の白色脂肪細胞は周辺のマクロファージが分泌するサイトカインの影響を受けて、褐色化が抑制され熱産生によるエネルギー消費が生じないことが示された（図５）。

従来、マクロファージによるサイトカインの分泌は免疫に関連したパスウェイ（反応経路）の一部であり、脂肪細胞の褐色化によるエネルギー消費のパスウェイとは無関係と考えられており、両者にまたがる論文報告は存在しなかった（図６）。

図５　マクロファージによるサイトカイン分泌を介した白色脂肪細胞の褐色化の抑制機構

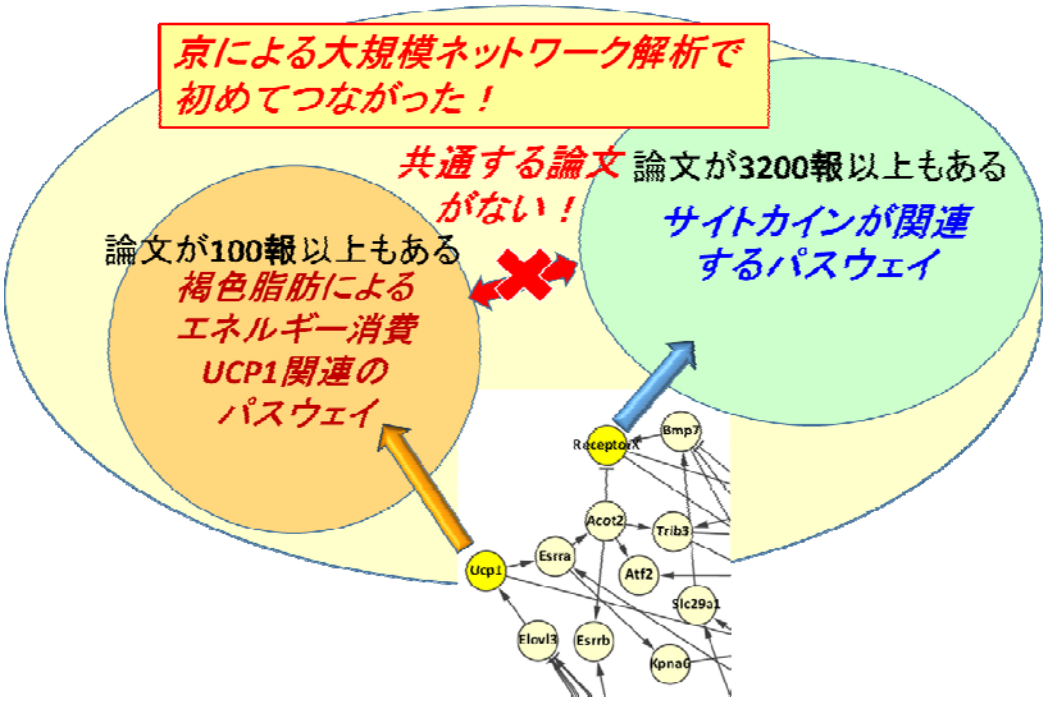

図６　大規模ネットワーク解析により新たに発見された、複数の生命現象間の関連性

「京」による大規模ネットワークの網羅的な解析により、無関係と思われていた両者が関連していることが初めて示された。本研究によって新たに明らかになった、白色脂肪細胞の褐色化を調節している新規の機構は、従来にない視点からの肥満是正の戦略につながることが期待される。

（2） 刺激に対するエネルギー消費に向けての脂肪細胞の状態変化での microRNA も含めた生体分子ネットワーク解析

最近の総説（M. Trajkovski et al. 2013, J. Y. Zhou et al. 2014 など）によると、褐色脂肪細胞における熱産生によるエネルギー消費において micro RNA の関与が報告されているが、その数は 10 個程度しかなく、しかも刺激なしでもエネルギー消費を行っている褐色脂肪細胞と刺激を受けてエネルギー消費に向けて転換するベージュ脂肪細胞での microRNA の機能の違いがほとんど明らかになっていなかった。

そこで、本研究では、マウス個体の褐色・白色・ベージュの 3 種類の脂肪細胞について、マウス遺伝子について取得したのと共通の 6 時点（寒冷刺激前と寒冷刺激後 0, 1, 2, 4, 12, 24 時間後）について各 3 サンプルずつ microRNA の時系列発現プロファイルをマイクロアレイ(Agilent Expression Array Mouse miRNA 8x60k Rel.19.0)を用いて取得した。

3 種類の脂肪組織ごとに、寒冷刺激下において発現が検出された microRNA の個数を調べたところ、褐色・白色・ベージュについて、それぞれ 291 個、232 個、351 個と多くの microRNA の関与が示唆された。これらの間で共通する microRNA の個数を調べたところ、図 7 のように脂肪細胞ごとに大きな違いが見られた。特に、白色脂肪細胞でのみ発現している microRNA の個数が 8 個にとどまっているのに対して、褐色脂肪細胞では 11 個に増え、ベージュ脂肪細胞になると 56 個にまで増大している。このことから、寒冷刺激が与えられたときのベージュ脂肪細胞でのエネルギー消費に向けての転換に microRNA が大きく関与していることが示唆された。

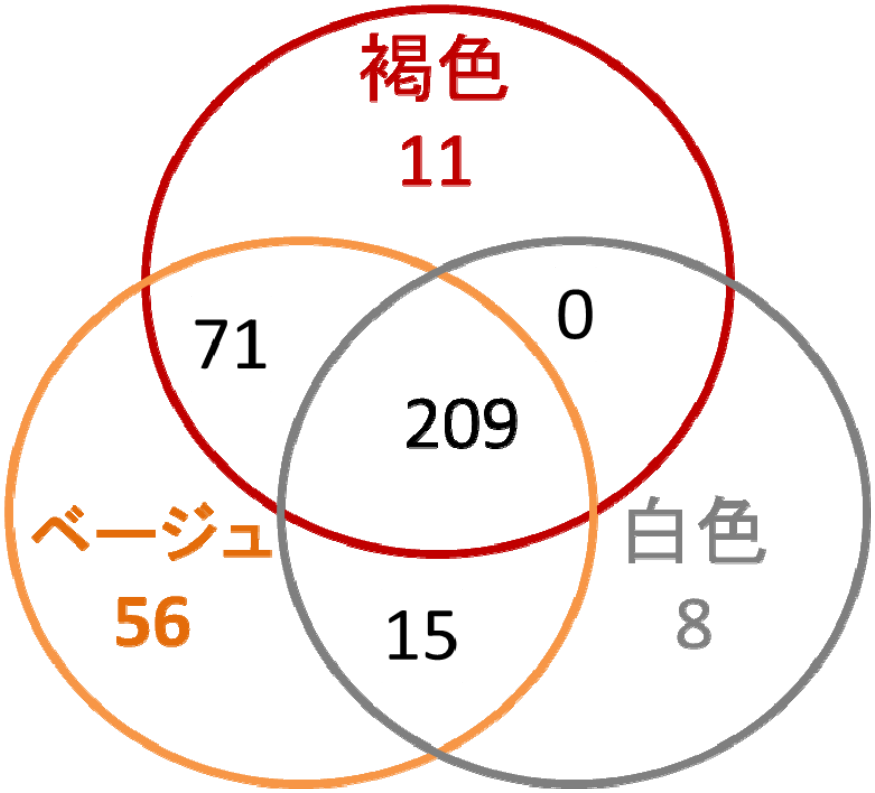

図 7　褐色・白色・ベージュの各脂肪細胞において寒冷刺激が与えられたときに発現が検出された microRNA の個数

そこで、褐色脂肪細胞とベージュ脂肪細胞について、これらの microRNA に加えて、同様に寒冷刺激を受けて発現が上昇した遺伝子約１万個に含めた時系列遺伝子発現プロファイルからダイナミックベイジアンネットワークモデルにより生体分子ネットワークを作成した（図８、図９）。図８と図９にあるように、寒冷刺激が与えられた時に働く生体分子ネットワークの中で機能している microRNA の個数は、褐色脂肪細胞と比べると明らかにベージュ脂肪細胞の方が多く、刺激による白色脂肪細胞からベージュ脂肪細胞へのエネルギー消費に向けての転換に、microRNA が重要な役割を果たしている可能性が示唆された。平成 27 年度は、これらの microRNA の機能をさらに詳細に解析して、脂肪細胞のエネルギー消費に向けての転換の機構への関連性の解明につなげていく。

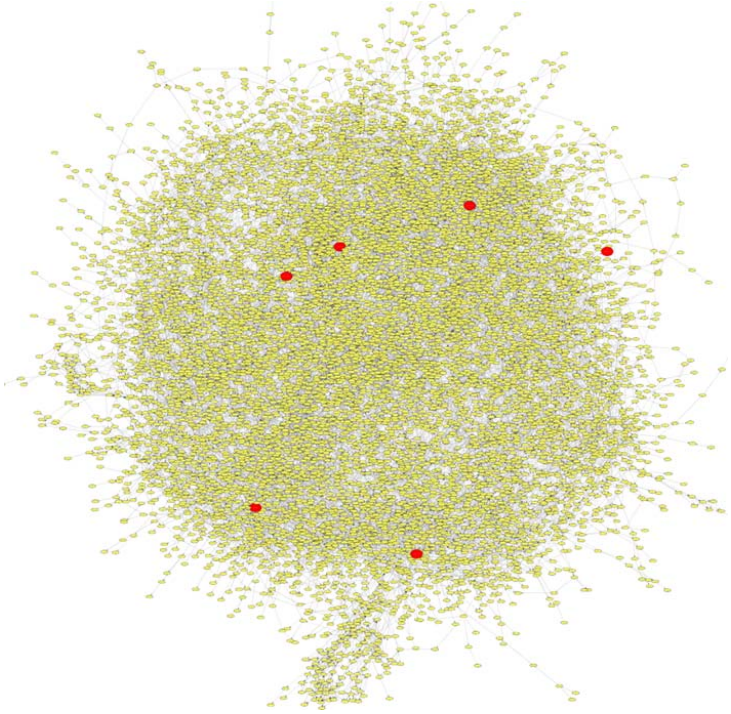

図８　寒冷刺激が与えられた時の褐色脂肪細胞での生体分子ネットワーク（赤い点が microRNA で黄色の点が遺伝子）

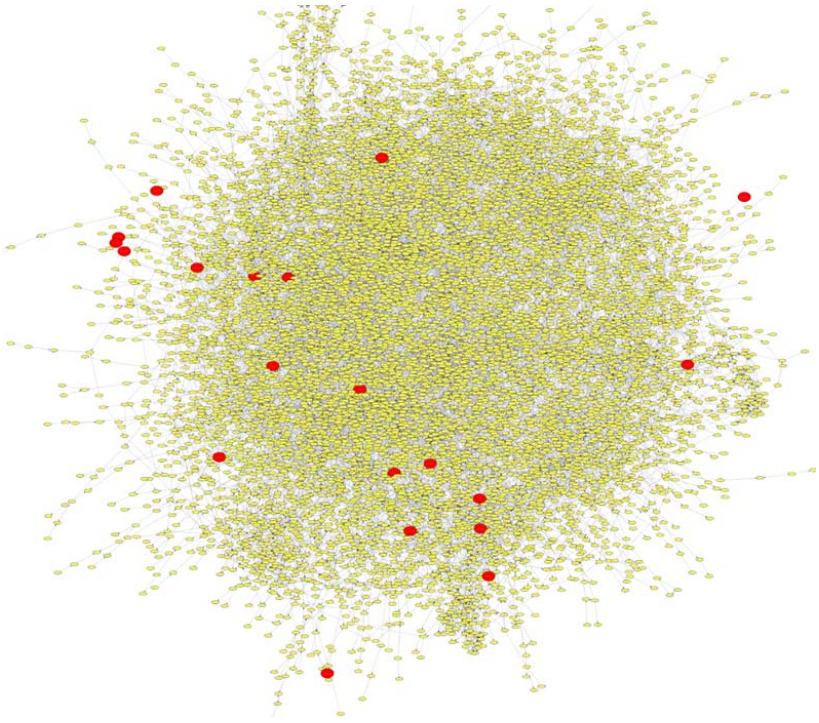

図９　寒冷刺激が与えられた時のベージュ脂肪細胞での生体分子ネットワーク（赤い点が microRNA で黄色の点が遺伝子）