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 の電気泳動像を図 4 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図 5 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、表 4 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。

3. 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表 6 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。

4. ProteoMap ソフトウエアによる Web 上でのタンパク質プロファイル情報提供

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 を開発した。

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。今回は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。

概要を図7に示す。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。

D．考察

今年度は、染色体転座などのゲノムリアレンジメントの検出および低頻度の突然変異の検出という観点において、次世代シークエンサーの応用に関して検討を行ってきた。既知の遺伝子転座を持つ細胞株の切断点の解析より、WGSデータ中に切断点を決定できるシークエンスリードの情報が存在することが確認できたが、通常のマッピング解析においては、こうした転座点をカバーするフラグメントはリファレンス配列への適合率が低く、アラインメントされない情報として利用できない。そこで、転座に特化したアルゴリズム、例えば全体としての合致率は高くないが、部分的には異なるゲノム上の配列に完全にマッチしている配列を拾い上げる方法により、転座のあるフラグメントのみを検出できる可能性がある。現在その有効性を検証している。

また、今回の結果から、別の方法として取りうるアプローチは、CGHデータより切断点を探す方法である。今回の検討からも、ホールゲノムシークエンスは、そのゲノムカバー率（冗長度）の利用により、より詳細なコピー数変化の解析が可能となることから、コピー数が変化する点を可能な切断点として、可能性のあるリアレンジメント配列を人為的に作成し、この配列に対してマッピングすることにより実際の切断点を含むフラグメントを検出できる可能性が示された。今後これまでに得られたモデル細胞でのデータを用いて、この手法の有効性を検証していきたい。

さらに、もうひとつのアプローチとしては、かなり長いフラグメントを解析できるPacBioシークエンサーの利用が考えられる。残念ながら現時点ではPacBioシークエンサーのスループットはそれほど高くないため、全ゲノムをカバーできるデータを取得するのは費用的にも現実的でないが、既存のスループットの高い方法との組み合わせや、染色体分離を含めたターゲットエンリッチメントの手法を組み合わせることにより、検出は可能であると考えられる。

がん遺伝子の活性化のように融合遺伝子の特定が細胞の機能解析に重要な情報を与えることが期待されるが、細胞の品質評価という観点からは、定性的にリアレンジメントの有無を解析できれば十分であるという考え方もできる。こうした観点から、現時点で取りうる最も効率的なアプローチは、ホールゲノムシークエンス解析データを使ったゲノムワイドCGHによるコピー数変化を指標とした解析である。リアレンジメントは染色体の部分的増減を伴うことが多く、WGSデータの利用によりかなり小さい領域の変化までが検出できることがわかっている。また、コピー数変化を伴わない相互転座のような場合にも、マッピングデータの部分的減少が観察されうることから、カバー率を高めれば、ある程度の信頼性を持って検出が可

能であると考えられる。

ただし、これらの検討を行う際に注意が必要なのは、ゲノム中に散在する繰り返し配列の取り扱いであり、こうした領域におけるコピー数変化の情報は、マッピング解析上のアーティファクトである可能性も高い。リピート配列をマスクした解析法や、その位置情報による検証が必要であると考えられる。(ただし、相同な繰り返し配列間でのりアレンジメントも当然起こりやすいと考えられるため、その検出は今後と課題となる)

低頻度突然変異の高感度検出に関する検討では、理論的には一分子シークエンサーを用いたamplification freeの方法が最適であると考えこのアプローチをとったが、シークエンスエラー率が他のシークエンサーに比べて高かったため、十分な感度が得られなかった。インサートサイズを下げることにより、シークエンスリードの冗長度を上げ、さらにコンセンサス配列に関する基準を厳しくすることによりこのエラー率を落とす必要があり、現在この観点からPacBioシークエンサーを用いた再検討を行っている。

また別のアプローチとして、スループットの高いIllumina シークエンサーを用いたPCR free 解析に、ペアエンドリードを用いたコンセンサス配列に利用よるシークエンスエラーの除去法組み合わせた方法を試みている。

これまでの検討から、自然状態でのDNAポリメラーゼによる変異誘発率は10^{-9}から10^{-10}であると考えられており、1細胞分裂あたり3×10^{10}ゲノム上数個であると予想されている。このことから数Kb程度の遺伝子レベルでの突然変異率は10^{-6}から10^{-7}程度であると予想され、今回のtk遺伝子を用いた自然突然変異頻度とも一致する。よって、遺伝子あたりこの頻度の突然変異を検出するためには、通常の方法では少なくとも10^{7}個の細胞が必要となるが、ホールゲノムのシークエンス情報が正確に得られれば、1ホールゲノムすなわち1細胞でも$3x10^{10}$bpのカバー率により、検出が可能となる。よって、クローニングした均一な細胞集団を用いることができれば、シングルコロニーアイソレーションをして、処理または培養前後での変異の誘発率を検討可能であるが、iPS等のクローニングが難しい細胞に関しては、現在我々が用いているアプローチが必要となる。

通常クローニングを行わないNGS シークエンス解析においては、得られる結果はheteroな細胞集団の平均値(majority)を反映しており、新たな生じた変異は検出できないが、リード数を増やして、エラー率を落とすことによりクローニングなしでどこまでこの変異の検出が可能とできるかが、今後の課題となる。

一方、NGS解析データでは均一に見える細胞集団も、ホールゲノムで考えれば、分裂が起こるごとに細胞間のバリエーションが起きることとなる。通常新たな変異は、ニュートラルか劣性となるため、細胞集団全体に拡大することはないが、過去に我々の経験したhMSC細胞の変異株の例のように、増殖性を獲得して変異を持った細胞が細胞集団全体を置き換えることも起こりうる。特にiPS細胞のように、培養が難しい細胞においては、培養環境からの選択圧により増殖性の変異細胞が選択されやすい状況が想定される。今回iPS細胞をシングルコロニーに近い状態でシークエンス解析した結果から、今後細胞間のheterogeneityと変異率に関して有益な情報が得られると期待される。(残念ながら本報告書作成までに解析が間に合わなかったため、データについては後ほど公表する予定である。)

以上のNGS解析に関する検討に加え、プロテオームの観点から細胞の品質評価、標準化を可能とするために、得られた質量分析データを可

視化して Web 上にて提供できるシステムを構築した。ユーザーがローカルに ProteoMap ソフトウエアを動作できれば、自分で取得したデータを本システムを使って Web 上で公開してリファレンスデータとすることも可能であり、今後フリーソフトウエアとしての提供とオンラインデータベースの構築をめざしたい。

E. 結論

以上の検討を基に、現段階での NGS によるシークエンスデータの利用に関しての提言を以下にまとめる。

1. CGH データを利用したゲノムワイドな品質評価
2. 細胞起源同一性の確認
3. がん遺伝子を含めた全遺伝子の変異確認
4. 培養過程における増殖性変異獲得のチェック

1. に関しては WGS データの取得が必要であるが、2-4 に関してはエクソーム解析のみでも検討が可能であり、今後目的に応じて品質評価への利用が期待される。4. に関してはあまりこれまでに意識されていなかったが、細胞の heterogeneity の維持という観点からも今後さらに検討が必要な課題である。

細胞のプロテオームデータに関して、PeroteomeMap ソフトウエアを利用して Web 上にて情報提供できるシステムを構築した。今後は、ユーザーからのプロテオームデータを受け入れることにより、細胞のプロテオームデータに関するリファレンスデータベースの構築が可能となった。

F. 健康危険情報

なし

G．研究発表

1．論文発表

1）Nishikawa K, Iwaya K, Kinoshita M, Fujiwara Y, Akao M, Sonoda M, Thiruppathi S, Suzuki T, Hiroi S, Seki S, Sakamoto T. Resveratrol increases CD68+ Kupffer cells co-localized with adipose differentiation-related protein (ADFP) and ameliorates high-fat-diet-induced fatty liver in mice. Mol Nutr Food Res. 2015

2）鈴木孝昌　コンパニオン診断薬の現状と課題　「最先端バイオマーカーを用いた診断薬/診断装置開発と薬事対応」　p271-275（技術情報協会）2015

2．学会発表

1）Suresh T., Maekawa K., Saito Y., Sato Y., .Suzuki T. Individual variations in the human urinary proteome in relation to rat. The 3rd International Confrerence on Personalized Medicine（2014.6）（Prague）

2）　スレッシュ　ティルパッティ、斎藤嘉朗、本間正充、佐藤陽治、鈴木孝昌 変異原暴露モニタリング手法としてのタンパクアダクトミクス日本環境変異原学会第 43 回大会　（2014. 12）（東京）

3）Suzuki T., Suresh T. Protein adductome analysis for the human exposure monitoring to mutagens. The 4th Asian Conference on Environmental Mutagens　（ 2014.12 ）（Kolkata）

4）　鈴木孝昌　医薬品開発においてヒト内在性物質を測定する際の定量分析法に関する 留意点(案)の概要:規制の重要性と今後の課題 第 6 回 JBF シンポジウム　（2015. 2）（東京）

H．知的財産権の出願・登録状況

H-1. 特許取得　なし

H-2. 実用新案登録　なし

H-3.その他　特記事項なし

表 1　HL60-RG 細胞における c-myc 増幅領域のジャンクションの配列

	breakpoint position (hg19)		Junction sequence		inserted sequence
Junction	left segment	right segment	left end	right start	
6b–4a	130086178	128689007	CCTCAGGGTC**TC**	CTGTTCCTGA	None
4b–5a	128772037	130000919	CTTCCTCC**CA**	GAGAAGCCTG	None
5b–7a	130215269	130367023	ACACACTTG**T**	AGAGGGTGGG	None
7b–8	130698147	136580808	CATTC**CAACAC**	TCTTAACTCC(r)	None
8–1a	136580616	126224548	ATG AATTTCG(r)	GAGACGTCTC	None
1b–3b	126547448	128344474	CACCT AATTA	AAGGCAGCAG(r)	ATAACTTG
3a–2a	128068264	126710881	ATGTGCCCCT(r)	GGAGGCTCTG	AAAACATA
2b–6a	not analyzed				
赤字：重複配列		(r)：逆向き配列			

表 2　TK6 細胞 14-20 番染色体転座切断点候補領域の検索

chr 14 55,382,000–55,38,2999				10箇所							
# read_name	read_num	hit	strand	start	end	unique	read_num	hit	strand	start	end	unique
MG00HS05:361:C3W91ACXX:4:2115:3017:28497	2	chr14	+	55382209	55382310	○	1	chr20	+	49,282,334	49282434	○
MG00HS05:361:C3W91ACXX:5:2205:14160:62557	2	chr14	−	55382209	55382309	○	1	chr20	−	49282957	49283057	○
MG00HS05:361:C3W91ACXX:6:2210:21129:96634	1	chr14	−	55382211	55382311	○	2	chr20	−	49282966	49283063	○
MG00HS14:443:C3YEDACXX:7:2104:14959:59907	1	chr14	−	55382213	55382311	○	2	chr20	−	49283436	49283493	○
MG00HS05:361:C3W91ACXX:4:1202:10455:11484	1	chr14	+	55382214	55382314	○	2	chr20	+	49282510	49282610	○
MG00HS05:361:C3W91ACXX:6:2303:16054:77158	1	chr14	+	55382214	55382305	○	2	chr20	+	49282514	49282598	○
MG00HS05:361:C3W91ACXX:7:2201:17991:62643	2	chr14	+	55382214	55382314	○	1	chr20	+	49282450	49282550	○
MG00HS05:361:C3W91ACXX:5:1215:17129:3600	1	chr14	+	55382215	55382315	○	2	chr20	+	49282417	49282517	○
MG00HS14:443:C3YEDACXX:7:2111:10449:86501	2	chr14	+	55382215	55382315	○	1	chr20	+	49282508	49282608	○
MG00HS05:361:C3W91ACXX:6:2315:17139:90283	2	chr14	+	55382217	55382317	○	1	chr20	+	49282361	49282461	○

表 3　TK6 細胞の変異原処理による tk 遺伝子突然変異試験

Table3
Mutation
frequency

検体名	Positive well no.			Total well no.	Negative well no.	PE3	Mut. freq. ($\times 10^{-6}$)			%SG
	N	S	Total				N-MF	S-MF	T-MF	
control	2	5	7	384	377	0.9	**0.1**	**0.4**	**0.5**	**71.6**
MMS (6ug/ml)	25	8	33	192	159	0.2	**14.3**	**5.0**	**19.3**	**26.0**
ENU (12ug/ml)	139	16	155	192	37	0.8	**38.3**	**10.7**	**49.0**	**21.8**
γ-ray 2Gy	2	82	84	192	108	0.5	**0.6**	**30.7**	**31.3**	**98.2**

Comments:

PE: Plating efficiency

RS: Relative survival

RSG: Relative suspension growth

RTG: Relative total growth

MF: Mutation frequency; N: Normally growing colony; S: Slowly growing colony; T: Total

%SG: Ratio of S-MF to T-MF

表 4　PacBio シークエンサーにて検出された mtDNA 変異

Sample	Mapped Read	Mapped bp	Mutations
TK6_cont_1st	9258	6,685,859	31
TK6_cont_2nd	6,647	5,102,372	33
TK6_ENU	14,018	12,469,403	32
TK6_MMS	1,272	1,089,609	28
TK6_γ −ray	2,361	2,409,918	36
HL60_RG	292	2,999,400	17

表 5　TK6/BLM 細胞で確認された突然変異の例

Location	Ref seq		TK6	hom/hetero	TK6/BLM	hom/hetero
chr17	T	::	G	hom	A	hom
15420960		call	(G37/C32)	G/C hetero	A27/G 25	A/G hetero
	周辺seq		ATCAC(G/C)TGCTT		ATCAC(A/G)TGCTT	
	変異		(G/C)から(A/G)に変化	C to A mutation		

表 6　TK6/BLM 細胞にて発現変化を示したタンパク

Protein Candidates after removing peptides with sore < 10								
Accession	Peptide co	Unique peptides	Confidence	Anova (p)	Max fold	Highest	Lowest	Description
CD20	8	7	446.44	0.0106	3.20	BLM	TK6	B-lymphocyte antigen CD20
BRWD3	2	1	32.15	0.0158	2.06	TK6	BLM	Bromodomain and WD repeat-containing protein 3
CCL3	1	1	17.37	0.0253	4.32	BLM	TK6	C-C motif chemokine 3
CHD3	9	1	226.66	0.0273	2.06	TK6	BLM	Chromodomain-helicase-DNA-binding protein 3
IGKC	8	8	433.31	0.0056	3.28	TK6	BLM	Ig kappa chain C region
KV402 and KV401	4	4	210.12	0.0305	3.18	TK6	BLM	Ig kappa chain V-IV region Len
IGHM;MUCB	15	13	630.79	0.0158	7.29	TK6	BLM	Ig mu chain C region
NUCB1	6	2	164.57	0.0204	2.42	TK6	BLM	Nucleobindin-1
SCAM1	2	2	54.25	0.0068	3.20	TK6	BLM	Secretory carrier-associated membrane protein 1
SPTB1	13	1	309.44	0.0485	2.03	TK6	BLM	Spectrin beta chain, erythrocytic
PUR2	3	2	56.74	0.0457	2.16	TK6	BLM	Trifunctional purine bi
CN166	3	2	118.78	0.0025	2.05	BLM	TK6	UPF0568 protein C14orf166

図 1　HL60-RG 細胞 c－ｍｙ ｃ増幅領域の概略

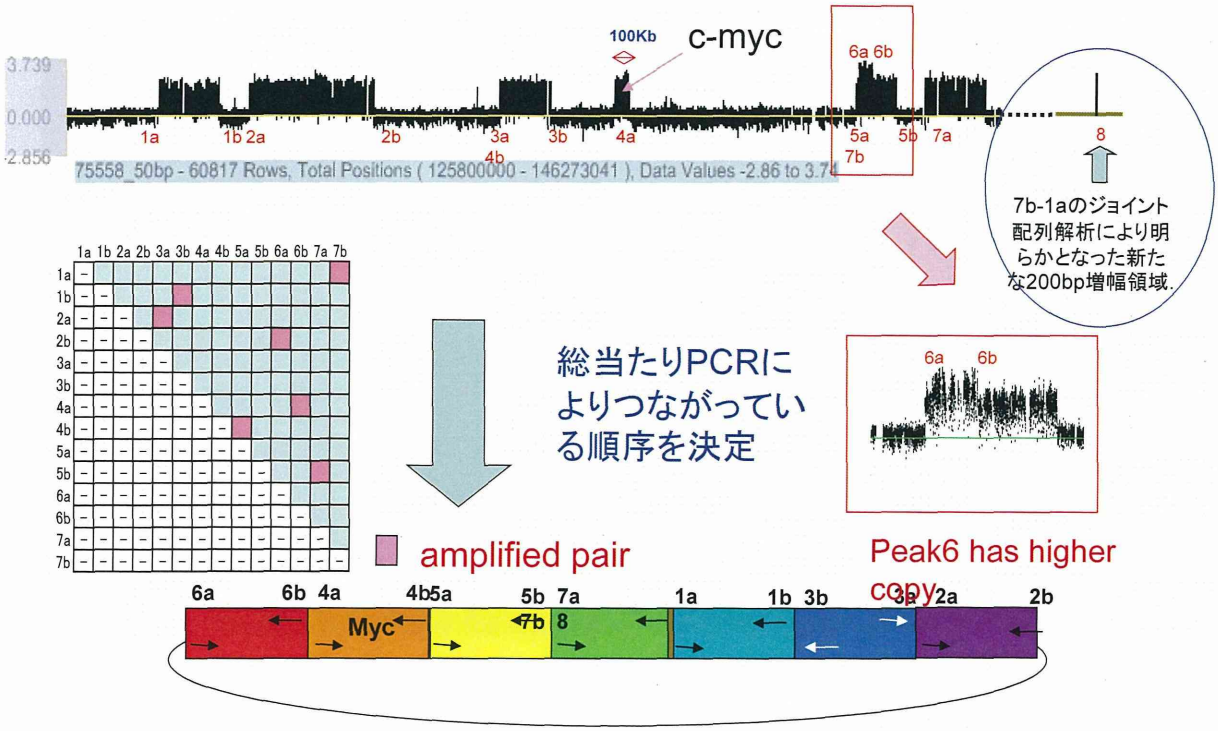


図 2　HL60-RG 細胞 c－ｍｙ ｃ増幅領域の融合点を含む配列の検出

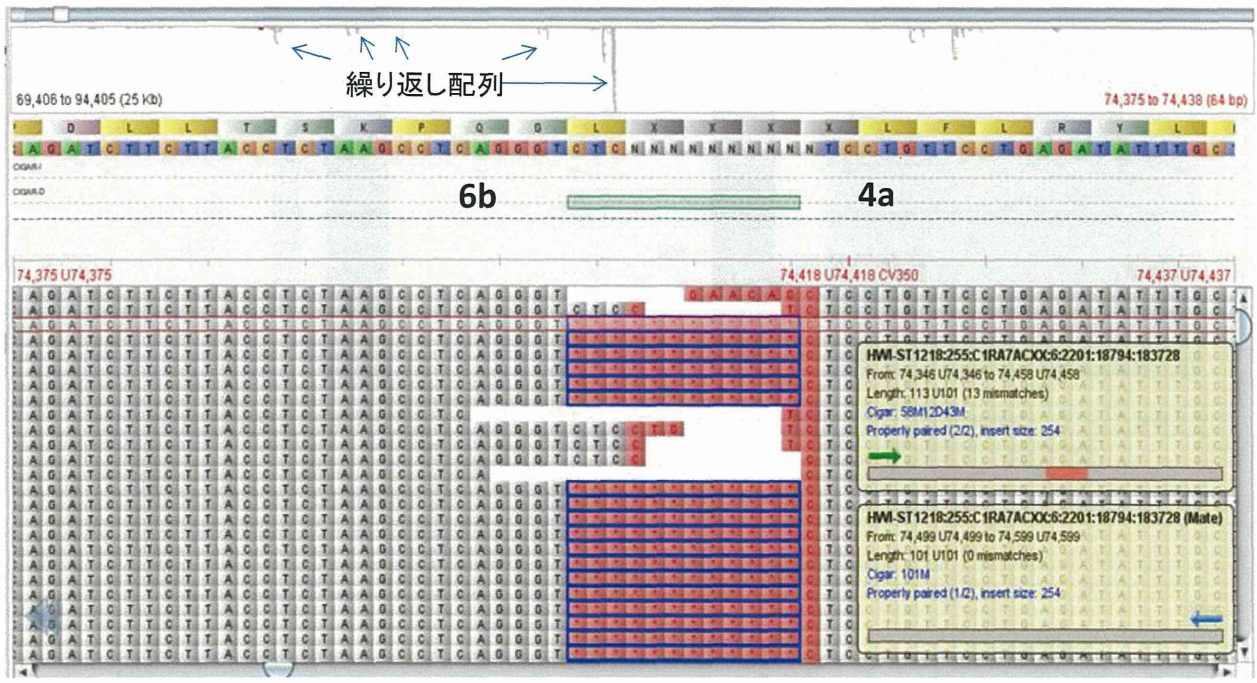


図 3　TK6 細胞の染色体解析

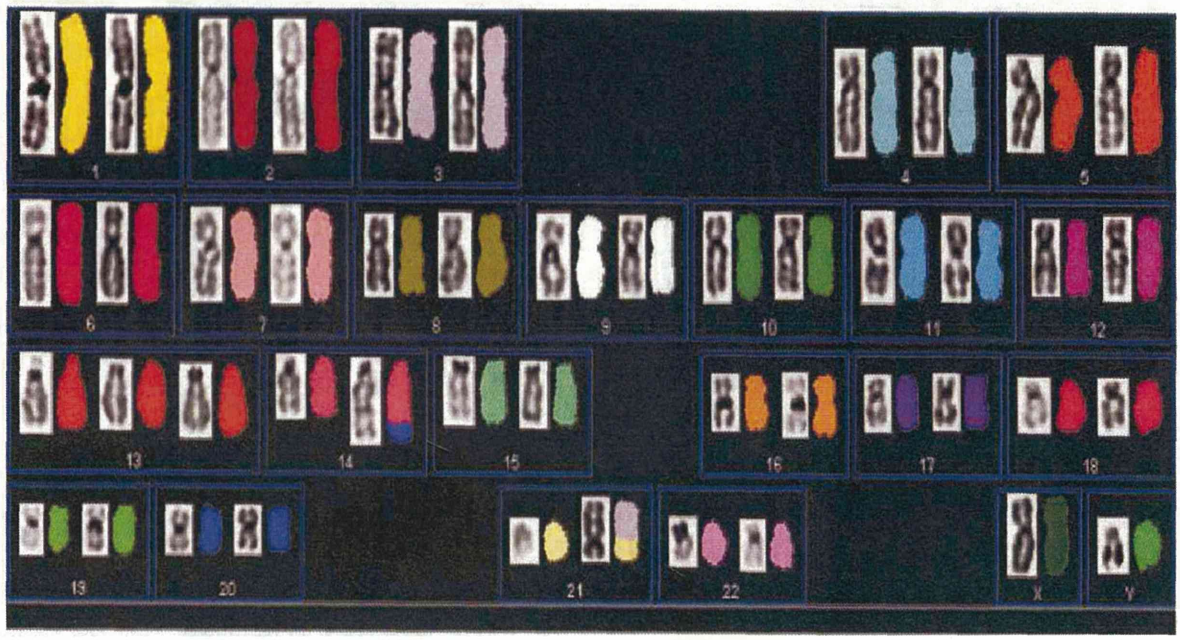
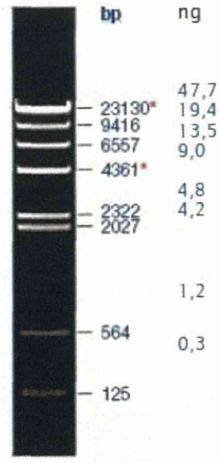
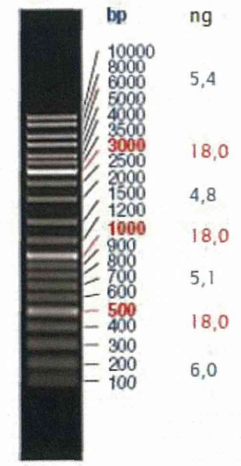


図 4　抽出した mtDNA の品質チェック





lane	sample name	µl
1	Gene Ruler DNA Ladder Mix	5
2	TK6_cont	2
3	TK6_ENU	3
4	TK6_MMS	5
5	TK6_MMS_2	3
6	TK6_RAY	2
7	HL60_RG	4
8	HL60_RG_2	2
9	Lambda Hind III	5
10		

Sample No.	Sample Name	Conc. [ng/µl]	Total volume[µl]
1	TK6_cont	11,4	100
2	TK6_ENU	7,13	100
3	TK6_MMS	2,23	100
4	TK6_MMS_2	6,06	100
5	TK6_RAY	11,2	100
6	HL60_RG	5,24	100
7	HL60_RG_2	13,7	100

図 5　PacBioRS シークエンサーによる解析パフォーマンス

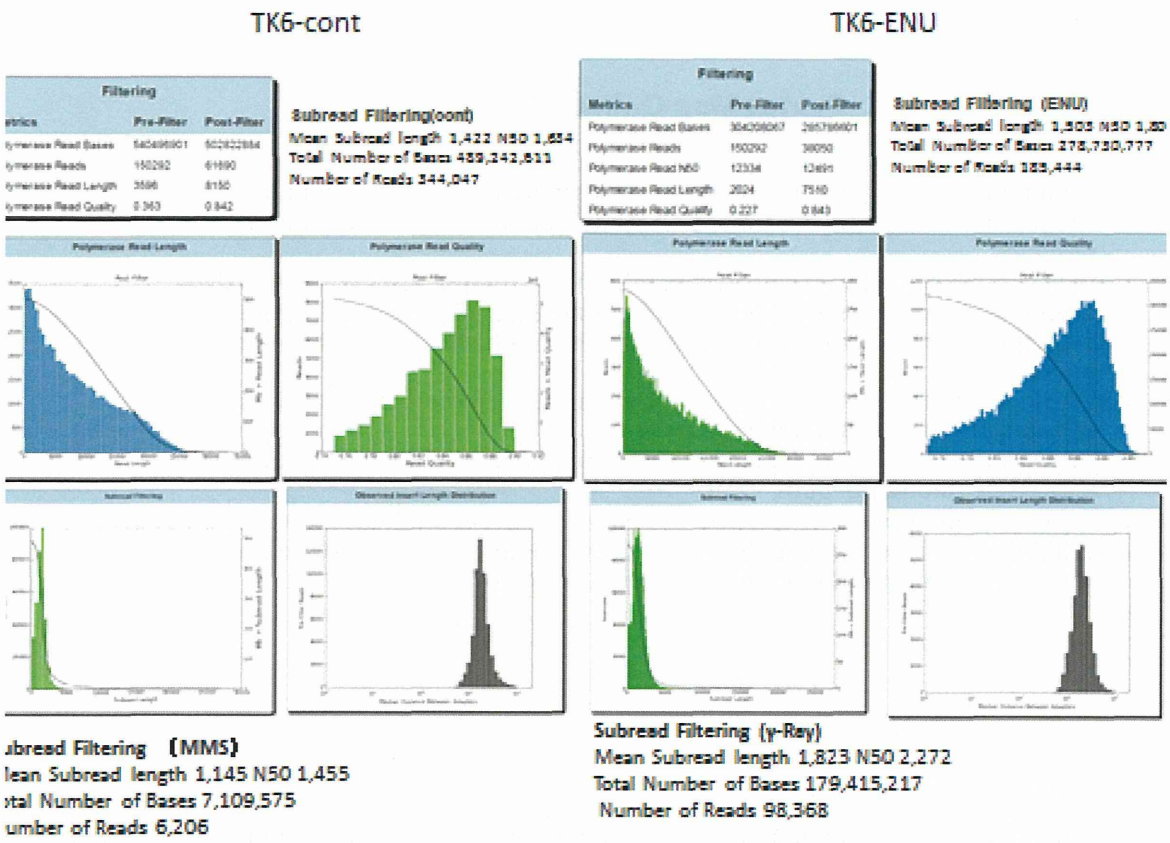


図 6　Tablet 上での PacBio シークエンサーデータの確認

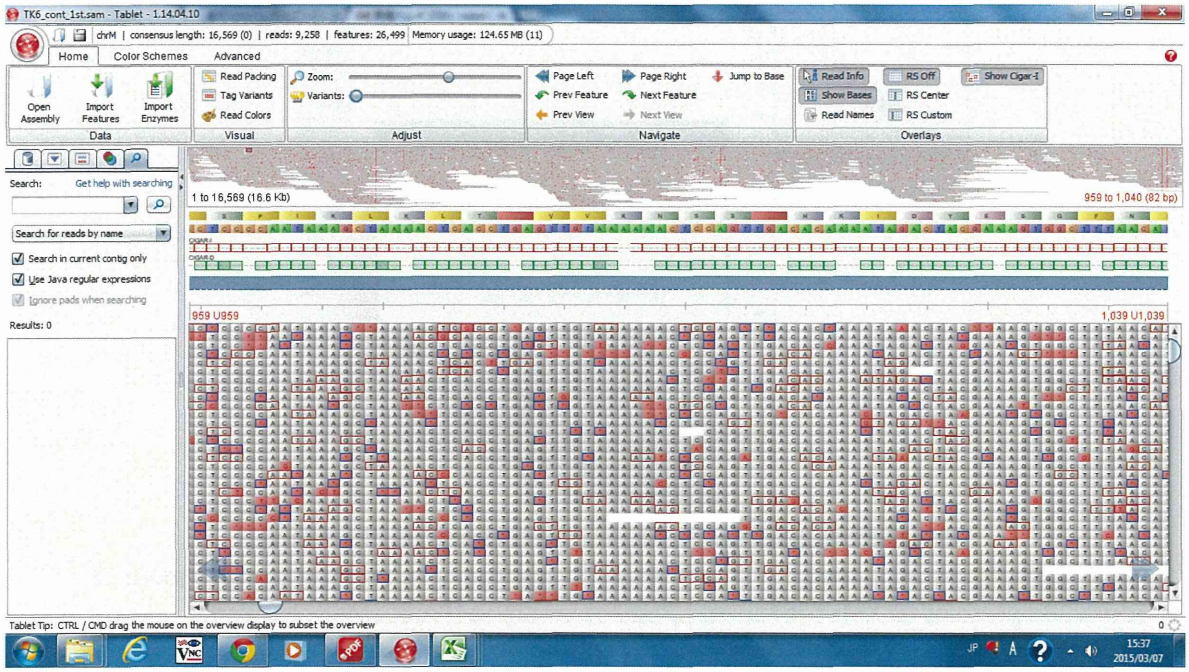


図 7　ProteoMap Online ソフトウエアによるプロテオームデータ公開

1)　ProteoMap Online ソフトウエア概略

ProteoMap Online 1.0

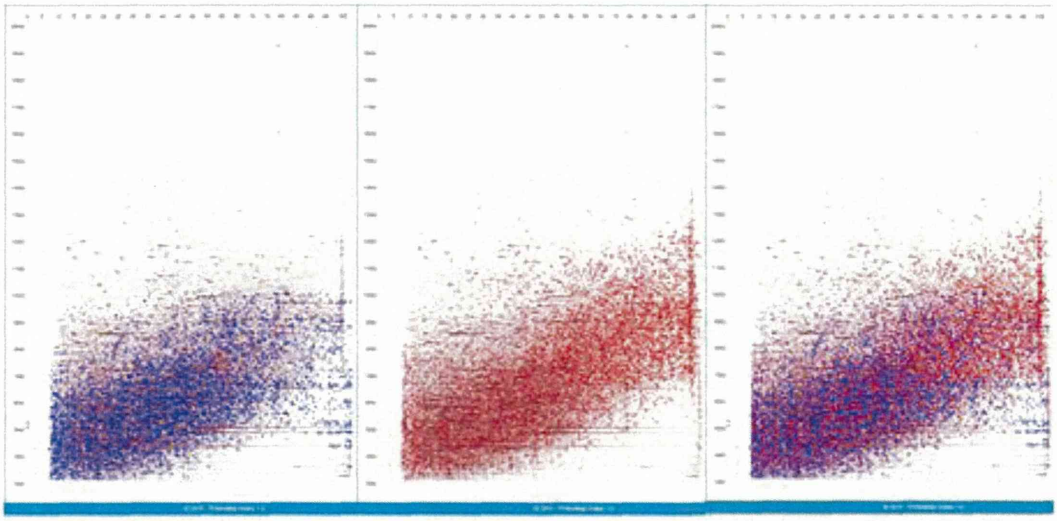
1. ProteoMap Online 1.0 is a web based tool for sharing basic LC-MS/MS data such as 2D image of the LC-MS data, MS/MS spectrum and list of identified proteins and its peptides.
2. User can upload the 2D image of LC-MS, MS/MS spectrum as .mgf file format and list of identified proteins as .xml file format (Desktop version of ProteoMap required).
3. ProteoMap Online will display the MS/MS spectrum and identified proteins on the 2D image.
4. The location of MS/MS data with and without a peptide/protein are shown in two different colours blue and red respectively.
5. User can choose to display only the identified proteins or only the MS/MS spectrum of unidentified proteins or both.
6. User can select a protein to display only the peptides of the selected protein.
7. Clicking on the blue cross the identified peptides sequence, its assigned protein and their score are displayed along with its MS/MS spectrum.

2）ProteoMap Online 操作画面

1. ProteoMap Online 1.0 is a web based tool for sharing basic LC-MS/MS data such as 2D image of the LC-MS data, MS/MS spectrum and list of identified proteins and its peptides.

2. User can upload the 2D image of LC-MS, MS/MS spectrum as a .mgf file format and list of identified proteins as .XML file format (Desk top version of ProteoMap required).
3. ProteoMap Online will display the MS/MS spectrum and identified proteins on the 2D image.

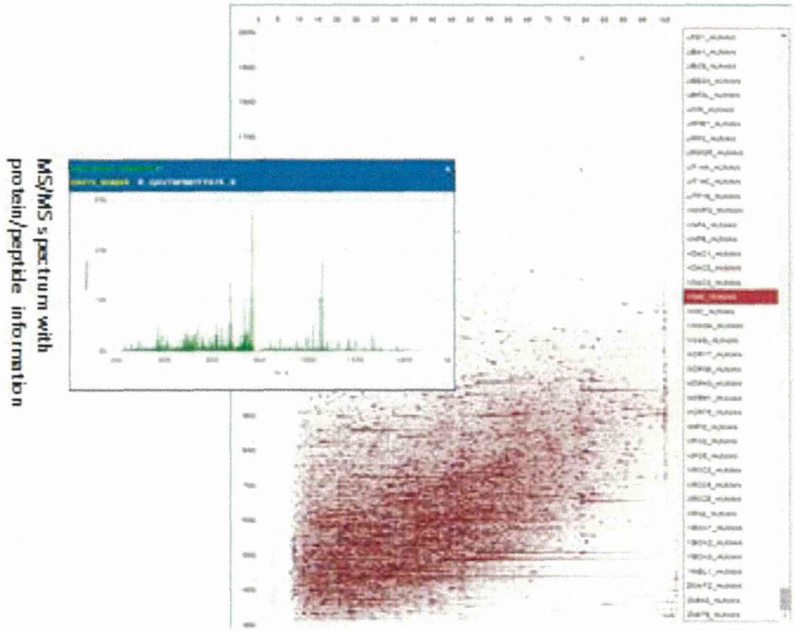
4. The location of MS/MS data with and without a peptide/protein hit are shown in two different colours, blue and red respectively.



Location of identified proteins | Location of MS/MS spectrum without protein hit | Location of MS/MS spectrum both with and without protein his

5. User can choose to display only the identified proteins or only the MS/MS spectrum of unidentified proteins or both.

6. User can select a protein to display only the peptides of the selected protein.



View of the location of peptides for the selected protein

7. Clicking on the blue cross the identified peptides sequence, its assigned protein and their score are displayed along with its MS/MS spectrum.

厚生労働科学研究費補助金（医薬品等規制調和・評価研究事業）
「細胞・組織加工製品の開発環境整備に向けたレギュラトリーサイエンス研究」
平成26年度分担研究報告書

（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）

研究分担者　三浦　巧　国立医薬品食品衛生研究所再生・細胞医療製品部第1室・室長

研究要旨【目的】再生医療は、人工的に培養した細胞や組織などを、損傷した臓器や組織に移植することにより、機能不全に陥った細胞や組織を再生し、再び機能を回復させるという新しい医療技術である。細胞ソースとして特に期待されている細胞は、分化多能性の特性をもつ胚性幹細胞（ES細胞）および人工多能性幹細胞（iPS細胞）であり、既に数例の疾患に対して、ES/iPS細胞を用いた臨床試験が国内外で実施されている。しかしながら、「生きた細胞」を製品として扱うこのような再生医療は、これまでの医薬品等とは異なる治療法でもあり、「生きた細胞」を製品化するための品質管理などには標準的な手法は未だ定まってなく、多くの課題が存在している。このような現状の中、最近、iPS細胞の品質評価系の指標として、iPS細胞の特性のひとつである「ゲノム不安定性」に着目し、次世代シーケンサーを用いたゲノム変異の包括的な解析が行われている。そこで本研究では、ゲノム不安定性の解析のために使用される「次世代シーケンサー」について、その機器としての性能および有用性を検証することにより、ゲノム不安定性の評価系を確立することを目指した。【方法】次世代シーケンサーを用いて遺伝子配列を解読する現在の方式では、0.1～10%程度の読み間違いエラーが生じることが知られている。そのため、品質管理や医療診断などに次世代シーケンサーを用いる場合には、各種解析ごとにエラー率を考慮する必要があり、そのエラー頻度を正確に予測する技術が要求される。そこで、次世代シーケンサーを用いた場合の測定誤差を正確に理解するためには、標準ゲノムDNAを用いて次世代シーケンサーの読み間違いエラー率を特定の塩基ごとに正確に計測する必要がある。そこで、本解析では実際の読み間違いエラー率を各変異箇所において検証し、それら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。【結果】ゲノム上の20箇所に人工的に1塩基変異を導入した標準ゲノムDNA（20μg）を準備し、4つの独立したライブラリーを作製した。イルミナ社シーケンサーを用いて標準ゲノムDNAにおける変異率の定量的解析を行ったところ、15～40Gbaseのシーケンスデータ量を取得した場合でも、各ライブラリー間でのバラツキが散見された。その一方で、複数回の独立した解析を行うことで、シーケンスデータのバラツキは抑えられることも確認された。【結論】シーケンスの精度を高めるためには、カバレッジを深くすることの他に、独立した複数の解析を実施することも重要であることが示唆された。

A．研究目的

再生医療は、難治性疾患や機能不全に陥った組織や器官の根本的な治療が可能となる革新的な医療システムとして大きな期待を集めており、総合科学技術会議の提言や「新成長戦略」、「日本再生戦略」などにおいても最重要研究課題の一つとされているが、再生医療は極めて新しい医療システムということもあり、その実現のためには解決すべき課題も多い。

平成25年公布の医薬品医療機器等法で定義された再生医療等製品のうち、再生医療・細胞治療に使用することを目的に生きた細胞を加工して製造される製品は細胞・組織加工製品と呼ばれ、国内外で活発に研究・開発が行われている。細胞ソースとしてはヒト体細胞に加え、近年ではヒト体性幹細胞、胚性幹細胞（ES細胞）などの幹細胞が対象とされてきている。また最近、生命倫理的な問題や免疫学的な拒絶をクリアできると考えられる人工多能性幹細胞（iPS細胞）が登場し、再生医療が社会的に大きな期待を集めている。しかしながら細胞・組織加工製品は、臨床使用経験が少ないために知見の蓄積も乏しく、国内指針やICH、WHOなどの生物製剤製造国際ガイドライン等にある従来の品質・安全性評価法が適用できないケースも頻出しており、新たに適切な評価技術を樹立することが火急の課題となっている。そこで今回我々は、iPS細胞がゲノム不安定性という性質を有することに着目し、ゲノム不安定性に基づいた品質評価系について、次世代シークエンサーを用いてその遺伝的安定性を解析する系を検証し、他のゲノム評価方法に対する簡便性、再現可能性、信頼性などについて総合的に比較検討することにより、ゲノム解析系の向上を目指した。

現在、次世代シーケンサーは、その解析能力が改善され、病気の診断、治療、創薬などの分野で幅広く使用されており、医学の分野のみならず分子生物学などの基礎研究分野においても利用されている装置である。また、再生医療の分野においても、iPS細胞由来の細胞・組織加工製品などにおいて、次世代シーケンサーによるゲノム変異解析がおこなわれ、それら細胞の特性を把握するために、次世代シーケンサーが利用されつつある。しなしながら、次世代シーケンサーから得られる情報量は膨大であり、細胞の品質評価を行う上で、データをどのように解釈すれば標準的な解析方法を確立することができるかなど、次世代シーケンサーを用いた品質評価系の開発には、課題も多く残されている。そこで我々は、次世代シーケンサーの機器としての性能について検証を行うこととした。なぜなら、次世代シーケンサーの技術開発は急速に進んでおり、数年前と比較すると得られるデータの精度は飛躍的に向上されているが、その一方で、サンガーシーケンス解析と比較して読み取りリード長の長さは短く、さらに0.1～10%程度の読み間違いエラーも依然として生じるという欠点もあるからである。そのため、細胞・組織加工製品などの品質評価に次世代シーケンサーを用いる場合には、事前に次世代シーケンサーの性能を詳細に検証する必要があり、個々の解析レベルにおいて、次世代シーケンサーによって得られたデータの質をモニターすることは極めて重要である。即ち、次世代シーケンサー

における読み間違いエラーを考慮することにより、次世代シーケンサーの機器（プラットフォーム）間、あるいは実験者間などの違いによるデータのバラツキというものを解決することができると考えられる。本研究では、標準ゲノムDNAを用いて次世代シーケンサーにおける読み間違いエラー率を検証し、次世代シーケンサーを用いた品質評価法の有用性について検討した。

B．研究方法

B-1　標準ゲノムDNAの調製

特定のゲノム領域上に変異をもつ細胞から得られたゲノムDNA、あるいは人工的に変異を導入したゲノムDNAと、それら変異箇所に変異がないゲノムDNAとを混合し、それらアレル比率が正確に測定されたゲノム溶液を標準ゲノムDNAとした（表1）（Horizon Diagnostics社）。

B-2　次世代シーケンサーによる解析

ゲノムDNAの品質を電気泳動および濃度測定により確認後、ゲノムDNAを4つに分け、それぞれ独立して数百bpに物理的に断片化を行い、二本鎖DNAの両末端にアダプターを付加したフラグメントライブラリーを作製した。Sureselect Target Enrichmentシステム（Agilent Technologies社）を用いてターゲット領域を濃縮し、タグ配列を有するプライマーを用いてPCR増幅を行い、シーケンスの鋳型となるDNAライブラリーを作製した。イルミナ社シーケンサーを用いて、シーケンスを行い、シーケンサー付属のソフトウェアにより塩基配列（リード配列）を取得し、得られたシーケンスデータを参照配列にマッピングした。それら標準ゲノムDNAのマッピングは、各ライブラリーあたり40Gbase、30Gbase、15Gbase程度のシーケンスデータを用いて行った（3種のシーケンス量 × 4種のライブラリー）。さらにそれら標準ゲノムDNAのマッピングをすべてまとめた後、160Gbase、120Gbase、80Gbase、40Gbase程度のデータを取得して解析を行った（4種のシーケンス量 × 1種のライブラリー）。情報処理については、標準ゲノムDNA中の塩基比率が保証されている塩基位置について、各マッピング結果から標準塩基と変異塩基のリード数のカウントを行った。

B-3　リードマッピングと塩基頻度カウント

B-3-1　リードクリーニング

シーケンスにより得られた塩基配列（以下リード）から、変異検出に影響を及ぼすと思われる低品質領域を除去した。まず、シーケンスアダプター配列由来の領域がリードの 3'末端にある場合は除去し、続いて塩基品質が低い領域がリードの 3'末端にある場合はその領域を除去した。なお、除去後のリード長が短くなり過ぎた場合はリード全体を破棄した。さらに、除去後のリードに一定の割合の低品質塩基が存在する場合はリード全体を破棄した。上記の処理は、ペアリードのリード1とリード2のそれぞれに対して行った。最後に、除去ならびに破棄後のリードの内、対応するリード1とリード2が存在するペアを抽出し、それをクリーンリードとした。

B-3-2　マッピング

クリーンリードを参照配列にマッピングし、マッピング結果を変異検出に適した状態に調整した。まず、クリーンリードを参照配列にマッピングした。次に、各マッピング結果のダウンサンプリングを実施した。ダウンサンプリングは、各マッピング結果を 40Gbase、30Gbase、15Gbase 相当のデータとなるようにダウンサンプリングした。また、全てのマッピング結果をひとつにまとめ、160Gbase、120Gbase、80Gbase、40Gbase 相当のデータとなるようにダウンサンプリングした。さらに、ダウンサンプリングしたマッピング結果のうち、アライメントの疑わしい領域に対して再アライメントを実施し、再アライメントしたマッピング結果を用いて変異塩基の検出を行った（プレコール）。最後に、再アライメント結果から、参照配列と各位置にアライメントされたリードの塩基配列を比較し、シーケンス時に得られた塩基品質をより正確な値へと再調整した。なお、塩基品質の再調整は、参照配列とリードの塩基配列の一致性に基づき処理が行われた。再アライメントしたマッピング結果から検出した変異塩基リスト（プレコール）ならびに既知変異データ（dbSNP など、存在する場合のに使用）を利用し、参照配列とリードの塩基配列が一致する領域を判断した。

B-3-3　塩基頻度の算出

サンプリングした各マッピング結果から、参照配列の既知変異位置におけるシーケンスされた塩基種類の頻度を数え上げた。短い挿入・欠失配列がある場合も同様に数え上げた。その内、最も頻度の高かった塩基種類と二番目に頻度の高かった塩基種類に関してはピックアップし、一覧表内の独立したカラムにまとめた（データ省略）。また、各検体の変異塩基の頻度一覧表を横並びにすることで、検体間比較に使用できる一覧表を作成した（データ省略）。

（倫理面への配慮）

ヒト由来の生体試料に関しては、試料提供者に一切不利益および危険性が伴わない、人権擁護を含めたインフォームドコンセントのもとに採取された試料を用いた。また、研究目的を含め、研究内容の倫理的、科学的妥当性について国立医薬品食品衛生研究所・研究倫理委員会による審査・承認を得た上で研究を実施した。

C．研究結果

C-1　HiSeq システム（イルミナ社）を用いたシーケンス解析

C-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（表 2）。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（表 3、図 1）。

C-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

度が表 1 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）（表 5）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-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（図 2）。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（表 4）。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（表4）。

C-2　シーケンサーの精度評価

C-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（図 3）。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（図 4）。

C-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。図 5 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、図 5 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。図 6 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、図 5 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。

D．考察

次世代シーケンサーは、従来のサンガー法とは異なり処理能力やコストの面で格段の進化を遂げ、たった一回のランで数億塩基以上も得ることが可能である。今現在も、次世代シーケンサーの性能は、すさまじい勢いで向上を続けており、今後数年の間で、個人のゲノム解読は 1 時間以内で終了し、数万円でできるようになると言われている。このような次世代シーケンサーの登場によって、ゲノム科学は加速度的に進展を遂げている。その一方で、リード（シーケンスの最小単位）の長さが短く、またベースコールのクオリティに問題があるなどの欠点も指摘されている。そのため、次世代シーケンサーによって得られたデータの質を評価するシステムを構築する必要がある。今回我々は、上述のような実態を把握するため、ショートリード配列の解析時に起こりうる読み間違いエラーに着目し（イルミナ社 HiSeq2500）、次世代シーケンサーの性能について検証することにした。まずエラー率の評価を厳密に行うために、ゲノム DNA のアレル頻度が正確に測定されている標準ゲノム DNA を準備し、量、質ともに次世代シーケンサー解析における条件を満たしていることを確認した。このような標準ゲノム DNA におけるライブラリーを、独立して 4 種類作製し、各ライブラリーについて、40Gbase、30Gbase、15Gbase 相当のシーケンスを行ったところ、シーケンスデータ量が多いほど、測定値のバラツキは低減されることが判明した。さらに、解析の精度を評価する目的で、すべてのライブリーから得られたシーケンスデータを統合し、それらシーケンスデータのサンプリングサイズの幅を 160Gbase、120Gbase、80Gbase、40Gbase に設定し、解析精度の閾値を検討したところ、複数回のシーケンスデータ