を取得することで、マッピング量とは無関係にバラツキは低減されることが判明した。つまり、図 4 で見られた各塩基間の CV 値のバラツキの程度は、図 5 では低減していることから、ライブラリー間のバラツキが大きく影響していることが考えられた。即ち、実際の検体（診断用組織や細胞加工製品など）について得られたシーケンスデータを解析する場合には、複数回の独立した解析を行うことも重要であると考えられる。以上、次世代シーケンサーのエラー率評価のために標準ゲノム DNA を用いることは、標準ゲノム DNA をポジティブコントロールとして精度管理をすることが可能となるだけでなく、検出限界（Limit of Detection (LOD)）の検証、複数のプラットフォーム間での比較評価に役立つものと思われる。

E．結論

本研究では、次世代シーケンサーの精度を理解するため、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した。まず、我々は、参照配列に対する変異塩基の頻度が既に測定されているゲノム標品を準備し、この標準ゲノムDNAを用いてシーケンスを行った。実際にシーケンスされた塩基種類の頻度を測定したところ、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。

一般的に、エキソーム解析における読み取り総塩基数は、通常5Gbase程度であるが、今回の実験のように、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が確認された。つまり、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。また、今回の解析ように、通常の8倍程度のカバレッジでシーケンスを行ったとしても、測定する塩基の位置においては、バラツキに差が現れることも観察された。このことは、現在のシーケンサー自体の性能の限界であると考えられるため、今後は、コスト面や解析速度など、目的に応じてシーケンサーの精度を理解し、解析方法を使い分けることが重要であると思われる。

F．研究発表

F-1　論文発表

1. Fukuda A, Tomikawa J, Miura T, Hata K, Nakabayashi K, Eggan K, Akutsu H, Umezawa A. The role of maternal-specific H3K9me3 modification in establishing imprinted X-chromosome inactivation and embryogenesis in mice. *Nat Commun.* 2014, **5**: 5464.
2. 三浦 巧, 佐藤 陽治. 再生医療・細胞治療に使用する細胞加工物の品質・安全性評価の原則と造腫瘍性の考え方. 谷本学校毒性質問箱. 2014, 16: 1-10

F-2　学会発表

1. 阿久津 英憲, 菅原 亨, 三浦 巧, 梅澤 明弘. mir-302 マイクロ RNA ファミリーによるヒト多能性幹細胞の中・内胚葉初期分化制御. 第 14 回再生医療学会総会, 横浜（2015 年 3 月 19〜21 日）
2. Takumi Miura, Tohru Sugawara, Atsushi Fukuda, Ryo Tamoto, Akihiro Umezawa, Hidenori Akutsu. Generation of Committed Neural Progenitors from Human Fibroblasts by Defined Factors. The 12th Annual Meeting International Society for Stem Cell Research, Vancouver, Canada（2014 年 6 月 18-21 日）

G．知的財産権の出願・登録状況

G-1　特許取得

なし

G-2　実用新案登録

なし

G-3　その他

特記事項なし

表1．標準ゲノムDNAのアレル頻度

Gene	Chromosome Number	Reference Base	Observed Base	Amino Acid Change	Allelic Frequency	Codon Change	Exon ID	Validated by Digital PCR	Validated by NGS	Class of Mutation
BRAF	chr7	A	T	V600E	10.5	GTG/GAG	Exon 7 140453075_140453193	Yes	Yes	Missense
KIT	chr4	A	T	D816V	10.0	GAC/GTC	Exon 17 55599236_55599358	Yes	No	Missense
EGFR	chr7	AAGGAATTAAGAGAAGCA	AA	E746 – A750	2.0	N/A	Exon 19 55174772_55174870	Yes	No	Deletion
EGFR	chr7	T	G	L858R	3.0	CTG/CCG	Exon 21 55191719_55191874	Yes	No	Missense
EGFR	chr7	C	T	T790M	1.0	ACG/ATG	Exon 20 55181293_55181478	Yes	No	Missense
EGFR	chr7	G	A	G719S	24.5	GGC/AGC	Exon 18 55241614_55241736	Yes	Yes	Missense
KRAS	chr12	C	T	G13D	15.0	GGC/GAC	Exon 12 25398208_25398329	Yes	Yes	Missense
KRAS	chr12	G	A	G12D	6.0	GGT/GAT	Exon 2 2539869_2539748	Yes	No	Missense
NRAS	chr1	C	A	Q61K	12.5	CAA/AAA	Exon 3 115256599_115256777	Yes	No	Missense
PIK3CA	chr3	G	A	E545K	9.0	GAG/AAG	Exon 10 178935998-178936122	Yes	No	Missense
PIK3CA	chr3	A	G	H1047R	17.5	CAT/CGT	Exon 3 178951882_178957881	Yes	Yes	Missense
CDX2	chr13	AC	A	V306fs	41.5	N/A	N/A	Yes	Yes	Deletion
ARID1A	chr1	GC	G	P1562fs	33.5	N/A	N/A	Yes	Yes	Deletion
CCND2	chr12	AT	A	N/A	32.5	N/A	N/A	Yes	No	Deletion
BRCA2	chr13	CA	C	A1689fs	33.0	N/A	N/A	Yes	Yes	Deletion
ALK	chr2	G	A	P1543S	33.0	CCT/TCT	Exon 2 29415640_29416788	Yes	Yes	Missense
CTNNB1	chr3	C	A	S33Y	32.5	TCT/TAT	Exon 3 41266017_41266202	Yes	Yes	Missense
FBXW7	chr4	TC	T	G667fs	33.5	N/A	N/A	Yes	Yes	Deletion
PDGFRA	chr4	G	A	G426D	33.5	GGC/GAC	Exon 4 55138561_55138687	Yes	Yes	Missense
APC	chr5	C	T	R2714C	33.0	CGT/TGT	Exon 5 112173250_112181936	Yes	Yes	Missense
NOTCH1	chr9	G	A	P668S	31.5	CCG/TCG	Exon 9 139409742_139409852	Yes	Yes	Missense
FLT3	chr13	GGA	G	S985fs	10.5	N/A	N/A	Yes	Yes	Deletion
FLT3	chr13	A	G	V197A	11.5	GTG/GCG	Exon 13 28626682_28626811	Yes	Yes	Missense
IDH1	chr2	G	A	S261L	10.0	TCA/TTA	Exon 2 209106718_209106869	Yes	Yes	Missense
CTNNB1	chr3	CCTT	C	S45del	10.0	N/A	N/A	Yes	Yes	Deletion
MET	chr7	GT	G	V237fs	6.5	N/A	N/A	Yes	Yes	Deletion
SH2D2A/NTRK1*	chr1	AC	A	N/A	8.5	N/A	N/A	Yes	Yes	Deletion
ABL2	chr1	TG	T	P986fs	8.0	N/A	N/A	Yes	Yes	Deletion
CDH1	chr16	A	G	N/A	7.5	N/A	N/A	Yes	No	None
FANCA	chr16	ACT	A	E345fs	7.5	N/A	N/A	Yes	Yes	Deletion
NF1	chr17	CT	C	L626fs	7.5	N/A	N/A	Yes	Yes	Deletion
NF2	chr22	AC	A	P275fs	8.0	N/A	N/A	Yes	Yes	Deletion
EP300	chr22	CA	C	K291fs	8.0	N/A	N/A	Yes	Yes	Deletion
MLH1	chr3	C	A	L187M**	8.5	CTG/ATG	Exon 3 37061801_37061954	Yes	Yes	Missense
FGFR1	chr8	G	A	P124L***	8.5	CCC/CTC	Exon 8 38285439_38285611	Yes	Yes	Missense

* This chromosome location was annotated to both SH2D2A and NTRK1 using hg19 and GRCH37.
** L187M is the correct annotation with reference to ENST00000383761. It may be observed as L323M in other transcripts.
** P124L is the correct annotation with reference to ENST00000335922. It may be observed as P150L in other transcripts.

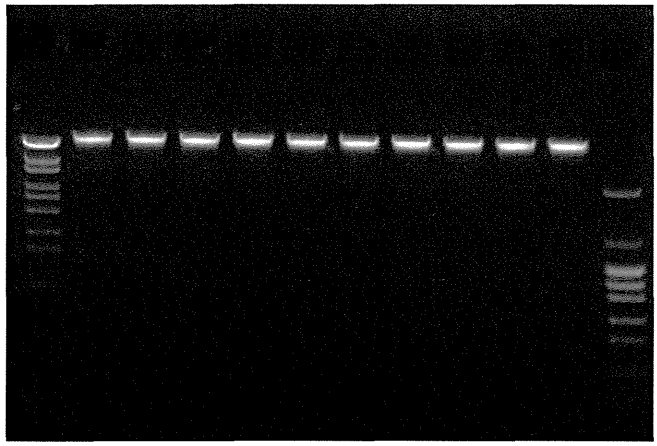
表2.　標準ゲノムDNAにおける二本鎖DNA定量結果

サンプル	液量(ul)	濃度(ng/ul)	総量(ug)	吸光定量値/蛍光定量値
Standard DNA1	20	59.8	1.2	1
Standard DNA2	20	57.9	1.2	1
Standard DNA3	20	64.8	1.3	0.9
Standard DNA4	18	61.2	1.1	0.9
Standard DNA5	18	59.3	1.1	1
Standard DNA6	18	61.2	1.1	1
Standard DNA7	18	64.2	1.2	0.9
Standard DNA8	18	63.6	1.1	1
Standard DNA9	18	58.4	1.1	1
Standard DNA10	18	61.8	1.1	0.9
Standard DNA11	18	50.6	0.9	1.1
Standard DNA12	18	59	1.1	1
Standard DNA13	18	58.7	1.1	1
Standard DNA14	18	56.9	1	0.9
Standard DNA15	18	58	1	1
Standard DNA16	18	60.6	1.1	0.9
Standard DNA17	18	59.6	1.1	0.9
Standard DNA18	18	54.9	1	1
Standard DNA19	18	59.7	1.1	0.9
Standard DNA20	18	61.2	1.1	0.9
Total	510	40.8	20.8	-

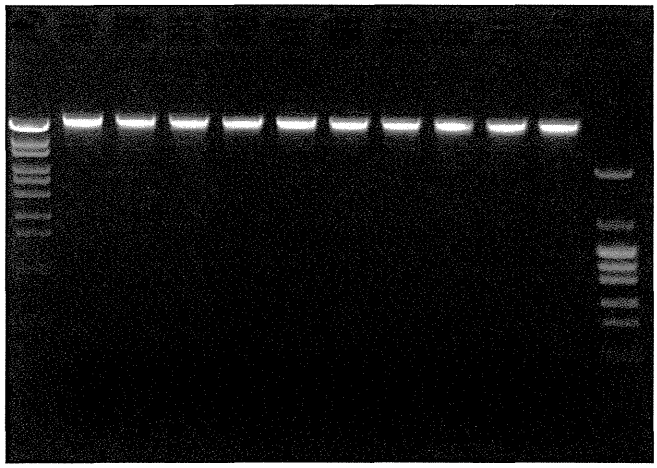
Quant-iT dsDNA BR Assay Kit（Thermo Fisher Scientific）を用いた蛍光定量が行われた。「吸光定量値/蛍光定量値」は核酸定量結果と二本鎖DNA 定量結果の乖離を示す（推奨値≤3）。評価項目として、解析に必要とされる核酸濃度、総量、及び核酸定量結果（表1）と二本鎖DNA定量結果の乖離データを指標にして評価した。

表3.　標準ゲノムDNAの定量結果（Nanodrop を用いた吸光定量）

サンプル	液量(ul)	濃度(ng/ul)	総量(ug)	260/280	260/230
Standard DNA1	20	60.3	1.2	2.2	1.3
Standard DNA2	20	57.6	1.2	2.1	1.3
Standard DNA3	20	57.9	1.2	2.2	1.3
Standard DNA4	18	55.8	1	2.2	1.2
Standard DNA5	18	56.4	1	2.2	1.3
Standard DNA6	18	59.7	1.1	2.2	1.2
Standard DNA7	18	60.3	1.1	2	1.3
Standard DNA8	18	62.1	1.1	2.2	1.3
Standard DNA9	18	57	1	2.1	1.2
Standard DNA10	18	55.8	1	2.1	1.2
Standard DNA11	18	57.3	1	2.2	1.3
Standard DNA12	18	56.1	1	2.2	1.3
Standard DNA13	18	56.4	1	2.2	1.3
Standard DNA14	18	54	1	2.1	1.3
Standard DNA15	18	56.4	1	2.2	1.2
Standard DNA16	18	54.9	1	2.2	1.2
Standard DNA17	18	52.5	0.9	2.2	1.2
Standard DNA18	18	52.8	1	2.3	1.2
Standard DNA19	18	54.9	1	2.2	1.3
Standard DNA20	18	56.1	1	2.1	1.2

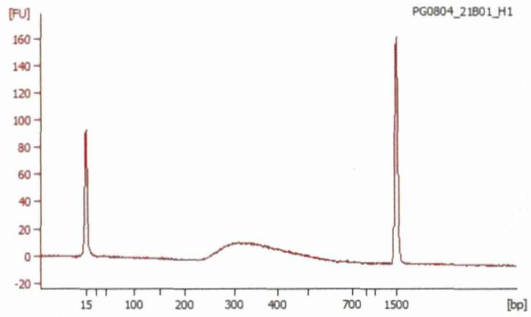


レーン（左から）	サンプル
1	λ-EcoT14 I digest
2	Standard DNA1
3	Standard DNA2
4	Standard DNA3
5	Standard DNA4
6	Standard DNA5
7	Standard DNA6
8	Standard DNA7
9	Standard DNA8
10	Standard DNA9
11	Standard DNA10
12	pHY Marker

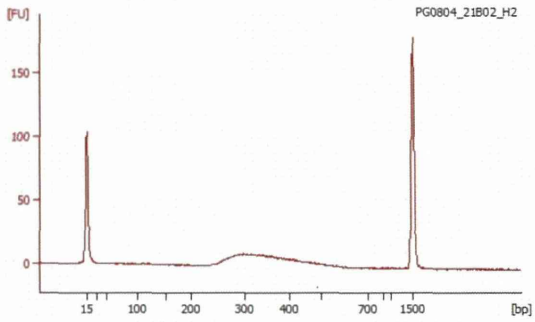


レーン（左から）	サンプル
1	λ-EcoT14 I digest
2	Standard DNA11
3	Standard DNA12
4	Standard DNA13
5	Standard DNA14
6	Standard DNA15
7	Standard DNA16
8	Standard DNA17
9	Standard DNA18
10	Standard DNA19
11	Standard DNA20
12	pHY Marker

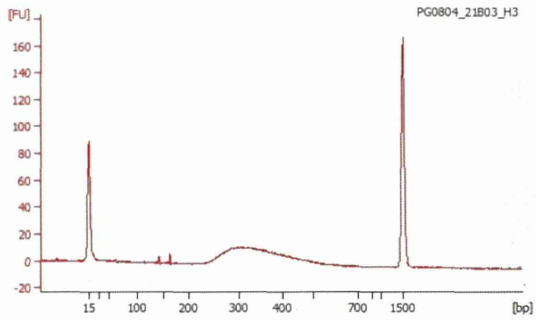
図1. アガロースゲル電気泳動によるゲノムDNAの品質確認
蛍光定量結果より50ng 分の二本鎖DNA をアプライした。



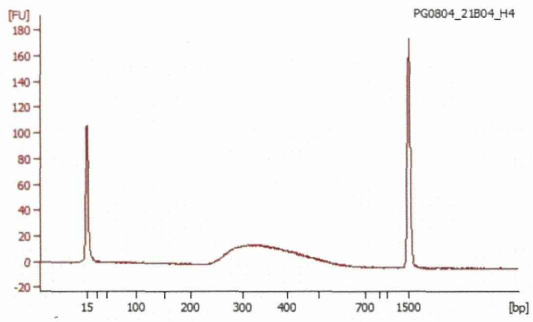
サンプル	Standard DNA #1
ピークサイズ	310 bp
濃度	25.9 nmol/L



サンプル	Standard DNA #2
ピークサイズ	305 bp
濃度	16.5 nmol/L



サンプル	Standard DNA #3
ピークサイズ	304 bp
濃度	25.0 nmol/L



サンプル	Standard DNA #4
ピークサイズ	312 bp
濃度	30.4 nmol/L

図2.　Agilent 2100 Bioanalyzer を用いたリード長の検定結果
作製されたシーケンスライブラリーの品質をAgilent 2100 Bioanalyzer を用いて測定した。また、ライブラリーにはアダプター配列を付加されているため、ピークサイズからアダプターサイズ（約100base）を除いたサイズがクローニングサイズとなる。
サンプルは、4種類のDepthについて解析を行うため、同サンプルで4解析分のライブラリーを独立して作製した。

表 4. Phred クオリティスコアに基づいたシーケンス解析の精度評価

サンプル名	リード数	塩基数	Q30R1	Q30R2
Standard DNA#1	474,895,722	47,489,572,200	95.3	93.6
Standard DNA#2	529,964,716	52,996,471,600	94.8	92.0
Standard DNA#3	440,178,726	44,017,872,600	95.5	94.2
Standard DNA#4	469,051,582	46,905,158,200	95.0	93.5

Q30R1: 片鎖・両鎖解析において 1 回目に読み取られるリードの品質

Q30R2: 両鎖解析において 2 回目に読み取られるリードの品質

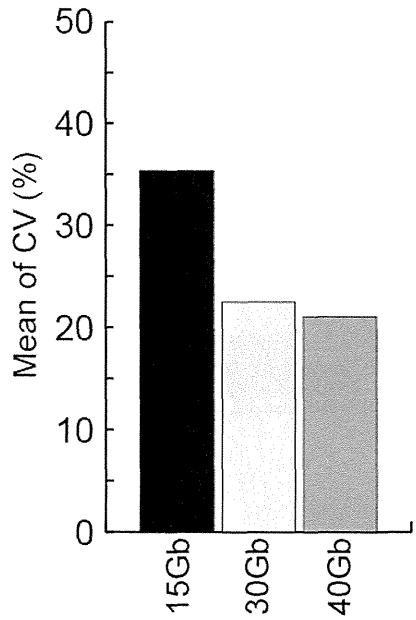


図 3. シーケンス量が 15Gb、30Gb、40Gb の場合における変動係数（CV）の平均

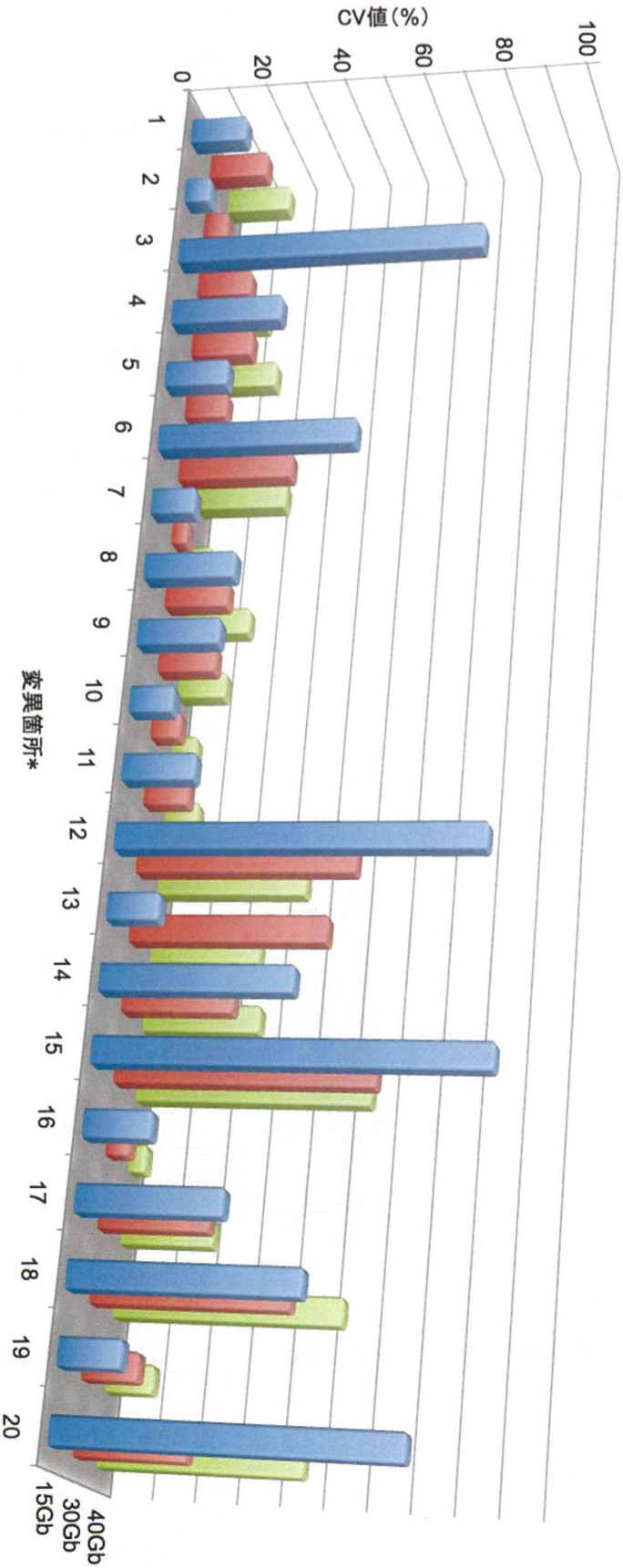


図 4. 15Gb、30Gb、40Gb の量でシーケンスされたときの各変異箇所における変異塩基の変動係数（CV）
CV 値は 4 回の独立したシーケンス解析から得られた変異頻度に基づいて算出された
*横軸の変異箇所については、表 5 に示した

表5. 標準ゲノムDNA上で確認された変異箇所

	リファレンスID	位置	遺伝子	アミノ酸変化	リファレンス配列（塩基）	変異塩基
1	chr1	115,256,530	NRAS	Q61K	G	T
2	chr2	29,416,326	ALK	P1543S	G	A
3	chr2	209,106,786	IDH1	S261L	G	A
4	chr3	37,061,883	MLH1	L187M**	C	A
5	chr3	41,266,101	CTNNB1	S33Y	C	A
6	chr3	178,936,091	PIK3CA	E545K	G	A
7	chr3	178,952,085	PIK3CA	H1047R	A	G
8	chr4	55,138,600	PDGFRA	G426D	G	A
9	chr4	55,599,321	KIT	D816V	A	T
10	chr5	112,179,431	APC	R2714C	C	T
11	chr7	55,241,707	EGFR	G719S	G	A
12	chr7	55,249,071	EGFR	T790M	C	T
13	chr7	55,259,515	EGFR	L858R	T	G
14	chr7	140,453,136	BRAF	V600E	A	T
15	chr8	38,285,611	FGFR1	P124L***	G	A
16	chr9	139,409,754	NOTCH1	P668S	G	A
17	chr12	25,398,281	KRAS	G13D	C	T
18	chr12	25,398,284	KRAS	G12D	C	T
19	chr13	28,626,706	FLT3	V197A	A	G
20	chr16	68,867,462	CDH1	N/A	T	C

変異頻度については、表1を参照。

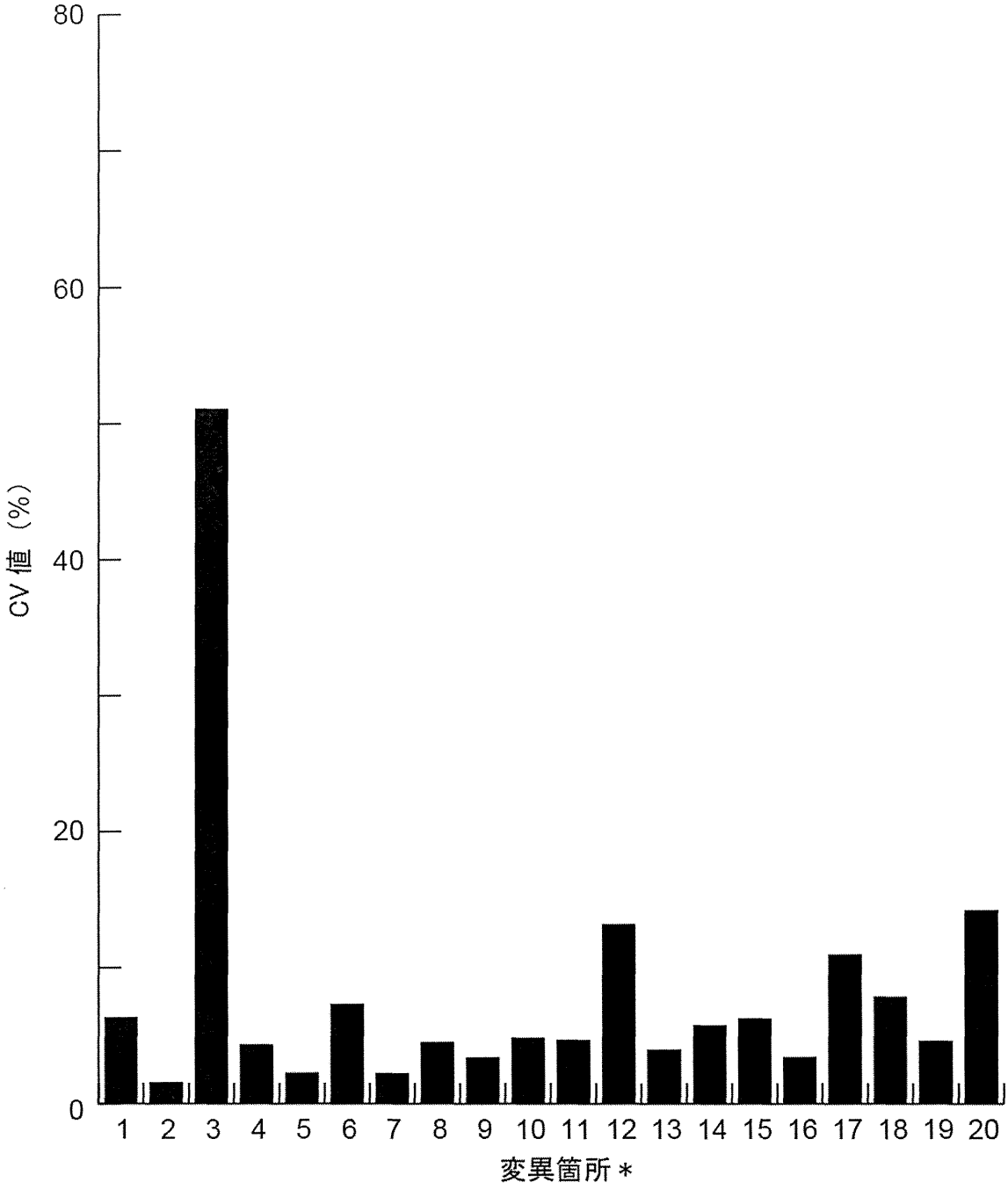


図 5. 全マッピング結果を合わせた時の各変異箇所における変異頻度の変動係数（CV）
全マッピング結果をまとめた後、40Gb、80Gb、120Gb、160Gb 相当のデータとなるようにダウンサンプリングし、各変異箇所におけるそれぞれのシーケンス量（40Gb、80Gb、120Gb、160Gb）に相当する塩基頻度を測定後、CV 値を算出した。
*横軸の変異箇所については、表 5 に示した

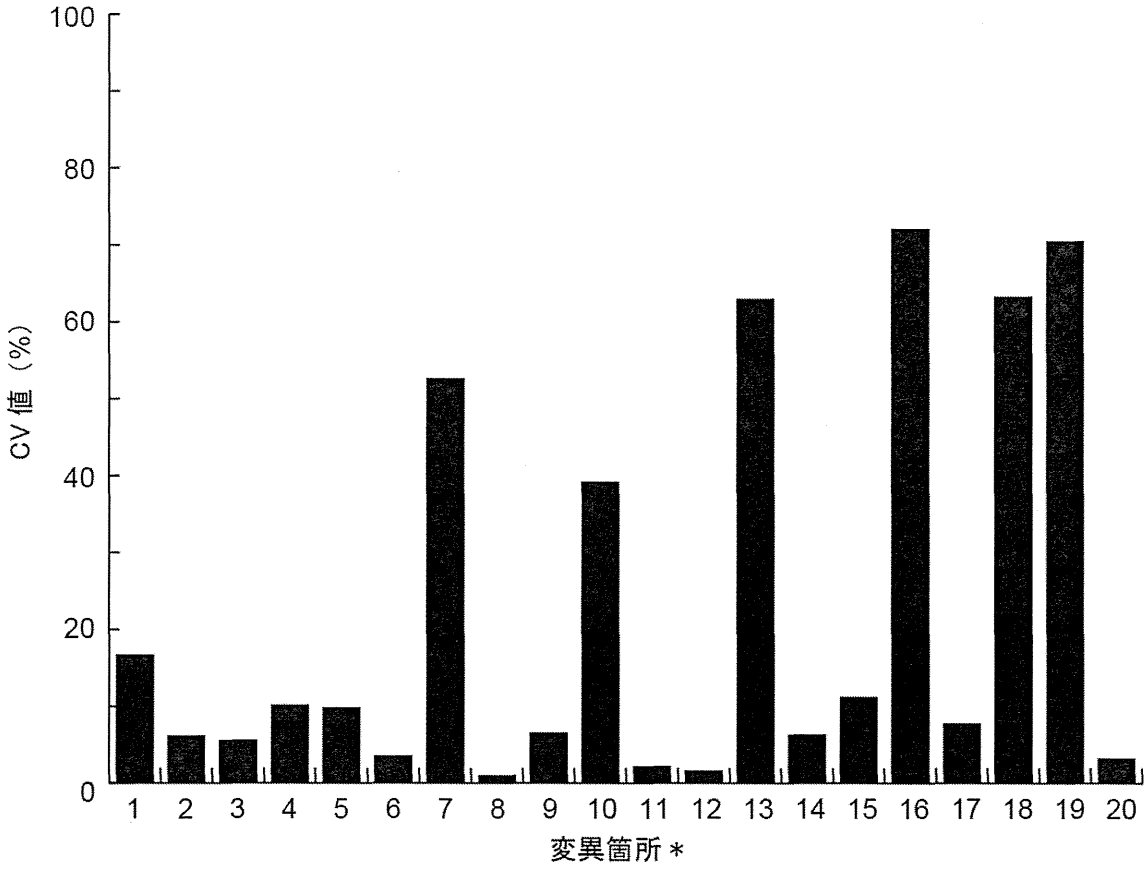


図 6. 全マッピング結果を合わせた時の各変異箇所における変異頻度の公表値に対する変動係数（CV）（公表値に対する実測値のバラツキ）
全マッピング結果をまとめた後、40Gb、80Gb、120Gb、160Gb 相当のデータとなるようにダウンサンプリングし、各変異箇所におけるそれぞれのシーケンス量（40Gb、80Gb、120Gb、160Gb）に相当する塩基頻度を測定後、CV 値を算出した。さらに、公表値を含めた CV 値を算出し、実測値および公表値の 2 群間の CV 値を求めた。
*横軸の変異箇所については、表 5 に示した

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」
分担研究報告書

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

研究分担者:
(独)国立成育医療研究センター 再生医療センター センター長
梅澤 明弘

研究要旨

人工多能性幹細胞(iPS 細胞)の発見から 10 年近くが経過し、培養細胞を用いた再生医療が現実味を帯びてきた。実用化に向け、その品質や安全性への関心が年々高まっており、遺伝的安定性評価、つまりゲノム塩基配列の変異を調べる手法自体への目も厳しくなっている。これまで、全ゲノムと比較して効率的という観点から主にエクソーム・シークエンシング解析および SNP アレイを用いた構造変異解析が採用され、細胞の安定性評価がなされているが、次世代シークエンサ技術の進歩と塩基当たりのコスト低減により、遅かれ早かれ全ゲノムを対象とした評価が求められることは間違いない。一方、一概に全ゲノムと言っても、我々日本人を対象とした研究、特に医療を行う場合には注意しなければならない点がある。それは主にヨーロッパ系やアフリカ系のヒトサンプルから構築されたヒトゲノムのリファレンス配列がどの程度、一般的な日本人のゲノム配列の比較対象として役立っているかという問題である。一塩基置換である SNP については多くのデータがそろっているが、それは配列の類似性が高く比較できた結果見つけられたヴァリアントであって、リードデータがリファレンス配列と比較して著しく異なる場合はマップされず、つまりそういったリードは無視されているのが現状である。これまで、こういった観点から網羅的な調査は行われてこなかった。それぞれの人種、また個々人は、さまざまな多型を保持していること知られ、既存のリファレンス配列とは大きく異なる構造多型等を検出するためには、リファレンス配列に頼ることなく、*de novo* アセンブリを行う必要がある。そこで本研究では、特に日本人を対象とした再生医療を念頭に、第一段階として一人の健常日本人男性を選び、既存のリファレンス配列を用いずに全ゲノム配列決定を行った。ペアエンド・ライブラリ、メイトペア・ライブラリをそれぞれ 3 つ、合計 6 ライブラリを用意し、次世代シークエンサでシークエンシング、4 テラバイトのメモリを搭載する最新の計算機でアセンブルを行い、さらに第 3 世代シークエンサから得られる 10 kb を超える長いリードデータを利用することで多くのギャップを埋めることができた。加えてさらに 5 人の健常日本人男性のサンプルからペアエンド・ライブラリ

を作製、シークエンシングを行い、本研究で決められたゲノム配列にマッピングを行い、その正確性および一般性を確かめた。新規配列、新規構造多型の同定はもちろんのこと、これまで日本人に共通して報告されていたコピー数多型の多くをブレークポイントとともに検出することができ、安全な再生医療に向けた網羅的な安定性評価法として、また将来のパーソナル医療を見据えた基礎データとして貢献し得るものであることが確認された。

研究協力者

(独)国立成育医療研究センター システム発生・再生医学研究部 組織工学研究室 室長
岡村 浩司

(独)国立成育医療研究センター 小児がん疫学臨床研究センター 登録データ管理室
井原 千琴

(独)国立成育医療研究センター 周産期病態研究部 周産期ゲノミクス研究室 室長
中林 一彦

(独)国立成育医療研究センター 周産期病態研究部 部長
秦 健一郎

A. 研究目的

再生医療に用いられる各種培養細胞の遺伝的安定性を評価する手法として、これまで主にエクソーム・シークエンシング解析(以下、エクソーム解析)、SNP アレイを利用した構造変異解析が行われてき。エクソーム解析は、ヒトゲノム中にわずか 1%にも満たない割合で存在するタンパク質に翻訳されるコーディング領域を回収し、全ゲノムと比較してきわめて効率良く重要な塩基配列を読み、一塩基置換や数塩基の挿入または欠失をヴァリアントまたは変異として検出する手法である。安定性の評価にはこのエクソーム解析でじゅうぶんなようにも思われるが、例えば染色体の広範囲に渡って欠失や、コピーニュートラル LOH などの構造変異が発生すると、一塩基置換が起こっていないにも関わらず、コールされるジェのタイプがヘテロからホモになってしまうため多数の変異が誤検出され、正しい評価を妨げることになる。本研究グループは、染色体構造変異を高頻度で起こす毛細血管拡張性運動失調症患者の線維芽細胞から iPS 細胞を樹立し、エクソーム解析および SNP アレイ解析を行い、実際に検出された近接する多数のヘテロからホモへの変異が、たった 1 つの構造変異に由来することを確かめている(Fukawatase et al. 2014)。単に蛍光強度の変化を測定する CGH アレイとは異なり、SNP アレイは UPD のようにコピー数が変化しない構造変異をも見つけることができ、エクソーム解析と組み合わせることで、遺伝的安定性を評価するための欠かせないツールとなっている。

しかしながらエクソーム解析が対象とす

る領域は、全ゲノムのうちほんの1%であり、SNPアレイにしても既知の30万SNPを用いての解析、つまり10 kbに1つのSNPで構造変異を検出しており、解像度はきわめて低いと言わざるを得ない。パイロット的な研究であれば問題はなかったが、次世代シークエンサの登場以来、塩基当たりの配列決定コストは年々下がっており、全ゲノム・シークエンシングが現実味を帯びてきた。医療に用いられる細胞の安定性評価という目的のためには全ゲノムを対象とする解析が望ましいことは改めて言うまでもない。

一方、エクソーム解析、全ゲノム解析に関わらず、特に日本人サンプルを扱う上で注意しなければならない点がある。ヒトゲノム配列のリファレンスの問題である。454、Solexa、SOLiDといった次世代シークエンサ登場以降、1000人ゲノムプロジェクトを含め、多くの個人のゲノム配列が決定されているが、これらはどれもリシークエンスと呼ばれる方法で、得られた数十億リードをGRCh37あるいはhg19などのリファレンス配列と比較し、その違いから変異やヴァリアントを検出する。それゆえ、得られたリードがヒトゲノム由来であったとしても、用いたリファレンス配列に似た配列がない場合は比較ができず、マップされなかったデータとして捨てられることになる。ミスアセンブリなどにより、同一配列がリファレンス内に複数存在した場合も、マルチヒットとなり、ヴァリアントコールの信頼性が落ちる。例えば転座を考えてみると、ブレークポイントとなっている場所は異なる染色体に存在すべき配列が偶発的に隣り合うわけで、ブレークポイントをカバーする配列がアラインメントされる可能性は低く、リシークエンシングのみによる構造変異の検出はきわめて困難である。仮に日本人独自の塩基配列から成る領域があったとすると、残念ながらその領域は解析対象になり得ない。

それぞれの人種、さらに個々人は、SNPに限らず、重複や欠失によるコピー数多型、逆位、転座、リピート伸張や短縮、トランスポゾンの挿入や欠失、遺伝子変換といったさまざまな多型を保持していることが知られている(Migita et al. 2014)。現在、広く使われているヒトゲノムのリファレンス配列は、主にヨーロッパ系およびアフリカ系の人種に由来するデータから構築されており、日本人のゲノム配列とどの程度合致するものなのか網羅的に調べられているわけではない。特に構造多型についてはリシークエンシングのみによって調べられるものではないため、リファレンス配列を用いずに、日本人ゲノムから得られたリードデータを独自に組み立て、つまり *de novo* アセンブリすることにより新規な配列を作り出して調べる必要性が生ずる。将来的に我が国において個々人の遺伝情報に基づいた医療を行う際には、きわめて重要な基礎データになるとも考えられる。

そこで本研究においては、その第一段階として、健常日本人男性一名を選び、そのゲノム配列を既存のリファレンス配列にマッピングすることなく、独自に全ゲノム配列の決定を行うことにした。一般的なリシークエンシングにおいては、ランダムに断片化された数百bpの両端配列を次世代シー

クエンサで読むだけであるが、本研究では、そういったペアエンド・ライブラリを 3 種類、さらに最大 9 kb 離れたゲノム断片を結合させたメイトペア・ライブラリを 3 種類作製、Illumina HiSeq でリードデータを取得し、最新のソフトウェアと 4 テラバイトの大規模メモリを搭載する最新の計算機を用いて *de novo* アセンブリを行うこととした。実際にはこれら 6 ライブラリのデータを駆使したとしても多くのギャップが生じることが予想されるため、10 kb 以上に渡って連続した塩基のリードデータを取得することができる第 3 世代シークエンサ PacBio RS II も用い、これらを組み合わせることで日本人ゲノム・リファレンス配列の構築に挑んだ。得られたデータは、他の 5 人のリードデータから高い正確性と日本人としての一般性が確認され、細胞の安定性評価に限らず、日本人、日系人、近隣アジア人の医療に役立つことはもちろんのこと、広く人類学への貢献も期待されるデータであることが確認された。

B. 研究方法

国立成育医療研究センター倫理委員会において承認を得た後、ポスター掲示等により検体提供のボランティアを募った。ボランティアの方々からは 3 日以上の禁欲期間後に精液を採取、リサーチコーディネーター宛てに提出していただいた。提供者にはできる限り 3 回の採取と提供、さらに採取日、先祖の国籍、先祖の出身地、実子がいる場合は全ての実子の年齢について記入をお願いした。実験に直接関与しないリサーチコーディネーターが連結不可能匿名化を行い、凍結した検体と回答情報を実験担当者に渡した。

1 回目の提供 6 検体(JOM001, JOM002, JOM003, JOM004, JOM005, JOM006)全てに対して DNA 抽出を行った。精液 250 µL に 10 mL の PBS を加えて 10 秒間激しく撹拌し、2000×g にて 10 分間遠心した後、約 1 ml を残してデカンテーションした。撹拌後、2 ml のチューブに移し、0.5 ml の PBS を加えてさらに撹拌し、遠心機の最高速度で 2 分間遠心、マイクロピペットを用いて注意深く液体を取り除き、Macherey-Nagel 社製の lysis buffer 325 µL、65℃、約 10 時間の Proteinase K 処理によりタンパク質分解を行った。その後の処理は Macherey-Nagel 社製のイオン交換樹脂カラムである NucleoBond AXG 100 を用い、説明書の指示に従い、その後の処理に用いるゲノム DNA を精製した。

アンケートの回答無関係に、6 検体から無作為に 5 検体を選び、Illumina 社製のビーズアレイ HumanOmni2.5-8 BeadChip を用いて SNP タイピングを行った。1000 人ゲノムデータから、全体と比較して日本人に稀な SNP を 17,201、欧州人と比較して日本人に稀な SNP を 1531、東アジア人と比較して日本人に稀な SNP を 4483、東アジア人にはまれだが日本人に比較的多い SNP を 827 選び、5 検体分のデータと比較した。

また、ヒト第 18 番染色体から 4000 の SNP を選び、主成分分析を行った。1000 人ゲノムから番号が 10 の倍数となっているヨーロッパ人(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、こ

られに今回得られた 5 検体のデータを加えた合計21人分のデータを用いて解析を行った。実際の解析はEIGENSTRAT 6.0.1 を Red Hat enterprise Linux 6.1 (Linux 2.6.32) にインストールしたものを用いた。また、SNP タイピングの結果は、最終的に得られた配列の検証にも用いた。

SNP解析を行った5人からJOM005を選び、Illumina HiSeq のプラットフォームでシークエンシングを行うためのライブラリ作製を行った。まず、ゲノム DNA をアコースティックソルビライザーCovaris を用いて物理的に数百 bp に断片化し、酵素処理により両末端の平滑化、およびリン酸化を行った。サイズ選別後に3'-dA突出末端処理を行い、インデックス付きアダプタを付加することで 3 種類のペアエンド・ライブラリを作成した。試薬は Illumina TruSeq DNA PCR-Free LT Sample Prep Kit を用いた。メイトペア・ライブラリは Illumina Nextera Mate Pair Sample Prep Kit を用い、最長 9 kb の 3 種類のライブラリを作製した。片側のリード長は150 nt とした。第3世代シークエンサは、タカラバイオ社が保有する PacificBio Science 社製 PacBio RS II を用いた。同検体から得られたゲノム DNA を AMPure XP を用いて精製後、g-TUBE により約 20 kb に断片化、両端を平滑化し、SMRTbell 一本鎖アダプタをライゲーションすることでライブラリを作成した。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、BluePippin を利用してサイズ選別、2 分してS1およびS2の2ライブラリを作製した。シークエンシングにはP6/C3試薬を用いた。その他5サンプル(JOM001, JOM002, JOM003, JOM004, JOM006)についても平均インサート長が 360 bp 程度のペアエンド・ライブラリを作製し、HiSeq にて片側 125 nt のシークエンシングを行った。

得られたリードデータは FASTQ 形式で出力し、独自のスクリプトによりフォーマットチェックとフィルタリングを行った後、Cutadapt 1.7.1 により、インサート長が長い場合に 3' 末端に現れるアダプタ配列を取り除いた。さらに独自のスクリプトを用い、5' および 3' 両端の低品質塩基(Q スコア 16 未満)を取り除き、対の少なくとも一方が 36 nt 以上のリードペアをその後の解析に用いた。

3つのペアエンド・ライブラリ、また3つのメイトペア・ライブラリの正確なインサート長を推定する目的で、まずミトコンドリア DNA のゲノム配列を決定した。決定には、ヒトとマウスで保存されている配列 5'-CCG TGC AAA GGT AGC ATA ATC ACT TGT TCC T-3' のみを利用し、本研究グループが独自に開発した DNA 配列アセンブラ GrepWalk 0.6 を用いた (http://epigenetics.nrichd.ncchd.go.jp/grepwalk/)。完全決定された 16,570 bp の JOM005 ミトコンドリア DNA をリファレンスとし、今回得られたペアエンド・ライブラリおよびメイトペア・ライブラリのリードデータを BWA 0.7.12 を用いてマッピングし、各ライブラリにおける平均インサート長およびその標準偏差を決定した。プログラムおよびスクリプトはCまたはPerl を用いて記述した。

本研究におけるデータ処理は主に、国立成育医療研究センターが所有する 35 ノード

からなる計算機クラスター Hitachi HA8000/RS210 を用いた。また、本研究の中心となる大規模な *de novo* アセンブリを遂行するため、4テラバイトの大規模メモリを搭載する計算機 Dell PowerEdge R920 を新規に購入した。最初、OS として Fedora 21(Linux 3.17.4)をインストールして運用を開始したが、*de novo* アセンブリのためのプログラム ALLPATHS-LG 52188 が動作しなかったため、後に OS を CentOS 7.1 (Linux 3.10.0)に入れ替えた。アセンブリは主に ALLPATHS-LG を用いて行ったが、他にも SOAPdenovo2 r240、ABySS 1.5.2、Trinity r20140413p1、GrepWalk 0.6 を組み合わせて行った。PacBio RS II から得れたロングリードは BLASR および PBJelly を用いて処理した。

C. 研究結果

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(Figure 1)。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 1)。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 2)。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(Figure 2)。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 3)。

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 4)。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 5)。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 6)。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 7)。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(Figure 3)。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(Figure 4)。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(Figure 5)。

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(Figure 6)。

D. 考察

3種類のペアエンド・ライブラリともう3種類のメイトペア・ライブラリを作製し、HiSeqを用いて片側150 ntのペアエンドデータを取得した。これらに対し、最新の大規模メモリ搭載型計算機で*de novo*アセンブリを行い、N50が45 kbのデータを取得することができた。総contig長は2,255 Mbに達し、反復配列を除いたゲノム領域をおおよそカバーすることができた。本研究ではまずmtDNAの完全長を決定し、*de novo*アセンブリを行う前に6種類の各ライブラリのインサート長を推定するという工夫を行った。実際に見積もったインサート長は、どれもライブラリ作製時に狙ったサイズよりも小さく、この正確さによりアセンブリの精度も上がったと考えている。その一方でまだ多くのgapが残っており、PacBioから得られたリードを、塩基配列そのものではなく、scaffoldの並び換え目的で使用することにより全ゲノム配列決定に向けて作業を進めて行く予定である。

本研究グループが高解像度SNPアレイを用いて日本人固有のコピー数多型を多数報告したが、欠失はともかく、コピー数獲得