意水準5%の条件で細胞株間の平均値に差がないという帰無仮説が棄却できたもの，すなわち10 細胞株の中で発現が飛びぬけて多いもしくは少ないものが少なくとも 1 つは存在する結果が出たProbe Setは次のフィルターをかけ，いずれの細胞株間でも有意差が認められなかったProbe Setは棄却した．

[フィルター3]

細胞株 10 種類の最低の平均値と最高の平均値の差が5倍以上あるProbe Setをスクリーニングし，差が5倍より小さいものは棄却した．

B-6-6 定量的PCR

胚葉体から抽出した RNA の逆転写反応は，High Capacity RNA-to-cDNA Kit（アプライドバイオシステムズ）を用いて，プロトコールに従って行った．1 μg total RNAから合成したcDNAを TaqMan Gene Expression Master Mix（アプライドバイオシステムズ）と混和し，Bock ら（*Cell.* 2011; 144: 439-52.）の報告を参考にして選択した 97 種類の遺伝子をターゲットとするTaqManプローブとプライマーの入った384ウェルTaqMan Array Micro Fluidic Cards（アプライドバイオシステムズ）にアプライした後，7900HT Fast Real-Time PCR System（アプライドバイオシステムズ）を用いて duplicate で測定した．PCR条件は，50°C, 2 min; 94.5°C, 10 min; 97°C, 30 sec, 59.7°C, 1 min, 40 cycles で行った．発現量の補正はGAPDHにより行い，$\Delta\Delta C_T$法（*Methods.* 2001; 25: 402-8.）により相対的な遺伝子発現量を算出した．平均値と標準偏差により標準化（z スコア化）した後，統計ソフトウェア SYSTAT（SYSTAT Software）により主成分分析し，外胚葉，中胚葉，内胚葉分化の指標となる主成分を得た．

B-6-7 倫理面への配慮

ヒト iPS 細胞を用いる場合は，国立医薬品食品衛生研究所「研究倫理審査委員会規程」を遵守した上で研究を実施した．遺伝子組換え実験に関しては，国立医薬品食品衛生研究所「遺伝子組換え実験安全管理規則」に基づき，遺伝子組換え実験計画書の承認を得た上で研究を実施した．

C．研究結果

C-1 新規免疫不全動物を用いた造腫瘍性試験法の開発

C-1-1 NOG-hr マウスにおけるヒト細胞生着能の定量的評価

移植後16週における NOG-hrマウスのHeLa細胞単体移植あるいはマトリゲルとの混合移植でのTPD50はそれぞれヌードマウスの 1/13（$3.74\times10^4/4.21\times10^5$），1/2040（$2.07\times10^2/4.21\times10^5$）であり，NOGマウスにおけるそれら（HeLa細胞単体移植：ヌードマウスの1/33（$1.29\times10^4/4.21\times10^5$），マトリゲルとの混合移植：ヌードマウスの 1/5431（$7.76\times10^1/4.21\times10^5$））よりも高く，異種細胞生着能は僅かに低いことが明らかになった．なお，マトリゲルによる生着性増強効果は，NOG-hrマウスで152倍であり，NOGマウスの165倍とほぼ同等であった（Table 1）．

C-1-2 NOG-hrマウスにおけるヒト臍帯血由来造血幹細胞（hCD34陽性細胞）移入後の血球球細胞分化能

NOG-hrマウスにおけるヒト臍帯血由来造血幹細胞移入後の hCD45陽性細胞への分化率は観察期間を通じてNOGマウスよりも低く，特に移入後13週までは有意であった（Fig. 3A）．しかし，分化した細胞中における B細胞および T細胞の出現パターンには NOGマウスとの差はなく（Figs. 3B and 3C），hCD4陽性細胞およびhCD8陽性細胞の比率にも系統間差はなかった（Figs. 3D and 3E）．

C-1-3　NOG-hrマウスのNK活性および補体溶血活性

NOG-hrマウスの脾臓重量（湿重量および比体重重量）はNOGマウスよりも重かったが（Figs. 4A and 4B），NK活性はNOGマウスと同様，認められなかった（Fig. 4C）．また，補体溶血活性もNOGマウスと同様，検出されなかった（Fig. 4D）．

C-1-4　NOG-hrマウスの外貌特性

NOG-hrマウスは全頭ともほぼ同じ周期で換毛を繰り返し，全身における被毛スコアは週齢の経過と共に低くなった（Fig. 5，中段グラフ）．第1ピークから第2ピークの間隔は 4.6±0.5週で，第2ピークから第3ピークは5.1±0.9週であった．また，部位別では第1～2ピークでは頭部，胸部，腹部ともに被毛の生え方は均一であったが，第3ピーク（19-20週齢時）時には腹部を中心に被毛が密生した（Fig. 5A and 5B，下段グラフ）．その他の外貌特性では全ての動物に過長爪が観察されたが（Fig. 6A），骨格に異常は認められなかった（Figs. 6B-6D）．その他，眼脂や眼球白濁化（Figs. 6E and 6F）が多くの動物に観察された．

C-1-5　BRG, BRG-nu および BRG-hrマウスにおけるヒト細胞生着能の定量的評価

移植後16週における BRGマウスのHeLa細胞単体移植あるいはマトリゲルとの混合移植でのTPD50はそれぞれヌードマウスの 1/7 ($1.00\times10^4/6.83\times10^5$)，1/2157 ($3.2\times10^1/6.83\times10^5$) であり，BRG-nu マウスでは 1/83 ($1.78\times10^3/1.47\times10^5$)，1/2163 ($6.8\times10^1/1.47\times10^5$) であった．また，移植後10週における BRG-hrマウスのHeLa細胞単体移植あるいはマトリゲルとの混合移植でのTPD50はそれぞれヌードマウスの1/10 ($3.17\times10^4/3.17\times10^5$)，1/682 ($4.65\times10^2/3.17\times10^5$) であった．

マトリゲルによる生着性増強効果は 3系統間で26～317倍の開きがあった（Table 2）．

C-2 幹細胞の *in vitro* 培養工程における遺伝子発現の動態解析による品質評価技術の開発

C-2-1　細胞・組織加工製品に用いられる間葉系幹細胞（hMSC）の品質評価—がん化の指標探索のための遺伝子発現解析

Ewing 肉腫 4 種類（Hs822.T, Hs863.T, RD-ES, SK-ES-1）を陽性対照として hMSC と比較検討することにより，細胞のがん化の指標となり得る候補遺伝子として CCND2, IGF2BP1 など 9 遺伝子を抽出した．（Table 3）そこで次に，Cyclin D2 及び IGF2BP1 発現組換えレンチウイルスベクターを作製し，hMSC に感染させた．感染後の hMSC から total RNA を抽出し，RT-PCR 法によって Cyclin D2 及び IGF2BP1 の発現量を測定した．その結果，Cyclin D2 は約 10,000 倍，IGF2BP1 は約 2 倍，発現が上昇していた．また感染後の細胞形態を位相差顕微鏡で観察した所，hMSC/Puro 及び hMSC/Neo, hMSC/IGF2BP1 は，感染前の細胞と形態は変わっていなかったが，hMSC/CyclinD2 の細胞形態は小さくなっていた．次に Cyclin D2 及び IGF2BP1 の強制発現によって細胞の増殖速度が変化するか調べた．その結果，hMSC/CyclinD2 は hMSC/Puro よりも明らかに増殖速度が上がっていた．一方，hMSC/IGF2BP1 の増殖速度は hMSC/Neo と比べ大きな変化はなかった．

Cyclin D2 発現組換えレンチウイルスベクター感染 14 日後の hMSC の total RNA を抽出し，細胞周期に関わる遺伝子（p16，p21，Bmi1）及びがん細胞の転移に関わる遺伝子（MMP2）の発現変化を調べた．Cyclin D2 導入 14 日後では p16，p21 遺伝子の発現に大きな違いは見られなかった．一方，Bmi1，MMP2 遺伝子の発

現は Cyclin D2 導入細胞で低くなっていた．hMSC を長期（3ヶ月）にわたり培養し，その間，2週間毎に total RNA を抽出して細胞周期に関わる遺伝子（p16，TGF-β2）の発現変化を RT-PCR 法により測定した．その結果，hMSC/CyclinD2 及び hMSC/Puro ともに継時的に p16 の発現が上昇したが，hMSC/CyclinD2 の方がその上昇率は高かった．一方，TGF-β2 の発現は全ての培養期間で hMSC/CyclinD2 の方が低かった．

Cyclin D2 強制発現により遺伝子の発現にどんな変化が起こっているか網羅的に調べるために，hMSC/CyclinD2 及び hMSC/Puro から total RNA を抽出し，mRNA のマイクロアレイ解析を行った．Cyclin D2 強制発現により遺伝子発現が 2 倍以上変化した遺伝子は 690 個あった．これらの遺伝子の発現変化について IPA によりパスウェイ解析を行い関与する細胞機能について検討したところ，「細胞増殖」や「細胞周期」が Activation z-score が 2 以上あり統計学的，生物学的に有意に亢進されることが示された．（Table 4）発現レベルが 2 倍以上変化した 690 個の遺伝子の中で，細胞増殖に関わる遺伝子 186 個のうち 94 個が細胞増殖を促進する方向に発現レベルが変化していた．また細胞周期に関わる遺伝子については，50 個のうち 19 個が細胞周期を進める方向に発現レベルが変化していた．（平成 24,25 年度）

C-2-2　hMSC におけるレトロトランスポジションの解析とその影響について

C-2-2-1　間葉系幹細胞における LINE-1s の発現について

これまで iPS 細胞やがん細胞で LINE-1s の発現が確認されていたが，hMSCs での発現については報告がなかった．そこで，ヒト iPS 細胞（201B7, 253G1, 409B2, R-1A, R-2A, R-12A, Ai100, Ai103, mc-iPS, TiC），HeLa 細胞を陽性対照とし，Lonza 社から購入した骨髄由来 hMSCs（8F3211, 8F3434, 8F3560）における LINE-1s 発現を qRT-PCR により調べた．その結果，解析した hMSCs のすべてのロットで LINE-1s は発現しており，その発現量は iPS 細胞や HeLa 細胞以上であった．（Fig. 7）（平成 25,26 年度）

C-2-2-2　hMSC における LINE-1s の発現に及ぼす APOBEC3B（A3B）の影響について

LINE-1s の転移を抑える細胞内因子として A3B が知られているが，この遺伝子には欠失多型が存在し，日本人にその割合が多いと報告されている．A3B を発現しない日本人由来 hMSCs は LINE-1s の転移によってゲノムの安定性が損なわれる可能性が考えられたので，A3B 遺伝子型と LINE-1s の発現量について解析を行った．

医薬基盤研究所から購入した日本人 25 人分の hMSCs において A3B 遺伝子型の解析を行ったところ，野生型ホモ個体（Ins/Ins）が 14 人，野生型/欠失型ヘテロ個体（Ins/Del）が 9 人，欠失型ホモ個体（Del/Del）が 2 人であり，欠失型アリル頻度は 26%であった．

それぞれの細胞から mRNA を抽出し，A3B mRNA の発現を定量解析したところ，A3B 遺伝子型と mRNA 発現量にある程度の相関は見られたが，それぞれの遺伝子型の発現量の差は統計学的に有意な差ではなかった．また，LINE-1s mRNA の発現量を定量解析し，A3B mRNA と LINE-1s mRNA の発現量を比較したが相関関係は見られなかった．（Fig. 8）

次に Ins/Ins（Yub633, PL523），Ins/Del（Yub621, Yub631, Yub 10F），Del/Del（Yub637b, PL523）の LINE-1s ORF2 領域の遺伝子配列を調べた．転移能力が残っていると報告されている LINE-1s（LINE-1.3, $L1_{\beta\text{-thal}}$, $L1_{RP}$）と比較した結果，Ins/Ins の変異頻度が 4.24%（884 mutations/20839 bp），Ins/Del が 3.46%（967

mutations/37942 bp), Del/Del が 0.794% (202 mutations/25410 bp) であった. また Ins/Ins や Ins/Del ではシトシン (C) →チミン (T) やグアニン (G) →アデニン (A) の変異が多く見られたが, Del/Del ではそれ以外の変異も見られた. (平成 25 年度)

次に, 次世代シーケンサーを用いて A3B 野生型ホモ (Ins/Ins) 由来の hMSCs (Yub633) 及び欠失型ホモ (Del/Del) 由来の hMSCs (Yub637b) における転移可能な LINE-1s の発現解析を行うため, まず, ヒトゲノム配列 (GRCh38) に対して, *in vitro* で転移活性が確認されている LINE-$1_{\square\text{-thal}}$ ($L1_{\square\text{-thal}}$; Genbank Accession No.AF148856), LINE-1_{RP} ($L1_{RP}$; AF149422) 及び LINE-1.3 (L1.3 ; L19088) 配列の BLAST 検索を行い, ヒトゲノム配列中に存在する転移可能な LINE-1s 配列領域を予測した. GRCh38 中の $L1_{\square\text{-thal}}$, $L1_{RP}$ 及び L1.3 様配列はそれぞれ, 58166 個, 58195 個, 58490 個であった. RNA シークエンス解析により取得した Yub633 (Ins/Ins) 及び Yub637b (Del/Del) から抽出した total RNA の配列データを GRCh38 にマッピングし, BLAST 検索により予想された $L1_{\square\text{-thal}}$, $L1_{RP}$ 及び L1.3 様配列にマッピングされたリード数をカウントすることで, 転移可能な LINE-1s 配列の発現量を算出した. 58166 個の $L1_{\square\text{-thal}}$ 様配列の normalized read count (各検体の総取得リード数を考慮し補正を行った検体間比較用の予想リード数) の合計は Yub633 (Ins/Ins) が 241.0167, Yub637b (Del/Del) が 249.5887 であった. また, 58195 個の $L1_{RP}$ 様配列の normalized read count の合計は Yub633 (Ins/Ins) が 240.8315, Yub637b (Del/Del) が 248.8003 であった. 58490 個の L1.3 様配列の normalized read count の合計は Yub633 (Ins/Ins) が 240.8571, Yub637b (Del/Del) が 249.2632 であった.

LINE-1s の変異の中で塩基の欠失はフレームシフトを起こし, LINE-1s の転移活性に負の影響を与えると考えられる. 本研究では転移活性の残った LINE-1s の発現解析を行うため, LINE-1 全長に占める割合 (qcov) が 99%以上のものを抽出し, さらに LINE-1 との一致度 (pid) がそれぞれ 95%未満, 95%以上 96%未満, 96%以上 97%未満, 97%以上 98%未満, 98%以上 99%未満, 99%以上における normalized read count の量を解析した. 塩基配列の一致度 95-98% までの LINE-1s の発現は Yub633 (Ins/Ins) に比べて Yub637b (Del/Del) の方が高く, 98-100%でそれが逆転していたが, 全体として両者における LINE-1s の発現量に大きな違いは見られなかった. (Fig. 9) (平成 26 年度)

C-2-2-3 hMSCs の未分化性と LINE-1s 発現との関連について

in vitro での細胞培養による LINE-1s 発現への影響を調べるために, 継代数 3 (P=3) から継代数 11 (P=11) まで hMSCs を培養し, 細胞数, 増殖速度及びその時の LINE-1s の発現を解析した. 細胞の増殖速度は徐々に低下しており, 一方 P=4, 6, 8,10 における LINE-1s の発現は P=4 から P=6 で大きく低下し, その後はロットによって異なっていたが, 平均して低下していた.

さらに, 細胞の分化による LINE-1s 発現への影響を調べるために, 脂肪分化培地で hMSCs を培養することにより hMSCs を脂肪に分化させた. 陰性対象として増殖培地で培養した hMSCs を用いた. BODIPY Lipid Probes により脂肪球を蛍光染色し, hMSCs 3 ロットにおける脂肪球蛍光面積を比較した. また, 脂肪分化培地及び増殖培地で培養した hMSCs 3 ロットにおける LINE-1s 発現を qRT-PCR で調べたところ, 脂肪分化させた hMSCs で LINE-1s の発現が低下していた. (Fig. 10) (平成 26 年度)

C-3 次世代シークエンサーを用いた細胞の遺伝的安定性評価指標の開発

C-3-1 ホールゲノムシークエンス解析による細胞の品質評価

遺伝子変異や欠失と増幅を含めたコピー数変化，さらには遺伝子転座等の染色体異常といったゲノムの異常を，どの程度シークエンス解析により検出できるかを確かめる目的で，既にSNPアレイを用いたCGH解析等により，遺伝子変異および増幅の起こっている場所に関する詳細な情報を得ているHL60-RG細胞を用いて，ホールゲノムシークエンス解析を行った．

イルミナ HiSeq2000 シークエンサーにて100bpリード2回のランにて読み取られたリード数は約17.5億で，塩基数としては1767億bpに達した．ヒトのゲノムは約30億塩基対であるため，平均重複度は約60となった．重複度（sequence depth）の分布をFig. 11に示す．

得られたシークエンスデータをヒトリファレンスゲノムhq19に対してマッピングした結果，最終的にマップ可能であったリードの割合は全体の93.1%と良好であり，リファレンスゲノム上の99.2%をカバーできた．ベースコールにある程度の信頼度を持てる重複度を10とした場合のカバー率は98.9%であった．よって十分な情報が得られなかったのはゲノム上わずか1%程度であり，遺伝子上のエクソン配列に関してはほぼ網羅できていると考えられる．

SAMTOOLSソフトウエアを用いてリファレンスゲノムからの変異を検出した結果，SNPの総数として3,545,099個，Indel（インサーションおよびデリーション）の総数565,658個，合計約400万箇所が抽出された．これらは，重複度10以下の信頼度の低い変化も含むため，一定の基準を設けてさらに吟味を行う必要があると考えられる．SAMTOOLSによるSNP qualityのスコアが参考になると考えられるが，暫定的には，ホモSNPの場合アレルあたり10call以上，変異率8割以上あれば確実であると考えている．また，ヘテロSNPの場合には，15 call以上，変異率4-6割程度が妥当であると考えられる．

HL60細胞には，ガン関連遺伝子に関する変異がすでにいくつか報告されている．これらが，ホールゲノムシークエンスで検出できたかどうかを検証してみた．塩基対置換の例としてCDKN2A((p16)遺伝子，およびNRAS遺伝子に関する結果をTable 5に示した．それぞれ，26および49の重複数で100%が変異塩基callであり，変異が確認できた．CDKNA2遺伝子の場合は，報告ではHomo変異であり今回の結果と一致していたが，NRAS遺伝子の場合には，報告ではHeteroであったが，今回の結果はHomoになっていた．これは，今回用いたHL60-RG細胞株がオリジナルなHL60細胞から派生した増殖性を示すsublineであることに起因している．これまでのSNPアレイを用いた検討により，RG細胞株では1番染色体でuniparental disomyが起きており，染色体全体においてLOHが起きていることがわかっていたが，それを反映して1番染色体で観察されたSNPおよびIndelは基本的に全てHomoであり，uniparental disomyが確認できた．RG細胞株にて残った側のアレルはNRAS変異を持っており，機能的に細胞の増殖性獲得との関連性が示唆された．

その他の部分のSNPとIndelについては，重複度は異なるものの，homoアレルに関してはほぼ100%でCallが一致しており，シークエンス解析の正確性がかなり高いことがわかった．これを，サイズの最も小さい22番染色体を例に染色体全体に関して検証してみると，Fig. 12に示すように，Homo callでは約95%が100%の一致率を示しており，90%以上の一致を示したものは約95%であった．一方，hetero callの

SNP サイトに関しては，変異 Call の割合を調べたところ，Fig. 13 に示すように，50%を中心に均等に分布しており，4 から 6 割の間に全体の 83%のデータが存在した．

次に，Homo call において 1 または 2 の別塩基コールがされた場合をシークエンスエラーと仮定した場合，この頻度は 22 番染色体全体において 2013/1039162=1.9%であり，約 2 %のエラー率であった．この程度のエラー率の場合，重複度 10 で読んだ際に，8 割以上ミスコールする確率は計算上 $(0.02)^{8}x(0.98)^{2}x_{10}C_{2}+(0.02)^{9}x(0.98)x_{10}C_{1}+(0.02)^{10}$=約 10^{-12} となり，ゲノム全体の $3x10^{9}$ 塩基中ほぼ 0 となる．同じく，hetero SNP を仮定して，10 リード中 4 から 6 リードミスコールをする確率は，$(0.02)^{4}x(0.98)^{6}x_{10}C_{4}+(0.02)^{5}x(0.98)^{5}x_{10}C_{5}+(0.02)^{6}x(0.98)^{4}x_{10}C_{4}=3x10^{-5}$ となり，無視できない数となる．重複度を 20 に増やして 4 から 6 割とすることで，この確率は 2.5×10^{-9} になり，ゲノム上 1 未満に押さえられることになる．

次に，このシークエンスリードの重複度を指標として，CGH 的にコピー数の変化を検出できるかどうかを検証した．HL60 細胞に関しては，すでに報告されているようにガン遺伝子 c-myc の増幅があることがわかっているが，この遺伝子増幅はシークエンス変化を伴わないため，リファレンスゲノムからの変化としては検出されなかった．ただし，リード数は遺伝子の存在量を反映するため，8 番染色体の c-myc 領域に関して，検出された SNP と Indel のリード数の変化をプロットした．その結果，Fig. 14 に示すように，従来 CGH アレイ等で得られていたのと同様に，複数の領域にまたがった遺伝子の増幅領域が再現できた．これにより，シークエンス解析データは CGH 解析として利用可能であることがわかった．

さらに，この遺伝子増幅領域に関しては，Fig. 15 に示すように，複数の増幅領域が複雑につなぎ合わさったリアレンジメントが起きていることが確かめられているが，通常の解析においてはシークエンスの変化として検出できなかった．転座タイプの変異の場合には，読まれた断片のアラインメントの際に，リファレンスゲノムにアラインすることが難しいと考えられ，アラインメントができない断片となったことが予測される．こうした，アラインメントが付かないフラグメントに注目し，独自のアルゴリズムによる転座配列の解析法の開発が必要であると考えられる．

HL60 細胞は p53 遺伝子の欠失変異をホモに持つが，通常のシークエンス解析においては，欠失領域が大き過ぎるため deletion 変異としては検出できなかった．比較的大きな挿入欠失変異に対しては CNV の解析からのアプローチが必要である．特に，ヘテロ欠失の場合には現存する正常アレルの影響によって，変化なしと判断される可能性が高い．1/2 のコピー数を正確に検出できるデータ量，および解析アルゴリズムが必要となる．

ホールゲノム解析に関しては，HL60-RG 細胞に続き，ヒト骨髄由来間葉系幹細胞（hMSC）に関しても，ホールゲノムシークエンス解析を行った．イルミナ HiSeq2000 シークエンサーにて 101bp リード 4 回のランにて読み取られたリード数は約 18.5 億で，塩基数としては 1871 億 bp に達した．ヒトのゲノムは約 30 億塩基対であるため，平均重複度は約 60 となった．重複度（sequence depth）の分布を Fig. 16 に示す．

得られたシークエンスデータをヒトリファレンスゲノム hq19 に対してマッピングした結果，最終的にマップ可能であったリードの割合は全体の 91．7%と良好であり，リファレンスゲノム上の 99.3%をカバーできた．ベースコールにある程度の信頼度を持てる重複度を 10 とした場合のカバー率は 99.1%であった．よって十分な情報が得られなかったのはゲノ

ム上わずか1%未満であり，遺伝子上のエクソン配列に関してはほぼ網羅できていると考えられる．

SAMTOOLSソフトウエアを用いてリファレンスゲノムからの変異を検出した結果，SNPの総数として4,475,876個，Indel（インサーションおよびデリーション）の総数687,240個，合計約500万箇所が抽出された．

今回解析に用いたhMSC細胞は，クローナルな染色体異常を持つことが既に確認されている細胞であり，今回得られたホールゲノムシークエンスデータの各染色体部位のリード数(冗長度)を用いた擬似的比較ゲノムハイブリダイゼーション(CGH)法による解析の結果，Fig. 17に示すように，既に得られている7番および17番染色体の部分的コピー数の増減が確認できた．コピー数変化に関しては，塩基配列レベルで高精細な情報が得られるため，今後さらに詳細なりアレンジメントの解析に応用できると期待できる．さらに，従来のアレイCGH法等によっても検出されていなかった微細なコピー数変化も検出され，その確認が必要となった．

次に，すでに得られたHL60細胞のホールゲノムシークエンスデータについて，遺伝子レベルでの詳細なコピー数変化の検討を加えた．コピー数変化の解析は，10kbを単位として冗長度の平均を取ったデータを使用したため，コピー数変化の最小単位は10kbとなっているが，実際にはさらに細かい単位での検討も可能である(25年度報告書データ参照)．

全体としては，欠失に比べて増加の方が多く，領域としては短いものから長いものまで幅が見られた．一つの特徴としては，比較的近傍に増加または減少の同じ方向の変化がまとまって見られた点であり，代表的なものが8番染色体のc-myc領域（8q24）における複雑な増幅であり，合計7つの領域が約16コピーという類似した増幅度で変化していた．このうち一つの領域に関してはコピー数が33と他の約二倍であり，この領域を2回含んだ増幅単位が1ユニットとして増幅したことを示している．

欠失については主にコピー数は1と片側のアレルのロスを示したが，部分的にはホモ欠失の領域も認められた．注目すべきは17番染色体の欠失領域であり，Table 7に示したように，ヘテロ欠失として短腕全体(17p)のロスが見られるが，さらにこのうちの二つの領域，それぞれ癌抑制遺伝子であるTP53とARHGAP44（Rho GTPase activating protein 44）を含む短い領域はホモに欠失していることがわかった．即ち染色体のFISH解析により7pのヘテロ欠失は簡単に観察されていたが，正常に見えた片側のアレルにも欠失領域が存在し，短いため染色体解析においては検出不可能であった変化が，シークエンス解析データにより検出できたことになる．

また増幅している領域には，MAP2K3やCDC27という細胞周期に関連した遺伝子が存在し，癌化との関わりが示唆された．17番染色体以外においても，コピー数変化の見られた領域には重要と思われる遺伝子が多く含まれており，HL60細胞の癌化のヒストリーを反映している可能性が考えられる．

さらに，コピー数変化のあった領域について見てみると，レトロトランスポゾンの挿入サイトを含めた繰り返し（相同）配列が特徴的に観察され，これらが染色体異常の生成に関与している可能性が示唆された．

C-3-2　エクソンシークエンスの利用

HL60細胞に関しては，ホールエクソンシークエンス解析も行ったが，データが間に合わなかったため，以前に行ったhMSCのホールエクソン解析の例を引用して，その利用法に関して考察を加えることにした．用いたhMSC細

胞に関しても，遺伝子増幅や転座などのリアレンジメントがあることがわかっていたが，エクソンシークエンスの結果からは，それらを完全に反映することは難しかった．これは，エクソンのみではゲノムの数パーセントを読むに過ぎないことから来る必然的な限界であり，エクソンシークエンスとしては，既存遺伝子のSNPおよび突然変異の検出という，遺伝子機能に根ざした利用法に重きが置かれると考えられる．検出された変異は遺伝子の機能に直結するため，細胞の機能的変化を解析する上では効率が良い．また，ゲノムワイドな遺伝子不安定性を検出する目的において，指標遺伝子として利用することを考える場合には，たとえ数パーセントでも十分に利用価値はあるといえ，そうした利用法も考慮できる．ただし，ホールゲノムシークエンスの場合も含め，通常の次世代シークエンサーから得られるシークエンス情報は，全体のポピュレーションのマジョリティーを反映するものであり，遺伝的不安定性の誘発によりマイナーな変異が頻発した場合にあっても，それを検出することが難しいと予想される．変異として検出されるためには，シークエンスの重複度にも依存はするが，通常重複度50程度でシークエンスをする場合には，シークエンスエラーの可能性を考慮すると，変異として確定するためには最低5程度のリード数（10%）が必要であり，ヘテロ変異の場合には，その倍即ち細胞のポピュレーションとして2割程度は必要であると考えられる．この感度を高めるためには，遺伝子を限定して重複度を増やしたディープシークエンスが必要となる．ただし，その場合にも，シークエンス解析の前処理にPCR反応を用いる場合には，そのエラーレベルがバックグランドノイズとり，感度的限界となる．そこで，その対策としては，PCR反応を解さず直接シークエンス解析が可能な，1分子シークエンサーの利用が必要となる．

C-3-3　ホールゲノムシークエンスデータによる細胞の遺伝子増幅およびリアレンジメントの解析

これまでに得られた，HL60細胞およびTK6細胞のホールゲノムシークエンスデータを用い，これら細胞における遺伝子増幅および染色体転座のシークエンスレベルでの詳細な解析を行った．

まず，HL60細胞においては，8番染色体のc-myc領域における複雑な遺伝子増幅と染色体のリアレンジメントに関して，CGHアレイを用いた解析から転座予想部位の同定を行ったが，このデータを元にして，WGSデータのアラインメントを行うことにより，転座点の配列情報と再配列の様式に関して確認を行った．

Fig. 15に示した各増幅単位の切断部位の予想配列（hg18由来）に相当するhg19でのリファレンスシークエンスを元に，転座リファレンス配列を合成した．この際，これまでの検討から，融合部位に余分な配列を含む場合があったため，各セグメント間に10個程度の未知塩基（N）を人為的に挿入した配列を合成し，この配列をリファレンス配列としてWGSデータのアラインメントを行った．

その結果，Fig. 18に示すようにこの配列に相補的なリードのアラインメントが得られたが，配列上で部分的に高頻度のアラインメントが得られている箇所が複数有り，これらは，LINE1等のゲノム上の単純繰り返し配列に由来することがわかった．通常のアラインメントにおいては，これら繰り返し配列は，ゲノム上に均一に分散しているために結果として，平均化して元来のコピー数（2）を反映した冗長度がいずれの部位でも得られるが，特定の部分的な配列のみを取り出してアラインメントをした場合には，全ゲノム上の繰り返し配列由来のリードが集中して張り付くため，見かけ上極端

に冗長度が高い部位となって現れることになる. これは元々のコピー数を反映するものではなくアラインメント上のアーティファクトであるため, 類似の解析を行う場合には注意が必要であることがわかった.

リファレンス配列にアラインメントされた配列は, 実際に切断点をまたがって異なるセグメント間に及ぶものが複数確認できるとともに, ペアーエンド同士のリードが隣接する異なるセグメント間にアラインメントされていた事実からも, 予想された部位と順序でリアレンジメントが起こっていることが確認された. 最終的に確認された切断点の配列を Table 6 に示す.

次に, 転座部位に関する正確な情報の得られていないTK6細胞株を用いてWGS解析により転座切断点の同定を試みた. TK6 細胞においてはこれまでの染色体解析より Fig. 19 に示すように, 正常に近い核型を持つが, 3 番と 21 番, および 14 番と 20 番染色体間に転座を有することが, multicolor-FISH 解析によりわかっている. この転座点の配列を WGS データより決定するために, 次のようなアルゴリズムを構築した. イルミナシークエンサーからのリードは, インサート配列を両端から読んでいるため, 同一インサートフラグメント約数百 bp に対して, 両端から読んだ約 100bp のリード情報がペアで存在し, これらをメイトペアと読んでいる. 通常メイトペアは染色体上の近傍にマップされるが, 切断店をまたぐフラグメントからのシークエンス情報は, それぞれ別の領域にマッピングされることになる. この点を利用して, メイトペアがそれぞれ, 3 番と 21 番, および 14 番と 20 番染色体上にマップされたリードの情報を抽出した. 切断点をまたぐリードは残念ながら通常のマッピング解析ではアラインメントされず捨てられてしまうが, その近傍のリードについてはそれぞれ異なる部位に完全にマッチするためマップされたリードとして存在するので, このアルゴリズムで検出可能となる.

このようにしてマッピングデータから該当するメイトペアを抽出したところ, 膨大な数の切断点の候補が得られた. ランダムなアラインメントエラーも考慮し, 抽出されたリードが集中している領域に関して, そのゲノム配列を UCSC ゲノムブラウザにて確認したところ, そのほとんどが, LINE-1, LTR などのゲノム上に散在する繰り返し配列の位置と一致した. 即ち, こうした繰り返し配列はその相同性のため本来とは別の染色体上の位置にミスアラインメントされる可能性が高く, こうした解析をする場合に偽陽性結果を与えることがわかった. ただし, 一部には各染色体に固有の領域にマッピングされるペアが存在し, これらは転座切断点の候補となる. 今後得られた位置情報よりプライマーを設計し, 切断点の増幅によるクローニングを行い, シークエンスの確認を行う予定である.

一方, 別のアプローチとして, 新たなアルゴリズムによるシークエンスデータ解析からの検討も行っている. 本来マッピング操作時に, リファレンスシークエンスにはアラインメントされずに捨てられているデータの中に切断点を含むリードが含まれているはずであるので, これかの中から, 部分的に完全にマッチして異なる部位にマッピングされる配列を拾い上げるためのアルゴリズムの構築を行っている.

C-3-4　1 分子シークエンサーを利用した高感度変異検出系の確立

TK6 細胞を代表的な変異原物質である ENU, MMS, γ 線にて処理し, tk 遺伝子を用いた遺伝子突然変異試験により, 変異の誘発を確認した. (Table 8). 変異原物質処理による細胞の生存率は, 2-8 割と処理により異なったが, tk 遺

伝子の変異頻度はコントロールに比較して40-100 倍と有意に増加した．これらの細胞から，今回は mtDNA エクストラクターCT キット（WAKO）を用いて，ミトコンドリア DNA の抽出を行った．得られた DNA の電気泳動像を Fig. 20 に示すが，ゲノム DNA 由来と思われる全体にスメアなバンドの中に，16kb くらいのミトコンドリアサイズに相当するバンドも確認された．

これら DNA サンプルを元にして，PacBio シークエンサー用のライブラリーを調整し，1 SMART cell 分のシークエンス解析をおこなった．

その時のパフォーマンスを Fig. 21 に示すが，1 SMART cell あたり，トータルで 7～500Mb に相当するシークエンス情報が得られた．組み込まれたインサートサイズに相当する Subread length の平均長は 1.5kb ぐらいであり，リードクオリティーによるフィルター後のポリメラーゼ読み取り長が 8ｋｂ程度であることから，インサートは 4-5 回程度繰り返して読まれていることになる．サブリードごとの配列情報をそのままヒトミトコンドリアリファレンスシークエンス(hg38)にマッピングをすると，概ね10%近くの変異があり，かなりエラーレートが高いことがわかった．

そこで，Pacific Bioscience 社の解析パイプライン smartanalysis version2.3.0 に含まれる BLASR(PacBio long read aligner)プログラムを用いて，重複リードを考慮したマッピングをし，samtools(version0.1.19)を用いて変異コールをした結果，Table 9 に示した数の変異箇所が同定された．マッピングデータを可視化可能な Tablet ソフトウエアを用いて詳細に検討をしたところ，いずれの部位においても raw data 上は変異の存在が確認できた．すなわち共通して変化している部位に関しては，TK6 に元来存在している変異として検出された．

次に，問題となる新たな低頻度の変異の検出に関しては，TK6_cont-1 および ENU のサンプルにおいて，1000，および 2000bp 周辺の 2560 bese call に対してリファレンスと異なる bese call の数を計算したところ，cont-1 ではそれぞれ 48+61＝109，ENU では 36+27＝63 と，コントロールの方が高かった．変異 Call の頻度は，109/2560＝0.043 と 5%弱であり，これはシークエンスエラーと考えられるため，誘発変異の検出のためには，よりエラー率を落とす必要があることがわかった．

C-3-5　BLM ノックアウト細胞を用いた細胞の遺伝的不安定性の評価

細胞の遺伝的不安定性を検出するためのモデル細胞として，国立医薬品食品衛生研究所変異遺伝部において開発された BLM 遺伝子ノックアウト細胞を使用した．BLM 遺伝子は染色体不安定性を示すブルーム症候群の原因遺伝子であり，DNA 二本鎖切断の修復酵素である DNA ヘリカーゼをコードしている．この遺伝子を破壊した TK6 細胞株は，親株に比べて高い染色体異常および突然変異誘発性を持つことが確かめられている．

この TK6 細胞 BLM 欠損株の遺伝子配列を，親株の TK6 細胞と比較して，突然変異およびコピー数変化を検出するため，ホールゲノムシークエンス解析を行った．また同時にミトコンドリアのシークエンス解析も行った．

WGS 解析により得られたリードをリファレンス配列 hg19 にマッピングすることにより TK6 および TK6/BLM 細胞の SNP 部位の抽出を行った．いずれも約 233Gb のデータから平均冗長度 73，カバー率 99.8%でマッピングされ，得られた SNP の数は 370 万箇所に及んだ．このうち，二つの細胞で異なる SNP コールがされた箇所を抽出し，内容を吟味したところ，ほとんどがリファレンスに対して hetero SNP と

なる箇所のコールの選択の差によるものであり，新たに生じた変異であると考えられる箇所は僅かであった．

今回の検討においては，TK6/BLM 細胞を分離，培養後にクローニングを行わなかったことより，遺伝的不安定性により変異の誘発率が上がっていても，NGS による検出が難しかったと考えられる．わずかに得られた真の変異はおそらく TK6/BLM 細胞の樹立過程でシングルコロニーアイソレーションが行われたことによる選択の影響であると考えられる．

TK6/BLM 細胞との比較とは別に今回得られた TK6 細胞の WGS データの解析から，CGH データの取得によるゲノム異常の検出，およびコピー数変化領域と SNP の高頻度領域が一致するという知見が得られた．17 番染色体ではゲノムコピー数が約 3 倍に増加している部分があり，この領域におけるマップデータを詳細に検討した結果，部分的に最大 6 種類の配列バリエーションがあることが判明した．このことは，3 倍に増加した計 6 本のアレルがすべて異なる配列を有していることを示している．この領域においては，SNP の発生頻度が高く，増幅を伴うゲノム異常との関連性が注目された．残念ながら，イルミナシークエンサーから得られる各リードは約 100bp と短いため，詳細な SNP の連鎖解析は難しかったが，今後 PacBio などロングリードのシークエンサーと組み合わせることにより，より詳細なアレル情報を取得したい．

TK6 細胞と BLM 欠損株の比較に関しては，プロテオーム解析からのアプローチも行っており，解析の結果得られた発現変化を示したタンパク質のリストを Table 10 に示す．

LC-MS/MS を用いたショットガンプロテオミクス解析で得られたペプチドピークの数は総数 85845 であり，同時に取得した MS/MS データよりデータベース検索ソフトウエア MASCOT により同定されたタンパク質の総数は 1,985 個であった．このうち，タンパク質レベルの解析において，MASCOT による同定結果の信頼性スコア 10 以上のペプチドに関して，タンパクレベルで親株に対して 2 倍以上の変化を示しかつ ANOVA 解析の p 値 0.05 未満で有意となるタンパク質の総数は 12 個であった．今後これらのタンパク質の機能と BLM 遺伝子破壊によるゲノム安定性との関連について検討を行いたい．

C-3-6　タンパク質プロファイル情報提供のための可視化ツールとしての「ProteoMap」および Web 公開「ProteoMap　Online」用ソフトウエアの開発

我々はこれまでに，LC-MS/MS を用いたショットガンプロテオーム解析により得られたデータの可視化に関する検討を行ってきたが，その経験を生かして，各種細胞のタンパク発現プロファイルに関するリファレンス情報の提供と細胞間のデータ比較を可能とするためのソフトウエア「ProteoMap」の開発を行った．

通常質量分析装置から得られるデータは，膨大な量の数値データであり，このままではその全貌および詳細をつかみにくいことから，リテンションタイムと質量数(m/z)を各軸に取った 2 次元マップ上にイメージデータとして変換して可視化を行うことにした．この際，各ペプチドピークに対してタンデムマス（MS/MS）測定が行われていた場合には，そのスペクトル情報が付随してくるが，これらも合わせて情報提供できるよう，クリッカブルマップとして，対応するペプチドピークをマップ上でクリックした際に，ピーク情報がグラフとして表示される機能を加えた．また，MS/MS 測定がされたピークに対しては，MASCOT によるデータベース検索でのタンパク質同定結果の取り込みを行い，MASCOT 検索結果を表示させる機

能も開発した. 実際のプログラミングに関しては外部の専門家に委託し, 修正と改良を加えつつソフトウエアが完成した. その内容を Fig. 22 に示す.

現在のところ使用している質量分析器が Thermo Scientific 社のものなので, ".raw""形式の生データを取り込めるよう設計した. ここでは, 細胞より得られたデータの例としてＴK6 細胞のプロテオームデータを使用した. 得られたペプチドピークの総数は 48,950 であり, これらを横軸にリテンションタイム, 縦軸に質量数（m/z）を取りプロットした.（2)）得られたペプチドピークの総数は 48950 であり, 全てのピークを 2 次元マップ上にスポットした. この 5 万近いペプチドのうち, 実際に MS/MS 測定が成されたのは約 2 割の 10,000 個であり, そのうち MASCOT 検索にて同定されたペプチドの総数は 2805 と全体の約 6 パーセントであり, ここからタンパク質が同定されたのは 931 個であった. 今後はさらに同定数を増やしていきたい.

2 次元マップ上 MS/MS データを持つペプチドピークは青または赤の印がマークされるが, 前者は MASCOT 検索にて同定結果が得られたピーク, 後者は未同定のピークをあらわす. 画面はズームイン機能を有し, それぞれのピークをクリックすることにより, MS/MS のスペクトルデータを表示させることができる.

本ソフトウエアは複数のサンプルのデータを取り込み, 相互に比較することが可能であり, 各ピークの濃度からおよその定量比較が可能であるが, 今後より定量的な比較が可能となるよう改良を加えてゆきたい.

さらに, このソフトウエアの機能を利用して可視化したプロテオームデータをリファレンス情報として Web 上にて提供できるシステムを開発した(Fig. 22). 2 次元マップ上 MS/MS データを持つペプチドピークは青または赤の印がマークされるが, 前者は MASCOT 検索にて同定結果が得られたピーク, 後者は未同定のピークをあらわす. 画面はズームイン機能を有し, それぞれのピークをクリックすることにより, MS/MS のスペクトルデータを表示させることができる.

本ソフトウエアは複数のサンプルのデータを取り込み, 相互に比較することが可能であり, 各ピークの濃度からおよその定量比較が可能であるが, 今後より定量的な比較が可能となるよう改良を加えてゆきたい.

現在は外部サーバーにて試験的に稼働を行っているが, 近日中に国立医薬品食品衛生研究所, 遺伝子医薬部の HP 上にて公開を開始する予定である.

C-4 遺伝的安定性評価ツールとしての次世代シークエンサーの性能評価

C-4-1 HiSeq システム（イルミナ社）を用いたシーケンス解析

C-4-1-1 標準ゲノム DNA の品質評価

標準ゲノム DNA の品質を評価するために, 二本鎖 DNA の量を測定した結果, DNA 総量は 20μg 以上であり, 複数の読み深度（depth）の条件でシーケンス（3μg 以上/1 解析）が行える量を取得することができた(Table 12). また, 核酸定量・アガロースゲル電気泳動を行い, 濃度及び純度の品質検定を行った結果, すべてのサンプルにおいて問題が無いと判断された（Table 13, Fig. 23).

C-4-1-2 標準ゲノム DNA のアレル頻度について

ゲノム上の 35 箇所において, デジタル PCR によって正確に測定されたアレル頻度が Table 11 に示されている. ただし, 欠失による変異に関しては, どの位置の塩基が欠失されたかを特定することが困難であると判断し, 一塩基置

換の箇所（20 箇所）（Table 15）のアレル頻度のみを指標にして，次世代シーケンサーの精度について評価することとした．

C-4-1-3 ライブラリーの品質評価

品質確認を行った標準ゲノム DNA について，SureSelect XT Human All Exon v5 を用いてライブラリー作製を行い，それらライブラリーの品質を Agilent 2100 Bioanalyzer を用いて測定した結果，すべての検体においてクローニングサイズが約 200 bp 長からなるライブラリーを作製することができた（Fig. 24）．さらに，これらライブラリーを用いてシーケンスを行った結果，一検体あたりのペアエンドリード数が約 5 億個，総塩基数に換算すると約 50Gb 相当の配列がシーケンスされていることが確認できた（Table 14）．また，塩基配列をシーケンスするときに発生するエラー率を，以下の数式を用いて Phred クオリティスコアの値を算出した．

$$Q = -10\log_{10}p$$

その結果，Q30（シーケンスエラーが生じる確率が 0.1%）のクオリティスコアが，いずれの検体についても 90%以上であることが確認できた（Table 14）．

C-4-2 シーケンサーの精度評価

C-4-2-1 シーケンスライブラリーごとのリード数に応じたエラー率評価

4 つの独立した標準ゲノム DNA の各ライブラリーについて，40Gbase，30Gbase，15Gbase 相当のシーケンスを行い，参照配列の既知変異箇所におけるシーケンスされた塩基種類の頻度を測定し，各変異箇所における 4 回の解析結果についてのバラツキの程度を CV 値（変動係数）によって評価した．その結果，各シーケンスデータ量（40Gbase，30Gbase，15Gbase）ごとに CV 値の平均を算出したところ，シーケンスデータ量が多いほど，バラツキの程度が低いことが確認された（Fig. 25）．つまり，リード数を多く読むことで，読み間違いを減らすことが可能であると考えられた．さらに，変異箇所ごとについて，同様に塩基種類の頻度を測定し，各変異箇所における 4 回の解析結果についての CV 値を求めた．その結果，すべてのシーケンス量（40Gbase，30Gbase，15Gbase）で CV 値が 20%以下であった塩基箇所は，測定された全塩基箇所 20 個のうち 10 箇所であった．一方，CV 値が 20%よりも極端に外れている測定箇所においては，4 回のシーケンス解析のうち 1~3 回の外れ値が測定されているためでると考えられた．さらに，実際の変異頻度が低い場合（例えば，塩基箇所#12 における変異頻度は 1%）においても，バラツキが大きくなることが確認された（Fig. 26）．

C-4-2-2 すべてのライブラリーを統合して得られたシーケンスデータのエラー率評価

前述の実験で取得された 4 回のシーケンスデータをすべて統合し，160Gbase，120Gbase，80Gbase，40Gbase 相当のシーケンス量になるようにダウンサンプリングした．これらサンプリングサイズごとにマッピングを行い，前述と同様に，参照配列の既知変異箇所におけるシーケンスされた塩基種類の頻度を測定し，各変異箇所におけるそれぞれのリード数（160Gbase，120Gbase，80Gbase，40Gbase）ごとのバラツキを CV 値によって評価した．Fig. 27 に示すように，4 回のライブラリーを統合した場合の CV 値は，ほとんどの測定箇所において 20%以下であった．このことは，複数回のシーケンスを行うことで，シーケンスデータ量に関係なく測定のバラツキを低減させることができることを示唆している．一方で，塩基箇所#3 においては，40Gbase における変異頻度の測定値が他と極端に違うために，Fig. 27 に示しているような CV 値が 20%から大きくは外れた結果と

なってしまった．次に，実測値と公表値とのバラツキの程度についても解析を行った．Fig. 28 に示しているように，測定箇所#7, 10, 13, 16, 18, 19 において，実測値と公表値のバラツキが著しく高い結果となった．しかしながら，Fig. 27 の結果からも解るように，これら測定箇所におけるリード間のバラツキが低かったことから，これらの箇所における実測値と公表値のバラツキについては，シーケンサーでは正確に読み取りにくいゲノム配列（構造）であることが推測される．または，変異頻度の公表値を見直す必要も考えられるため，今後は，他のゲノム標準品を用いた解析も実施する必要があると思われる．

C-5 遺伝的安定性評価リファレンスとしての日本人ゲノムの *de novo* 配列決定

2014 年 12 月末日の段階で，6 名の健常日本人男性(JOM001, JOM002, JOM003, JOM004, JOM005, JOM006)から提供の申し出があり，これまで2回あるいは3回の射精分の検体提供を受けた．本人が把握できる範囲内で先祖の国籍は全て日本であり，出身地は本州または九州本土であることが確認された．健康状態を正確に知ることはできないが，6 名全てに実子がいることが確認された．それぞれの精液 250 µL からゲノム DNA を抽出したところ，それぞれ 3.2 µg, 1.1 µg, 2.5 µg, 7.7 µg, 3.1 µg, 4.0 µg の収量であった(Fig. 29)．吸光定量を行い，ライブラリ作製のための品質に問題ないことを確認した．1 回の射精分の精液量はおおよそ 1 mL から 2 mL であった．

これら6サンプルから無作為に5サンプルを選び，市販のアレイを用いた 2,294,794 サイトに対する SNP タイピングを行った (Table 16)．その一方で 1000 人ゲノムプロジェクト (http://www.1000genomes.org/) のデータを利用し，性染色体を除き，refSNP 番号が付けられている SNP から，アセンブリ対象とするサンプルを決定するための参考になる SNP を選び出した．現在利用可能なデータは 26 人種，合計 2504 人となっており，104 人の日本人が含まれていた．全データと比較して日本人に稀な SNP，ヨーロッパ人と比較して日本人に稀な SNP，東アジア人として比較して日本人に稀な SNP，また東アジア人には稀だが日本人に多い SNP を，それぞれ 22,172, 50,609, 5879, 1022 選び，今回調べた 5 サンプルがこれらをどのように持つかをまとめた(Table 17)．また，第 18 番染色体から無作為に 4000 の SNP を選び主成分分析を行った．比較サンプルとして，1000 人ゲノムデータからヨーロッパ人(CEU)4 名，アフリカ人(YRI)4 名，中国人(CHB)4 名，日本人(JPT)4 名を選び図に表した(Fig. 30)．以上の結果から，特に，他の東アジア人と区別されやすい JOM005 を選び，ペアエンド・ライブラリおよびメイトペア・ライブラリを作成し，*de novo* アセンブリを行うサンプルとした．このサンプルからはさらに 1750 µL の精液を用い，73.0 µg のゲノム DNA を得た．

ペアエンド・ライブラリはアダプタの配列を除いたインサート長がそれぞれ 260 bp, 360 bp, 660 bp の 3 種類を作成した．Illumina HiSeq を用いてシークエンシングを行い，それぞれ 2691 億塩基，1379 億塩基，1616 億塩基から成るリードデータを FASTQ 形式で取得した．メイトペア・ライブラリはインサート長がそれぞれ 2 kb, 5 kb, 9 kb の 3 種類を作成し，やはり Illumina HiSeq を用いてシークエンシングを行い，それぞれ 1022 億塩基，452 億塩基，461 億塩基から成るリードデータを FASTQ 形式で取得した(Table 18)．

260 bp のペアエンド・ライブラリのデータについて，トリミング処理を行い，ミトコンドリア DNA (mtDNA)の全ゲノム配列をアセンブリにより決定した．決定には，ヒトとマウスで保

存されている 5'-CCG TGC AAA GGT AGC ATA ATC ACT TGT TCC T-3'から両鎖に伸張させる，本研究グループが独自開発した GrepWalk を用いた(Hayashi et al. 2014). その結果，16,570 bp から成る JOM005 mtDNA の完全長配列を得ることができた. リファレンス配列 NC_012920 との比較では，39 の SNP が見つかり，その全てがプリン間またはピリミジン間の transition で，transversion は一つもなかった. indel は 3 ヶ所に見られ，そのうちの 2 つは C/G ホモポリマーの長さの違いであった.

ペアエンド・ライブラリのデータのトリミングは，Cutadapt を用いてアダプタの除去，GrepWalk を用いて，両端の低品質塩基の除去，独自スクリプトでトリミングにより短くなったリードの除去を行った. メイトペア・ライブラリのデータのトリミングは，Cutadapt を用いて 3'末端にアダプタ配列が現れるリードは除去後，Cutadapt によりジャンクション・アダプタの処理を行い，両端の低品質塩基の除去，独自スクリプトでトリミングにより短くなったリードの除去を行った. シークエンサから得られた塩基数に対し，アセンブリに用いることができる塩基数の割合は 2 kb，5 kb，9 kb のライブラリそれぞれについて 58.2%，59.5%，59.8%であった(Table 19). 今回新規に決定した mtDNA の完全配列をリファレンスとし，トリミング処理後のライブラリデータをマッピングすることでより正確なインサート長およびその偏り(標準偏差)を見積もり，*de novo* アセンブリにはこれらの値を用いた.

全てのライブラリデータを用いた *de novo* アセンブリには数ヶ月かかると予測されたため，現段階では一部のデータに限定し，ALLPATHS-LG を実行した結果，一倍体のゲノムサイズは 2,742 Mb，反復配列の存在比は 27.0%であると推定された(Table 20). 合計 2,255 Mb の contig が得られたので，ユニークな配列を持つ領域のほぼ全てをカバーすることができたと考えられる(Table 21). gap を含む scaffold の N50 は 45 kb に達した. 最長 scaffold は第 8 番染色体の一部と考えられる 1,300,683 bp であった.

第 3 世代シークエンサのためのライブラリは，同サンプル JOM005 から得られたゲノム DNA を約 20 kb を目安に断片化を行い，二本鎖両端にペアピンループ構造を持つ一本鎖アダプタを結合させることにより行った. パルスフィールド電気泳動によってサイズ分布を確認の上サイズ選別，2 分して S1 および S2 の 2 ライブラリを作製した. それぞれのピークサイズは約 35 kb および 24 kb であった. ともに P6-C4 試薬を用い，PacBio RS II でシークエンシングを行った. S1 および S2 について，それぞれ 10 セルおよび 46 セル分のデータを取り，合計 44,840 Mb におよぶリードデータを取得した(Table 22). それぞれの平均リード長は 12,066 bp および 9716 bp であり，各リード長の分布をライブラリごとにヒストグラムに示した(Fig. 31). 先に完全長を決めた mtDNA の塩基配列と比較することにより，PacBio RS II から得られるロングリードは確かに mtDNA の全長を超える長さであることが確認された. これまでの報告では 10%程度のランダムなエラーが入るとされており，このことを確かめるため，mtDNA に由来すると考えられるいくつかのリードを選んで BLASTN でアラインメントしてみたところ，mtDNA 全体で 92%の identity が得られており，最新試薬 P6-C4 により，良好な配列データが得られていることが確認できた(Fig. 32). 各ロングリードの由来すると思われる染色体は，PacificBio Science 社が提供する BLASR を用いて調べた.

HiSeq によって得られたショートリードのアセンブル結果は，これらロングリードを用いて scaffold あるいは contig 間の gap を埋める作

業を進めている. 最終的な結果は論文発表時に一般公開予定で, 現在は次のサイトより限定公開を行っている(Fig. 33).

http://epigenetics.nrichd.ncchd.go.jp/refjpn/

本研究グループは, 正常出産を経験した411 人の日本人女性のゲノムを高解像度 SNP アレイで調べ, 多くの日本人が共有するコピー数多型を報告している(Migita et al. 2014). その中でも第2染色体短腕側 52.77 Mb 辺り(2p16.2)にあると考えられる約 26 kb の欠失は 132 人の日本人女性で認められている. この欠失の有無を日本人男性 JOM005 で調べたところ, ヘテロで保持しており, かつ HiSeq によって得られたリードのアセンブルにより, 欠失部位は GRCh37/hg19 の chr2:52,749,687-52,785,272 で欠失長は 35,586 bp であることが判明した(Fig. 34).

C-6 分化プロペンシティを指標とした細胞特性解析法の開発

C-6-1 iPS 細胞株の三胚葉への分化プロペンシティ

多能性幹細胞を三胚葉へ自発的に分化させるために, 浮遊条件下で細胞を凝集させ胚葉体（embryoid body）を形成させる方法が広く行われている. また胚葉体形成は初期の *in vivo* 胚発生を模倣するモデルとしても用いられる. 内胚葉は消化管, 肺, 膵臓, 肝臓などを, 中胚葉は筋肉, 心臓, 血液, 骨などを, 外胚葉は脳, 皮膚などを形成することが知られている.

ヒト iPS 細胞株 10 株を, 無血清培地と超低接着プレートで 16 日間培養し, 胚葉体を形成させた. 胚葉体で発現した内胚葉マーカー遺伝子 27 種類, 中胚葉マーカー遺伝子 56 種類, 外胚葉マーカー遺伝子 45 種類の mRNA 量を定量 PCR で測定した. しかしながら測定した三胚葉マーカー遺伝子は多岐にわたるため, 各遺伝子の発現を分化プロペンシティと関連付けて解釈する際に, 個々のマーカー遺伝子にどの程度の重み付けをした上で解釈してよいかは明らかでない. そこで得られた多くのデータをより分かりやすくするために, 主成分分析を用いてデータ解析を行い, これによって算出された第一主成分得点から細胞株による分化プロペンシティの違いについて比較検討した.

外胚葉, 中胚葉および内胚葉マーカー遺伝子について, それぞれの胚葉ごとに主成分分析を行った. この結果から, 算出された主成分が資料の情報をどれくらい説明しているのかの目安を与える, すなわち主成分の妥当性を表す数値である寄与率を調べた. すると, 外胚葉での第一主成分は 45 個ある資料のうち約 22 個の情報, すなわち資料の本質の約 48.6%を説明していることが認められた. また, 中胚葉での第一主成分は 56 個ある資料のうち約 21 個の情報, すなわち資料の本質の約 38.2%を説明しており, 内胚葉での第一主成分は 27 個ある資料のうち約 11 個, すなわち資料の本質の約 41.3%を説明していることが認められた（Fig. 35）. また外胚葉, 中胚葉および内胚葉の主成分分析から算出された第一主成分係数を Table 24 に示した. 第一主成分係数は, 主成分の分散が最大になるようにかかる重みである.

さらに iPS 細胞株 10 種類における外胚葉, 中胚葉および内胚葉マーカー遺伝子の第一主成分得点ランキングを作成した. iPS 細胞株 10 種類の標準化した遺伝子発現量の値と, 外胚葉, 中胚葉および内胚葉マーカーのそれぞれの主成分分析から得た第一主成分係数の値を掛け合わせた値の総和である第一主成分得点を得た. この第一主成分得点について, 各細胞株における外胚葉, 中胚葉および内胚葉それぞれのマーカー遺伝子の合計点を算出し, ランキングを作成した（Table 25）. その結果, 外胚葉は mc-iPS が最も第一主成分得点が高く, R-2A が最も第一主成分得点が低かった. また中胚葉お

よび内胚葉はR-2Aが最も第一主成分得点が高く，mc-iPSが最も第一主成分得点が低かった．すなわちmc-iPSは外胚葉に分化しやすく，中胚葉および内胚葉には分化しにくい傾向にあることを示し，同様にR-2Aは中胚葉および内胚葉に分化しやすく，外胚葉には分化しにくい傾向にあることを示唆している．

C-6-2 iPS細胞株の分化プロペンシティ予測マーカー候補の選別

未分化状態における発現量が各胚葉への分化プロペンシティと相関する遺伝子を探索するために，スピアマンの順位相関分析を行った．ここで相関係数として順位相関を選択した理由は，未分化状態における特定の遺伝子発現が外胚葉，中胚葉および内胚葉への分化に影響する場合，未分化状態の遺伝子発現量と分化効率は必ずしも線形の相関を示すとは限らないためである．また，いくつかの飛離れた値が存在する場合には，たとえ相関が線形であったとしても，ピアソンの積率相関係数などの間隔尺度や比尺度間の相関係数の場合はそれらの値に引きずられて不当に高い相関係数が得られてしまうからである．

まず，マイクロアレイ解析によるmRNAおよびmiRNAの発現量シグナルと，各細胞株における外胚葉，中胚葉，内胚葉それぞれのマーカー遺伝子の第一主成分得点ランキングを用いて，スピアマンの順位相関係数を算出した．順位相関係数が正の場合は1に近いほど相関が高いことを表しており，分化しやすい細胞株に高発現していることを示している．また順位相関係数が負の場合は，-1に近いほど相関が高いことを表しており，分化しにくい細胞株に高発現していることを示している．算出した順位相関係数について，有意水準5%の条件（相関係数が0.648より大きい，もしくは-0.648より小さい）で有意差を検討した（Table 26-29）．外胚葉分化プロペンシティと正に相関するprobe setの数はmRNAで136，miRNAで3，負に相関するprobe setの数はmRNAで92，miRNAで3であった．同様に中胚葉では正に相関するprobe setの数は，mRNAで35，miRNAで12，負に相関するprobe setの数はmRNAで7，miRNAで1であった．内胚葉では正に相関するprobe setの数はmRNAで9，miRNAで23，負に相関するprobe setの数はmRNAで29，miRNAで0であった．これらのmRNAおよびmiRNAを分化プロペンシティ予測マーカー候補とした．

C-6-3 IPAを用いたiPS細胞分化プロペンシティ予測マーカーの絞り込み

上記の分化プロペンシティ予測マーカー候補の中から，細胞分化に機能的に関与しているmRNAおよびmiRNAを絞り込む目的で，外胚葉，中胚葉，および内胚葉で相関がみられたmRNAおよびmiRNAのprobe setについて，IPA (Ingenuity) を用いて遺伝子ネットワーク・パスウェイ解析を行った．IPAでは複数のデータベースを基に，miRNAのターゲットとなることが配列上予想される（またはターゲットであることが検証された）mRNAを，複数の予測データベース (Target Scan Human) をもとにして調べることができる (miRNA - mRNAペアリング解析)．この解析によって選別された，ペアをつくるmiRNA - mRNAは，機能的に細胞分化に影響を及ぼしている可能性が高いと考えた．そこで，各胚葉において相関のあったmRNAとmiRNAの間で，miRNA - mRNAペアリング解析を行った (Tables 31-33)．その結果，外胚葉マーカーの第一主成分得点と相関のあったmiRNA - mRNAペアとして3種類のmiRNA (miR-373，miR-371-5p，miR-371-3p)，中胚葉マーカーの第一主成分得点と相関のあったmiRNA-mRNAペアとして1種類の

miRNA (miR-524-5p)，内胚葉マーカーの第一主成分得点と相関のあったmiRNA - mRNAペアとして6種類のmiRNA (miR-4739, miR-4505, miR-4521, miR-520g, miR-3714, miR-367) が選別された.

D．考察

D-1 新規免疫不全動物を用いた造腫瘍性試験法の開発

D-1-1 NOG-hrマウスに関する考察

HeLa細胞を用いたTPD50の検討において観察された異種細胞生着能における NOG-hrマウスと NOGマウスとの差(NOG-hr＜NOG)は，異種造血幹細胞の分化動態（hCD45陽性細胞率）においても同様の結果（NOG-hr ＜NOG）となった．ただし，分化細胞の出現パターンに差はなく，NK活性や補体溶血活性のいずれも両系統ともに検出されなかったことから，この差が極めて小さな要因により生じていることが推察された.

生着に関わる因子としてマクロファージを介した機能が存在するが，これは遺伝的背景に影響を受けるとされる．したがって，NOG-hrマウス開発の際の理論上の適切な戻し交配数（スピードコンジェニック 5回）や次交配動物選抜時のマーカー検査では捉えきれなかった遺伝的差異が 一連の結果に反映されていると考えられた．外貌観察における周期的発毛や過長爪はオリジナル系統（BALB/ c -hr）の特性であり，それを裏付けるものと考えられた.

眼脂や眼球白濁化の発症原因は不明であったが，造腫瘍性試験実施に直接的に影響を与えるものではないと考えられた.

D-1-2 BRG, BRG-nu および BRG-hrマウスに関する考察

BRGマウスにおけるHeLa細胞単体移植時のTPD50は NOGマウスと同等であり，BRG-nuマウスにおいては NOGマウスを凌ぐ成績であった．BRG-hrマウスにおけるHeLa細胞単体移植時のTPD50は NOG-hrマウスに近似であり，異種細胞生着感度の点からヌード化（nu遺伝子導入）した系統が優位であること考えられた.

なお，マトリゲル増強効果は NOG および NOG-hrマウスよりも低く，背景遺伝子の違い（NOD/Shi と BALB/c) による影響が示唆された.

D-2 幹細胞の *in vitro* 培養工程における遺伝子発現の動態解析による品質評価技術の開発

D-2-1 細胞・組織加工製品に用いられる間葉系幹細胞（hMSC）の品質評価—がん化の指標探索のための遺伝子発現解析

本研究では, hMSC と Ewing 肉腫を用いた遺伝子発現の網羅的解析により Ewing 肉腫で発現が高い事が明らかとなった Cyclin D2 及び IGF2BP1 が hMSC のがん化のマーカーとしての妥当性について検討した.

Cyclin D2 は細胞増殖などを制御する Cell Cycle に関わる遺伝子の1つであるが，がん細胞や悪性度の高い腫瘍などでは, その発現が上昇している事が報告されている．一方で，Cyclin D2 の発現により細胞増殖が止まるという報告や Cyclin D2 遺伝子のサイレンシングによりがん化が進行するという報告もあり, 細胞によって Cyclin D2 の働きも変わると考えられている. また, IGF2BP1 は mRNA の核外輸送, 局在性, 安定性, 翻訳などに影響を与える RNA 結合因子で細胞増殖などにも関わるが, 肺がん患者の腫瘍の悪性度の上昇とともに発現が上がると報告されている.

IGF2BP1 は組換えレンチウイルスベクターで強制発現させたにも関わらず，RT-PCR による mRNA の定量の結果，2 倍程度しか発現が上昇していなかったことから，次に hMSC への Cyclin D2 の過剰発現による遺伝子発現の変

化について網羅的に解析した．その結果，Cyclin D2 の強制発現によって hMSC の「細胞増殖」や「細胞周期」に関わる機能が有意に亢進されることがわかった．このことから，Cyclin D2 は細胞増殖やがん化等に関わる遺伝子群の発現に影響を及ぼす事によって hMSC の増殖亢進に寄与する事が示唆された．

D-2-2　hMSC におけるレトロトランスポジションの解析とその影響について

LINE-1s は動く遺伝子，レトロトランスポゾンの一種で，ヒトゲノムの約 17%を占めている．現在，ヒトゲノム中の多くの LINE-1s は遺伝子の一部欠失や変異を受けており，転移活性を失っているが，80-100 コピーは転移活性が残っていると考えられている．LINE-1s は胚細胞において活性化しており，発生初期に重要な役割を担っていると考えられている．体細胞の多くは LINE-1s がメチル化により不活化していると考えられているが，ヒトの脳が発達する過程において神経前駆細胞で LINE-1s が活性化し，ゲノムの他の領域に転移する事が明らかとなっている．LINE-1s の転移は個性や進化などの多様性の獲得に関わっていると想定されるが，一方でランダムな LINE-1s の転移は遺伝子の機能を壊す可能性があり，実際に血友病 A，B や筋ジストロフィーなどいくつかの遺伝性疾患を引き起こしている．

これまで，ヒト ES 細胞や iPS 細胞では LINE-1s が発現しており，転移を起こしやすい環境である事が知られていた．LINE-1s の転移は genome integrity を脅かし，転移部位によっては遺伝子の機能を壊す可能性がある．一方，正常組織においては発現がほとんど認められていないことから，細胞の未分化性に関与する可能性も考えられる．本研究により hMSCs でも LINE-1s が発現しており，その発現量が iPS 細胞や HeLa 以上であることが明らかとなった．

LINE-1s の転移を抑える細胞内因子として A3B が知られている．A3B はシチジン脱アミノ化酵素で C→U (DNA 複製時に T に変わる) に変え (相補鎖では G→A)，この遺伝子変異によって LINE-1s は転移活性を失うと考えられている．A3B の欠失型ホモ（Del/Del）は欧米人には稀で，日本人に多い事が示されている．A3B を発現しない日本人由来 hMSCs は LINE-1s の転移によってゲノムの安定性が損なわれる可能性が考えられたので，A3B 遺伝子型と LINE-1s の発現量について解析を行った．hMSCs において A3B mRNA 発現量と LINE-1s mRNA の発現量を比較したところ，両者に相関関係は見られなかった．このことから，ES 細胞では A3B が LINE-1s の転移を抑制していると考えられているが，hMSCs ではあまり影響を与えていない可能性が考えられた．しかし，LINE-1s mRNA の ORF2 領域の配列に及ぼす A3B の遺伝子型による影響について検討したところ，A3B を発現する野生型ホモ（Ins/Ins）と野生型/欠失型ヘテロ（Ins/Del）では同程度の変異が見られたが，Del/Del ではほとんど変異が見られなかった．したがって，Del/Del では遺伝子配列が保存された転移可能な LINE-1s が多く残存している可能性が示唆された．Ins/Ins 及び Ins/Del では C→T や G→A の変異が多かったため，A3B による変異の可能性が強く示唆された．

次に，次世代シーケンサーを用いて RNA 配列を網羅的に解析した．しかし，A3B の Ins/Ins と Del/Del で LINE-1s の発現量に大きな違いが見られなかった．今回の次世代シークエンス解析では，100 bp 程度の RNA 配列を網羅的に解読して繋ぎ合わせているため，LINE-1s の全長（約 6 kbp）をもった配列のみを解析した結果ではない．したがって，全長では違いがあるのに，全長を持たない RNA が多く存在するために，その違いが見られていない可能性は十分に

考えられる．今後，次世代シーケンサーのロングリードが可能になった段階で，Ins/Ins と Del/Del 由来の hMSC における LINE-1s 配列をそれぞれ確認し，A3B 遺伝子型の違いによってゲノムの安定性に影響を与えるのかどうか再度検討する必要があるかもしれない．しかし，現在のところ，A3B mRNA と LINE-1s mRNA の発現に相関が見られない点や，さらに技術的な検出の限界はあるものの，転移活性の残った LINE-1s の発現量において Ins/Ins と Del/Del で大きな違いが見られなかった点から総合的に考えて，日本人に多いとされる A3B 欠失により hMSC における LINE-1s の転移によるゲノムの安定性を損なう危険性は示されなかった．

hMSCs は培養により，増殖能や分化能が低下することが知られている．本研究でも hMSCs の増殖速度は培養により徐々に低下しており，LINE-1s の発現量も低下していた．増殖能と LINE-1s の発現との相関については，LINE-1s の発現が P=4 から P=6 の間で低下する一方，増殖はその間は低下おらず，また，P=6 から P=11 で増殖速度が低下していたが，LINE-1s の発現は大きく変化していなかった．以上の結果から，細胞の増殖能に依存した LINE-1s の発現量の変化は見られず，両者には明確な相関がないと示唆された．

次に，iPS 細胞や hMSC など幹細胞における LINE-1s の発現は確認できたが，分化した正常組織ではその発現がほとんど認められないことから，hMSC における LINE-1s の発現に及ぼす分化の影響について検討した．脂肪へ分化させた hMSCs は分化させていない hMSCs と比較して LINE-1s の発現が低下していた．また脂肪球面積が小さい（脂肪分化能が低い）hMSC では他の脂肪分化能が高かった 2 ロットに比べて LINE-1s の発現が低かった．以上の結果から，LINE-1s の発現と分化能には関連が見られ，LINE-1s の発現が hMSCs の分化能を示すマーカーの一つとになり得る可能性が示唆された．さらに脂肪分化だけでなく，骨，軟骨などの分化との相関も解析し，LINE-1s の発現が多分化能を示すマーカーになり得るか検討する必要があるだろう．

D-3 次世代シークエンサーを用いた細胞の遺伝的安定性評価指標の開発

DNA シークエンサーによる細胞の遺伝的安定性の評価に関して基礎的検討を行ったが，シークエンス解析の正確性に関しては，次世代型のシークエンサーは複数リード間のベースコールの一致率が高く，予想以上のパフォーマンスを示した．従来型のシークエンサーでは，波形として判定が付きにくいケースも経験していたが，次世代シークエンサーの場合には，デジタルデータを重複して読み取ることによりデータの信頼度は増している．今回得られたエラー率より，概ね 10 以上の重複度を持って同じベースコールがされれば，ほぼ正確な結果が得られたと考えてよい．

再生医療における，DNA シークエンサーの細胞の品質評価への利用に関しては，主に細胞の同一性および安定性（培養環境の影響をを含めた）という二つの視点が考えられる．同一性という観点に関しては，例えば個人識別に用いられる STR マーカーでも十分な結果が得られるが，SNP チップを用いればさらに確実度が上がり，さらに次世代シークエンサーを用いれば，ほぼ完璧であるといえる．この観点では，エクソンシークエンスでも十分であり，細胞を取り間違えは起きないと考えられるが，例えばマスターバンクに登録した汎用性の高い細胞株の場合には，さらにホールゲノムで情報を取得しておくことは有意義であると考える．

現在では 1000 ドルでホールゲノムが読める程度のコストになり，マスター細胞については解析は一度のみでよいことも考慮すると，品質