T2R2 東京工業大学リサーチリポジトリ
Tokyo Tech Research Repository

論文 / 著書情報
Article / Book Information

題目(和文)	ハイスループ　ットDNAシークエンスデ　ータによる異質倍数体種Saccharomyces pastorianusの比較ゲ　ノム解析
Title(English)	
著者(和文)	奥野未来
Author(English)	Miki Okuno
出典(和文)	学位:博士（理学）, 学位授与機関:東京工業大学, 報告番号:甲第10088号, 授与年月日:2016年3月26日, 学位の種別:課程博士, 審査員:伊藤 武彦,黒川 顯,中島 信孝,山田 拓司,小寺 正明
Citation(English)	Degree:, Conferring organization: Tokyo Institute of Technology, Report number:甲第10088号, Conferred date:2016/3/26, Degree Type:Course doctor, Examiner:,,,,
学位種別(和文)	博士論文
Type(English)	Doctoral Thesis

Powered by T2R2 (Tokyo Institute Research Repository)

平成 27 年度　博士論文

ハイスループット DNA シークエンスデータによる

異質倍数体種 *Saccharomyces pastorianus* の比較ゲノム解析

指導教官　伊藤　武彦　教授

東京工業大学大学院
生命理工学研究科生命情報専攻

奥野　未来

目次

第 1 章　序論 1

1.1　ゲノムシークエンス，配列決定の歴史 1

1.2　異質倍数体種のゲノム解析の取り組み 2

1.3　異質倍数体種 *S. pastorianus* のゲノム解析の現状 3

1.4　*S. pastorianus* の地理的起源に基づく分類と系統関係 5

1.5　本研究の目的 6

1.6　本論文の構成 6

第 2 章　本論文で用いたサンプル及びデータ 8

2.1　使用株 8

2.2　シークエンスデータ 8

2.3　32-mer 頻度分析 10

2.3.1　方法 10

2.3.2　結果 11

第 3 章　*de novo* アセンブリに基づいた *S. pastorianus* 及び祖先種 *S. eubayanus* ゲノムの決定 14

3.1　方法 14

3.2　結果・考察 15

3.3　まとめ 22

第 4 章　*S. pastorianus* ゲノムデータの再解析 23

4.1　ミトコンドリアの配列決定，系統解析 23

4.1.1　方法 23

4.1.2　結果・考察 24

4.2　倍数性の推定 30

4.2.1　方法 30

4.2.2　結果・考察 31

4.3　染色体間転座の検出 40

4.3.1　方法 40

4.3.2　結果・考察 41

4.4　SNV 検出，系統解析 50

4.4.1　方法 50

4.4.2　結果・考察 51

4.5　まとめ 56

第 5 章　**地理的起源に基づく分類と系統関係の再考察** **57**
5.1　*S. pastorianus* Sc type における hetero SNV 57
5.2　Group 1 と 2 の間に存在する SNV 61
5.3　仮説の検証（homo/hetero サイト）..... 65
5.3.1　方法 65
5.3.2　結果 70
5.4　仮説の検証（homo/homo サイト）..... 84
5.5　考察 89
5.6　まとめ 91
第 6 章　**結論と展望** **92**
参考文献 **94**
謝辞 **100**

第1章　序論

1.1　ゲノムシークエンス，配列決定の歴史

1996年に真核生物として初めて出芽酵母 *Saccharomyces cerevisiae* のゲノム解読が完了し（Goffeau et al. 1996），その 2 年後にミトコンドリア配列が決定された（Foury et al. 1998）．12 Mbp（当時の完全解読されたゲノムとしては最長）のゲノム配列に対してヨーロッパ，北アメリカ，日本の研究者約 600 人が関わる大きなプロジェクトであった．染色体 III（315 kbp）の配列決定から全ゲノム解読まで 4 年を要したが（Oliver et al. 1992），1 年で 6 Mbp の配列決定が可能になるなど，当時のゲノム配列決定において革新的な出来事であった．

2000 年代に入り，Roche 社，Illumina 社製を代表とするハイスループット DNA シークエンサの登場により，一塩基あたりのシークエンスコストが低下するとともに，容易に大量のシークエンスデータが取得可能となった（Metzker 2010）．Illumina 社製のシークエンサが発表された当初はリード長が 36 bp と短かったため，あらかじめ構築されたゲノムに対してリードをマッピングし，変異を検出することが主な利用目的であった．しかし，シークエンスリード長の伸長と，paired-end, mate-pair ライブラリ（DNA 断片の両端をシークエンスすることで得られる 2 箇所の DNA 配列とその配列間の距離及び向きの情報を持ったリードのペア）の登場により，ハイスループット DNA シークエンサの短所であるリード長の短さを克服したゲノムの構築の可能性が示唆され始めた（Glenn 2011）．さらに，大量のシークエンスデータを扱うためのアルゴリズムの開発が進み，高速かつ高精度のマッピングやゲノムアセンブリが可能になった（Li and Homer 2010; Miller et al. 2011）．このように，シークエンサ性能が向上するとともにと様々な解析ツールが開発され，さらには計算機能力の向上とも相まって，モデル生物以外のゲノム解析も一般的に行われるようになってきた．

さまざまな生物のゲノムが解読され，データベースが充実していくにつれて，生物種間の比較ゲノム解析や系統解析と同時に種内の株間の配列違い，個体差に注目した比較ゲノム解析も行われるようになった．一塩基多型（Single nucleotide variant, SNV）や，大規模な挿入・欠失（InDel），染色体間転座やコピー数多型（Copy number variation, CNV）などのゲノムの構造の違いにも着目され，表現型との関係が明らかになりつつある（Alkan et al. 2011）．ヒトゲノムを用いた解析がもっとも進められているが（Durbin et al. 2010），*Saccharomyces cerevisiae* でも，多様な地域から単離した株やワイン酵母，清酒酵母，エタノール生産酵母などの産業用酵母のゲノム解析（Bradbury et al. 2006; Curtin et al. 2012; Akao et al. 2011; Babrzadeh et al. 2012），比較ゲノム解析が行われている（Liti et al. 2009; Legras et al. 2007; Borneman et al.

2011）.

1.2 異質倍数体種のゲノム解析の取り組み

近年のさまざまな生物に関するゲノム解析により，植物の 25%，動物の 10%程度で異種交配または他の種からの潜在的な遺伝子移入が起こっていることが示唆され（Mallet 2005），酵母でも頻繁に異種交配が起きていることが明らかになった（Dunn and Sherlock 2008; Nakao et al. 2009; Querol and Bond 2009; Sipiczki 2008）．異質倍数体種は，2 つ以上の異なる種の交雑により得られ，2 種以上のゲノムを 1 つの細胞内に持つ．また，異種交配後のゲノム再構成の過程において，染色体間転座，染色体欠失・重複などが繰り返し起き，祖先種と異なるゲノム構造を持っている．

近年行われた異種交配株の酵母のゲノム解析を表 1-1 にまとめた．対浸透圧性酵母 *Pichia sorbitophila* は，祖先種 2 種が単離されていなかったため，Sanger 法でシークエンスし，配列を構築した後，ドラフトゲノムを GC 含量に従って祖先種 2 種へ分類し，染色体間転座の領域の推定を行っている（Louis et al. 2012）．高酢酸耐性酵母 ISA1307 株は，ドラフトゲノムの構築後，祖先種 *Zygosaccharomyces bailii* のドラフトゲノムとの比較解析を行っているが，他方の祖先種が単離されていないために，乳酸発酵能や胞子形成を行わないといった形質が祖先種に由来するものであるのかを明らかにすることができなかった（Mira et al. 2014）．異種交配後のゲノム再構成を理解するには，染色体レベルで構築された祖先種 2 種のゲノムと比較することが理想的である．しかし，実際には祖先種が単離されてない場合やドラフトゲノムであることが多く，上の例のように祖先種のゲノムが利用できない場合には解析作業の負担が大きくなることが予想される．

Saccharomyces 属内の異種交配株は，工業用の発酵において重要な役割を果たしており，また祖先種が単離されゲノムが利用可能な場合が多いことから実験的手法のみならずゲノム解析を中心にした研究も進められている（Borneman et al. 2012; Peris et al. 2012）．*Saccharomyces* 属内の異種交配株（ワイン酵母やビール酵母など）について，アレイ比較ゲノムハイブリダイゼーション（array Comparative genomic hybridization, array-CGH）やハイスループット DNA シークエンサによるゲノムシークエンスがさまざまな研究グループによって行われている（Peris et al. 2012; Borneman et al. 2012; Scannell et al. 2011）．

表 1-1 異質倍数体種の酵母のゲノム解析

異質倍数体種	祖先種		(参考文献)
Pichia sorbitophila 対浸透圧性酵母	-	-	(Louis et al. 2012)
ISA1307 高酢酸耐性酵母	*Zygosaccharomyces bailii*	-	(Mira et al. 2014)
S. bayanus CBS380, NBRC1948	*S. uvarum*	- (*S. eubayanus*)*	(Nguyen et al. 2011)
VIN7 ワイン酵母	*S. cerevisiae*	*S. kudriavzevii*	(Borneman et al. 2012)
HA1841 など 12 株 ワイン酵母，ビール酵母	*S. cerevisiae*	*S. kudriavzevii*	(Peris et al. 2012)

* 当時，*S. eubayanus* は未単離であり，*S. bayanus* の祖先種として知られていたのは *S.uvarum* のみであった．

1.3 異質倍数体種 *S. pastorianus* のゲノム解析の現状

ラガービールはビール市場でも大きなシェアを占めており，ラガービール酵母 *Saccharomyces pastorianus* は何世紀にも渡って研究が行なわれてきた．15 世紀頃に現在のミュンヘンでラガービールに相当する飲料ができたとされており（Corran 1975），現在使われているラガービール酵母は，19 世紀に Hansen らが単離したものであると考えられている（Wendland 2014）．19 世紀以降，ラガービール酵母は *S. pastorianus* として分類され，さらに地理的起源に基づいて大きく 2 つのグループ（Saaz type, Frohberg type）に分けられている．Saaz type（Group 1）はチェコのボヘミアに由来し，Frohberg type（Group 2）はドイツに由来するとされている（P 1909; Gibson et al. 2013）．*S. pastorianus* は多くの株が単離されており，代表的なものに *S. monacensis*（CBS1503; Group 1），*S. carlsbergensis*（CBS1513; Group 1），*S. pastorianus*（CBS1538; Group 1），Weihenstephan 34/70（W34/70; Group 2）が存在する．

ラガービール酵母のゲノム解析に関する主な研究を表 1-2 にまとめた．1987 年に *S. pastorianus* は *Saccharomyces* 属の 2 種からなる異質倍数体種であること，その祖先種の 1 種が *S. cerevisiae* であることが明らかなった（Nijhoffpublishers 1987）．しかし，2011 年まで *S. pastorianus* のもう 1 種の祖先種に相当する同質 2 倍体種（non Sc type）は単離されず，*S. pastorianus* の解析には，当時 non Sc type にもっとも近縁であった同質 2 倍体種 *Saccharomyces uvarum* が主に用いられた．2009 年にサンガー法によりラガービール酵母で初めて W34/70 株のドラフトゲノム配列が決定された．

S. pastorianus は祖先種である *S. cerevisiae* (Sc type) と non Sc type に由来するそれぞれ 16 本の染色体及び Sc – non Sc type 間の染色体間転座を含む染色体から構成され，異種交配後のゲノム再構成による複雑なゲノム構造を持つことがシークエンスデータから明らかになった（Nakao et al. 2009）.

表 1-2 ラガービール酵母ゲノムに関する主な研究

年	内容 (参考文献)
1987	ラガービール酵母が異質倍数体種であることが分かる (Nijhoffpublishers 1987)
2005	*Saccharomyces* 属 112 株の LTR 型レトロトランスポゾンの比較解析 (Liti et al. 2005)
2006	*S. pastorianus* を含む 12 株 48 サイトでの PCR/RFLP 解析 (Rainieri et al. 2006)
2008	*S. pastorianus* とその近縁種のミトコンドリアの系統解析 (Rainieri et al. 2008)
	Array CGH により *S. pastorianus* Group 1 と 2 の間に倍数性の差異があることを示す (Dunn and Sherlock 2008)
2009	Weihenstephan 34/70 の Sanger 法によるドラフトゲノム解読 (Nakao et al. 2009)
2011	パタゴニアでの *S. eubayanus* の単離 (Libkind et al. 2011)
2014	Group 1 と 2 の共通の染色体間転座の確認 (Hewitt et al. 2014)
	Array CGH による CNV の解析と mRNA の発現レベルの比較 (Tadami et al. 2014)
	S. carlsbergensis CBS1513 のゲノム解析 (Walther et al. 2014)
	チベットでの *S. eubayanus* の単離 (Bing et al. 2014)
2015	PCR によるラガービール酵母特有遺伝子の比較解析 (Monerawela et al. 2015)
	実験的に作った *S. cerevisiae* x *S. eubayanus* hybrid 株のゲノム解析 (Hebly et al. 2015)
	S. eubayanus と *S. pastorianus* の比較ゲノム解析 (Baker et al. 2015)
	S. pastorianus CBS1483 のゲノム解析 (Van den Broek et al. 2015)

2011 年に入りようやくパタゴニアにおいて *S. pastorianus* の non Sc type の配列と高い相同性（99.5%）を持つ種が単離され，*Saccharomyces eubayanus* と命名された．そして，*S. eubayanus* が *S. pastorianus* のもう 1 種の祖先種（non Sc type）であり，*S. pastorianus* がエールビール酵母 *S. cerevisiae* と比較して低温での生育，発酵能に秀でているのは *S. eubayanus* の耐寒性に由来することが明らかになった（Libkind et al. 2011）．その後，アジアや北アメリカでも *S. eubayanus* が単離され，現在ではアジア（チベット周辺地域）から持ち込まれた *S. eubayanus* が *S. pastorianus* の祖先種である可能性が高いと考えられている（Bing et al. 2014）．*S. eubayanus* が単離されたことで異質倍数体種 *S. pastorianus* は祖先種 2 種のゲノムとの比較が可能になったが，当時の *S. eubayanus* シークエンスデータは（Libkind et al. 2011），Illumina 社製シークエンサによる 36 bp single end リードであり，シークエンス量もゲノムサイ

ズの 13 倍程度しかなかったため，*S. eubayanus* の *de novo* アセンブリは行われておらず，本研究を実施した時点では *S. eubayanus* のドラフトゲノムは存在しなかった．

異質倍数体種の祖先種 2 種のゲノム配列は，異種交配後のゲノム再構成を理解する上で大きな助けになると考えられるが，実際に，*S. pastorianus* のように祖先種 2 種が単離され，配列が利用可能な異質倍数体種は少ない．ゲノム再構成の理解という観点からも祖先種 2 種のゲノム配列を基盤とした *S. pastorianus* の比較ゲノム解析は有用であると考えられる．

1.4 *S. pastorianus* の地理的起源に基づく分類と系統関係

S. pastorianus における地理的起源に基づく分類が系統的な違いを意味しているのか，つまり，Group 1 と 2 が異なる祖先株による別々の異種交配に由来するものであるか，あるいは，共通の祖先株による異種交配に由来するものであるか，については現在，相矛盾する結果が得られており，議論が続いている．先行研究で得られた知見とそれぞれの考察を以下に示した．

- 倍数性

2006 年には array-CGH を用いた研究により 2 つのグループ間で倍数性が異なることが明らかとなり，地理的起源に基づいた分類がゲノムにおいても差異が存在することが示唆された（Dunn and Sherlock 2008）．具体的には，Group 1 (Saaz type) は 1 倍体の *S. cerevisiae* 由来の染色体（Sc type）と 2 倍体のもう一種に由来する染色体 (non Sc type) を持ち，Group 2 (Frohberg type) は両祖先種とも 2 倍体（異質 2 倍体) であることが示された．しかし，当時，*S. eubayanus* が単離されておらず，array CGH 解析は non Sc type と最も相同性の高い同質 2 倍体種であるとされた *S. uvarum* の配列を基に行われていた．このため，non Sc type の倍数性の推定は Sc type に比べて解像度が低く，倍数性の違いや染色体間転座を明確に見ることができなかった．

- 系統解析

Sc type において一部の遺伝子の配列を基にした系統解析の結果から，*S. pastorianus* は系統的に 2 つに分類され，それが地理的起源に基づく分類と一致することが明らかになった（Dunn and Sherlock 2008）．倍数性の差異と合わせて，Group 1 と 2 は別々の異種交配に由来することを示唆する結果と考えられている．

- ミトコンドリア

異種交配株のミトコンドリアは祖先種 2 種のうちどちらか片方のミトコンドリアに由来することが知られている．実験的に *S. cerevisiae* と *S. bayanus* を異種交配させた場合，異種交配株として *S. cerevisiae* のミトコンドリアを持つ株と *S. bayanus* のミトコンドリアを持つ株が得られるが，実際のラガービール酵母 *S. pastorianus* 22 株の制限酵素断片長多型解析において，*S. pastorianus* は *S. cerevisiae* ではない祖先種に由

来するミトコンドリア配列を持つことが確認された（Rainieri et al. 2008）．しかし，*S. eubayanus* のミトコンドリア配列が未決定であるため，*S. pastorianus* のミトコンドリア配列の由来となる同質体種が明らかにされていない．

- 染色体間転座

2 つのグループの株間においてブレイクポイントが一致する染色体間転座の存在が明らかになった（Hewitt et al. 2014）．これらの染色体間転座は，別々の系統で起きたイベントではなく，共通の異種交配後に起きたイベントである可能性を示唆しており，Group 1 と 2 が共通の異種交配に由来することを支持する結果であると考えられている．

以上のように，倍数性と系統解析からは Group 1 と 2 は別々の異種交配に由来することを支持する結果が得られ，一方，ミトコンドリア配列と染色体間転座からは共通の異種交配に由来する可能性が示唆されている．しかし，これらの解析は，*S. eubayanus* が単離される前に行われたものが中心であるため，祖先種 2 種のゲノム配列に基づいた解析により，倍数性の推定や染色体間転座についてはより解像度の高い結果が得られると考えられる．また，ハイスループット DNA シークエンスデータを用いたゲノムワイドな解析により，地理的起源に基づいた分類と系統関係に見られる相矛盾する結果を満足させるシナリオが明らかになることが期待される．

1.5 本研究の目的

以上の背景を踏まえ，ハイスループット DNA シークエンスデータを用いた全ゲノムレベルでの解析により，異質倍数体種 *S. pastorianus* の地理的起源に基づく分類の系統関係の解明を目的として本研究を実施した．特に Group1, 2 の分類とゲノム配列に基づいた系統関係との間にどのような関係があるかに着目し，両グループが同一祖先由来／異なった祖先由来どちらであるのか議論になっている状況に対して全ゲノム情報からの結論を導くことを目的とした解析を試みた．

1.6 本論文の構成

本論文は異質倍数体種 *S. pastorianus* の地理的起源に基づく分類（Group 1 及び Group 2）の系統関係の解明を目的としており，次章にて使用サンプル及びデータについて説明した後，以下に示す 3 章において本論文の主要部分を述べる．ここではその概要を示す．

第 3 章では，*S. pastorianus* 及びその祖先種の一つである *S. eubayanus* CBS12357 の配列決定について説明する．*de novo* アセンブリの結果，*S. eubayanus* CBS12357 については染色体レベルでの配列の構築に成功した．一方，*S. pastorianus* 4 株につい

ては，異種交配後のゲノム再構成における倍数性の変化及び染色体間転座がアセンブリの障害となっている可能性が示唆された．このため，以降の *S. pastorianus* の比較解析は，祖先種 2 種（*S. cerevisiae* S288c 完全ゲノム配列と今回構築した *S. eubayanus* CBS12357 のドラフトゲノム）を基盤としたマッピングベースで実施した．

第 4 章では，*S. pastorianus* Group 1，2 に属するそれぞれ 5 株，計 10 株及び祖先種 2 種の比較ゲノム解析の結果を示す．具体的には，ミトコンドリア，倍数性，染色体間転座，及び，一塩基多型（Single nucleotide variants, SNV）に基づいた系統解析について，ハイスループット DNA シークエンスデータを用いて得られた新たな知見と系統関係についての考察を述べる．祖先種のゲノム配列を基盤としたことでより解像度の高い結果が得られたが，これらの結果も地理的起源に基づく分類と系統関係について相矛盾する仮説をそれぞれ支持しており，結論を出すことはできなかった．しかし，Sc type における系統解析に関して，Group 1 と 2 における倍数性の違い，及び Group 2 に属する株で観察された hetero SNV を考慮する必要があると判断し，さらに詳細に解析を実施することとした．

第 5 章において，再び SNV に着目した時に得られた系統関係に関する仮説とその検証結果を示す．Group 2 に属する株の Sc type において，異種交配後のゲノム再構成の中で loss of heterozygosity（LOH）が起きた可能性が高いことが示唆された．また，Group 2 の Sc type における hetero SNV に着目すると，Group 1 (1 倍体)，Group 2 (2 倍体) のハプロタイプが共通祖先株に由来する可能性が示唆された．この仮説を検証するために，ハプロタイプ別ゲノムの構築と系統解析を実施し，その結果，Group 1 と 2 のハプロタイプの片方の相同性が高いことが示された．また，Group 1 – 2 間に存在する homo SNV は，祖先株が異なることに起因するものではなく，Group 2 に属する株における LOH によって生じた可能性が示唆された．これらの結果は仮説を支持するものであり，染色体間転座及び Se type の系統樹の結果も踏まえ，最終的に，地理的起源に基づいた分類（Group 1 及び Group 2）は，共通の異種交配に由来すると結論を出した．

最後に第 6 章おいて本論文の総括としてまとめと展望を述べた．

第2章　本論文で用いたサンプル及びデータ

2.1　使用株

本研究では表 2-1 の通り，*S. pastorianus* とその祖先種（*S. cerevisiae*，*S. eubayanus*）と近縁種を含む合計 5 種 16 株を使用した．*S. cerevisiae* S288c は *Saccharomyces* Genome Database (SGD) から完成ゲノム配列を, また, *S. eubayanus* BaiFY1 (Bing et al. 2014) は NCBI の Sequence Read Archive (SRA) からシークエンスリードをそれぞれダウンロードした (Cherry et al. 1998)．それ以外の 14 株は本研究において新たにシークエンスを実施した．シークエンスに際し，*S. pastorianus* W34/70 は Fachhochschule Weihenstephan (ドイツ フライジング) から，それ以外の株は，CBS-KNAW Culture Collection (オランダ ユトレヒト) から入手した．

表 2-1 本研究で使用した株

種名		株名	
S. cerevisiae		S288c	
S. eubayanus		CBS12357	type strain
		BaiFY1	
S. uvarum		CBS7001	
S. pastorianus *S. cerevisiae* x *S. eubayanus*	Group 1	CBS1503	*S. monacensis* type strain
		CBS1513	*S. carlsbergensis* type strain
		CBS1538	*S. pastorianus* type strian
		CBS1174	
		CBS2440	
		W34/70	
	Group 2	CBS1483	
		CBS1484	
		CBS2156	
		CBS5832	
S. bayanus *S. eubayanus* x *S. uvarum*		NBRC1948	
		CBS380	type strain

2.2　シークエンスデータ

用意したライブラリと得られたシークエンスデータを表 2-2 に示した．ライブラリ

は DNA 断片の長さによって paired-end ライブラリと mate-pair ライブラリの 2 種類に大きく分けられる．ライブラリ調製時に PCR 増幅を行うと，増幅のバイアスによりシークエンス量からの染色体の倍数性の推定に影響があると考えられたため，本研究では paired-end ライブラリの調製には PCR 増幅を必要としない TruSeq PCR-free DNA sample Prep kit (Illumina 社) を使用した．paired-end ライブラリの DNA 断片長（インサートサイズ）は 600 bp, リード長は 300 bp とし, シークエンスには Illumina 社製 MiSeq を使用した. Mate-pair ライブラリの作成には Nextera Mate Pair Sample Prep kit (Illumina 社) を使用し，アガロースゲルから 3,000 bp から 15,000 bp までの断片を切り出し，各サンプルのライブラリを用意し，シークエンスは paired-end ライブラリと同様に MiSeq で行なった. SRA からダウンロードした *S. eubayanus* BaiFY1 のリードは, 151 bp の paired-end ライブラリで, 使用したシークエンサは HiSeq2000 であった．以降の解析では，シークエンスリードからアダプタ配列と低クオリティ配列を取り除いたデータを使用している（使用ツール Platanus_trim v.1.0.7）.

表 2-2 シークエンスデータ

strain		Paired end	Mate pair (3k)	Mate pair (6k)	Mate pair (9k)	Library Mate pair (12k)
***S. eubayanus* strains**						
CBS12357		1,408,342,206	670,138,052	748,047,601	718,992,262	680,496,916
BaiFY1		1,570,289,028	-	-	-	-
***S. pastorianus* strains**						
Group1	CBS1503	5,975,575,441	913,568,499	773,105,824	-	-
	CBS1513	5,231,503,163	882,938,463	923,494,068	-	-
	CBS1538	5,400,469,519	915,278,945	766,304,051	-	-
	CBS1174	3,277,582,182	-	-	-	-
	CBS2440	3,126,795,966	-	-	-	-
Group2	W34/70	3,947,713,936	915,990,175	824,153,049	-	-
	CBS1483	2,388,387,396	-	-	-	-
	CBS1484	2,559,669,269	-	-	-	-
	CBS2156	2,493,813,010	-	-	-	-
	CBS5832	3,113,287,910	-	-	-	-
***S. bayanus* strains**						
NBRC1948		1,343,621,555	560,999,165	545,686,503	-	-
CBS380		3,778,865,952	558,089,990	531,198,248	-	-
***S. uvarum* strain**						
CBS7001		1,481,062,097	-	-	-	-

各株で作成したライブラリとシークエンス量 (bp)

2.3 32-mer 頻度分析

2.3.1 方法

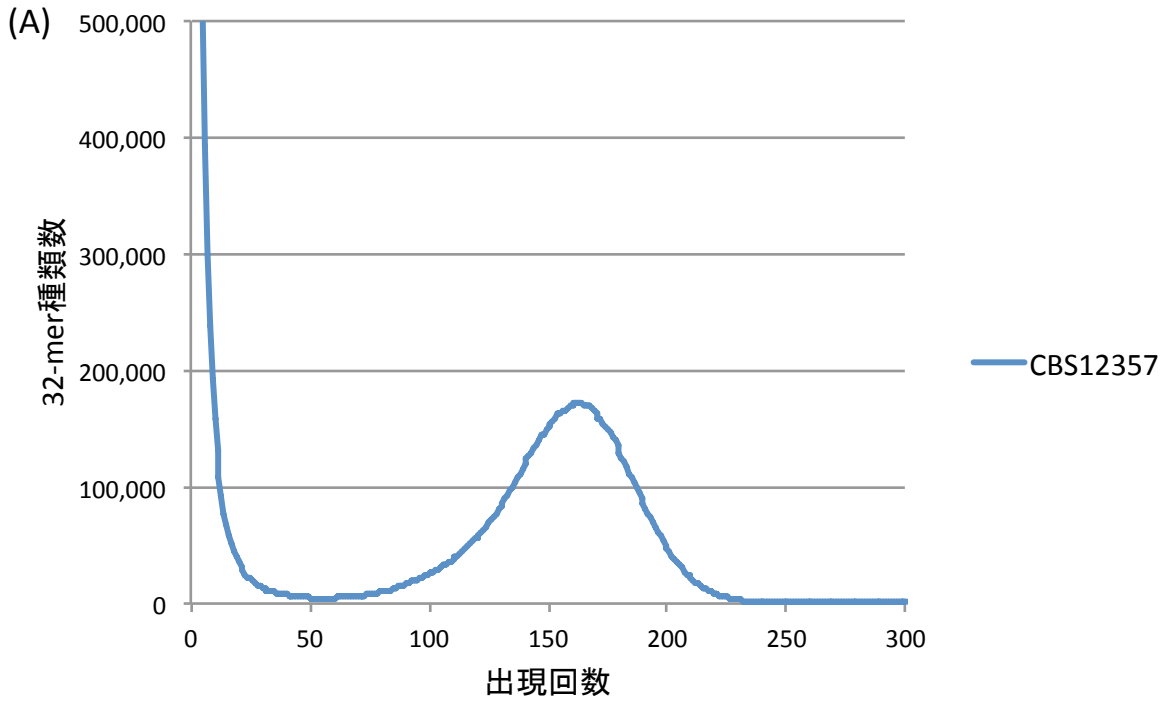
詳細な解析の前に，シークエンスデータの *k*-mer 頻度分布を用いた解析によりサンプルのおおまかな特徴を捉えることとした．*k*-mer 頻度分布とは，長さ *k* の配列

(*k*-mer) を各リード配列から 1 bp ずつ移動しながら取得し，全シークエンスリードから得られる *k*-mer とその出現回数を集計し，出現回数に対する *k*-mer の種類数を示したヒストグラムである．例えば 4 種類の文字（塩基）からなる長さ *k* (ここでは $k = 32$ とする）の文字列は，$4^{32} = 1.8 \times 10^{19}$ 通り存在し，この値はヒトのゲノムサイズ（3.0×10^9 bp）と比較してもはるかに大きい．このため，理論上は，リピート領域などを除き 32-mer はゲノム中の 1 カ所に由来し，その出現回数はシークエンスカバレッジに相当すると考えられる．1 倍体，または低ヘテロ接合性の同数体ゲノムのシークエンスデータに対して *k*-mer 頻度分布を作成した場合，シークエンスエラーを含む *k*-mer が存在するため出現回数の低い *k*-mer の種類数が多く観察されるが，平均のシークエンスカバレッジにピークを一つ持つ分布が得られる．*k*-mer 頻度分布から複数のピークが観察される場合，そのサンプルは高ヘテロ接合性ゲノムあるいは異数体ゲノムである可能性が示唆される．高ヘテロ接合性ゲノムの場合，相同染色体間の変異が存在するローカスからは 2 種類の *k*-mer が得られ，それらの出現回数はシークエンスカバレッジの半分程度になるため, *k*-mer 頻度分布は平均のシークエンスカバレッジ（相同染色体間で差異がない領域由来）とその半分の値（相同染色体間で変異が存在する領域由来）の 2 つをピークに持つ．異数体ゲノムの場合は，倍数性の違いがシークエンス量に反映されるため，平均のシークエンスカバレッジ以外にピークを持つ分布が得られることが予想される．

2.3.2 結果

S. eubayanus CBS12357 及び *S. pastorianus* 10 株の paired-end リードより得られた 32-mer 頻度分布を図 2-1 に示した. *S. eubayanus* CBS12357 の頻度分布は（図 2-1 A），出現回数 160 付近にピークを一つ持つことから，低ヘテロ接合性ゲノムであると考えられる．一方，*S. pastorianus* の 32-mer 頻度分布は．シークエンスを実施した 10 株すべてで複数のピークを示した．Group 1 に属する 5 株の 32-mer 頻度分布は 3 つのピークを持ち（図 2-1 B），一番左にあるピークの出現回数の 2 倍，3 倍の値に他のピークが存在していることから，各ピークは左から順に 1 倍体，2 倍体，3 倍体の染色体に由来するシークエンスリードから得られた 32-mer によって構成されていると考えられる．この 32-mer 頻度分布から Group 1 に属する 5 株は，1 倍体から 3 倍体を含む異数体であることが予想される．また，各ピークの高さ（32-mer の種類数）の違いは，それぞれの株の構成している倍数体の割合の違いを示している．*S. pastorianus* Group 2 に属する 5 株においても同様に複数のピークが存在し（図 2-1 C），異数性によるものであると考えられる．詳細に見ると，CBS1483 株と CBS2156 株は 1 倍体と 2 倍体に由来する 2 つのピークが同程度であるのに対して，他の 3 株（W34/70 株，

CBS1484 株, CBS5832 株）は 2 倍体に由来するピークが他のピークよりも高かった．この 32-mer 頻度分布から Group 2 に属する株も Group 1 に属する株と同様に異数性があり，主に 1 倍体と 2 倍体の染色体によって構成されていると予想される．*S. eubayanus* CBS12357 の 32-mer 頻度分布において異数性は観察されていないことからも，少なくとも祖先種 *S. eubayanus* は同数性ゲノムである可能性が高く，*S. pastorianus* の異数性は異種交配後のゲノム再構成における染色体の欠失・重複に起因すると考えられる．すべての株で一番左に存在するピークが出現頻度 30 から 70 の間に位置することから，各株の 1 倍体の染色体のシークエンスカバレッジが 30 以上あると推定され，*S. pastorianus* のシークエンス量は十分であると判断した．



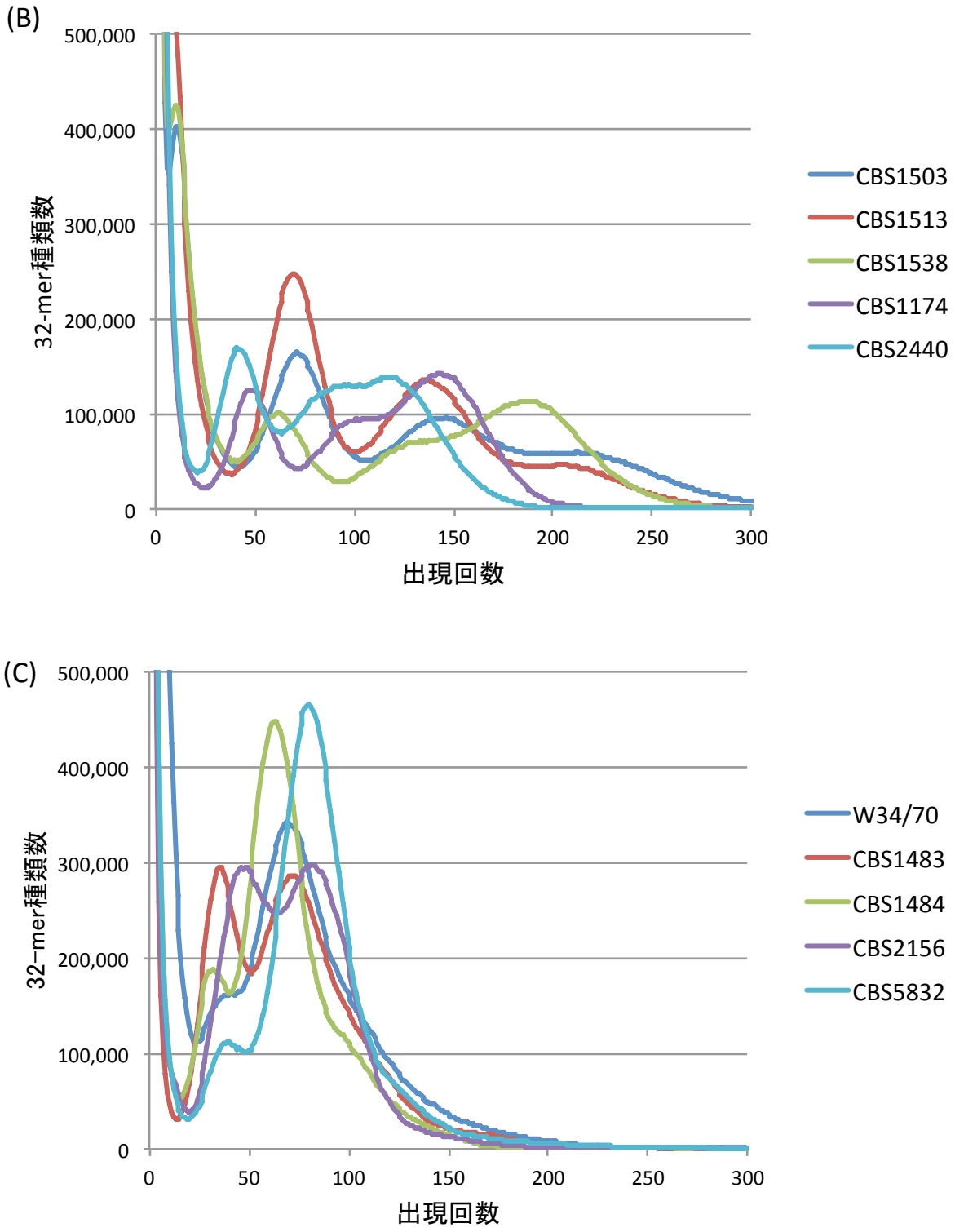


図 2-1 32-mer 頻度分布

Paired-end リードから算出した 32-mer の頻度分布．(A) *S. eubayanus* CBS12357 (B) *S. pastorianus* Group 1 に属する 5 株 (C) *S. pastorianus* Group 2 に属する 5 株

第3章　*de novo* アセンブリに基づいた *S. pastorianus* 及び祖先種 *S. eubayanus* ゲノムの決定

3.1　方法

はじめに，*S. pastorianus* 4 株（CBS1503, CBS1513, CBS1538, W34/70）及び *S. eubayanus* CBS12357 のゲノム決定を実施した．使用したシークエンスデータは，表 2-2 に示した通りであり，Platanus v.1.2.4 を用いて *de novo* アセンブルを実施した (Kajitani et al. 2014)．　Platanus によるアセンブリは以下の 3 つの工程に分かれている．

- Contig assembly
  de Bruijn グラフと呼ばれるデータ構造を用いてシークエンスデータから構築された配列（contig 配列）を出力する．
- Scaffolding
  ペアリードのマッピング結果から contig 配列のレイアウトを決定し，contig 配列を繋ぎ合わせた配列（scaffold 配列）を出力する
- Gap closing
  scaffold 配列中に存在する配列未決定領域の配列を構築する

Contig assembly では paired-end ライブラリのみを用いて配列を構築し, scaffolding と gap closing では mate-pair ライブラリを含む全ライブラリを使用した. Scaffolding の前に，mate-pair リードを contig 配列にマッピングし，マップされたポジションから PCR 重複由来と考えられるリードの除去を行なった．

S. eubayanus CBS12357 のドラフトゲノムを以降の解析に用いるために，染色体レベルでの配列の構築を目指した．アセンブル後に得られた配列を *S. cerevisiae* S288c 完全ゲノムと blastn で相同性検索をし（Altschul et al. 1997），各配列を染色体に対応させた．*S. cerevisiae* と *S. eubayanus* との間に存在する既知の染色体間転座（*S. cerevisiae* における染色体 II-染色体 IV，染色体 VIII-染色体 XV）が存在するため，その 2 箇所についてはアセンブル結果である *S. eubayanus* の配列を維持させることとした．

S. pastorianus CBS1503，CBS1513，CBS1538，W34/70 のアセンブリも *S. eubayanus* CBS12357 と同様の手順で実施した．*S. pastorianus* は 32-mer 頻度分布から異数性が確認されており，低カバレッジ領域（1 倍体）から高カバレッジ領域（多倍体）の配列構築に対応するために contig assembly 時に-n 15 (W34/70 株)，-n 20 (その他 3 株)，　scaffolding 時に-u 0 オプションをつけてアセンブルを実行した．精度の評価のために，構築した配列に対してインサートサイズ 6 kbp の mate-pair リードを

マップし，配列全体をリードがカバーしているかを確認した．最後に，ミトコンドリア配列とコンタミ配列を除去して，500 bp 以上の配列を各株のドラフトゲノムとした．ミトコンドリア配列とコンタミの特定には NCBI Bacteria DB, RefSeq viral DB と相同性検索を行い（blastn），identity 90%以上，scaffold に対するカバレッジ 50%以上のアライメントが取れた配列を除去した．なお，ミトコンドリアゲノムに関しては染色体ゲノムとは別に配列の構築を行い，解析を実施した．

3.2 結果・考察

S. eubayanus CBS12357 株のアセンブルは paired end ライブラリ及びインサートサイズが 3 kbp から 12 kbp からなる mate-pair ライブラリを使用し実施した．結果として 5 kb 以上の scaffold 配列が 17 本得られ，16 本の染色体のうち 15 本をそれぞれ 1 本の scaffold 配列として構築することができた．残り 2 本の scaffold 配列は *S. cerevisiae* S288c の染色体 XII に相当する配列であり，rRNA 領域を挟む形で分断されていた．rRNA 領域は，rRNA が非常に多コピーに並んだ領域であり，正確な繰り返し回数を推定することが不可能である．そのため 5,000 bp の gap（N 配列）を間に入れる形でその 2 本の配列を *S. cerevisiae* S288c の配列の向きを考慮して，つなぎ合わせた．これにより最終的に *S. eubayanus* CBS12357 の染色体 16 本に相当する super scaffold（合計長 11.67 Mbp）を構築することに成功した．配列中の gap（N 配列）は 30 箇所（22,137 bp）であった．配列長，本数，gap の箇所数と領域から，以降の *S. pastorianus* の解析のためのリファレンス配列として十分なドラフトゲノムを構築することができたと考えられる．最後にアセンブルされた配列を染色体に割り当てるため，*S. cerevisiae* S288c complete genome と相同性検索を行った．*S. cerevisiae* S288c と構造が異なる場合も基本的に *S. cerevisiae* の染色体にならい，染色体 II - IV，IV – II，VIII – XV，XV - VIII として *S. eubayanus* CBS12357 の染色体とした（図 3-1）．ミトコンドリア配列の構築については後述する（4.1.1 参照）．

次に *S. pastorianus* 4 株の mate-pair ライブラリのインサートサイズの分布を図 3-2，アセンブル結果を表 3-2 に示す．Group 2 に属する W34/70 株のアセンブル合計長が 22.5 Mbp となり，*S. cerevisiae* S288c（12.1 Mbp）と *S. eubayanus* CBS12357（11.7 Mbp）のゲノムサイズの合計とほぼ等しいサイズとなった．一方，Group 1 に属する 3 株のアセンブル合計長は 14.4 Mbp から 19.2 Mbp となり，*S. cerevisiae* S288c と *S. eubayanus* CBS12357 のゲノム合計長よりも小さく，これは染色体欠失によるものであると考えられる．しかし，アセンブリによって配列が構築されなかった領域が染色体欠失によるものかアセンブリの過程で消失したかを判断することは，この結果のみでは明らかにできない．また，ゲノムサイズに対して scaffold 数が多く，*S. pastorianus* のアセンブリにおいて何かしらの障壁があることが示唆される．このため，染色体間

転座の検出，倍数性の推定は祖先種（*S. cerevisiae* S288c 及び *S. eubayanus* CBS12357）のゲノムをベースとしたマッピングに基づいて解析することとした．

S. pastorianus におけるアセンブリの障害を明らかにするために，株ごとに得られた scaffold を祖先種 *S. cerevisiae* S288c と *S. eubayanus* CBS12357 にアライメントし（blastn），各 scaffold を祖先種のゲノムに対応させた．その結果，scaffold が長く構築できた領域とそうでない領域に区分されることが明らかになった（図 3-3）．CBS1513 株においては scaffold が比較的長く構築されており，数本の scaffold で 1 本の染色体が構成されている．しかし，CBS1503 株では 1 本または数本の scaffold で構成される染色体がある一方で，数十本の短い scaffold に配列が断片化されている染色体が確認された．CBS1538 株や W34/70 株においても，他の領域と比較して配列が極端に断片化されている箇所が見られた．この結果から，異数性によるシークエンスカバレッジのばらつきがアセンブリの障害となっている可能性が示唆された．今回は 1 倍体領域（低カバレッジ領域）の配列構築を重視したパラメータでアセンブリを実行したため，多倍体の領域（高カバレッジ領域）において配列が断片化していると考えられる．高カバレッジ領域ではエラー由来の 32-mer の出現回数が相対的に多くなるため，部分的に存在する高カバレッジ領域においてエラー由来の配列の除去が十分に行われず，エラー由来の配列との分岐で配列が分断されて出力されたと推察される．それぞれのカバレッジ領域に対応したパラメータでアセンブルすることで，異数性によるアセンブリ結果の改善が期待される．異質倍数体ゲノムでは，同質体のゲノムや同数性のゲノムと比較してアセンブルの障壁が多く存在し，複数のプロセスを経てゲノム構築していく必要があると考えられる．

表 3-1 *S. eubayanus* CBS12357 のドラフトゲノム

染色体	配列長 (bp)
chr. I	196,426
chr. II-IV	1,269,399
chr. III	306,932
chr. IV-II	984,473
chr. V	580,751
chr. VI	257,810
chr. VII	1,050,894
chr. VIII-XV	833,488
chr. IX	402,821
chr. X	746,565
chr. XI	635,207
chr. XII	1,035,605
chr. XIII	957,126
chr. XIV	766,570
chr. XV-VIII	744,082
chr. XVI	903,844
合計	**11,671,993**

染色体番号は *S. cerevisiae* S288c に基づいている.

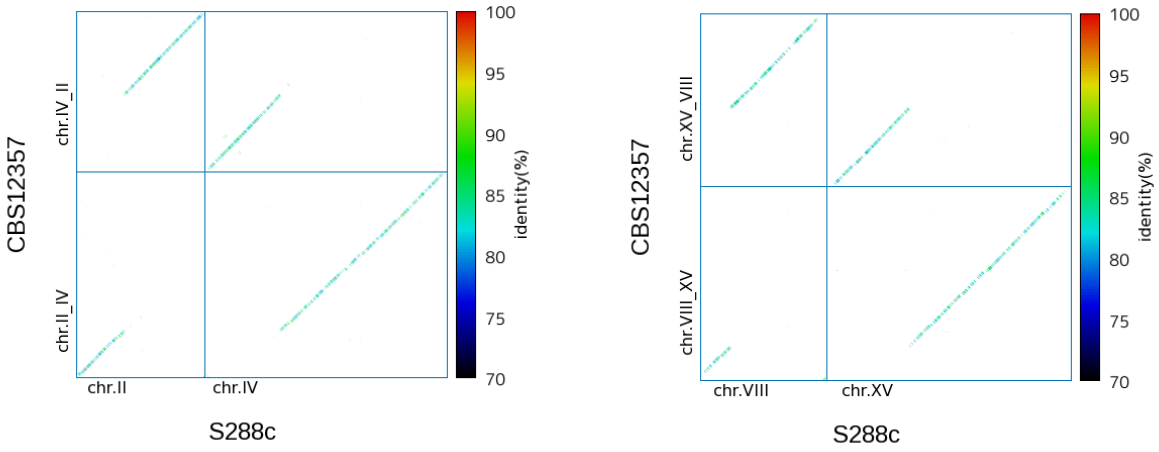
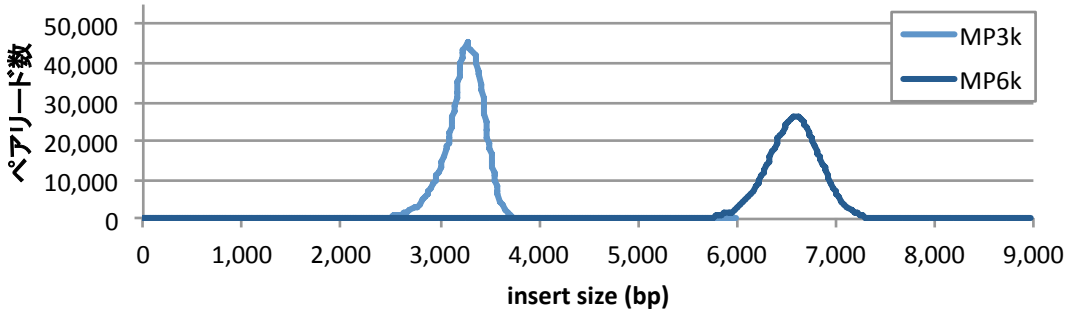


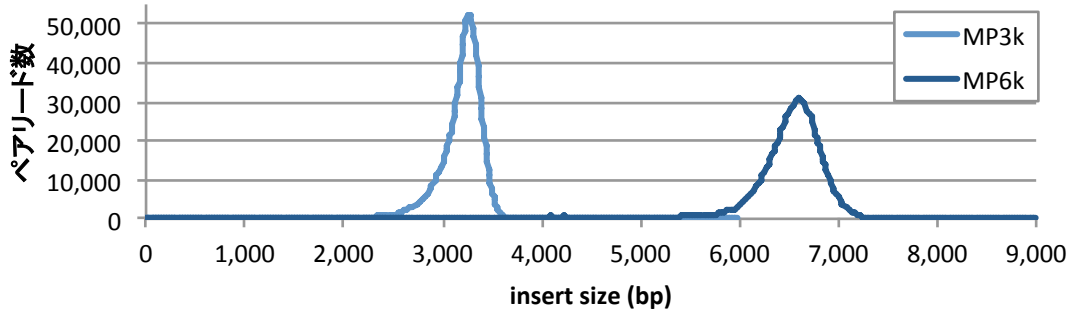
図 3-1 *S. cerevisiae* と *S. eubayanus* 間に存在する染色体間転座

左：*S. cerevisiae* S288c 染色体 II と IV とそれに対応する *S. eubayanus* CBS12357 の染色体 2 本のアライメント図. 右：S288c 株の染色体 VIII と XV とそれに対応する CBS12357 株の染色体 2 本のアライメント図.

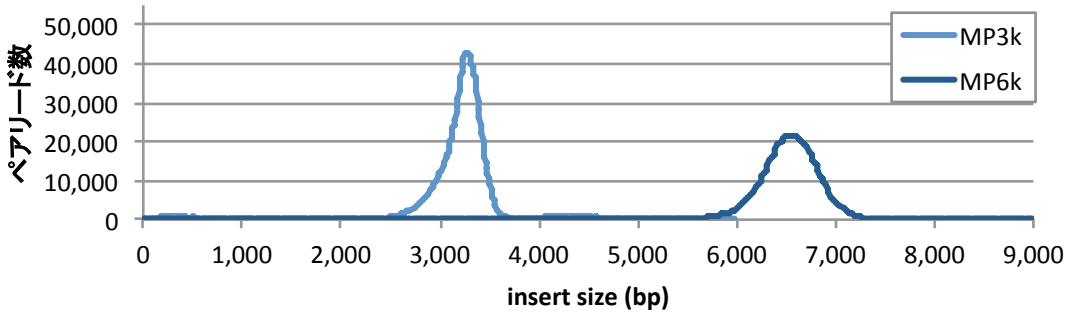
CBS1503



CBS1513



CBS1538



W34/70

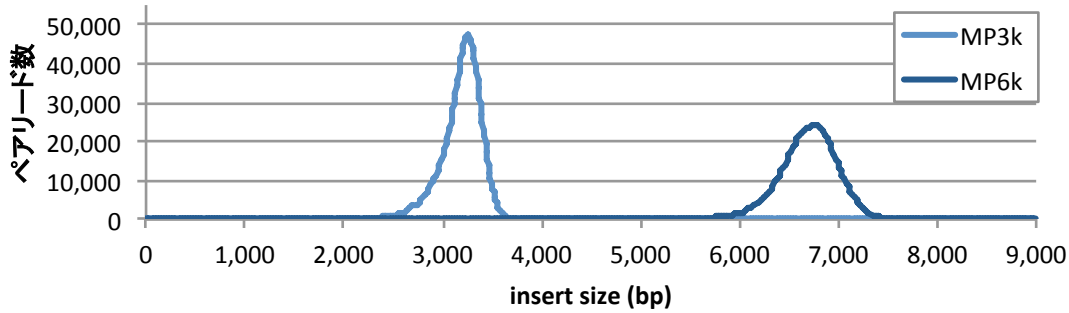


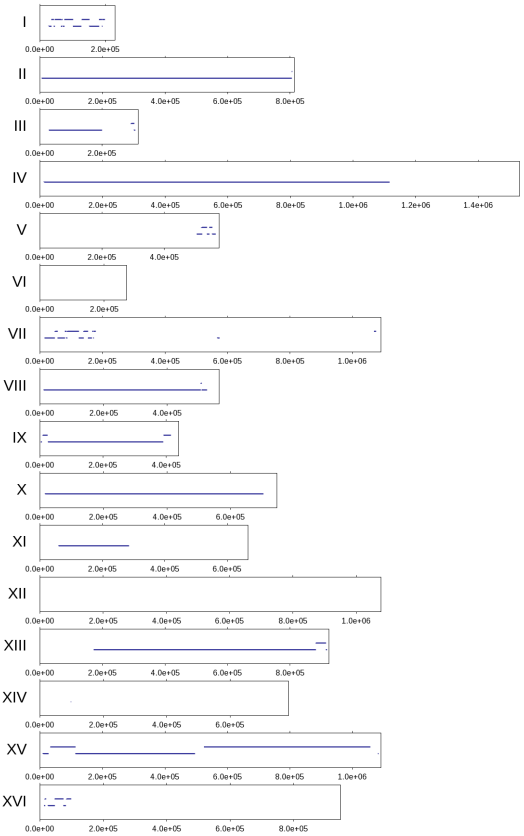
図 3-2 mate-pair ライブラリのインサートサイズの分布

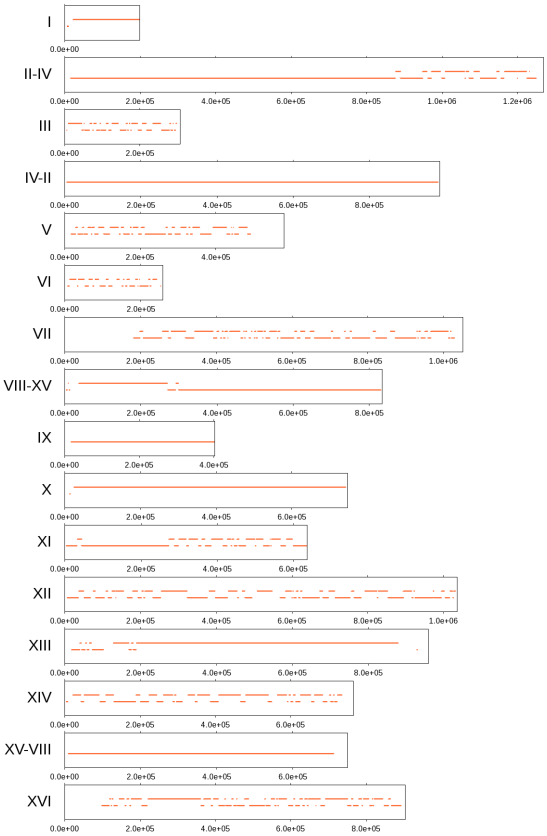
表 3-2 *S. pastorianus* のアセンブル結果

	株名	合計長 (bp)	本数	N50長 (bp)	N50 #	最小長 (bp)	最大長 (bp)
Group1	CBS1503	17,195,167	631	484,478	12	500	1,060,739
	CBS1513	19,248,212	178	644,406	12	502	1,050,489
	CBS1538	14,404,124	277	428,791	13	501	760,567
Group 2	W34/70	22,500,926	495	723,289	13	500	1,455,873

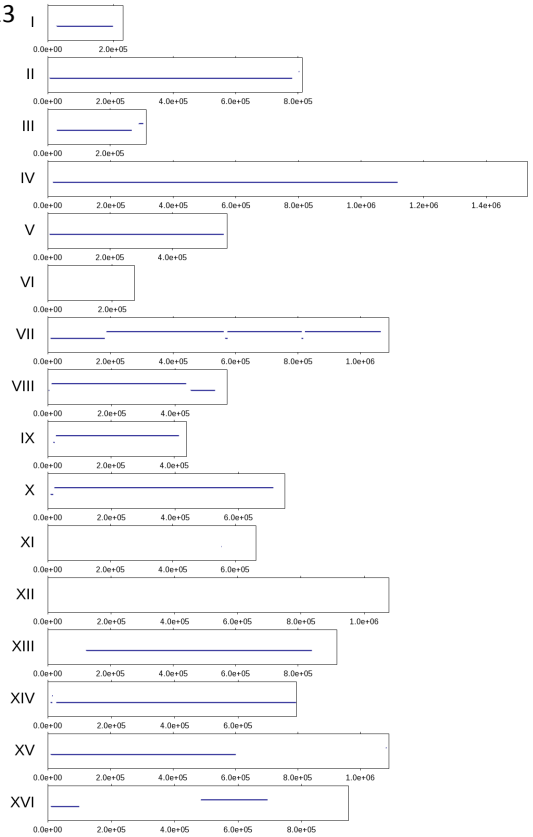
N50 長（bp）はアセンブル結果の scaffold 長を長い方から順に足し合わせた時に，合計長の 50%を越えた時の scaffold 長を意味し，N50#は N50 長に達した時の scaffold 数を表している．

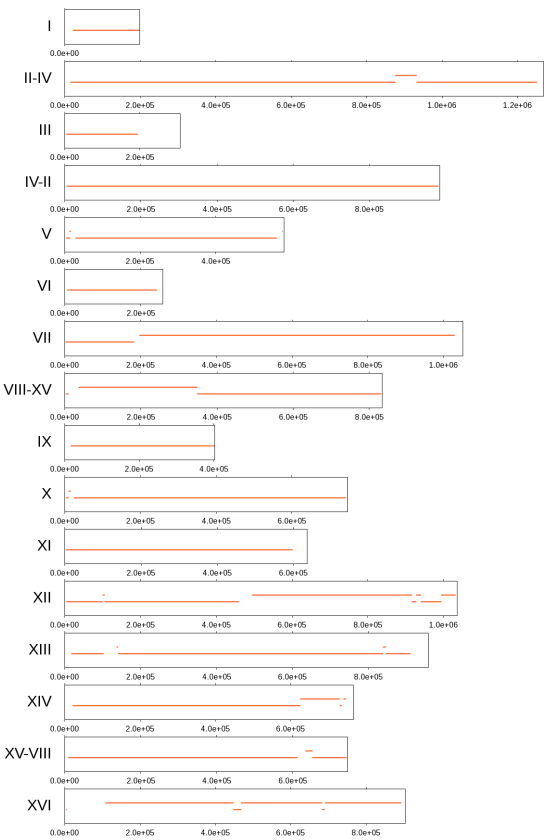
CBS1503





CBS1513





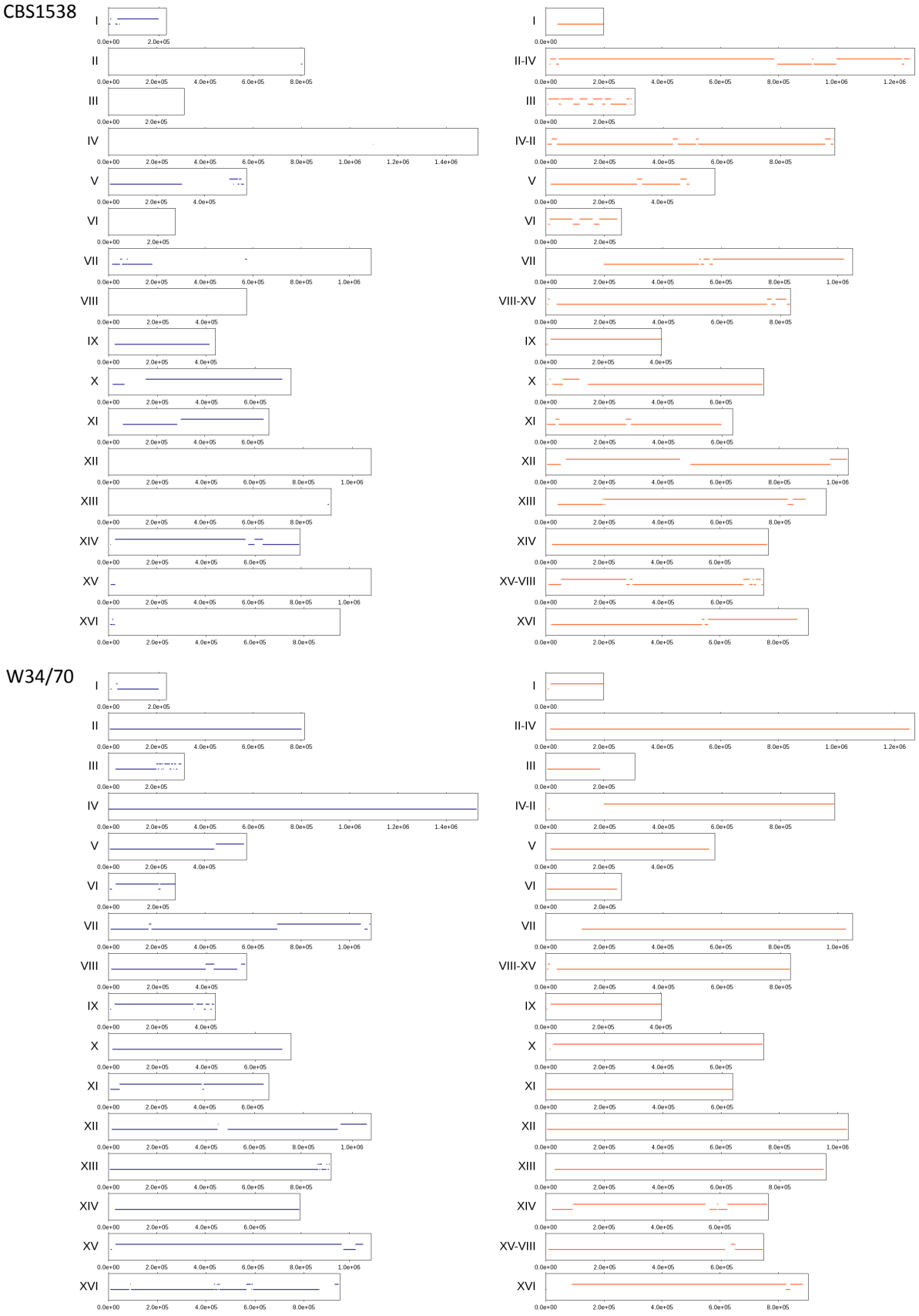


図 3-3 *S. pastorianus* の scaffold と祖先種 2 種の染色体との対応関係

S. pastorianus 4 株（CBS1503, CBS1513, CBS1538, W34/70）の scaffold 配列

を祖先種 *S. cerevisiae* S288c と *S. eubayanus* CBS12357 にアライメントした結果の模式図．祖先種の染色体に対してアライメントされた領域と scaffold の長さを線で表している．Sc type 由来の scaffold を青（左），Se type 由来の scaffold をオレンジ（右）で示し，1 本の線が 1 本の scaffold に対応している．

3.3 まとめ

本章において，*S. pastorianus* 4 株及び祖先種 *S. eubayanus* の配列決定を実施した．*S. eubayanus* CBS12357 においては染色体の本数と同数の 16 本の配列（合計長 11.67 Mbp）を構築することに成功した．一方，*S. pastorianus* 4 株のアセンブリ結果からは配列が断片化している領域が存在することが確認され，異種交配後の倍数性の変化及び染色体間転座がアセンブリの障害となっている可能性が示唆された．これらの結果を踏まえ，以降では，祖先種 2 種のゲノム配列を利用した *S. pastorianus* 10 株の比較解析を実施することとした．

第4章　*S. pastorianus* ゲノムデータの再解析

4.1　ミトコンドリアの配列決定，系統解析

4.1.1　方法

S. pastorianus のミトコンドリアの由来を明らかにするために，*S. pastorianus* 及びその近縁種を含む合計 4 種 9 株のミトコンドリアゲノム配列の決定と，*S. cerevisiae* S288c を加えた 5 種 10 株における系統解析を実施した（表 4-1）．ミトコンドリアゲノム配列は染色体よりもコピー数が多いため，アセンブリは出現頻度の高い 32-mer だけを利用した．具体的には，32-mer の頻度分布のピークより 2 倍の値を初期の *k*-mer カットオフ値（-n）に設定して Platanus assembly を実行し，その後，すべての paired-end リードを用いて scaffolding, gap closing を行なった．scaffold 配列のうち *S. cerevisiae* S288c ミトコンドリア配列とアライメントが得られた配列をミトコンドリアのドラフトゲノムとした．さらに，ドラフトゲノムの環状化を以下の手順で行った．ドラフトゲノムに対して paired-end リードをマップし，全ペアリード中から，片方のリードが配列の両端 600 bp (paired-end のインサートサイズ) 以内にマップされ，他方がマップされなかったペアのみを回収し，そのリードを使ってアセンブリを実施し，構築した配列によってドラフトゲノムの両端を伸長させた（図 4-1 A）．具体的には伸長させた配列が実際にミトコンドリアのドラフトゲノムの両端と 100%の相同性で一致することを blat (Kent 2002) によるアライメント結果から確認した．そして，環状化に不足している配列の挿入または冗長な配列の除去を行い，最終的に得られた配列をミトコンドリアの完全ゲノムとした（図 4-1 B）．

ORF の配列に基づいて 5 種 10 株のミトコンドリアの系統解析を行なった．構築したミトコンドリア配列から ORF を取り出すために．*S. cerevisiae* S288c のミトコンドリアの ORF のタンパク配列と blastx で相同性検索を行なった．すべての株において得られた ORF をつなぎ合わせた配列を株ごとに用意し，マルチプルアライメントを実行した（使用ツール ClustalW2）. マルチプルアライメントの結果から gap (InDel) を除外した結果を用いて系統距離を最尤法により計算し（使用ツール phyml –b 1000），樹形を FigTree で作図した．

表 4-1 ミトコンドリアの配列決定，系統解析に使用した株

種名		株名	
S. pastorianus	Group 1	CBS1503	*S. cerevisiae* x *S. eubayanus*
	Group 1	CBS1513	
	Group 1	CBS1538	
	Group 2	W34/70	
S. bayanus		NBRC1948	*S. eubayanus* x *S. uvarum*
		CBS380	
S. eubayanus		CBS12357	
		BaiFY1	
S. uvarum		CBS7001	
S. cerevisiae		S288c	SGDより配列を取得

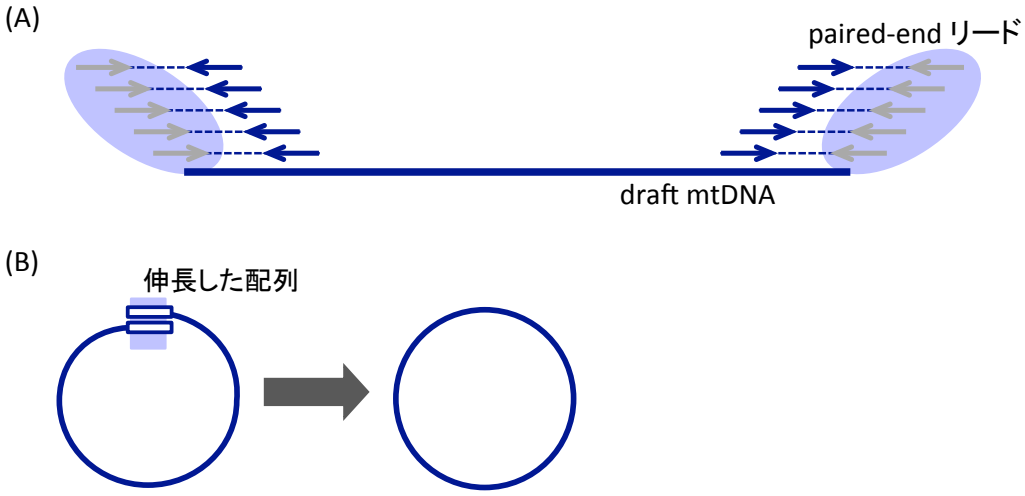


図 4-1 ミトコンドリアのドラフトゲノムの環状化

(A) アセンブル結果から得られたドラフトのミトコンドリア配列に paired-end リードをマップし，配列の両端にマップされたリードのみでアセンブルを行い，両端の配列を伸張させる．(B) 両端の配列の相同性を調べ，両端が重なり合う部分を除去して，環状化させた配列をミトコンドリア配列とした．

4.1.2 結果・考察

各株のミトコンドリア配列のアセンブル時に用いたパラメータとアセンブル結果を表 4-2 に示した．*S. pastorianus* W34/70 のミトコンドリア配列はすでに決定しており (Nakao et al. 2009)，その配列長が 69 kbp であることを考え，アセンブル結果のうち

S. eubayanus CBS12357 以外の 8 株において最大長の scaffold 配列はミトコンドリア配列のほぼ全長に相当すると判断した．*S. eubayanus* CBS12357 は，-n 250 のアセンブル結果ではミトコンドリア配列の全長を構築できなかったため，-n の値をデフォルトに設定したアセンブリ結果からミトコンドリア配列由来の scaffold を取り出し，2 つのアセンブリ結果を統合し 1 本の scaffold 配列を構築した．各株で得られた scaffold 配列と *S. cerevisiae* S288c と blastn で相同性検索を行ない，すべての株で構築された最大長の配列がミトコンドリア配列であることを確認した．環状化したミトコンドリア配列の配列長と，GC 含量を表 4-3 に示した．*S. pastorianus* 4 株のミトコンドリアの配列長は約 69 kbp で，*S. cerevisiae* S288c のミトコンドリア配列（85,779 bp）に比べて小さく，*S. eubayanus* CBS12357 (64,290 bp)，BaiFY1 (67,205 bp) より大きいことが分かった．

S. pastorianus W34/70 及び *S. eubayanus* CBS12357，BaiFY1 の 3 株のミトコンドリア配列全体の相同性を，図 4-2 に示した．*S. pastorianus* W34/70 と *S. eubayanus* BaiFY1 はゲノム全体にわたって 95%以上の相同性があった．しかし，*S. eubayanus* CBS12357 株と BaiFY1 株の間ではシンテニーは保存されているものの，ミトコンドリア配列全体での相同性は 80～95%程度であり，核染色体の相同性 99.42 %に比べてはるかに低いことが分かった．なお，核染色体の相同性は，BaiFY1 株のシークエンスデータを CBS12357 株ドラフトゲノムにマップし，検出された SNV の数（67,599 個）から算出している．

S. eubayanus CBS12357 と BaiFY1 のミトコンドリア配列の相同性が核染色体同士の相同性を大きく下回ったこと，また CBS12357 株のみ配列構築に手作業を伴ったことから，CBS12357 株のミトコンドリア配列の精度について検証する必要があると考えた. 別のグループ（Baker et al. 2015）が構築した CBS12357 株のミトコンドリア配列と比較したところ，完全にアライメントが取れたことから（図 4-3 A），CBS12357 株と BaiFY1 株との相同性の違いが CBS12357 株のミトコンドリア配列のミスアセンブリによるものではないことが確かめられた．また，CBS12357 株のミトコンドリア配列にシークエンスリードをマップしたところ，シークエンスカバレッジのばらつきが大きいことが分かり（図 4-3 B），これがアセンブリの障害になった可能性が示唆された．*S. pastorianus* W34/70 や *S. eubayanus* BaiFY1 のミトコンドリア配列のシークエンスカバレッジは比較的均一であるのに対して（図 4-3 C, D），*S. eubayanus* CBS12357 のミトコンドリア配列のカバレッジは 100 未満の領域から 15,000 以上の領域まで存在した．これは GC 含量に起因したものでもなく，*S. eubayanus* CBS12357 株のミトコンドリア配列においてのみシークエンスカバレッジのばらつきが生じた理由については明らかにできなかった．

ミトコンドリアの ORF 配列に基づいた系統解析の結果，*S. pastorianus* のミトコン

ドリアと最も相同性の高いミトコンドリアを持つ同質体種は *S. eubayanus* BaiFY1 であることが分かった．今まで，*S. pastorianus* のミトコンドリア配列の由来となる同質体種が見つかっておらず，一部の遺伝子配列による系統解析と制限酵素断片長多型解析から *S. cerevisiae* 由来ではないこと，*S. uvarum* と *S. eubayanus* の異種交配株である *S. bayanus* NBRC1948 のミトコンドリアが *S. pastorianus* のミトコンドリア配列と相同性が高いことが知られていた（Rainieri et al. 2008）．今回の解析によって，*S. pastorianus* のミトコンドリアが *S. eubayanus* に由来し，さらにチベットで単離された BaiFY1 株との相同性がパタゴニアで単離された CBS12357 株と比較して高いことが明らかになった．この結果から *S. pastorianus* の Se type がチベットまたはその周辺に由来することも示唆された．

異種交配株 *S. bayanus* NBRC1948 と CBS380 とその祖先種 *S. eubayanus* と *S. uvarum* のミトコンドリア配列を決定したことで，*S. bayanus* NBRC1948 と CBS380 のミトコンドリアの由来も明らかにすることができた．NBRC1948 株のミトコンドリアは，*S. eubayanus* 由来であるのに対し，CBS380 株は *S. uvarum* に由来することから，*S. bayanus* はミトコンドリアの由来となる祖先種が株ごとに異なり，片方の祖先種（*S. eubayanus*）にのみ由来する *S. pastorianus* とは異なる傾向が見られた（図 4-5）．

表 4-2 使用したパラメータとアセンブル結果

種名	株名	パラメータ (-n)	合計長 (bp)	scaffold数	最小長 (bp)	最大長 (bp)	N50 (bp)
S. pastorianus	CBS1503	300	103,884	70	274	68,770	68,770
	CBS1513	300	82,352	26	139	68,998	68,998
	CBS1538	300	79,818	11	138	69,013	69,013
	W34/70	200	80,460	13	151	68,862	68,862
S. eubayanus	CBS12357	250	73,824	40	119	28,054	18,744
	BaiFY1	150	236,770	87	102	67,144	19,171
S. bayanus	NBRC1948	100	86,855	35	147	65,522	65,522
	CBS380	250	75,151	9	143	64,917	64,917
S. uvarum	CBS7001	150	74,435	3	258	64,920	64,920

-n は Platanus assembly のパラメータで，de Bruijn グラフを構築する時に使用する 32-mer の出現回数の最小値を決める値である．核ゲノムに対してコピー数の多いミトコンドリア配列を構築するために高い値を設定した．ミトコンドリア配列に相当する scaffold が得られた株の最大長は青で示した．

表 4-3 ミトコンドリア配列の長さと GC 含量

種名	株名	配列長 (bp)	GC含量 (%)
S. pastorianus	CBS1503	68,790	18.97
	CBS1513	69,294	19.41
	CBS1538	68,699	19.24
	W34/70	68,862	19.04
S. eubayanus	CBS12357	64,290	17.48
	BaiFY1	67,205	18.43
S. bayanus	CBS380	64,736	16.32
	NBRC1948	65,795	19.51
S. uvarum	CBS7001	64,920	16.28
（参考）*S. cerevisiae*	S288c	85,779	17.11

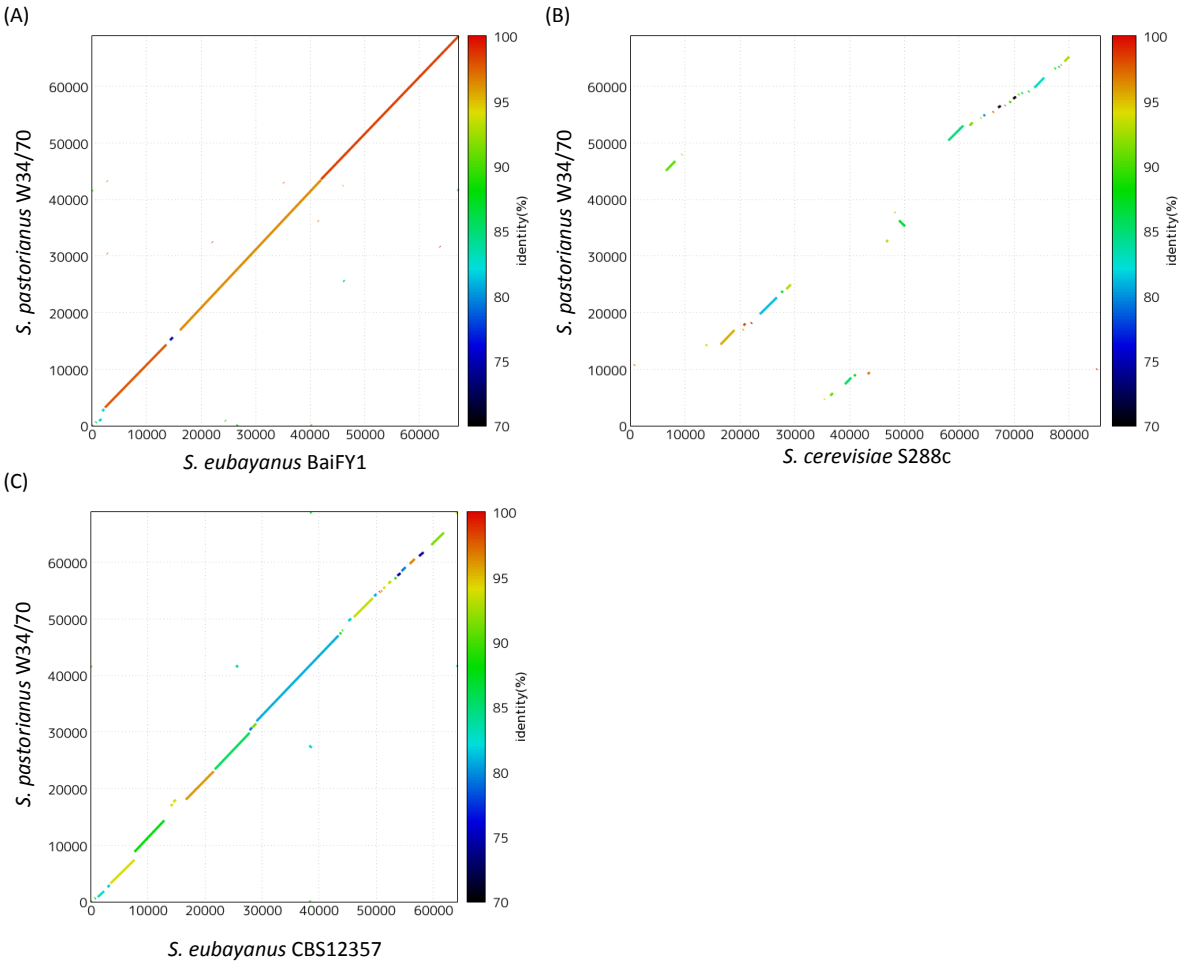


図 4-2 ミトコンドリアの相同性

S. pastorianus W34/70 に対する祖先種 *S. eubayanus* BaiFY1 (チベット由来) , CBS12357 (パタゴニア由来) , *S. cerevisiae* S288c との配列同士のアライメント図

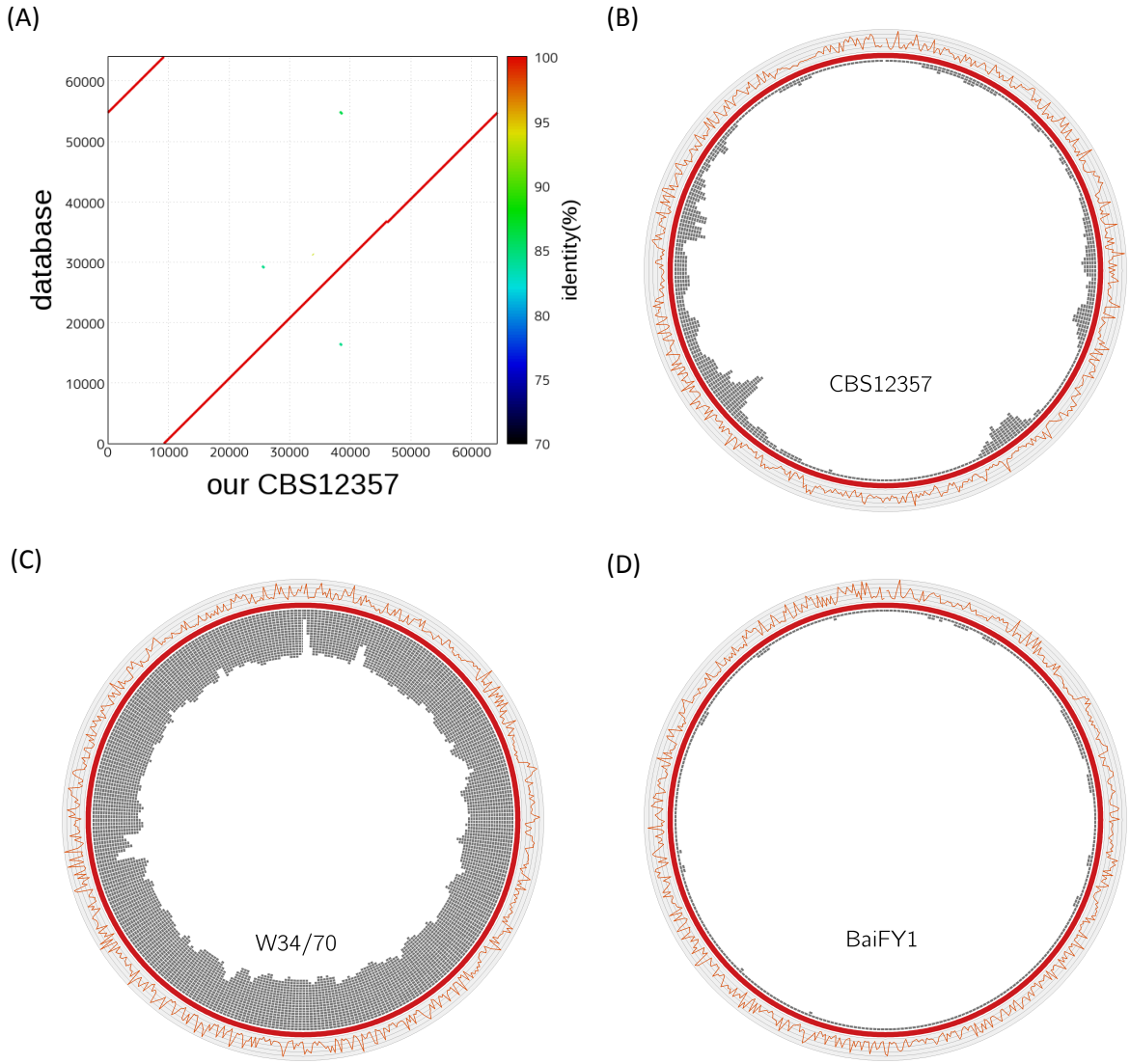


図 4-3 シークエンスカバレッジの分布

(A) データベースに登録されている *S. eubayanus* CBS12357 のミトコンドリア配列との比較. (B) – (D) 各株のミトコンドリア配列の GC 含量とシークエンスカバレッジ. 外側の折れ線グラフは 100 bp ごとの GC 含量を示している. 縦軸の範囲は 0 から 50 %. 内側のタイルプロットは 1 kbp ごとのシークエンスカバレッジの平均を示し, 1 タイルあたりのカバレッジは 1,000 となっている.

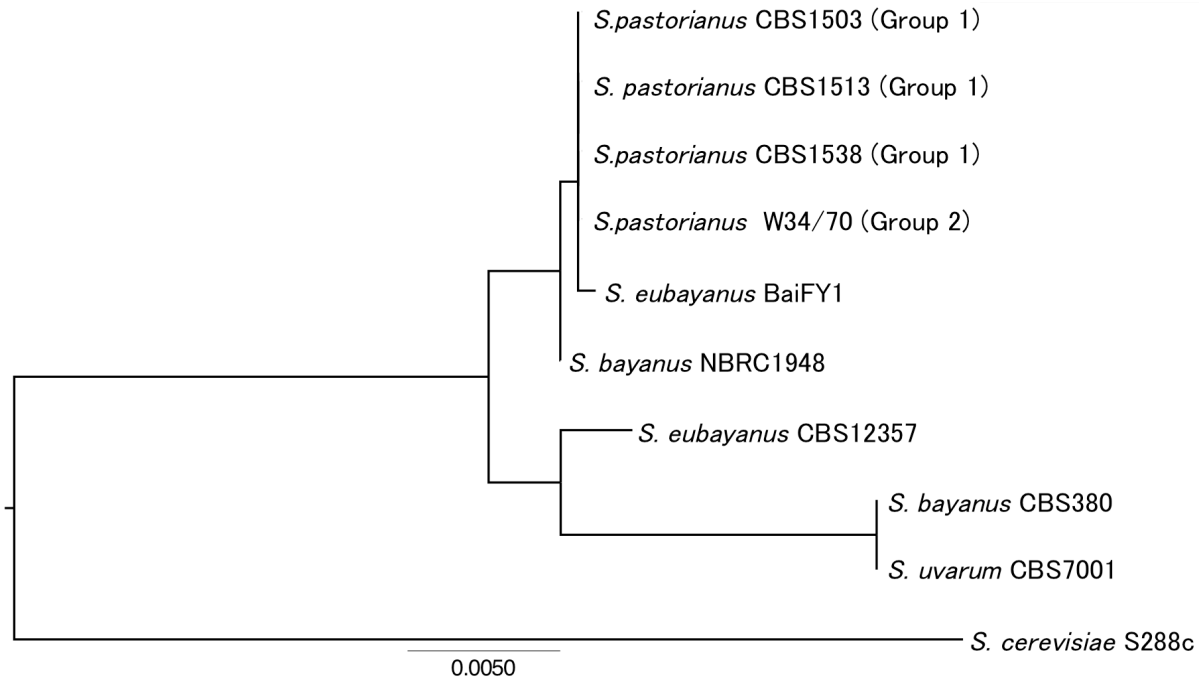


図 4-4 ミトコンドリアの系統樹

今回配列を決定した 4 種 9 株及び *S. cerevisiae* S288c のミトコンドリア配列の系統樹．異質倍数体種 *S. pastorianus*, *S. bayanus* のミトコンドリアの由来が明らかになった．

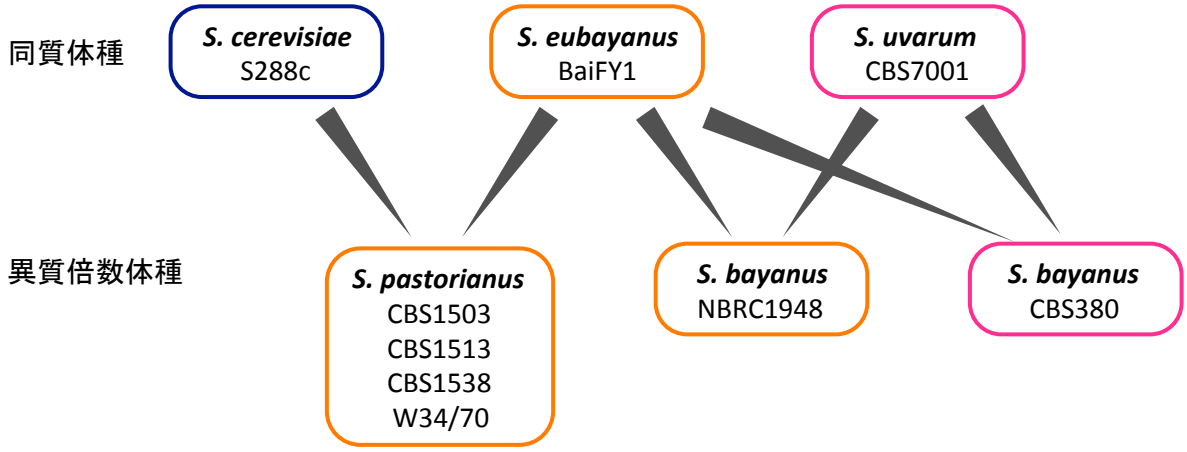


図 4-5 異種交配株のミトコンドリアの由来

異質倍数体種とその祖先種（同質体種）の関係．異質倍数体種の祖先種 2 種を線で結んでいる．異質倍数体種を囲んでいる線の色は，ミトコンドリアの由来となっている祖先種を表している．*S. pastorianus* 4 株及び *S. bayanus* NBRC1948 のミトコンドリアは *S. eubayanus* に由来し，*S. bayanus* CBS380 は *S. uvarum* 由来のミトコンドリアを持つ．

4.2 倍数性の推定

4.2.1 方法

S. pastorianus 10 株の各染色体の倍数性及びゲノムサイズの推定を以下の方法で実施した．この解析は *S. pastorianus* 株ごとの倍数性の違い，異種交配後の染色体の欠失の検出を目的としているため，*S. pastorianus* の祖先種である *S. cerevisiae* と *S. eubayanus* のゲノム配列を基盤とし，マッピングによる解析を実施した（図 4-6）．まず，*S. cerevisiae* S288c と *S. eubayanus* CBS12357 ゲノム，各株のミトコンドリア配列をリファレンスゲノムとして，各 *S. pastorianus* 株から得られた paired-end リードをリファレンスゲノムにマップした（使用ツール BWA) (Li and Durbin 2009）．マッピング結果から以下のいずれかの条件を満たすリードを除外し，ゲノム上にユニークに（一意に）マップされたペアリードのみを以降の解析に利用した．

- リファレンスゲノムとの相同性が 90%未満のリード
- ベストヒットスコア（sam format 中の AS 値）とセカンドヒットスコア（sam format 中の XS 値）との差が 5 未満のリード
- リード全長がアライメントされなかったリード

フィルタリングの後，samtools mpileup (Li et al. 2009）の結果からローカスごとにマップされたリードの本数を集計し，一倍体あたりのシークエンスカバレッジを推定した．株ごとにシークエンス量とゲノムサイズが異なるため，それぞれの株で一倍体当たりのシークエンスカバレッジを求め，シークエンス量を正規化することで株間の倍数性の比較を可能にした．window size 10 kbp, step size 1 kbp ごとのカバレッジの移動平均に対する一倍体あたりのカバレッジの比から倍数性の推定し，結果を図示した．

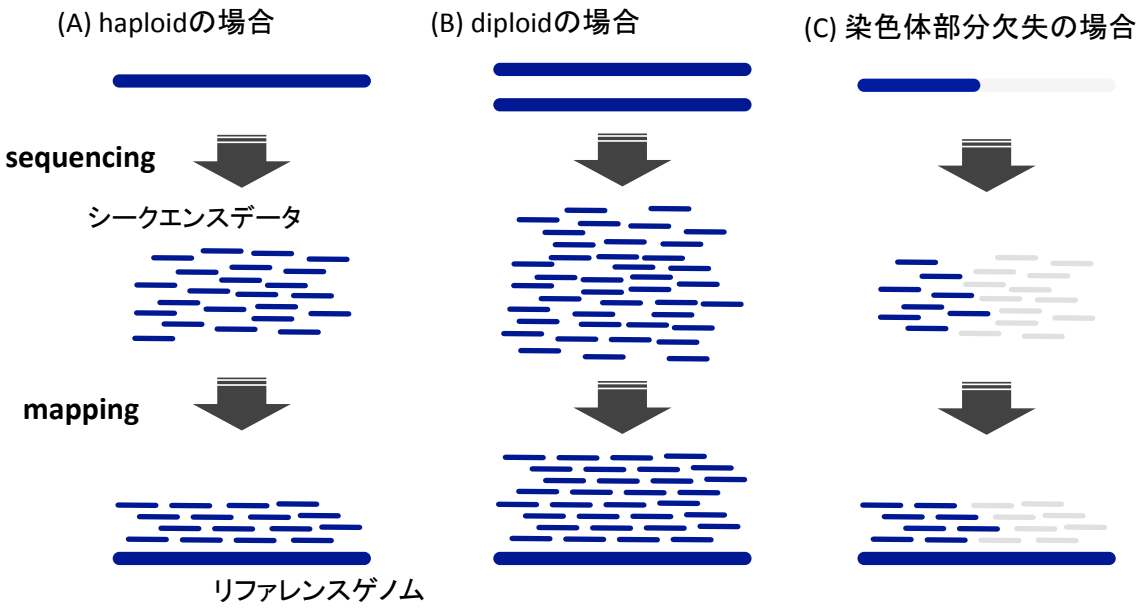


図 4-6 倍数性の推定の模式図

シークエンス量の違いから倍数性の違い（A, B），染色体欠失（C）を検出することができる．

4.2.2 結果・考察

まず，paired-end リードのリファレンスゲノムへのマッピング結果を示す．*S. pastorianus* 10 株をリファレンスゲノム（祖先種である *S. cerevisiae* S288c と *S. eubayanus* CBS12357，各株のミトコンドリア配列）にマップし，フィルタリング後のリードが張り付いた割合の内訳を図 4-7 に示した．いずれの株も 93%以上のリードをマップすることができたことから，祖先種の配列をリファレンスとしたマッピングベースの解析が可能であったと判断される．マップされなかったリードは，シークエンスエラーを多く含んだリード，複数箇所に同一スコアでマップされたリード，リファレンスゲノム中に存在しない *S. pastorianus* 特有の配列などが考えらえる．マップされたリードの内訳を見ると，Group 1 と Group 2 で Se type と Sc type が占める割合に大きく違いが見られた．Group 1 に属する株では，Se type の染色体にマップされたリード数が Sc type にマップされたリード数よりも 3〜5 倍程度多く，一方，Group 2 に属する株では，Se type よりも Sc type の方がマップされたリード数が多かった．この結果から，Group 1 と 2 で構成している祖先種（*S. cerevisiae*, *S. eubayanus*）のゲノムの割合が異なることが示唆された．

マップされたリード数（カバレッジ）とポジション数の分布を図 4-8 に示した．32-mer 頻度分布と同様に，カバレッジに複数のピークが観察され，一番左側にあるピークの値の，2 倍，3 倍の値に他のピークが存在していることから，異数性によるシー

クエンスカバレッジのばらつきが存在すると考えられる．また，Group 1 に属する株ではカバレッジの値が 0 の領域が多数存在することから，染色体欠失領域の存在が示唆された．株ごとのシークエンス量のばらつきを補正するために，各株の一倍体当たりのカバレッジ（図 4-8 中の一番左にあるピークの値）で正規化した分布図を図 4-9 に示した．すべての株のピークが整数値の箇所に現われることから，Group 1 に属する株は異数体であり. 1 倍体から 3 倍体の染色体により構成されていることが示唆された．一方，Group 2 に属する株は，ピークから 1 倍体と 2 倍体の染色体で主に構成され，3 倍体の領域は Group 1 に属する株と比較して少ないことが予想される．

シークエンスカバレッジから推定した *S. pastorianus* 10 株の倍数性を染色体毎にプロットしたものを図 4-10 に示した. Group 1 と Group 2 の間で明確な違いが見られ，Group 1 に属する株の Sc type は 1 倍体あるいは染色体の一部分または全体の欠失が観察された．具体的には，染色体 I, IX, X 及び染色体 VII の左腕は Group 1 に属する 5 株すべてで存在しているが，染色体 IV の右腕，染色体 XIII の左腕，染色体 XII はすべての株において欠失が確認された．その他の領域については，染色体の有無は株ごとに異なっていた. Group 1 に属する株の Sc type では 1 倍体あるいは欠失であるのに対して，Se type は多くの染色体で 2 倍体あるいは 3 倍体であった．一方，Group 2 に属する株では Sc type が 1 倍体あるいは 2 倍体，Se type は 2 倍体あるいは 3 倍体であった．同一染色体内で倍数性の変化が起きている領域は，Sc type と Se type 間での染色体間転座が起きている領域であると考えられる．一見すると，Sc type，Se type 内では異数体のようだが，CBS1503, CBS1513, CBS1538，及び W34/70 株の染色体の倍数性を考慮した染色体構造の模式図（図 4-11）の通り，Sc type と Se type 間の相同染色体の合計数は株ごとにおおむね一定であることが示唆された. Group 1 に属する 5 株と CBS2156 株（Group 2）は相同染色体の合計が 3 であり，CBS2156 株以外の Group 2 に属する 4 株は 4 であった．祖先種は 2 倍体または 1 倍体の同数体であると考えられ，異種交配後に片方のタイプで染色体の欠失が起きた時，その欠失を他方のタイプが補完するということを繰り返して，現在のような染色体の倍数性になったと推察される．このことから異種の相同染色体間で，染色体の合計数を一定に保つためのメカニズムが存在していると考えられる．

シークエンスリードがマップされた領域から推定される各株のゲノムサイズを表 4-4 に示した. Group 1 に属する株においてゲノムサイズは 14.1 Mbp (CBS1538 株）から 18.7 Mbp (CBS1513 株）と株ごとに異なり，これは Sc type の染色体欠失に起因していると考えられる．一方，Group 2 に属する株は染色体欠失が見られないため，ゲノムサイズは祖先種である *S. cerevisiae* と *S. eubayanus* の合計におおむね等しく, 株ごとの差異も小さかった．

以上述べたように，祖先種 2 種の染色体レベルの配列を基盤としたハイスループッ

トDNAシークエンスデータによる倍数性の推定によって，*S. pastorianus* の倍数性についての詳細な情報を得ることができた．*S. eubayanus* が単離される以前の倍数性の推定は，*S. cerevisiae* を Sc type，*S. uvarum* を non Sc type（現在の Se type）とした array-CGH が中心であったため，特に non Sc type の解像度が低く，染色体の有無以上の倍数性に関する情報を正確に得ることが困難であった．本研究においては *S. eubayanus* CBS12357 の染色体レベルでの配列を利用したことで Se type の倍数性も高い解像度での推定ができた．*S. pastorianus* 10 株の倍数性の推定結果から Group 1 と 2 において Sc type の倍数性に大きな差異があり，Group 1 に属する株において異種交配後に Sc type の染色体欠失が頻繁に起きていることが確認された．

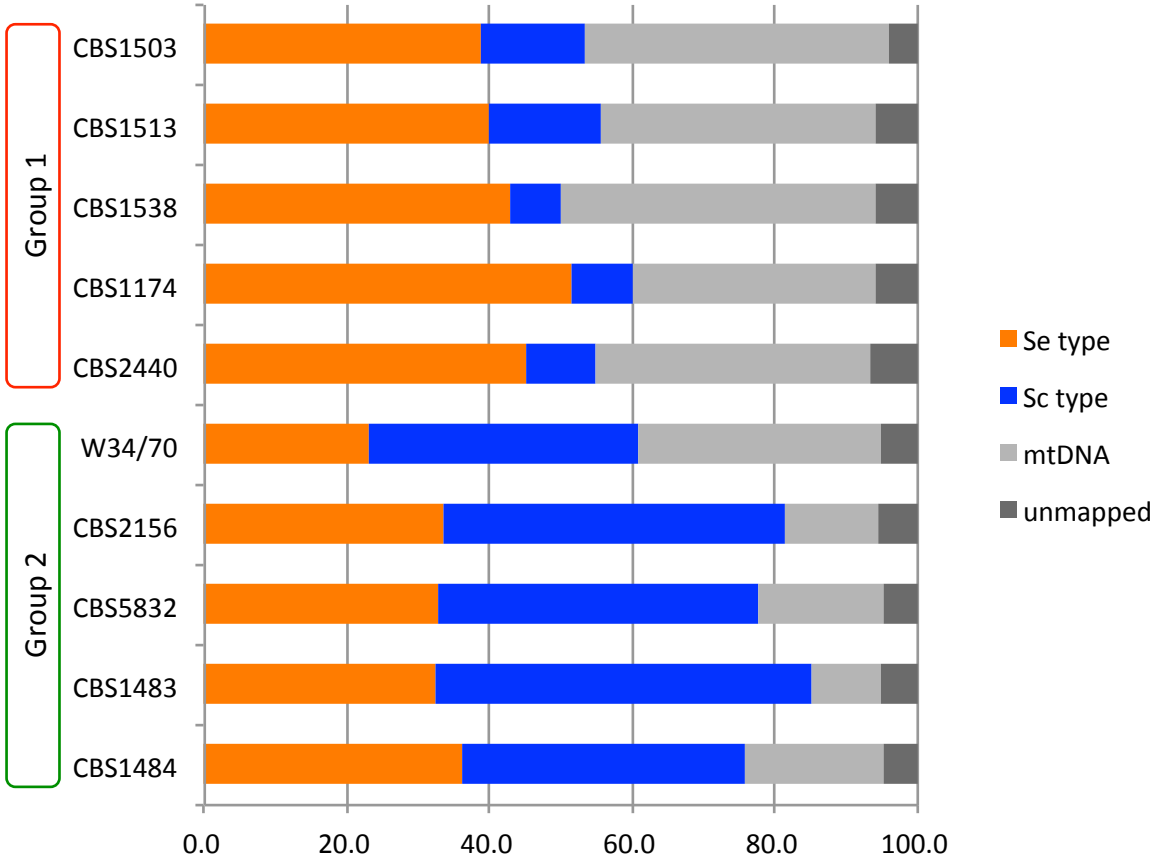


図 4-7 *S. pastorianus* 10 株のマッピング結果

S. pastorianus 10 株をリファレンスゲノム（*S. cerevisiae* S288c，*S. eubayanus* CBS12357 及び各株のミトコンドリア配列）にマップし，フィルタリングした後のリードの内訳．すべての株で 93%以上のリードがリファレンスゲノムにマップされた．

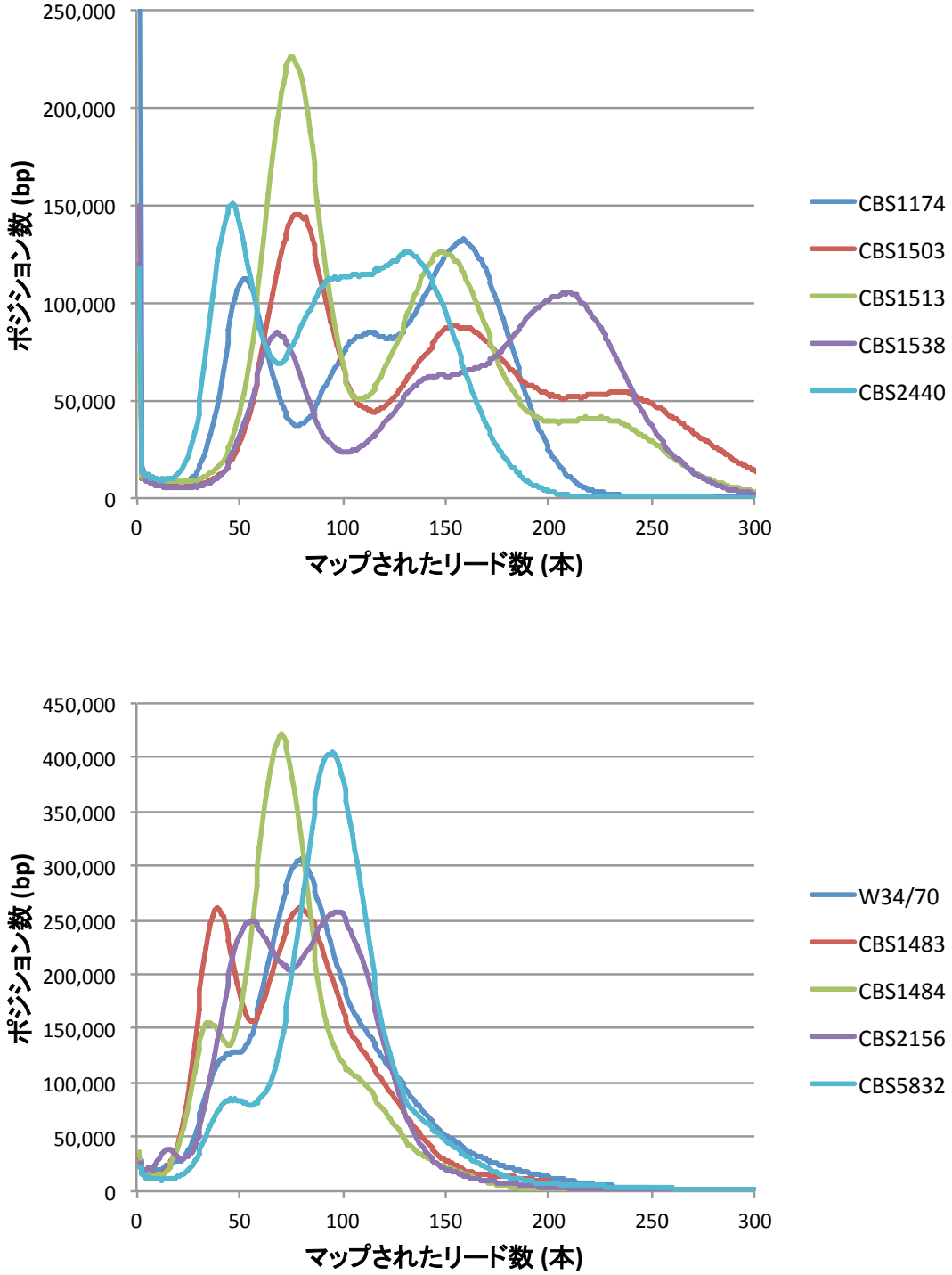
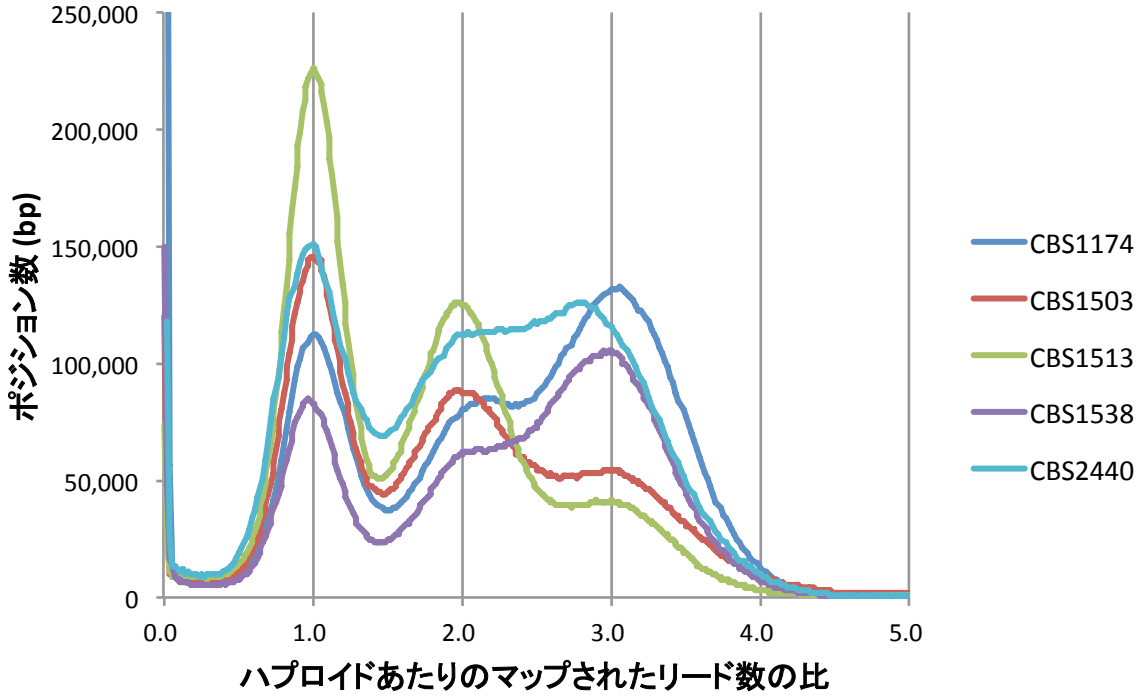


図 4-8 *S. pastorianus* のシークエンスカバレッジの分布

マッピング結果から得られたシークエンスカバレッジの分布（上図：Group 1, 下図：Group 2）．異数性により複数のピークが観察される．



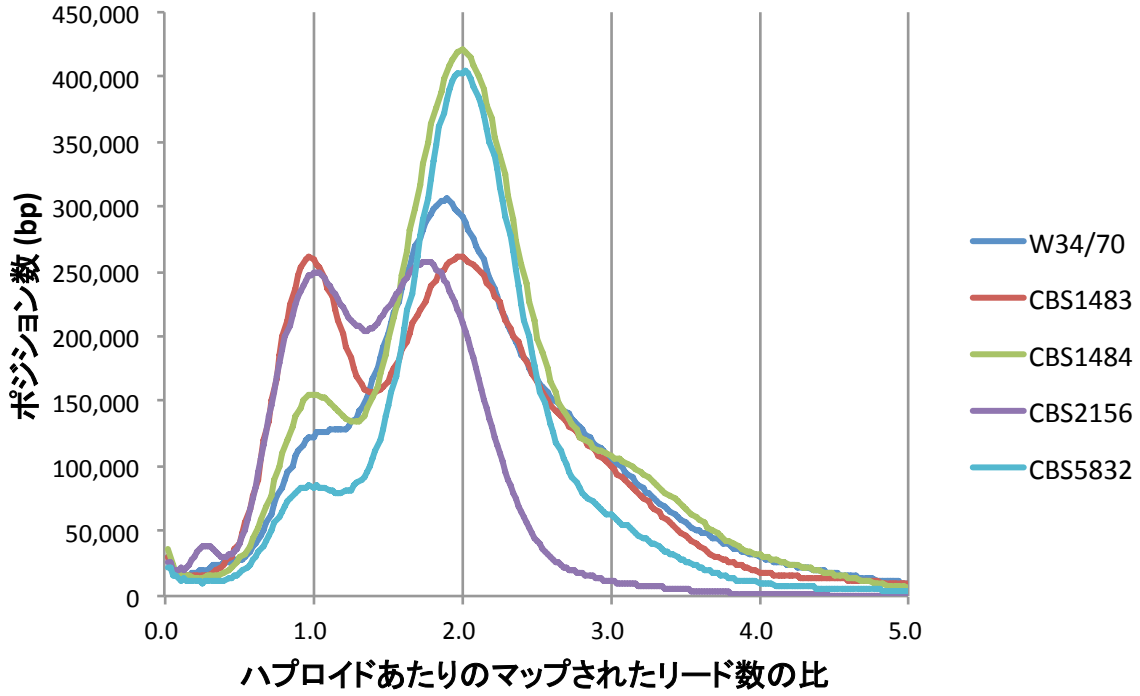


図 4-9 ハプロタイプあたりのカバレッジの分布図

1倍体あたりのシークエンスカバレッジを1としたときの *S. pastorianus* 各株のシークエンスカバレッジの分布図（上図：Group 1，下図：Group 2）.

表 4-4 *S. pastorianus* の推定ゲノムサイズ

	S. pastorianus	**Sc type (bp)**	**Se type (bp)**	**total (bp)**
Group 1	CBS1503	5,996,745	10,851,032	16,847,777
	CBS1513	7,643,871	11,089,689	18,733,560
	CBS1538	3,060,643	11,007,548	14,068,191
	CBS1174	3,142,642	10,996,606	14,139,248
	CBS2440	4,100,146	10,988,045	15,088,191
Group 2	W34/70	11,119,766	10,813,878	21,933,644
	CBS1483	10,912,447	10,814,177	21,726,624
	CBS1484	10,696,441	10,832,878	21,529,319
	CBS2156	11,072,798	10,802,714	21,875,512
	CBS5832	11,116,063	10,846,338	21,962,401

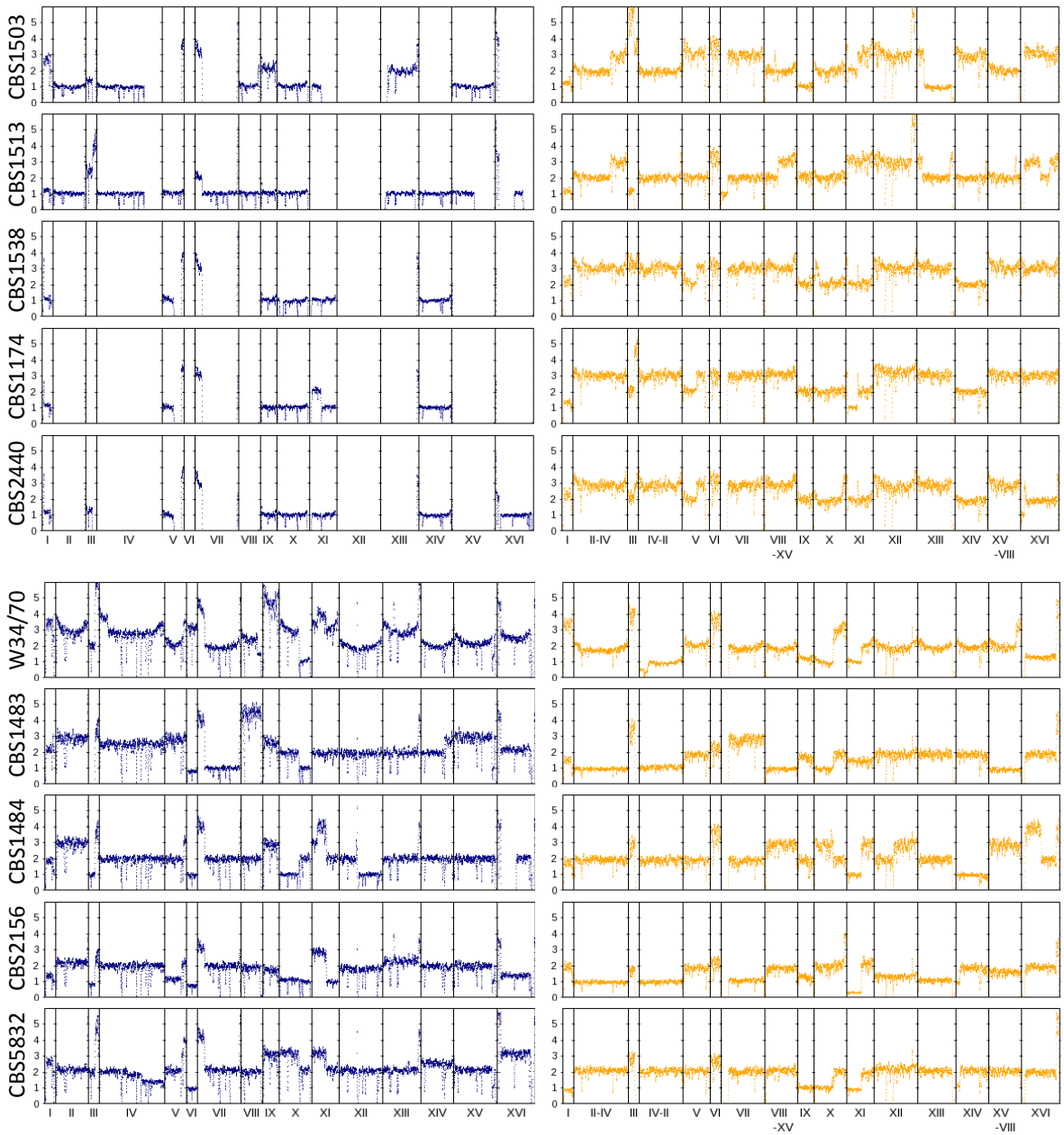
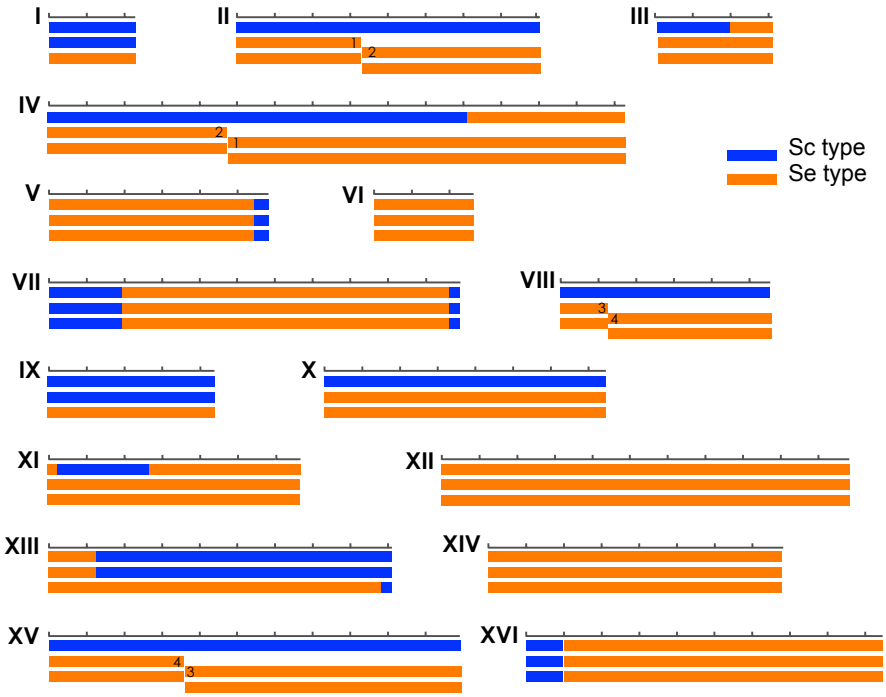


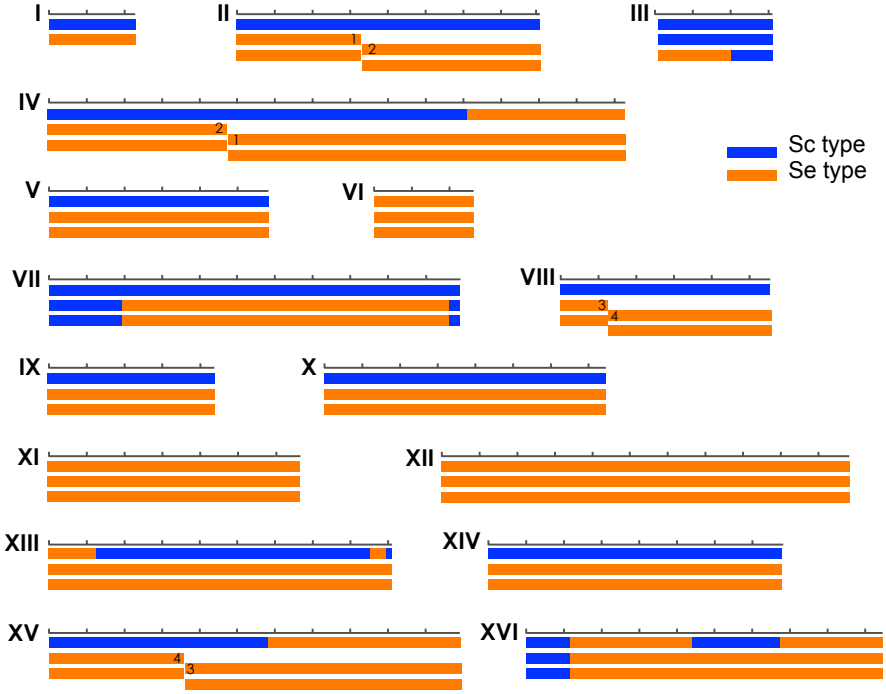
図 4-10 *S. pastorianus* 10 株の倍数性の推定結果

シークエンスカバレッジから推定した *S. pastorianus* Group 1 に属する 5 株（上）と Group 2 に属する 5 株（下）の染色体毎の倍数性．左（青色）が Sc type の染色体，右（オレンジ色）が Se type の染色体を表している．縦軸は推定される倍数性で，横軸は染色体の位置を示している．プロットが存在しない領域は，染色体欠失している領域を表している．

CBS1503



CBS1513



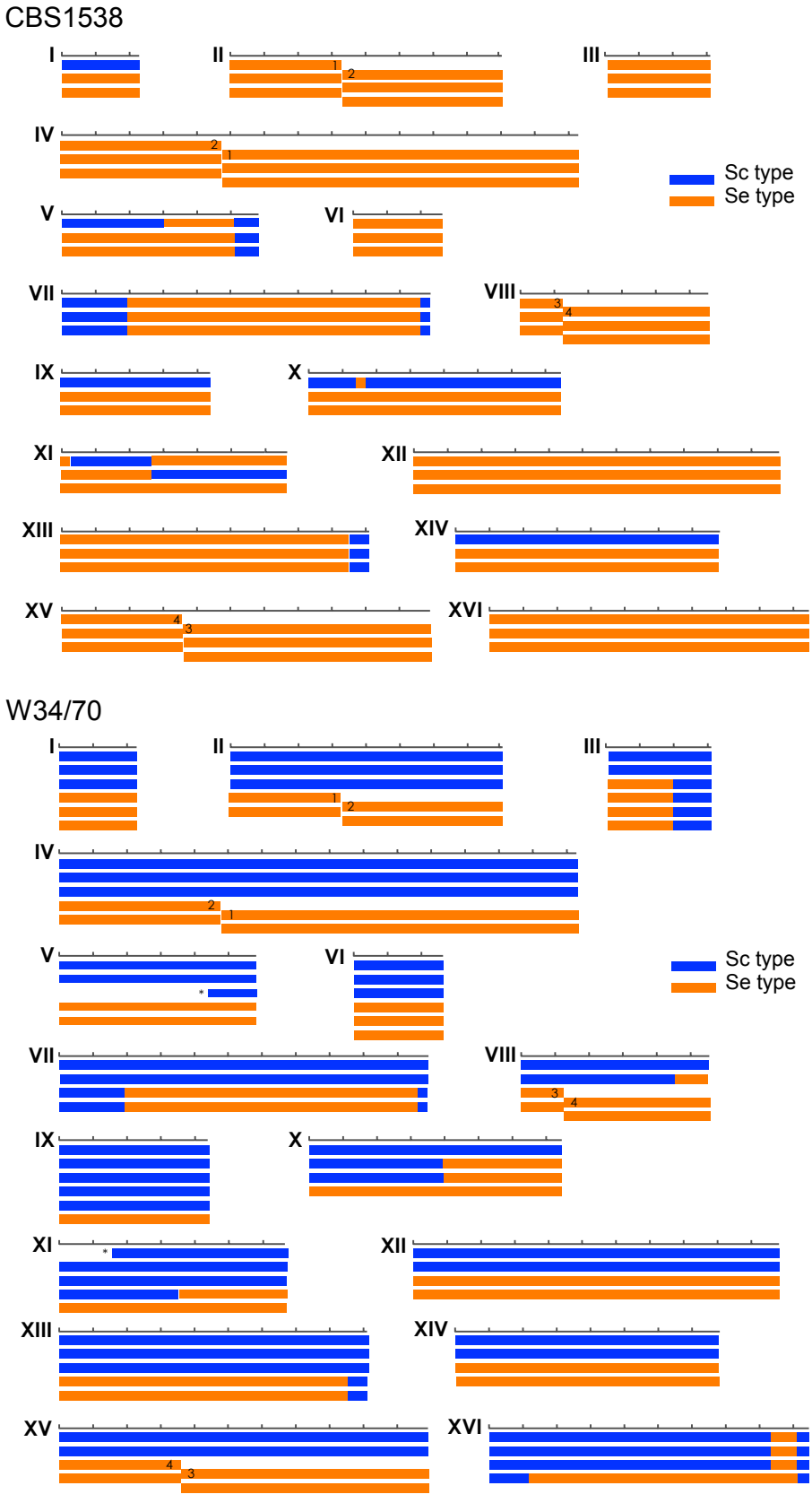


図 4-11 *S. pastorianus* 4 株の染色体構造の模式図

倍数性と染色体間転座を考慮した CBS1503 (Group 1), CBS1513 (Group 1), CBS1538 (Group 1), W34/70 (Group 2) の染色体構造の模式図．*S. cerevisiae* の染色体（I～XVI）をベースにレイアウトしている．Sc type の染色体を青，Se type の

染色体をオレンジで示している．染色体の端にある数字は，染色体 II-IV 間，染色体 VIII-XV 間に存在する Se type の染色体間転座箇所を示し，同じ数字の染色体が Se type の染色体に対応している．

4.3 染色体間転座の検出

4.3.1 方法

mate pair リードを用い, 各 *S. pastorianus* 株に存在する異種交配後の染色体間転座の検出を行なった（図 4-12）．まず，Mate-pair リード（インサートサイズ 6 kbp）をリファレンスゲノム（4.2.1 と同様）にマッピングし（使用ツール BWA），マッピング結果から相同性が 95%未満，または，ベストヒットスコアとセカンドヒットスコアとの差が 5 未満のリードを除外した．次に染色体を 10 kbp ごとのブロックに分割し，それぞれのリードがマップされた向きを考慮した上で，ペアリードが結ぶブロックとブロック間のリンクの本数を集計した．Sc type と Se type のブロック間で 100 組以上のペアリードが支持するリンクを Circos plot で図示した（Krzywinski et al. 2009）．

また, Mate-pair リードから推定された染色体間転座のブレイクポイント近傍領域のローカルアセンブリを行ない，塩基レベルでのブレイクポイント特定を試みた．染色体間転座が相同染色体の片方で起きた場合，全シークエンスリードを用いてアセンブリを実行してもブレイクポイントで配列が分岐する構造を取るため，目的としたブレイクポイント近傍の配列を構築できるとは限らない．このため，ブレイクポイントを含むリードのみを回収しアセンブルすることでブレイクポイントの配列の構築を行なった. しかし, ブレイクポイントを含むリードは *S. pastorianus* 株特有の配列であり，祖先種であるリファンレスゲノム（*S. cerevisiae* S288c，*S. eubayanus* CBS12357）に存在しないため，リード全長がマップされない．そこで，ブレイクポイント近傍でペアのうち片方のリードがマップされ，他方がマップされなかったリードがブレイクポイントを含むリードであると考え，そのようなリードを回収しアセンブルすることでブレイクポイント周辺の配列を構築した．最後に，祖先種 *S. cerevisiae* S288c，*S. eubayanus* CBS12357 の該当する配列を切り出し，アセンブリで得られた contig と clustalW2 (Li 2003) でマルチプルアライメントをし，株ごとのブレイクポイントを確認した．

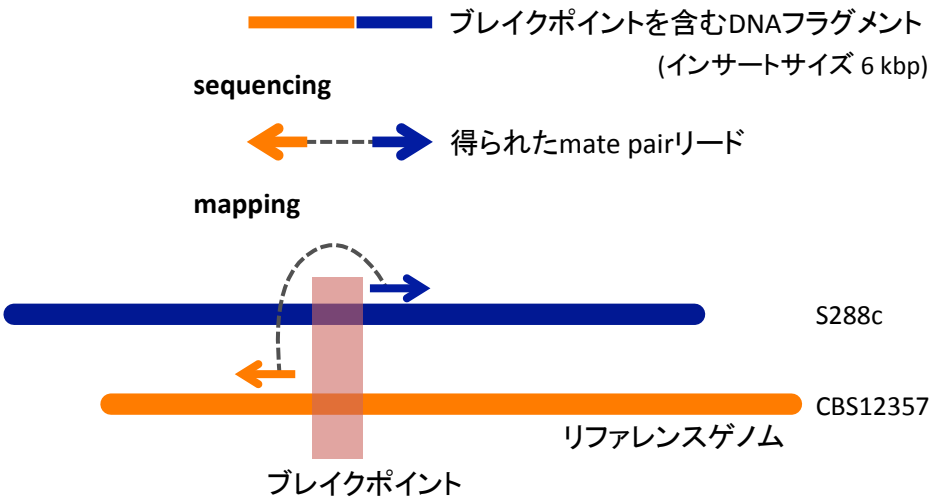


図 4-12 mate-pair リードを使った染色体間転座の検出

染色体間転座のブレイクポイントを含む DNA フラグメントから得られた mate-pair リードは，リファレンスゲノム上の異なる染色体にペアがマップされる．マッピング結果から染色体間転座のブレイクポイントがピンクのハイライト部分に存在することが推測される．

4.3.2 結果・考察

S. pastorianus 4 株（CBS1503, CBS1513, CBS1538, W34/70）において，異種交配後に起きた Sc type と Se type 間の染色体間転座を mate-pair リードを使って検出した．はじめに，mate-pair リードのマッピング結果からブロック間を結ぶリンクの本数について調べた．染色体を 10 kbp のブロックに分割しているため，mate-pair リード（インサートサイズ 6 kbp）のほとんどが同一のブロックまたは隣接するブロック同士にリンクし，染色体間転座のブレイクポイントを含む DNA フラグメント由来のペアリードや一部のキメラリードが離れたブロック同士をリンクすると考えられる．染色体間転座のブレイクポイントのペアリードは，リファレンスゲノム上では離れたブロックにリンクされるが，同一ブロック内または隣接するブロック同士のリンクの本数と同等のリンク数が得られると予想される．同一ブロック内と隣接するブロック間のリンクの本数の分布を図 4-13 に示した．倍数性の違いの影響でリンクの本数にばらつきが見られたが，4 株すべてで 100 本以上のリンクが全体の 96%以上を占めることから，Sc – Se type 間でこの閾値以上のリンクがあった場合，染色体転座の候補として扱うことにした（図 4-14）．

CBS1503 株，CBS1513 株，W34/70 株は，先行研究で Sc/Se 間の染色体転座が特定されており，それぞれ 11, 9, 8 カ所あることが PCR 増幅からも確認されている（Hewitt et al. 2014; Nakao et al. 2009）．今回の手法でもこれらの既知の転座が確認され，

mate-pair を用いた染色体間転座の検出が有効であることが示された（図 4-15）．これらの転座のうち，*HPS82*（染色体 XVI）と *KEM1*（染色体 VII）に存在する転座は，Group 1（CBS1503 株, CBS1513 株）と Group 2（W34/70 株, CBS1260 株）で共通していることが先行研究から知られている．今回，新たに mate-pair リードで確認できた CBS1538 株に加えて，paired-end のみシークエンスした他の 6 株においてもそのブレイクポイントが塩基レベルで一致していることが確認された（図 4-16）．

今回の手法によって CBS1503 株と CBS1513 株から 1 カ所ずつ，W34/70 株から 2 カ所新規の染色体間転座が検出された．そのうち一つは，これら 3 株で共通の転座であり，Sc type 染色体 VII の *ZUO1* 中にブレイクポイントがあることが確認された（図 4-16）．さらに，このブレイクポイントは塩基レベルで一致しており，*S. pastorianus* 10 株で共通していた．W34/70 株で検出されたもう一つの染色体間転座は，*PBR8*（Sc type 染色体 VIII − Se type 染色体 XV-VIII）上にブレイクポイントが存在した（図 4-17）．*S. eubayanus* が単離される以前は，*S. uvarum* を *S. pastorianus* の non Sc type の祖先種と考え，array-CGH の解析を行なっていたため（Dunn and Sherlock 2008），特に non-Sc type での array-CGH の解像度が低く，*ZUO1* や *PBR8* での転座が見落とされたと考えられる．*S. eubayanus* CBS12357 株のドラフトゲノム配列を構築し，リファレンスとして利用したことで，今回新たな転座を検出することができたと考えられる．

S. pastorianus 10 株における染色体間転座の検出結果から，ブレイクポイントが塩基レベルで一致する染色体間転座が既知のものを含めて 3 箇所確認された．Group 1 と 2 で共通の異種交配があった可能性を示唆する結果であると考え，ゲノムワイドに株間の比較が可能な SNV についてさらに解析を実施した．

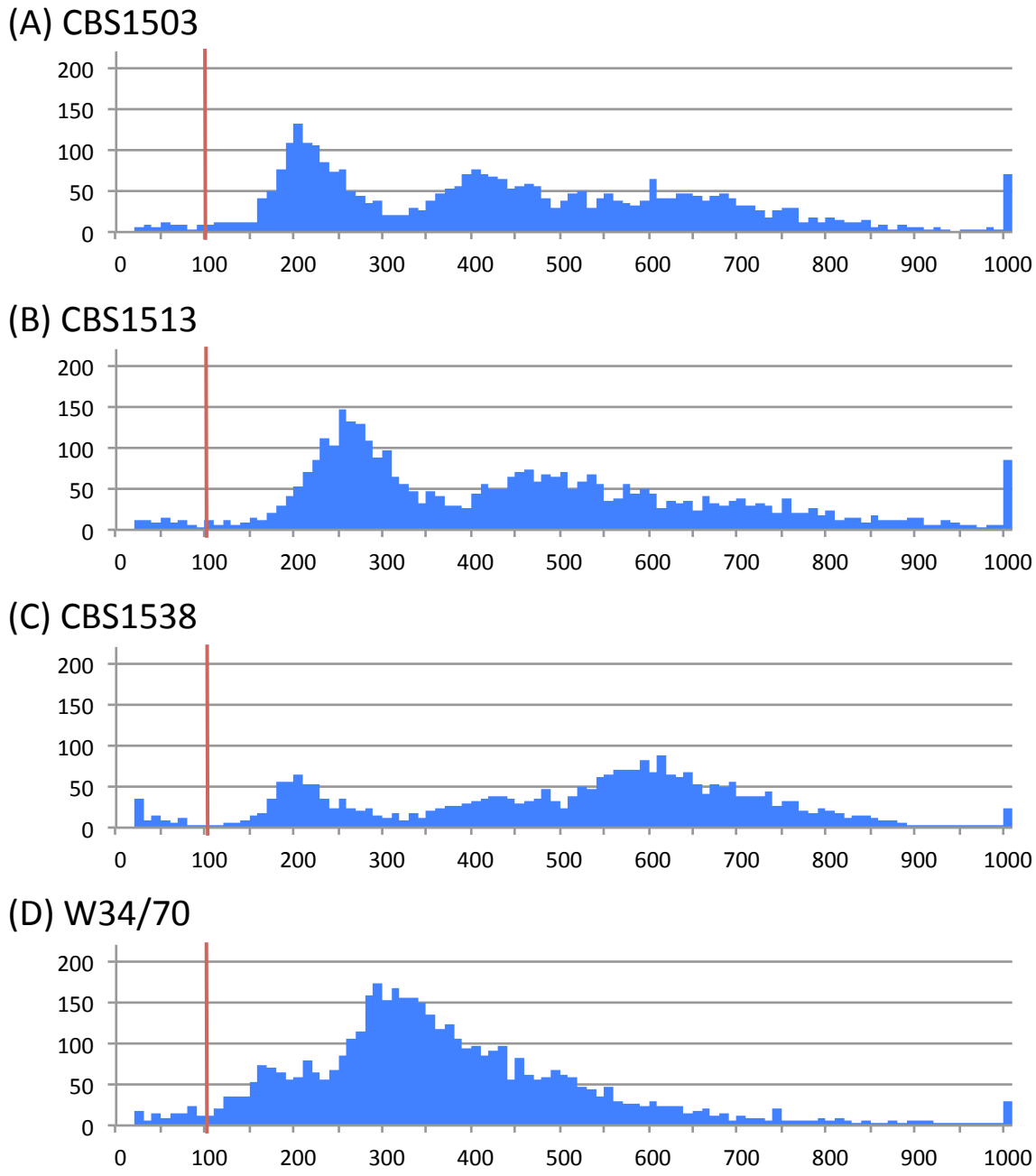
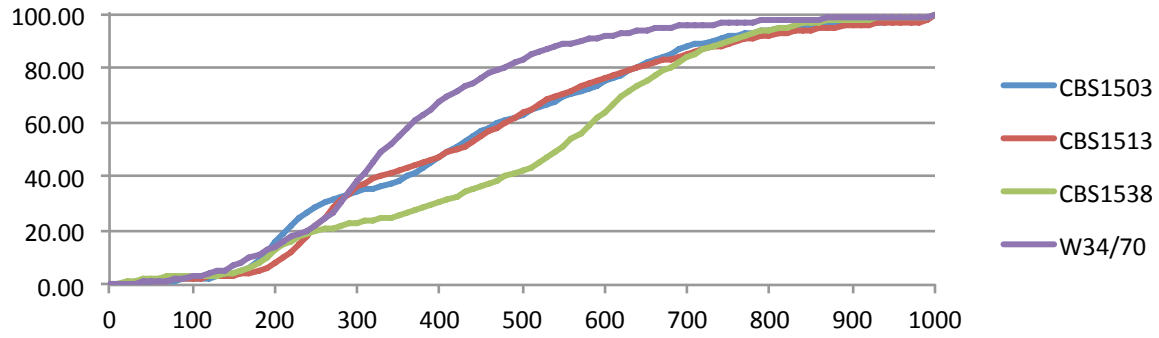


図 4-13 ブロック間のリンクの本数の分布

同一染色体の隣接するブロック間のリンクの本数．縦軸は箇所数を表し，横軸は mate-pair リードが支持するリンクの本数で，上限を 1,000 本としている．

(A)



(B)

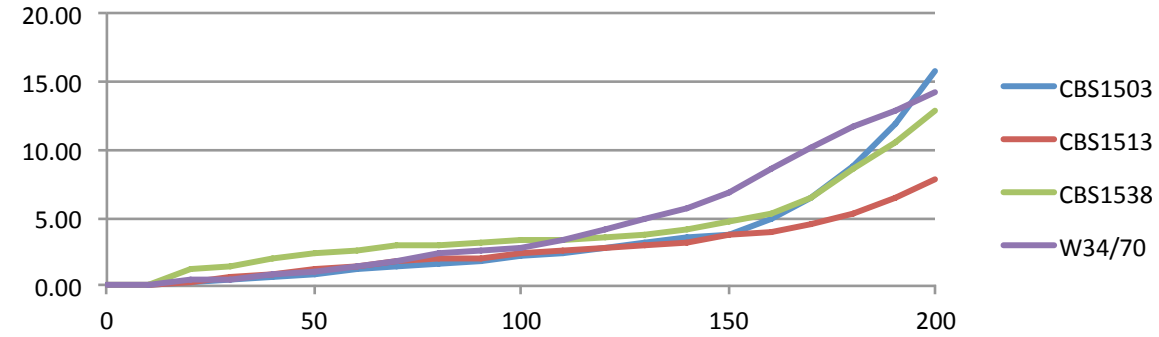
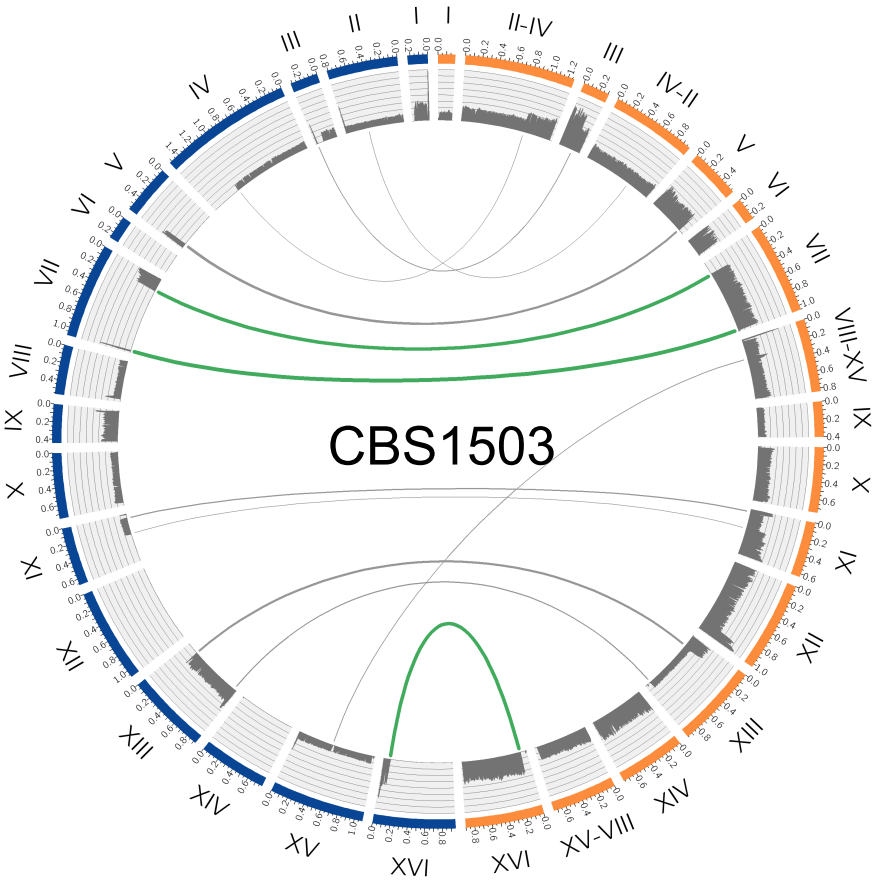
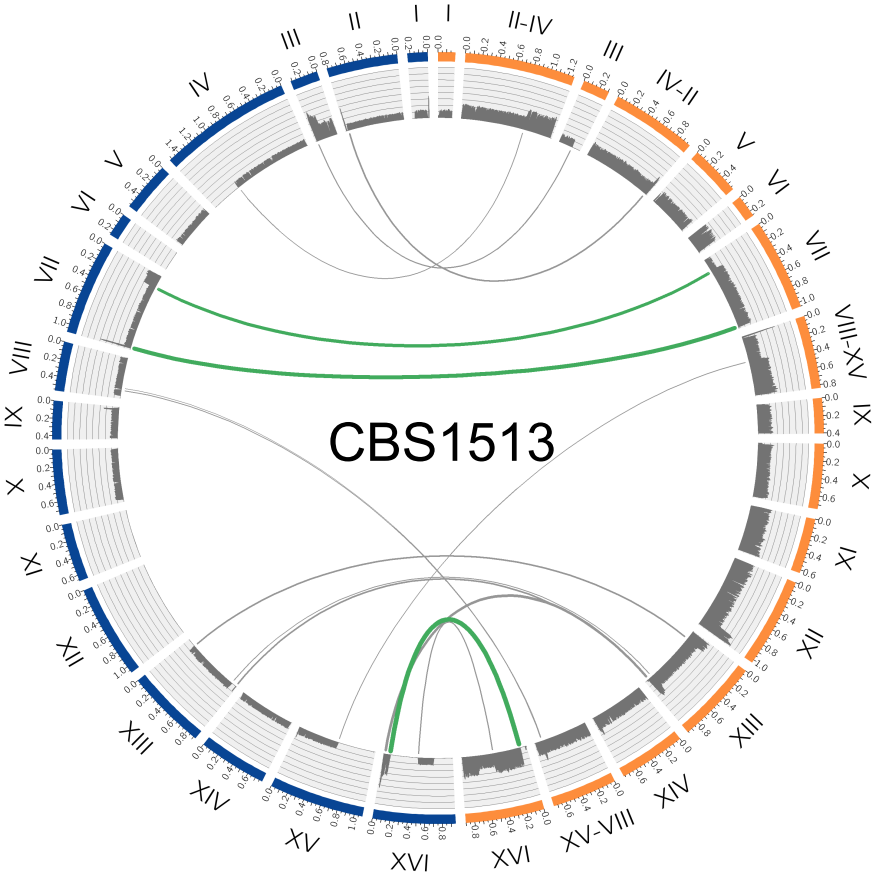
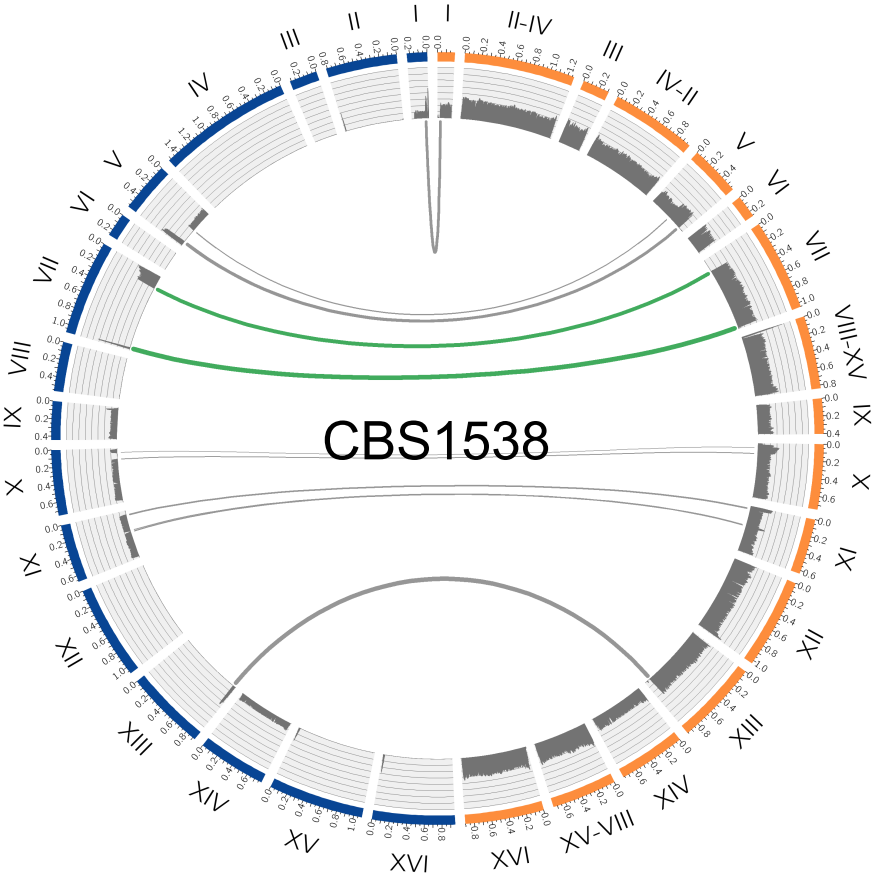


図 4-14 リンク数の積算

横軸はリンクの本数，縦軸は全箇所数に占める割合．横軸の値以下のリンク数が全体に対して占める割合を示している．(A) 全体の様子．(B) 一部拡大図．隣接するブロック間でリンクが 100 以下の領域は全体の 3〜4%程度となっている．


CBS1503


CBS1513


CBS1538

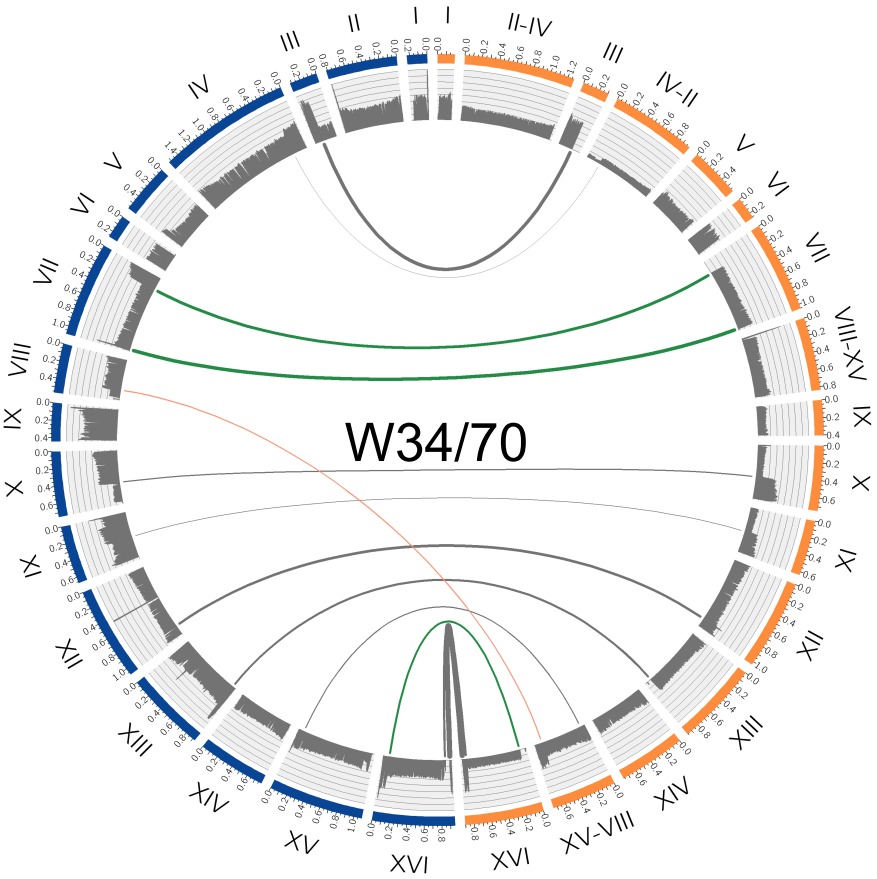
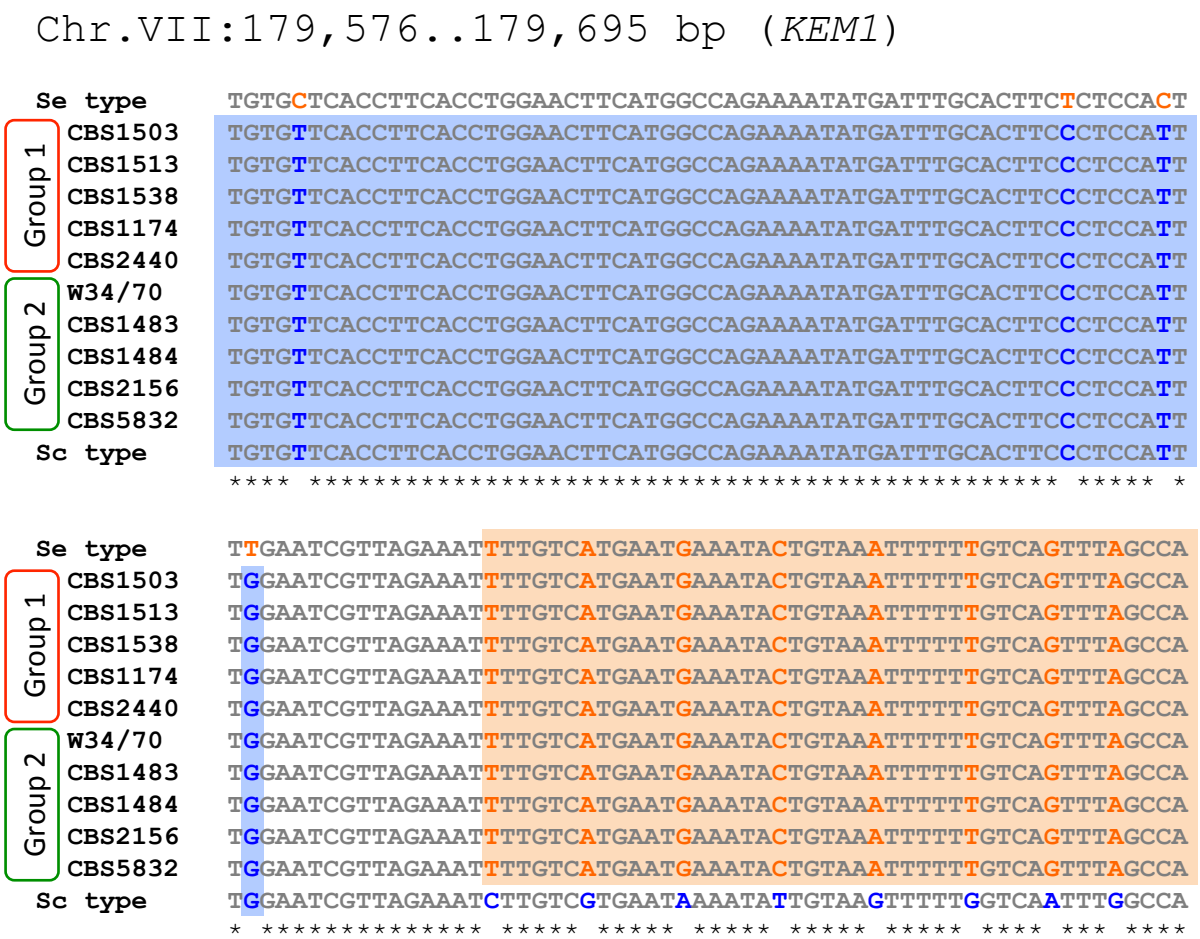
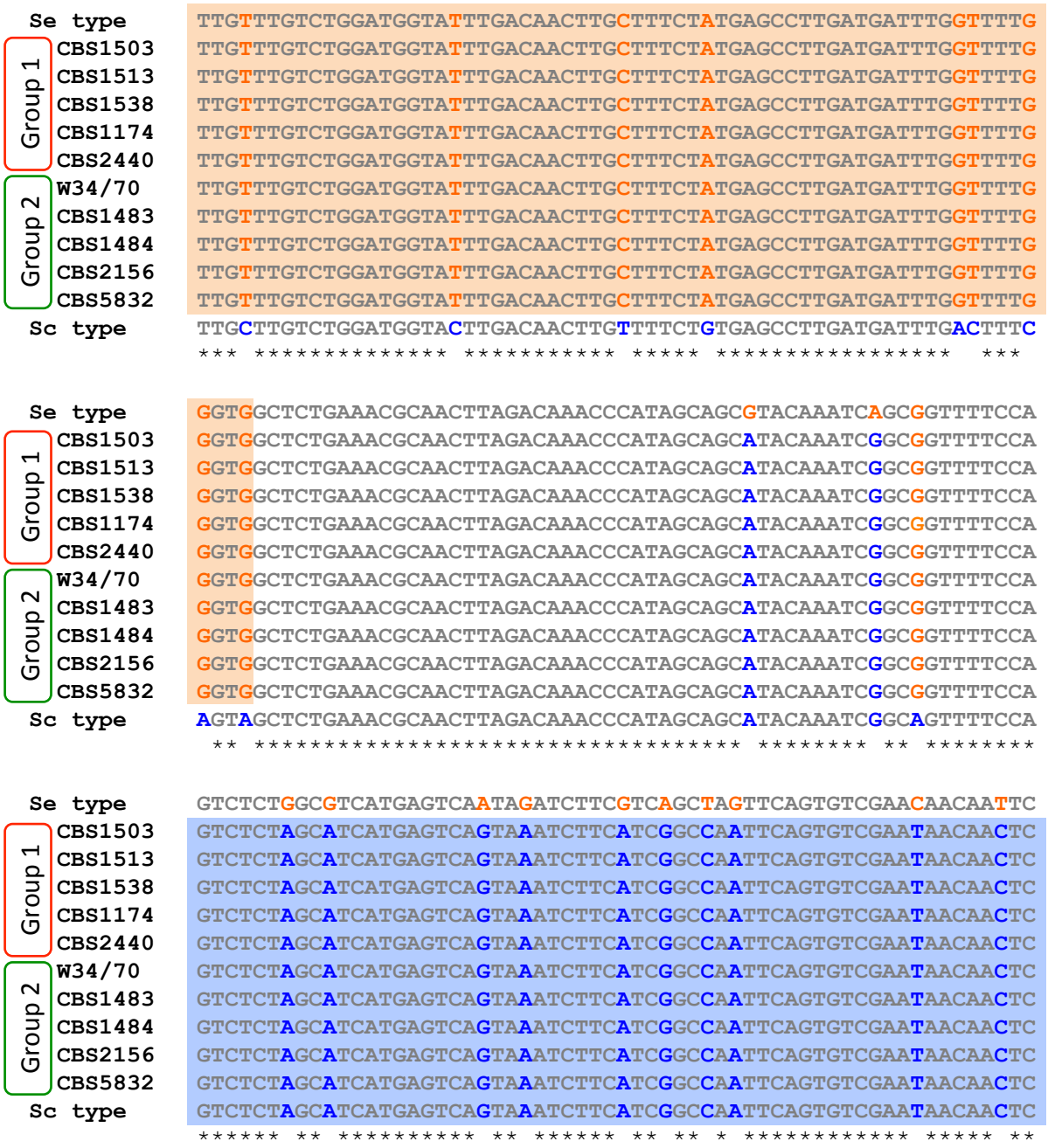

W34/70

図 4-15 Sc type と Se type 間の染色体間転座

mate-pair リードを使って検出した Sc –Se type 間の染色体間転座．環状に並んだ青とオレンジの太線はそれぞれ Sc type (左) と Se type (右) の染色体を表している．Mate-pair リードによって支持された染色体間転座を Sc - Se type の染色体を結ぶ線で示し，線の太さは支持しているリンクの本数を反映している．緑の線は，Group 1 と 2 で共通の転座（*KEM1, ZUO1, HSP82*）を示している．W34/70 株中のピンクの線は，今回新規で見つかった *PRP8* での染色体間転座（Sc type chr. VIII – Se type chr. XV-VIII）を示している．グレーのヒストグラムは，シークエンスカバレッジを表しており，ヒストグラム中にバーの存在しない箇所は染色体欠失を表している．



Chr.VII:1,062,758..1,062,937 bp (*ZUO1*)



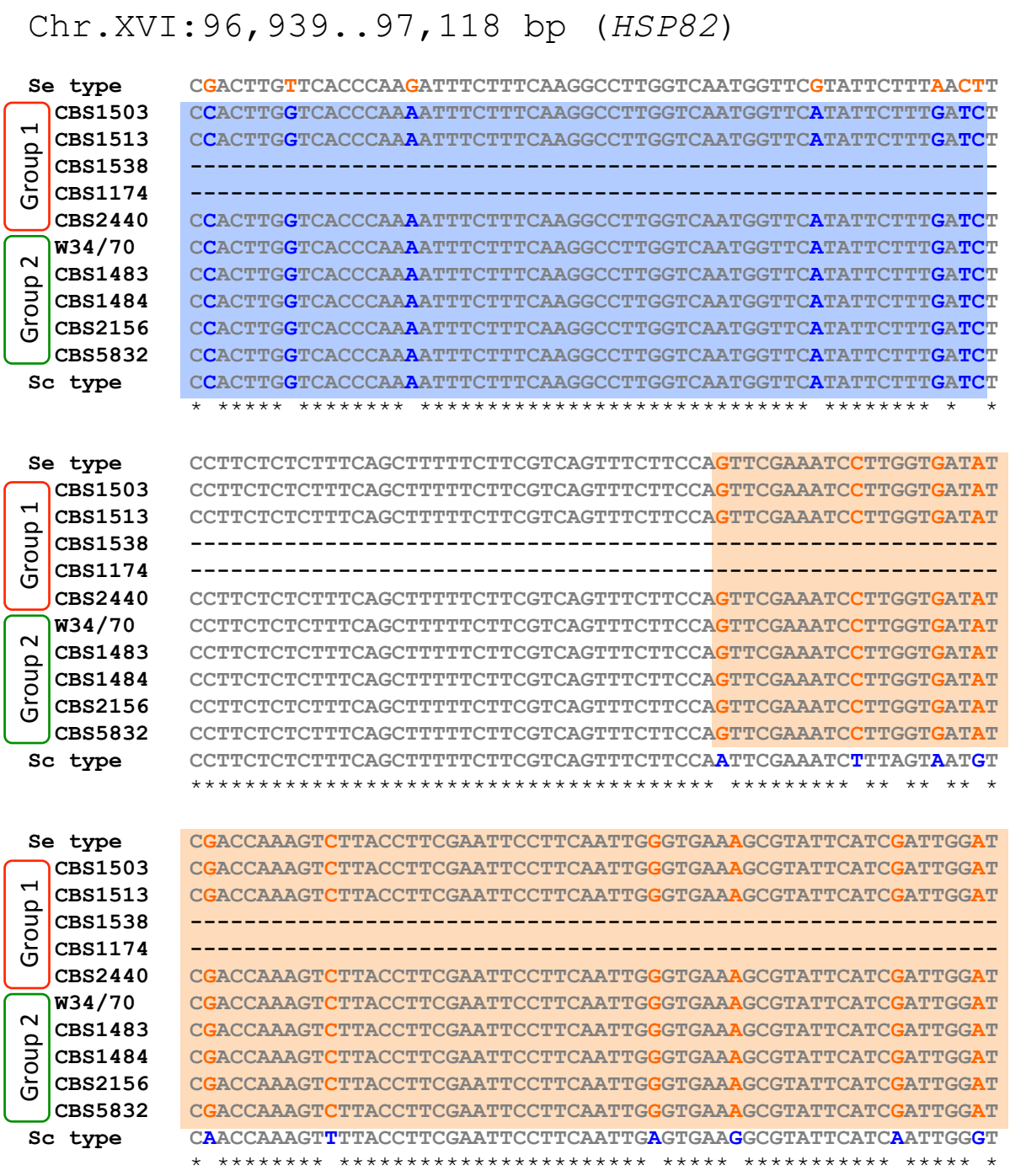


図 4-16 共通する染色体間転座のブレイクポイント周辺のアライメント図

S. pastorianus Group 1 と 2 において共通する染色体間転座のブレイクポイント周辺の配列及び祖先種 *S. cerevisiae* S288c (Sc type) , *S eubayanus* CBS12357 (Se type) の配列のマルチプルアライメントの結果. S288c と CBS12357 間の変異箇所をそれぞれ青とオレンジで示している. 青とオレンジのハイライトは, 由来する祖先種 (Sc type / Se type) を示している. CBS1538 株, CBS1174 株は Sc type chr. XVI が欠失しているため, ブランク (-) で表記している.

図 4-17 W34/70 株で見つかった新規のブレイクポイント周辺のアライメント図

S. pastorianus W34/70 の染色体間転座のブレイクポイント周辺の配列と *S. cerevisiae* S288c (Sc type) と *S eubayanus* CBS12357 (Se type) の配列のマルチプルアライメントの結果. S288c と CBS12357 間の変異箇所をそれぞれ青とオレンジで示している．青とオレンジのハイライトは，W34/70 株の由来する祖先種（Sc type / Se type）を示している．

4.4 SNV 検出，系統解析

4.4.1 方法

S. cerevisiae S288c と *S. eubayanus* CBS12357 に対しての一塩基多型（Single Nucleotide Variant, SNV）の検出方法を以下に示す．Paired-end リードのマッピングからフィルタリングまでを 4.2.1 の通りに行い，GATK realigner を使って挿入・欠失（InDel）領域周辺にマップされたリードの再アライメントを行なった（McKenna et al. 2010）．その後，samtools mpileup を実行し，各ローカスのマップされたリード数，リードのマップされた向き及び支持する塩基，アレル頻度（allele frequency, AF）を求

めた（自作スクリプト）．ここでのアレル頻度（AF）とは，そのローカスにマップされたリードの本数に対する変異を支持するリードの本数の比を意味する．

以下の条件を満たすローカスを以降の解析に利用した．

- マップされたリードが 20 本以上存在する．
- 順鎖，逆鎖方向にマップされたリードがそれぞれ 2 本以上存在する．
- リードの先頭及び末端 1 塩基が支持する変異は除外する．
- 挿入・欠失は除外する．

検出された SNV は AF の値に従って homo SNV（0.8 <= AF）と hetero SNV（0.2 < AF < 0.8）に大別した．Homo SNV はリファレンスゲノム（*S. cerevisiae* S288c と *S. eubayanus* CBS12357）に対する株特有の変異であり，hetero SNV は株の相同染色体間の変異を意味している．

S. pastorianus 10 株，祖先種 *S. cerevisiae* S288c，*S. eubayanus* CBS12357 の系統解析を SNV に基づいて行なった．今回は，*S. pastorianus* のゲノムを Sc type 由来と Se type 由来に分けて系統樹を作成することとした．ある株において染色体欠失している領域あるいはカバレッジが不十分な領域は対象から除外し，*S. pastorianus* 10 株すべてでアレルが存在する領域を対象とし，その中から祖先種（*S. cerevisiae* S288c または *S. eubayanus* CBS12357），*S. pastorianus* 10 株のいずれかで変異が存在するローカスを取り出し，phylip format で出力した．Hetero SNV を含むサイトも系統樹に反映させるために，以下に説明する 2 つの方法を用いた．一つ目は hetero SNV を IUPAC で表記し（表 4-5），phylip format にまとめる方法であり，二つ目は，homo SNV のアレルを 2 塩基分，hetero SNV はそれぞれ 1 塩基ずつ表記することで phylip format にまとめる方法である．系統樹推定は最尤法により計算し（使用ツール phyml –b 1000），樹形を FigTree で作図した（Guindon et al. 2005）．

表 4-5 IUPAC による表記方法

表記	ATGC	
R	G or A	puRine
Y	T or C	pYrimidine
M	A or C	aMino
K	G or T	Keto
S	G or C	Strong interaction(H-bonds x 3)
W	A or T	Weak interaction(H-bonds x 2)

4.4.2 結果・考察

S. cerevisiae S288c，*S. eubayanus* CBS12357 をリファレンスとして，*S.*

pastorianus 10 株から検出された homo / hetero SNV 数の集計結果を表 4-6 に，その密度を表 4-7 に示した．SNV の密度を求めるために用いたゲノムサイズはマップされたリードの領域から算出し（表 4-4），ゲノムサイズに対する homo SNV および hetero SNV の割合を示している．Group 1 の Sc type における染色体欠失によりゲノムサイズが株ごとに異なるため，homo SNV 数にばらつきがあるように見えるが，SNV の密度について比較すると株による大きな違いは見られなかった．また Group 1 に属する 5 株の Sc type における 2 倍体以上の領域が限られているため hetero SNV 数が少ないと考えられる．Se type の hetero SNV 数は Group 2 に属する株よりも Group 1 に属する株で多く，Sc type とは異なる傾向が確認された．

表 4-6 *S. pastorianus* 株で検出された SNV 数

		Sc type		Se type	
	S. pastorianus	**homo SNV**	**hetero SNV**	**homo SNV**	**hetero SNV**
Group 1	CBS1503	29,698	465	62,382	663
	CBS1513	36,371	258	62,520	1,636
	CBS1538	15,233	283	62,521	2,340
	CBS1174	15,671	264	62,044	2,359
	CBS2440	19,283	325	61,720	2,180
Group 2	W34/70	48,892	10,900	59,226	979
	CBS1483	49,352	6,459	58,932	683
	CBS1484	47,551	7,269	59,479	887
	CBS2156	50,819	5,871	58,920	507
	CBS5832	49,849	8,022	60,466	780

表 4-7 *S. pastorianus* の SNV 密度（%）

		Sc type	Se type
Group 1	CBS1503	0.503	0.581
	CBS1513	0.479	0.579
	CBS1538	0.507	0.589
	CBS1174	0.507	0.586
	CBS2440	0.478	0.582
Group 2	W34/70	0.538	0.557
	CBS1483	0.511	0.551
	CBS1484	0.513	0.557
	CBS2156	0.512	0.550
	CBS5832	0.521	0.565

今回の系統解析では，*S. pastorianus* を Sc type と Se type に分けて行い，それぞれの系統樹を検証することとした．*S. pastorianus* 10 株及び祖先種 1 株（*S. cerevisiae* S288c または *S. eubayanus* CBS12357）合計 11 株のうち少なくとも 1 株において変異が検出されたサイトは，Sc type で 9,840 サイト，Se type で 58,189 サイトであった．Sc type は，Group 1 に属する 5 株における染色体欠失により，Se type と比較して得られたサイト数が少ないが，系統解析を行うには十分であると判断した．Se type は，*S. pastorianus* 10 株において *S. eubayanus* CBS12357 に対する homo SNV が 58,920～62,521 サイトあるにもかかわらず（表 4-6），得られた領域が 58,189 サイトと少なかった．これは，上の条件で取り除かれたサイト以外では，ほとんどの SNV が 10 株で同一のローカスに重複して存在しているためと考えられる．

上記サイトに基づいて作成された Sc type と Se type の系統樹において *S. pastorianus* Group 1 と 2 間の距離関係に明確な違いが認められた（図 4-18）. Sc type において，*S. pastorianus* は大きく 2 つのクレードに分かれ，その 2 つのクレードが地理的起源に基づく分類（Group 1，Group 2）と一致しており（図 4-18 A），先行研究と同様の結果が得られた（Dunn and Sherlock 2008）．つまり，Sc type において地理的起源に従った分類には系統的な差異が存在し，Group 1 と 2 は別々の祖先に由来することが示唆された．一方，Se type では，*S. pastorianus* は地理的起源に基づく分類に関係なく 1 つのクレードに位置する樹形が得られた（図 4-18 B）. Se type の系統樹について詳細に見ると，*S. pastorianus* 10 株において Group 1 に属する CBS1513 株がまず分岐し，その後 Group 1 の 4 株と Group 2 に属する 5 株に分岐している．しかし，Sc type と比較して，Group 1 と 2 の系統的な距離は小さく，*S. pastorianus* 10 株間の Se type における変異率は 0.063%で，これは Sc type における変異率（0.348%）

と比較してはるかに低い．Se type においては，2 つのグループは系統的な差異は見られず，共通祖先に由来している可能性が示唆された．この 2 つの系統樹は，地理的起源に基づいた系統関係について相異なる可能性を示唆している．

仮に，2 つのグループが共通祖先に由来する場合，異種交配の直後のゲノムはすべての株において同一であり，異種交配後に起きた変異が SNV として検出されることになる．この時，同一核内に存在する染色体が祖先種によってその変異率が異なるということは考えにくく，Sc, Se 両タイプで同じ樹形の系統樹が得られると予想される．しかし，実際に得られた系統樹では Sc type における Group 1 と 2 が含まれるそれぞれのクレード間の枝の長さが Sc type のものよりも長いことから，Sc type が Se type と比較して高い変異率を持つことを意味することとなる．一方，2 つのグループが別々の異種交配に由来する場合，Sc type と Se type で異なる樹形の系統樹が得られても矛盾は生じない．Sc type は別々の祖先に由来し，異種交配前の祖先株間の変異が系統樹に反映されることで Group 1 と 2 が 2 つのクレードに分かれ，一方 Se type は 2 つのグループで共通の祖先に由来するため 1 つのクレードに 10 株が位置した系統樹が得られたと考えることができる．この考察は，(Baker et al. 2015) でもされており，彼らは，SNV による系統解析と倍数性の違いから 2 つのグループが別々の異種交配によるものであると結論を出している．

しかし，Sc type と Se type における地理的起源に基づく分類と系統関係に関して，特に Sc type の Group 1 と 2 の系統的な差異については，以上述べてきた様々な解析結果から Sc type における Group 1 と 2 間の倍数性と hetero SNV の密度の違いを考慮に加える必要があるのではないかと考えられた．つまり，Sc type において祖先種から今までに蓄積された変異以外の別の要因が系統樹に影響している可能性についての考察の必要性である．まず，Sc type の倍数性について，系統樹に反映されている領域では Group 1 に属する 5 株は 1 倍体であるのに対し，Group 2 に属する 5 株では 2 倍体あるいは 3 倍体であり，要するに 1 対 2 本以上の染色体間の変異を比較している．ここで，その株における同質の相同染色体間の SNV (hetero SNV) が少なければ，倍数性が系統解析に影響を与えることはない．しかし，hetero SNV が多い株の場合，2 倍体における「平均的なアレル」を系統解析に用いることになり，その株が持つ本来のハプロタイプとは異なる配列同士を比較することになる．今回の *S. pastorianus* の Sc type において Group 1 に属する株はほとんどの領域が 1 倍体であるため hetero SNV を考慮する必要はないが，Group 2 に属する株からは hetero SNV が 5,871 から 10,900 サイト検出されており，系統解析において相同染色体間の SNV を無視することはできない．一方，Se type においても倍数性は株ごとあるいは染色体ごとに差異は存在しているが，hetero SNV 数が Sc type と比較してはるかに少ないためその影響は小さいと考えられる．以上の理由から系統関係について再検討する必要があると考え，

特に Group 2 に属する株の Sc type における hetero SNV に注目して次章の解析を行なった.

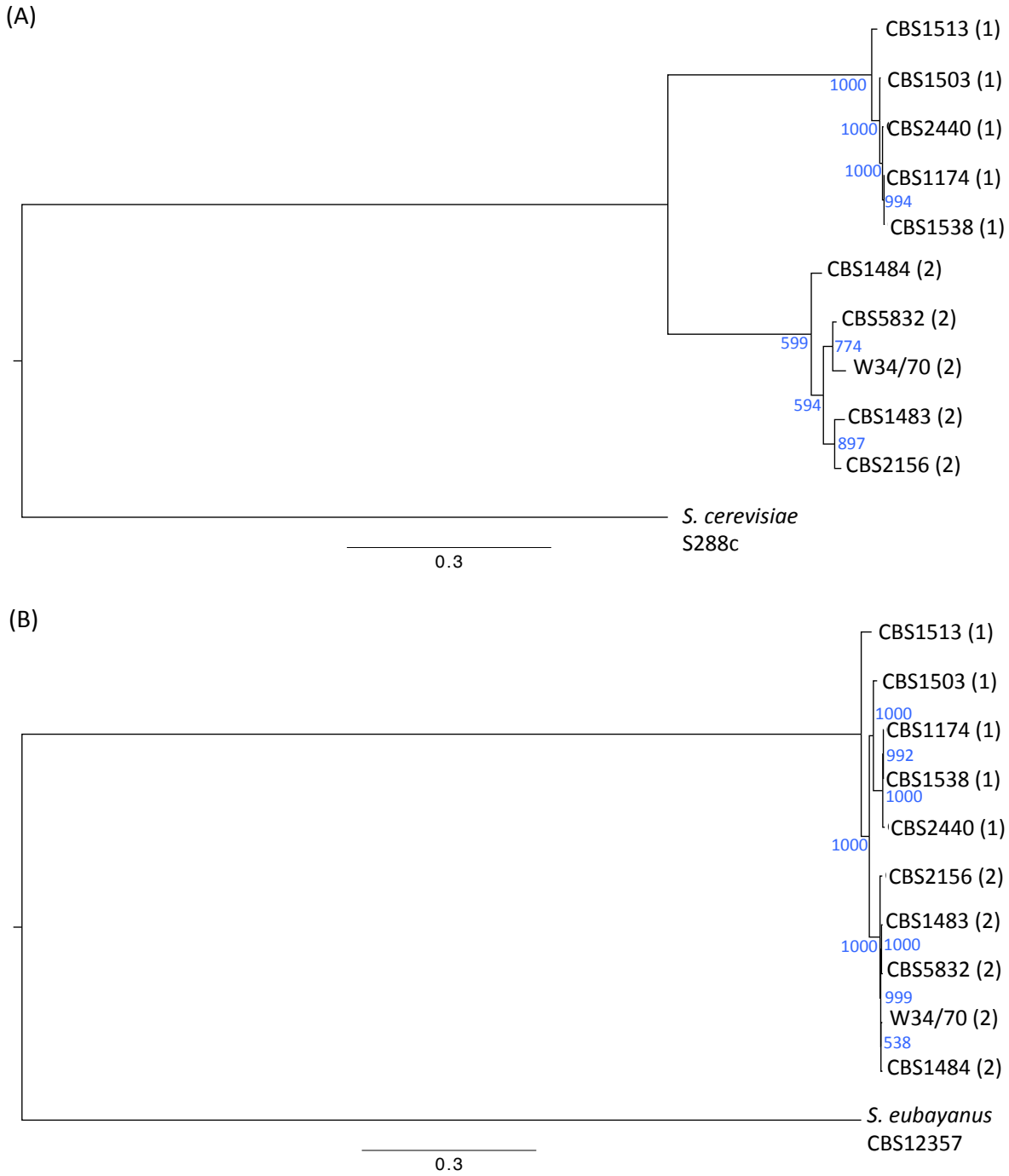


図 4-18 SNV に基づいた *S. pastorianus* 10 株と祖先種の系統関係

(A) *S. cerevisiae* S288c と *S. pastorianus* 10 株の系統樹.

(B) S. eubayanus CBS12357 と S. pastorianus 10 株の系統樹.

Hetero SNV を含む変異サイトで計算した. S. pastorianus の株名の後ろにある () 内の数字は, 分類されているグループを表している. 系統樹の内の青の数字はブートストラップ値 (n = 1,000) を示している.

4.5 まとめ

本章において *S. pastorianus* 10 株のゲノム解析を実施した．得られた結果を以下に簡潔に示す．

- ミトコンドリア

  S. eubayanus 2 株，*S. pastorianus* 4 株，近縁種 2 種 3 株のミトコンドリア配列の決定と系統解析を実施した．9 株すべてにおいてミトコンドリアの完全ゲノムを決定することに成功した．また系統解析の結果，*S. pastorianus* のミトコンドリアは *S. eubayanus* に由来し，チベットで単離された *S. eubayanus* BaiFY1 と高い相同性を持つことが明らかになった．

- 倍数性

  シークエンス量から *S. pastorianus* 10 株の倍数性の推定を実施し，解像度の高い結果を得ることができた．Group 1 に属する株の Sc type において染色体欠失が多く観察され，Group 2 に属する株の Sc type の倍数性（2 倍体あるいは 3 倍体）との明確な差異が確認された．また，異種交配後に染色体重複・欠失を繰り返しているにも関わらず，異種間の相同染色体の倍数性は一定に保たれていることが示唆された．

- 染色体間転座

  mate-pair リードを利用した Sc – Se 間の染色体間転座の推定を実施した．本手法により既知のブレイクポイントはすべて検出され，さらに今回新たに染色体間転座を 2 箇所で検出した．既知の染色体間転座も含めて，Group 1 と 2 においてブレイクポイントが一致する染色体間転座が 3 箇所存在することが確認された．

- 系統解析

  Sc type において *S. pastorianus* は系統的に 2 つに分類され，地理的起源に基づいた分類と一致すること確認され，先行研究と同様の結果が得られた．一方，Se type においては Group 1 と 2 に系統的な差異は確認されず，10 株すべてが同一のクレードに位置する系統樹が得られた．

ハイスループット DNA シークエンスデータを使用し，祖先種の染色体レベルの配列を基盤にして解析を実施したことで，新規の染色体間転座や Se type の倍数性，系統樹などの新たな知見を得ることができた．しかし，地理的起源に基づく分類と由来の関係の結論を出すためには，Sc type と Se type における系統解析の結果を再考察する必要があると判断し，次章に示した解析を実施することとした．

第5章　地理的起源に基づく分類と系統関係の再考察

5.1　*S. pastorianus* Sc type における hetero SNV

系統解析の結果を見直すにあたり，SNV の検出結果 – 特に，Group 2 に属する 5 株において Sc type の hetero SNV の密度が Se type と比較して 10 倍近く高いこと – に注目した（表 4-6）．Group 2 に属する 5 株の倍数性は Sc, Se type 共におおむね 2 倍体であるため（図 4-10），Sc - Se type 間における hetero SNV の密度の差異は倍数性の違いよるものではなく, 別の要因が存在すると考えられる. ここで, *S. pastorianus* の Sc type の祖先種と考えられているエールビール酵母 *S. cerevisiae* に着目すると，エールビール酵母が高ヘテロ接合性ゲノムであるという報告が確認された (Borneman et al. 2011）．この論文によると，エールビール酵母 *S. cerevisiae* Foster's B 株と Foster's O 株のゲノムシークエンスにより, 配列中にそれぞれ 37,784 個, 32,600 個の hetero SNV が検出されている．つまり，*S. pastorianus* Sc type の hetero SNV は，祖先株であるエールビール酵母 *S. cerevisiae* (高ヘテロ接合性ゲノム）に由来していると考えらえる．しかし，*S. pastorianus* Group 2 に属する 5 株の Sc type における hetero SNV はエールビール酵母 Foster's B 株と Foster's O 株と比較すると 3 分の 1 以下と少ないことから（図 5-1），異種交配後のゲノム再構成の過程で hetero SNV が減少したことが示唆される.

さらに染色体上における hetero SNV は, ゲノム上に不均一に分布していることが明らかになった（図 5-2）．hetero SNV の平均密度は hetero SNV がもっとも多く検出された W34/70 株においても 0.1%に満たないが, 分布からは hetero SNV の密度が 1% 近い領域も確認された．このような hetero SNV の不均一な分布は，清酒酵母 *S. cerevisiae* Kyokai no. 7 や産業用燃料エタノール酵母 CAT-1 でも報告されており (Akao et al. 2011; Babrzadeh et al. 2012），loss of heterozygosity (LOH）に起因することが示唆されている．ここで LOH とは，同質の相同染色体間の SNV (hetero SNV) が消失し，homo 化することを意味する．同質の相同染色体のうち片方のハプロタイプを欠失する，あるいは異質体において異種間の相同染色体の片方の相同染色体を欠失する場合も LOH と捉えるケースもあるが, 今回はそれらは染色体欠失として扱うこととする.

この 2 つの特徴 – つまり，異種交配後の hetero SNV 数の減少及び染色体上での hetero SNV の不均一な分布 – から *S. pastorianus* Group 2 に属する株の Sc type において異種交配後に繰り返し LOH が起きたことが示唆される．また，図 5-2 から Group 2 に属する 5 株の hetero SNV の分布を比較すると, 分布が一致している領域と各株で異なる領域が存在することが確認された．このことから，株に分化する前の段

階で起きた LOH 及び株ごとに独立に起きた LOH が存在すると考えられる．

LOH は，不稔性生物における遺伝的多様性を獲得するための原動力であると考えられている．*S. cerevisiae* 以外では，病原性真菌 *Candida albicans*，ウリ科の植物を宿主とする病原菌 *Phytophthora capsici* において LOH がストレス耐性，環境適応に重要な役割を果たしていると考えられている例が報告されている（Forche et al. 2011; Lamour et al. 2012; Bennett et al. 2014）．胞子形成能のない *S. pastorianus* においても LOH によって集団として多様性を保持している可能性が示唆される．

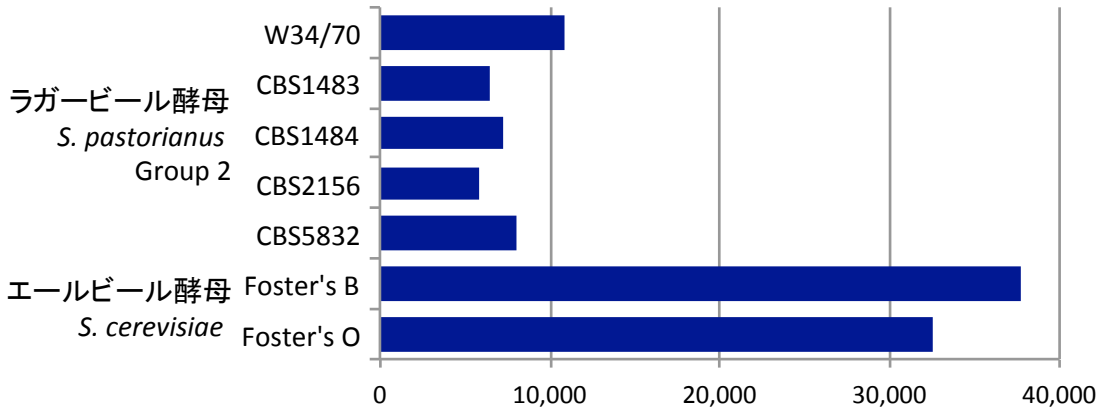


図 5-1 ラガービール酵母とエールビール酵母の hetero SNV 数

ラガービール酵母 *S. pastorianus* (Group 2 に属する 5 株）の Sc type とエールビール酵母 *S. cerevisiae* 2 株における hetero SNV 数.

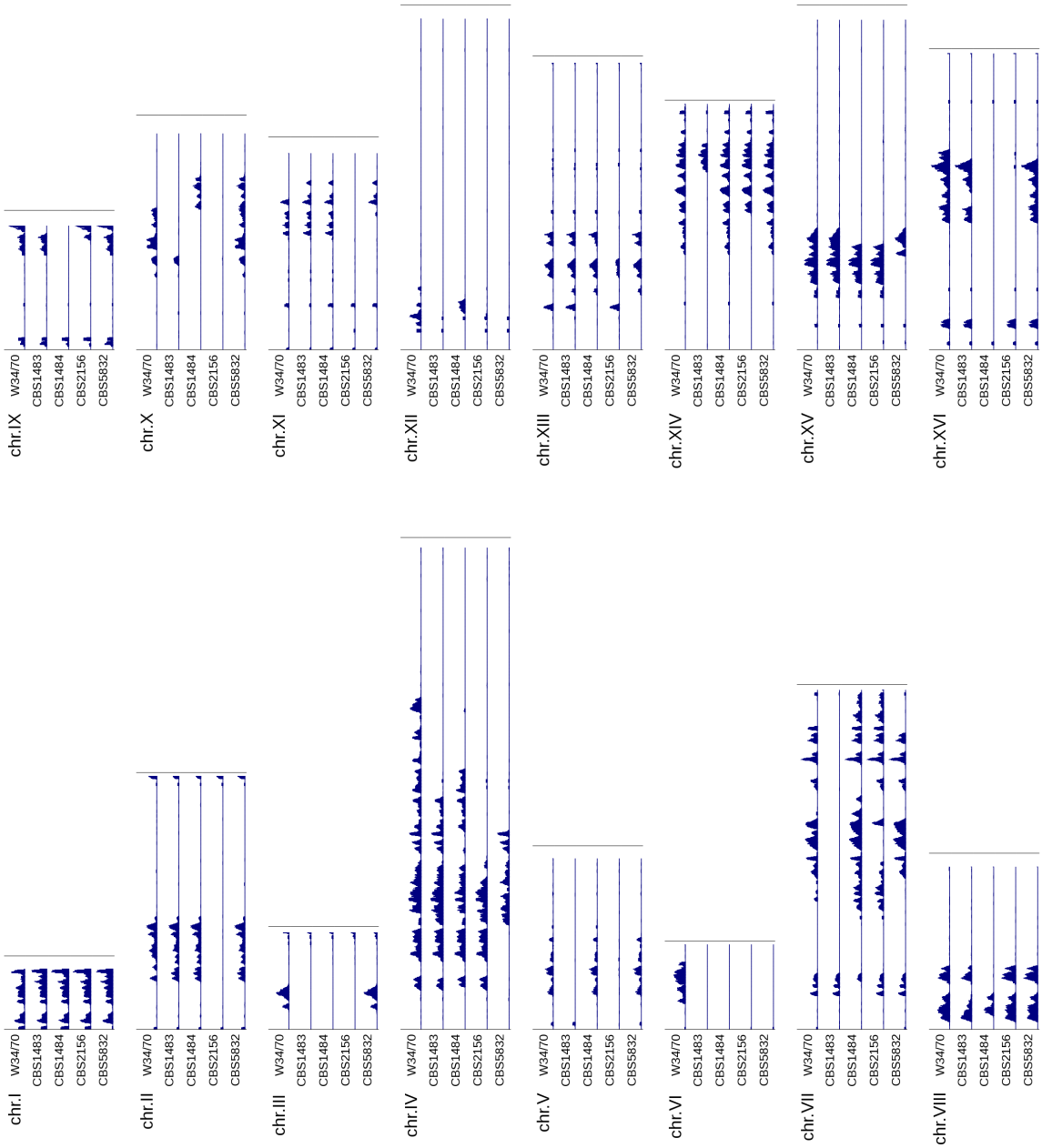


図 5-2 Group 2 に属する 5 株の Sc type 染色体における hetero SNV の分布

S. pastorianus Group 2 Sc type 染色体上での hetero SNV の分布．Window size 10 kbp, step size 1 kbp で hetero SNV の数をプロットした．縦軸の範囲はすべて 0 から 100 としている．

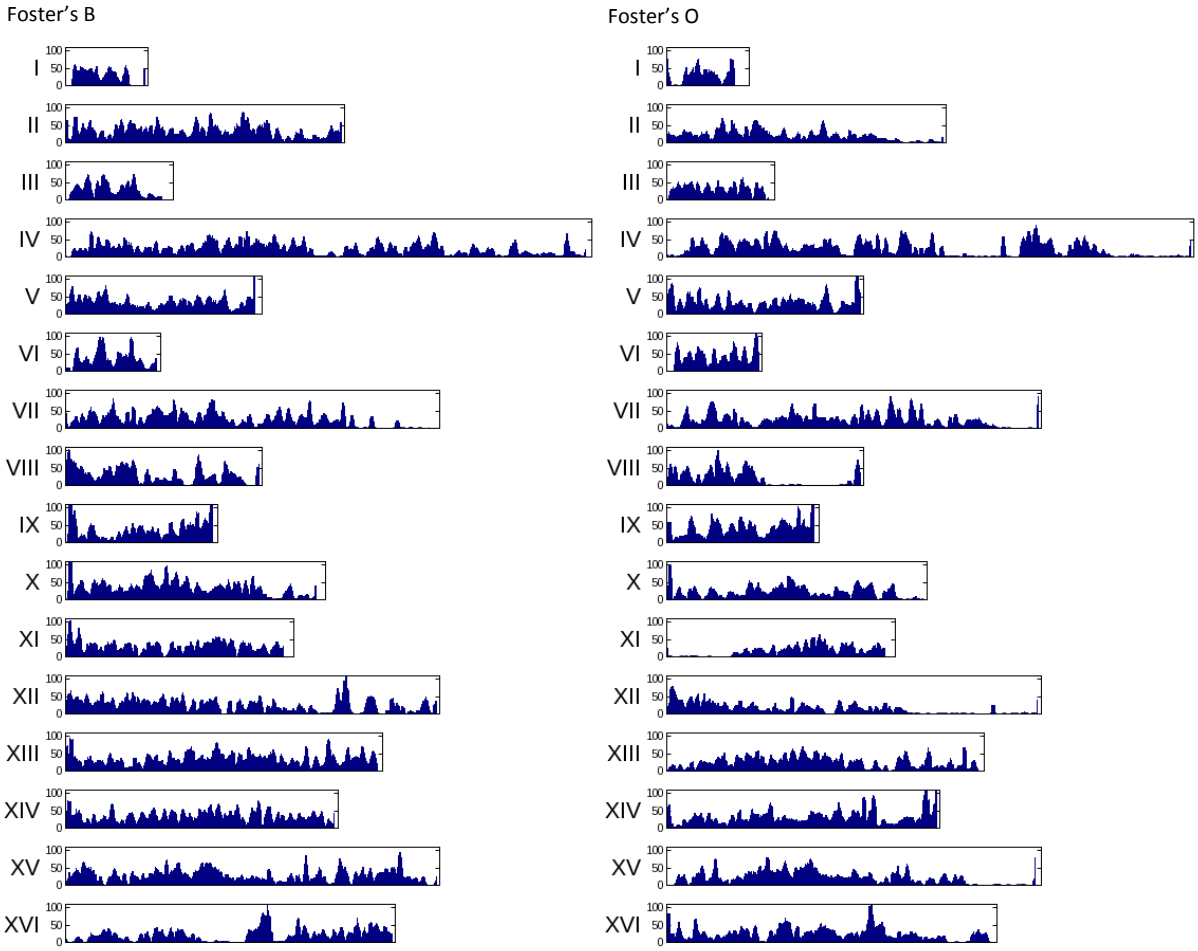


図 5-3 エールビール酵母の hetero SNV の分布

エールビール酵母 *S. cerevisiae* Foster's B と Foster's O の hetero SNV の分布．*S. pastorianus* の Sc type はエールビール酵母に由来すると考えられている．

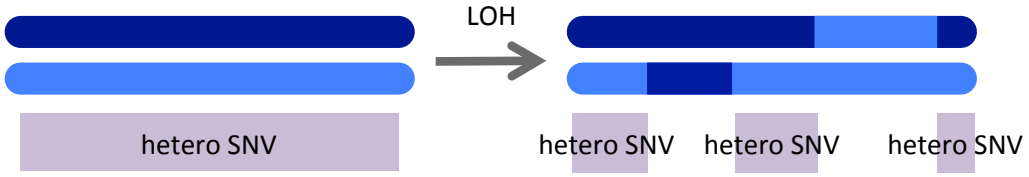


図 5-4 LOH による hetero SNV の不均一化の模式図

高ヘテロ接合性ゲノムにおいて LOH が起きた場合の hetero SNV の分布を模式図で示した．図中の青と水色のバーは同質の相同染色体を表している．LOH が起きた領域においては，どちらかのハプロタイプのアレルが消失し，もう片方のアレルに置換するため，LOH 前に存在した相同染色体間の SNV (hetero SNV) は観察されない．LOH が繰り返し起きた場合，hetero SNV は減少するとともに，不均一に観察されると考えられる．

5.2 Group 1 と 2 の間に存在する SNV

はじめに，Group 1 – 2 間の SNV (Group 1 に属する株と Group 2 に属する株の間に存在する SNV) を大きく 2 つに分類し，その内訳を求めた．Sc type において Group 1 に属する株はおおむね 1 倍体であるのに対して，Group 2 に属する株が 2 倍体あるいはそれ以上であるため，Group 2 に属する株における相同染色体間の SNV (hetero SNV) が存在するサイトとしないサイトに分類した．具体的なそれぞれの定義を以下に示した．

- homo/homo サイト

  Group 1 に属する株と Group 2 に属する株においてそれぞれ homo SNV が検出され，それらが異なるサイト

  例) Group 1：A – Group 2：C
- homo/hetero サイト

  Group 1 に属する株において homo SNV，Group 2 に属する株において hetero SNV が検出されたサイト

  例) Group 1：A – Group 2：A/G，Group 1：A – Group 2：C/G

Group 1 に属する株ごとに Group 2 に属する 5 株それぞれに対する SNV 数 (homo/homo，homo/hetero) を図 5-5 左に示した．SNV 数が Group 1 に属する株ごとに大きく異なるのは，Group 1 に属する株における染色体欠失により株間で比較の対象領域が異なることに起因している．Group 1 に属する株の中でもっとも染色体欠失の少ない CBS1513 株と Group 2 に属する 5 株間における SNV を比較すると，20,000 サイト以上の SNV が存在した．また，Group 1 と 2 の間に存在する SNV のうち，homo/hetero SNV が 1/4 から 1/3 程度占めていることが明らかになった．

さらに，homo/hetero SNV を，以下に示した定義に従って「一致するサイト」と「矛盾するサイト」の 2 つに分類した．

- homo/hetero サイト
  - 一致するサイト

    Group 2 に属する株の hetero SNV のアレルのうち片方のアレルと Group 1 に属する株のアレルが一致するサイト

    例) Group 1：A – Group 2：A/G
  - 矛盾するサイト

    「一致するサイト」に該当しないサイト

    Group 2 に属する株の hetero SNV のいずれとも異なるアレルが Group 1 に属する株から検出されたサイト

    例) Group 1：A – Group 2：C/G

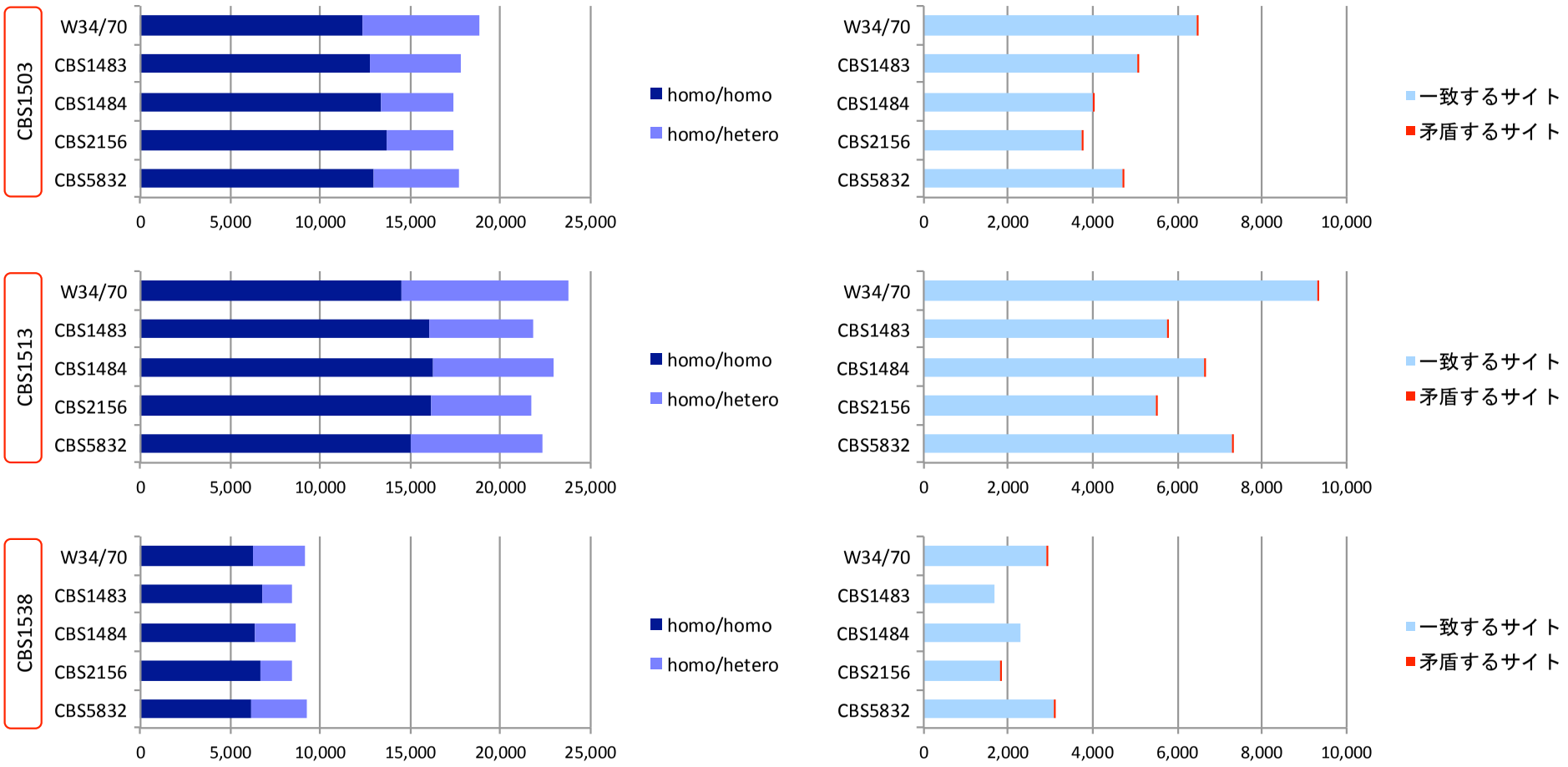
つまり，「一致するサイト」とは，Group 1 に属する株のアレルと Group 2 に属する株における片方のハプロタイプのアレルが一致するサイトを意味し，「矛盾するサイト」とは Group 1 に属する株と Group 2 に属する株のハプロタイプのアレルがすべて異なることを意味している．

S. pastorianus Sc type における Group 1 － 2 間の homo/hetero SNV を上記の通りに分類すると，「一致するサイト」が homo/hetero SNV のほとんどを占めており，「矛盾するサイト」は数カ所しか存在しなかった（図 5-5 右）．この結果から Group 1 – 2 間の SNV のうち 1/4 から 1/3 程度を占めている homo/hetero SNV は，Group 2 に属する株のハプロタイプにおいては Group 1 に属する株のアレルと一致しており，SNV は観察されないことが明らかとなった．

この結果を踏まえ，以下のような仮説を立てた．その仮説とは，Group 1 と 2 の系統関係について，Group 1 と 2 は共通の異種交配が 1 回存在し，Sc type においてハプロタイプが共通祖先株に由来するというものである．この仮説は，Se type の系統関係及びグループ間でブレイクポイントが一致する染色体間転座が存在することとも矛盾しない．しかし，homo/hetero サイトにおいて「一致するサイト」が大半を占めるという結果のみでこの仮説を支持するには証拠が不十分であり，Group 1 – 2 間の SNV について，以下の 2 点を検証する必要があると考えられた．

homo/hetero サイトについては，上の結果で「一致するサイト」が大半を占めていたことから，ローカスごとに比較すると Group 1 と 2 でハプロタイプが共通祖先株に由来すると考えても矛盾は生じない．しかし，実際には隣接する hetero SNV の連鎖関係が存在するため，仮説が正しいのであれば，Group 2 に属する株の hetero SNV のうち「Group 1 に属する株と一致するアレル」が連続して同一のハプロタイプ上に存在することが予想される．この仮説を検証するために Group 2 に属する株の Sc type の染色体（2 倍体あるいは 3 倍体）におけるハプロタイプ別のゲノムを構築し，それらと Group 1 に属する株の配列を比較した．

ここで，仮に homo/hetero サイトから Group 1 と 2 が共通祖先種に由来することが示唆された場合，Group 1 - 2 間における homo/homo サイトは何に起因するのかという疑問が残る．これについてもグループ間の倍数性の違いと Group 2 に属する株における hetero SNV と LOH を含めて検証した．


CBS1503
W34/70
CBS1483
CBS1484
CBS2156
CBS5832
0
5,000
10,000
15,000
20,000
25,000
homo/homo
homo/hetero
0
2,000
4,000
6,000
8,000
10,000
一致するサイト
矛盾するサイト
CBS1513
CBS1538

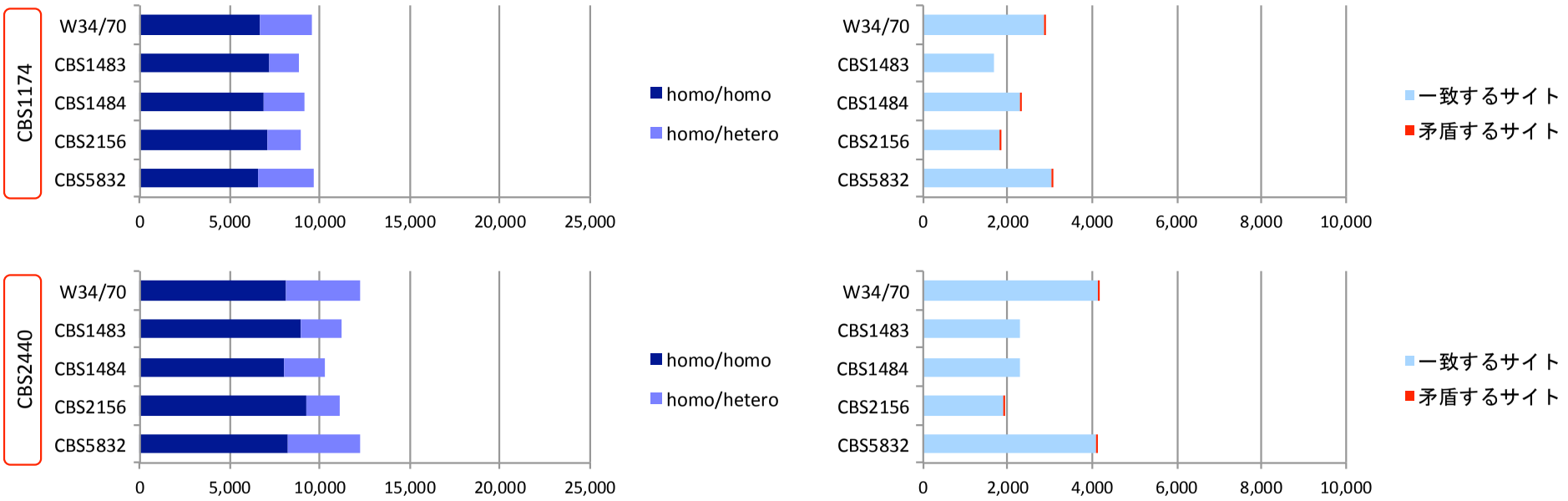


図 5-5 Group 1-2 間の SNV の内訳

Group 1 に属する株と Group 2 株に属する株の間に存在する SNV の内訳（左）．赤枠で囲んだ株名は Group 1 に属する株を示し，それに対する Group 2 に属する 5 株それぞれの SNV 数を示した．ここで homo/homo とは，Group 1，2 において異なる homo SNV が支持されたサイトを意味し，homo /hetero とは Group 1 に属する株において homo SNV，Group 2 に属する株において hetero SNV が支持されたサイトを意味している．

homo/hetero サイトの内訳（右）．Group 2 に属する株の hetero SNV の 2 つのアレルのうちの片方を Group 1 に属する株が支持しているサイトを「一致するサイト」，それに対して，Group 2 の hetero SNV とは異なるアレルを Group 1 が支持しているサイトを「矛盾するサイト」と定義している．

5.3 仮説の検証（homo/hetero サイト）

5.3.1 方法

2 倍体（3 倍体）である Group 2 に属する株の Sc type において隣接する hetero SNV の連鎖関係を解き，2 本のハプロタイプ別ゲノムの構築（phasing）を試みた．SNV 検出により hetero SNV のローカスとそのアレルの情報を取得できるが，ハプロタイプの各ローカスにおいて hetero SNV のどちらのアレルを持っているか，というアレルの連鎖関係まで知ることはできない（図 5-6 A）．しかし，多倍体のゲノムをシークエンスした場合においてもシークエンスリードはいずれかのハプロタイプに由来するため，シークエンスリードのアレルに従って隣接する hetero SNV の連鎖関係を明らかにすることで，その領域でハプロタイプの構築が可能である（図 5-6 B, C）．

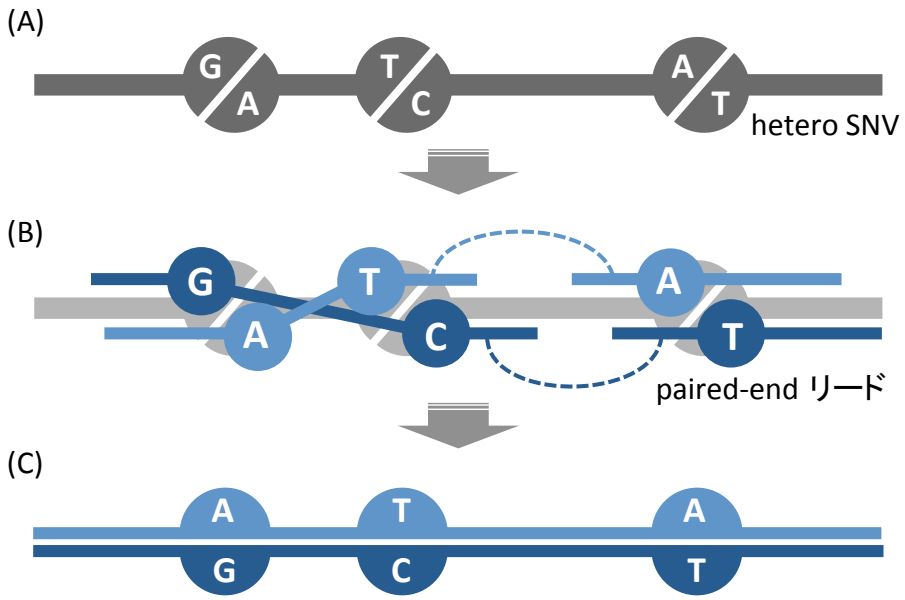


図 5-6 paired-end リードを使った phasing の模式図

(A) SNV 検出により hetero SNV のローカスとアレルは分かるが，ハプロタイプ上でのアレルの連鎖関係の情報は得られない．(B) シークエンスリードの配列をもとに隣接する hetero SNV のアレルの組み合わせを明らかにする（phasing）．ペアリードを考慮することで，隣接する hetero SNV 間の距離がリード長より離れている場合も連鎖関係と解くことが可能になる．2 つのハプロタイプ（図中の青色と水色）で連鎖関係を支持するリンクがそれぞれ存在する場合のみ，phasing を行う (C) phasing の最終結果．

はじめに，paired-end リードを用いて隣接する hetero SNV の連鎖関係を解く方法

を示す. 隣接する 2 つの hetero SNV のローカスを $locus_a$ と $locus_b$, それぞれの hetero SNV のアレルを $allele_{a0}$, $allele_{a1}$, $allele_{b0}$, $allele_{b1}$ とする．シークエンスリードが支持するアレルの組み合わせは 4 通り（$allele_{a0}$ - $allele_{b0}$, $allele_{a0}$ – $allele_{b1}$, $allele_{a1}$ – $allele_{b0}$, $allele_{a1}$ – $allele_{b1}$）が考えられる（図 5-7 A）．ハプロタイプの配列としては, $allele_{a0}$ - $allele_{b0}$ と $allele_{a1}$ - $allele_{b1}$（$links_{ab0}$）と，$allele_{a0}$ – $allele_{b1}$ と $allele_{a1}$ – $allele_{b0}$（$links_{ab1}$）の 2 つの組み合わせが考えられるので，シークエンスリードから $locus_a$ と $locus_b$ 間でこの 2 つのどちらのリンクを持っているのかを明らかにした（図 5-7 B, C）．まず，$locus_a$ と $locus_b$ を含むすべてのリード（ペアリード）からそれぞれのリードが支持するリンクを特定した．具体的には，リファレンスゲノムへのマッピング結果（sam format）及び hetero SNV のポジションとアレルの情報（vcf format）をもとに，$locus_a$ と $locus_b$ を含むリードを抽出した．sam format 中の CIGAR から，マップされたリードのリファレンスゲノムに対する InDel を考慮し，シークエンスリード上の $locus_a$ と $locus_b$ のポジション特定し, 配列からそのアレルを取り出した. $locus_a$ と $locus_b$ 間のアレルの組み合わせごとにリンクを支持するリードの本数を集計した. その後, $links_{ab0}$ と $links_{ab1}$ のうちリンク数が多い方を major links, 他方を minor links と定義したときに, major links 中の 2 つのリンクが 2 本ずつ以上存在し, かつ, minor links の占める割合が全リンクの 10%未満であるときに，$locus_a$ と $locus_b$ において phasing を実施した．すべての隣接する hetero SNV 間でこの操作を行い，隣接する hetero SNV の連鎖関係が 5 つ以上連続して解けた領域をブロックとして，結果を出力した.

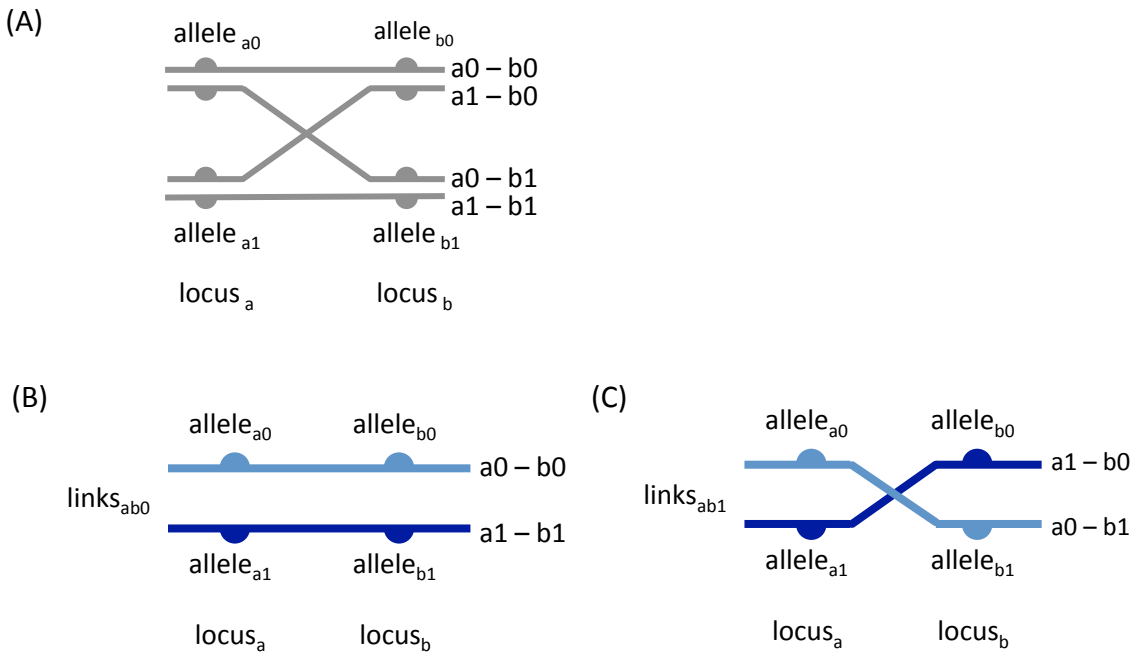


図 5-7 phasing の詳細

(A) $locus_a$ と $locus_b$ を含むシークエンスリードが取りうるアレルの組み合わせは

4 通り考えられる. 隣接する locus$_a$ と locus$_b$ 間で (B) と (C) のどちらのリンクを支持するリードが多いかを判定する. これをすべての隣接する hetero SNV 間で行い, phasing をする.

さらに, W34/70 株においては mate-pair リードを用いて paired-end で解いたブロック間の連鎖関係を解き, より広範囲の phasing を実施した (図 5-8). Paired-end リードのみの場合, 理論上, 隣接する hetero SNV 間の距離がインサートサイズ以上になると, 連鎖を解くことができない. しかし, インサートサイズの大きいペアリードを利用することで, 距離の離れた hetero SNV 間の連鎖の情報を得ることが可能になる. Mate-pair リードの場合も, paired-end リードの時と同様に, 2 つ以上の hetero SNV を含むペアリードから, そのアレルの組み合わせと支持するリードの本数を集計した. この時, 隣接するブロックの連鎖だけでなく, すべてのブロック間のリンクを利用し, 隣接するブロックの連鎖を解いた (図 5-8 B).

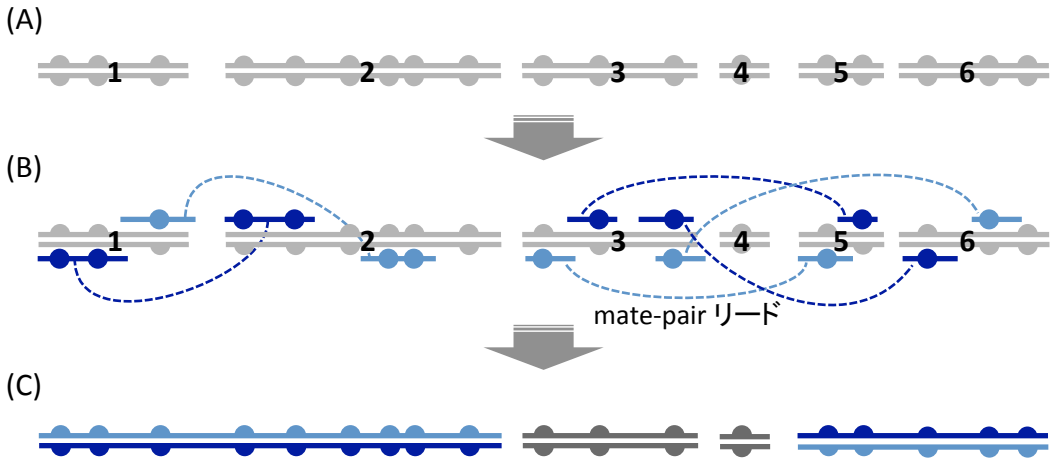


図 5-8 mate-pair リードを使った phasing の模式図

(A) pair-end リードから得られた phasing の結果. ここでは例として 6 つのブロックを示す. (B) mate-pair リードを使ってブロック間のハプロタイプ上における連鎖関係を解く. 隣接していないブロックの連鎖の情報も利用する. ここではブロック 3 – 5, 3 – 6 のリンクから相対的に 5 - 6 の連鎖関係が明らかになる. (C) phasing の最終結果. mate pair の情報から新たにブロック 1 – 2 と 5 – 6 の間の連鎖関係が得られた.

以上の解析により得られたハプロタイプが正しいものであるかを確認した方法を以下に説明する. 今回はもっとも長く phasing できた領域 2 箇所に対して確認を行った. まず, 対象とする領域にマップされた paired-end リードを回収し, 上記で示した方法により構築したハプロタイプのアレルに従ってリードをハプロタイプ a, ハプロタイプ b に分類した. 具体的には, ハプロタイプ間での変異 (hetero SNV) をアンカーと

して，ペアリードのハプロタイプを判定した．この時，アンカーとなるアレルを含まないペアリードやアンカーがハプロタイプと矛盾するペアリードは分類から除外した．そして，(i) 対象領域にマップされた全リード ; (ii) – a ハプロタイプ a ; (ii) – b ハプロタイプ b の 3 つのリードの集合に対して Platanus assembly (-u 0) でアセンブルを実行した．ここで, -u とはヘテロ接合性由来の配列（相同染色体間の変異を含む領域）の統合する時に許容する配列間の違いを指定するパラメータであり，デフォルトでは 0.1 となっている．-u 0 でアセンブルを実施するとヘテロ接合性由来の配列が構築された時に配列の統合を行わず，それぞれの配列を出力する．つまり，(i) マップされた全リードを使って-u 0 でアセンブルした場合は，相同染色体間の変異を含むため複数本の配列が出力される（図 5-9 i）．一方，正しく phasing ができた場合，(ii) - a, (ii) - b はそれぞれのハプロタイプに由来するリードの集合であり，相同染色体間の変異を含まない．このため，(ii) – a, b に対して-u 0 でアセンブルを実行した場合，領域長とほぼ等しい長さの 1 本の配列が得られることが予想される（図 5-9 ii）．ただし，phasing が誤っている場合，リードが十分に回収できない，あるいは，相同染色体間の変異を含むことが原因で，(ii) – a, b においても 1 本の配列が構築できず，複数の配列が出力されると考えられる．以上の方法により phasing が正しく行われたことを確認して，以降の解析を行なった．

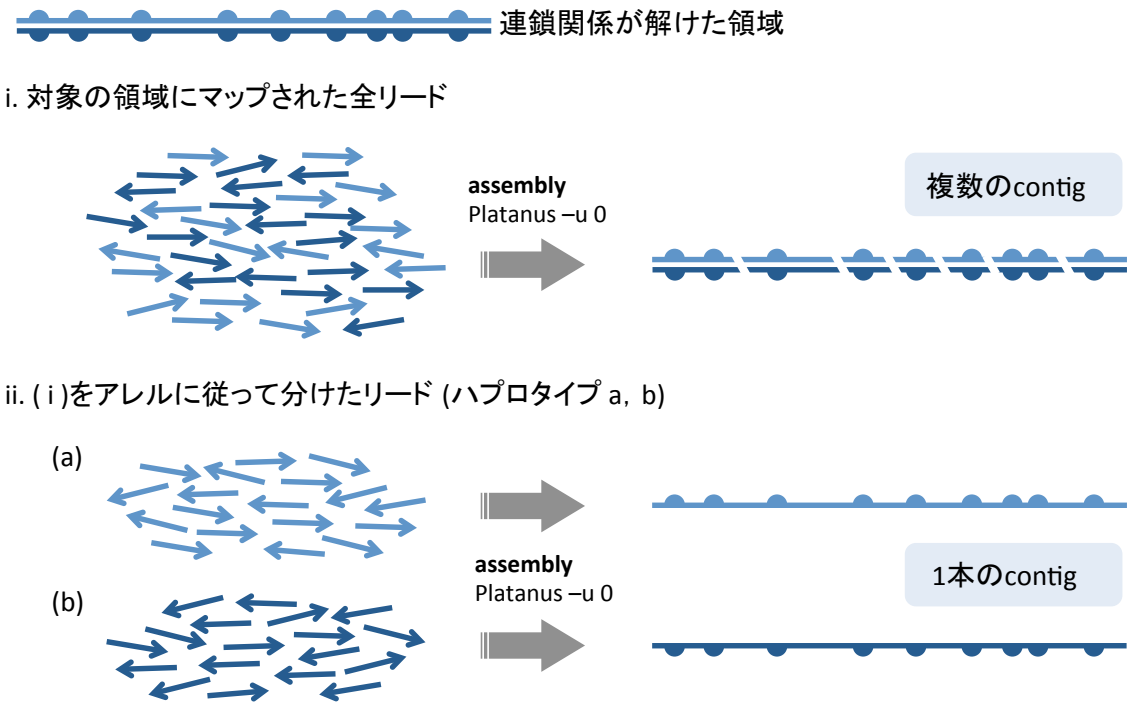


図 5-9 phasing の評価方法の模式図

(i) Phasing したある領域にマップされたリードを回収し，アセンブルを行う．このときのアセンブルのパラメータでは，hetero SNV の領域で配列が分断され，複数本の contig が出力される．(ii) phasing から得られたハプロタイプのアレルをアンカーとして，リードをハプロタイプ a と b に分けてアセンブルをする．Phasing が正しく行われていれば，1 本の contig がそれぞれ構築されることが予想される．

Group 1 に属する 5 株のコンセンサス配列と phasing 後の W34/70 株（Group 2）のハプロタイプをブロックごとに比較し，Group 1 株と Group 2 株の全ゲノムレベルにおける系統解析を試みた．Group 1 に属する 5 株は Sc type において染色体欠失が確認されており，5 株すべてが保持している染色体が少ないため，W34/70 株のハプロタイプと比較可能領域を確保するために Group 1 のコンセンサス配列を用意した上で比較することとした．はじめに，Group 1 のコンセンサス配列の構築を行なった．Group 1 の 5 株のうち少なくとも 1 つの株で染色体が存在する領域を対象にし，複数の株で保存されている染色体では多数の株が支持するアレル（メジャーアレル）をコンセンサス配列とした．メジャーアレルが存在しない場合（5 株すべてで欠失している，メジャーアレルが 2 つ以上存在する，等）は，比較対象から除外した．また，比較に用いた W34/70 株のハプロタイプは，5 つ以上連続する hetero SNV の連鎖関係が明らかになった領域とした．W34/70 株を使用した理由は，一つ目に mate-pair ライブラリがあるため phasing できている領域が多い，2 つ目に Group 2 株の中で hetero SNV 率が最

も高かったためである．

ブロックごとにGroup 1のコンセンサス配列とW34/70株のハプロタイプを比較し，Group 1のコンセンサス配列と相対的に相同性の高いハプロタイプを”W34/70_a”，他方を”W34/70_b”として，すべてのブロックにおいてハプロタイプの相同性を集計した．比較のために，同じ領域に対してアレルをランダムに分けたハプロタイプを作り，同様にして，Group 1との相同性にしたがってW34/70_a，W34/70_bに分類し，ブロックごとのGroup 1のコンセンサス配列との相同性を集計した．この試行を100回行ない，その平均を求めた．

最後にW34/70株のハプロタイプ（W34/70_a，W34/70_b），Group 1のコンセンサス配列，*S. cerevisiae* S288cの系統解析を行なった．それぞれのブロックで得られたハプロタイプ（W34/70_a，W34/70_b）の集合をW34/70株のハプロタイプと仮定し，Group 1のコンセンサス配列とS288c間におけるSNVに基づく系統距離をphymlで計算し，系統樹を作成した．

5.3.2 結果

Group 2に属する5株のSc typeにおいてphasingを実施し，連続する5つ以上のhetero SNVで連鎖関係が解けた領域をブロックとした時に，ブロック中に含まれるhetero SNV数，ブロック数，ブロック長の合計を表 5-1に示した．phasingできた領域の割合は5.3～15.4%と一部分ではあるが，5株すべてにおいてSc typeから検出されたhetero SNVの79%以上がブロックに含まれた．Hetero SNVがSc typeの染色体上に不均一に分布しているため，hetero SNVが高密度で存在している領域を中心にphasingすることができたと考えられる．W34/70株は*S. pastorianus* 5株中もっともhetero SNVの密度が高いため，phasingされるhetero SNV数が多いと予想していたが，ブロック中に含まれるhetero SNVの割合は5株中，最も低い値に留まった．隣接するhetero SNV間の距離の分布と距離ごとのphasingできた割合を図 5-10に示した．この結果よりphasingできなかった領域のほとんどがインサートサイズである600 bpよりも距離の離れたhetero SNV間であることがわかる．従って，計算上可能な範囲でのphasingは行うことができたと考えられる．株ごとの結果は図 5-11に示した．

表 5-1 Group 2 に属する 5 株における phasing の結果

株	使用ライブラリ	Sc type内の hetero SNV数	ブロック中に含まれる hetero SNV数 (%)	ブロック数	ブロック長の合計 (bp) (%)
W34/70	PE	10,900	8,369 (76.8)	595	969,637 (8.0)
W34/70	PE, MP	10,900	9,357 (86.6)	337	1,858,964 (15.4)
CBS1483	PE	6,459	5,102 (79.0)	344	636,402 (5.3)
CBS1484	PE	7,269	6,020 (82.8)	400	792,355 (6.6)
CBS2156	PE	5,871	4,928 (83.9)	271	637,705 (5.3)
CBS5832	PE	8,022	6,630 (82.6)	383	867,886 (7.2)

ブロック中に含まれる hetero SNV 数の () 内の値は，その株の Sc type に存在する hetero SNV 中の割合（%）を示している．ブロック長の合計の () 内の値は，ゲノム中に占める割合（%）を示している．

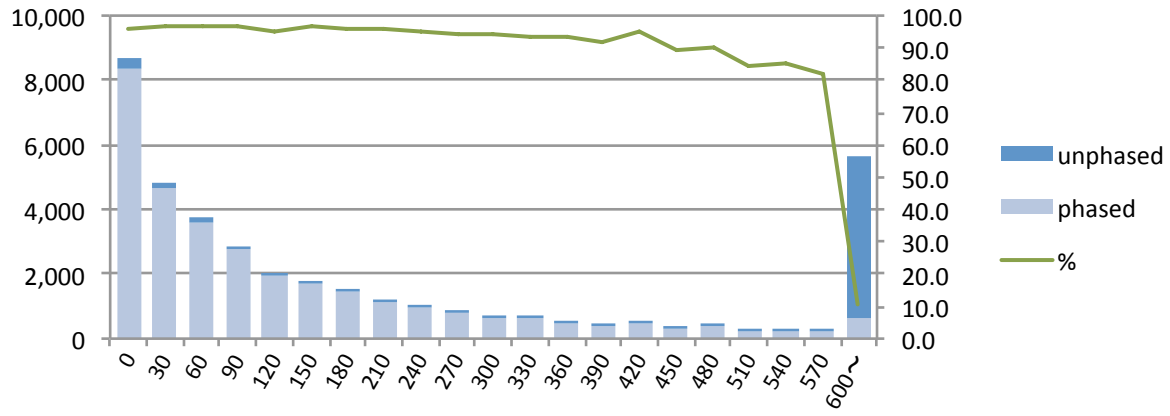


図 5-10 隣接する hetero SNV 間の距離の分布と phasing の割合

S. pastorianus Group 2 に属する 5 株の隣接する hetero SNV 間の距離の分布と phasing ができた割合．棒グラフは隣接する hetero SNV 間の距離の分布を示している．横軸は隣接する hetero SNV 間の距離（bp），縦軸（左側）は箇所数．隣接する hetero SNV のうち phasing できたものを水色，phasing できなかったものを青色としている．折れ線グラフはそれぞれの距離における phasing できた箇所数の割合（右側の縦軸 %）を示している．株ごとの結果を図 5-11 に示した．

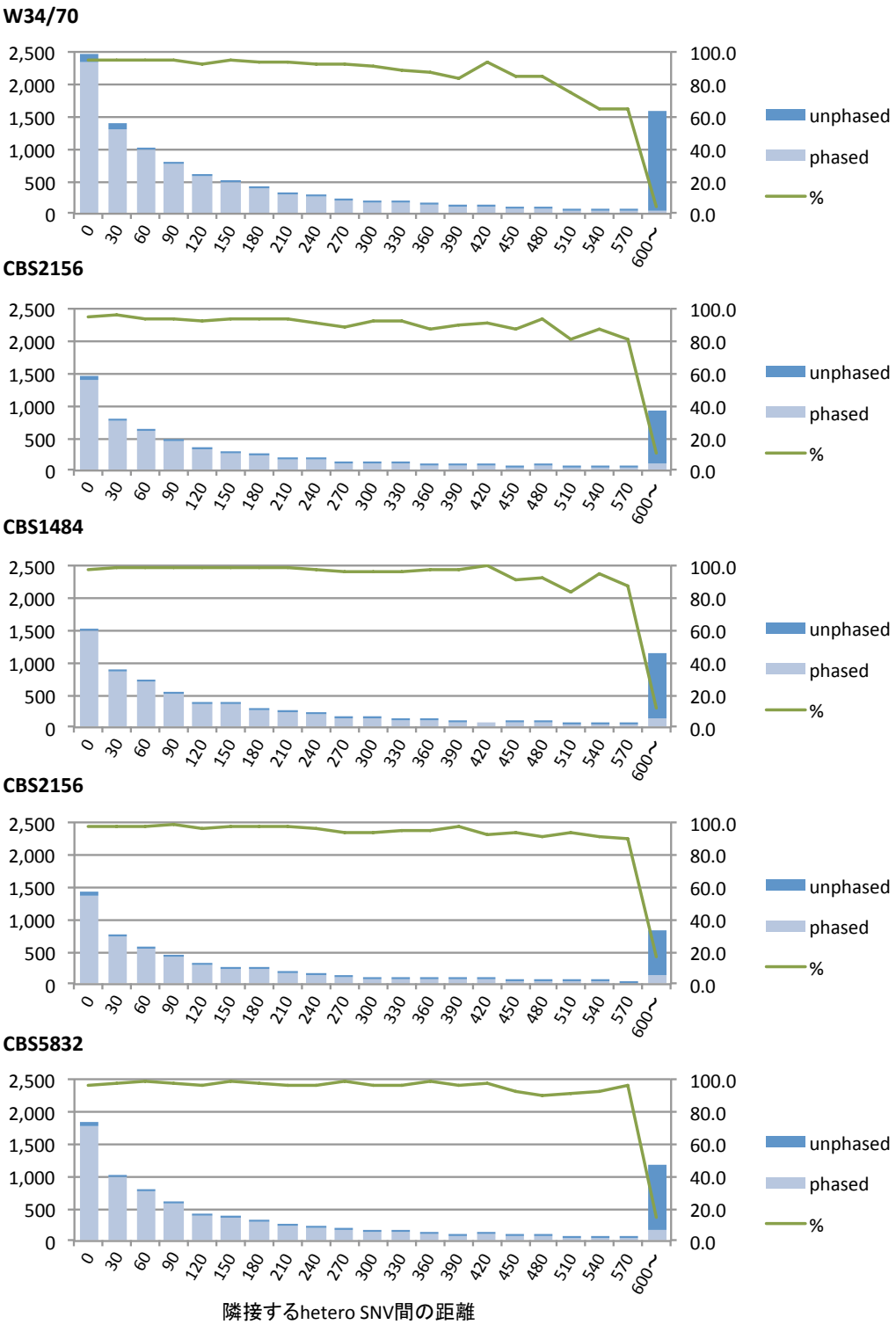


図 5-11 隣接する hetero SNV 間の距離の分布と
paired end リードを用いて phasing できた割合の分布

ブロック長の長い 2 つの領域でハプロタイプごとのアセンブリを実施し，phasing の正確性の評価を行なった．対象とした領域は，CBS2156 株の染色体 XV の 270,132～293,126 bp 及び CBS5832 株の染色体 VIII の 163,765～185,587 bp で，それぞれの hetero SNV の数と密度は 183 個（0.80%），173 個（0.82%）であった．この 2 箇所の領域について，(i) 対象領域にマップされた paired-end リード，(ii) phasing により構築したハプロタイプのアレルに従って分類したリードの本数と総塩基数をそれぞれ表 5-2 左に示した．マップされたリードのうち 95%以上をハプロタイプのアレルに従ってハプロタイプ a または b に分類することができた．ハプロタイプに分けることができなかったペアリードは，アンカーとなるアレルを含んでいない，あるいは，phasing 結果と矛盾するアレルの組み合わせを持っていたと考えられる．ヘテロ接合性領域の統合を無効化した Platanus によるアセンブルを実行した結果，20 kbp 以上にわたり phasing された 2 つの領域についてそれぞれ 1 本の配列を構築することができた（表 5-2）．この結果は，本手法によって，それぞれのハプロタイプに由来するリードに分類できたことを意味しており，phasing は正しく行われたと考えられる．

上の 2 つの領域に対して，それぞれ 3 つのリードの集合に対する 32-mer 頻度分布を図 5-12 に示した．対象領域にマップされた全リードから得られた 32-mer 頻度分布からは 2 つのピークが観察された．2 つのピークのうち出現回数の少ない方のピークは，相同染色体間の変異により 2 種類の 32-mer が得られたローカスに由来し，他方のピークは相同染色体間の変異の存在しないローカスに由来すると考えらえる．それに対して，それぞれのハプロタイプに分類されたリードの集合から得られた 32-mer 頻度分布は 1 つのピークのみが観察され，それぞれのハプロタイプに由来するリードに分類できたことが示唆される．

この 2 つの領域の *S. cerevisiae* S288c ゲノムに対するマッピング結果を IGV で可視化したものを図 5-13 及び図 5-14 に示す．phasing 前の回収した全リードのマッピング結果では，hetero SNV が領域中に多数確認されたが，ハプロタイプに分けた後のそれぞれのマッピング結果では hetero SNV は確認されず，全リード中で hetero SNV が検出されたローカスにおいて homo SNV を示していることからも phasing が正しく行なわれたと判断した．

表 5-2 phasing 結果に基づいたハプロタイプごとのアセンブリ結果

例1) CBS2156 chr. XV: 270,132…293,126 bp183 hetero SNV (0.80%)

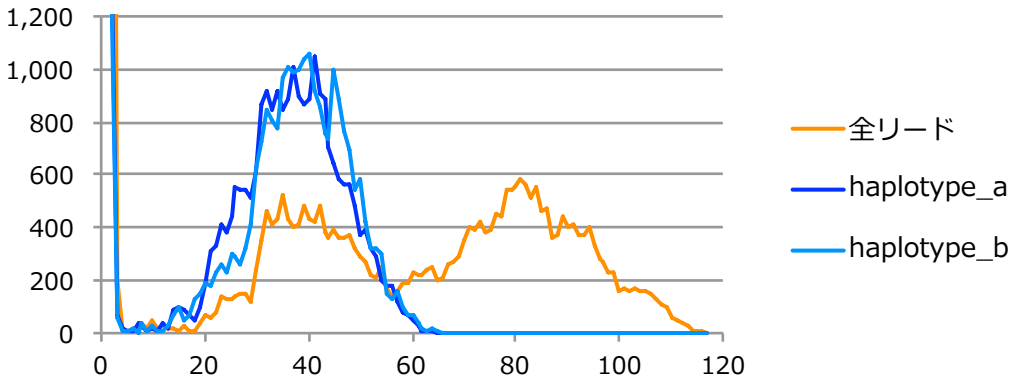
	回収したリード		アセンブル結果 (platanus –u 0)	
	本数	総塩基数 (bp)	本数	総塩基数 (bp)
i 全リード	8,834	2,253,578	166	58,912
ii-a ハプロタイプ_a	4,152	1,054,548	1	23,172
ii-b ハプロタイプ _b	4,244	1,092,203	1	23,141

例2) CBS5832 chr. VIII: 163,765…185,587 bp 178 hetero SNV (0.82%)

	回収したリード		アセンブル結果 (platanus –u 0)	
	本数	総塩基数 (bp)	本数	総塩基数 (bp)
i 全リード	7,622	1,932,996	138	44,185
ii-a ハプロタイプ_a	4,070	1,038,260	1	22,428
ii-b ハプロタイプ _b	3,198	8111,740	1	22,331

phasing の正確性の評価を行なった 2 領域について，(i) 対象の領域にマップされた全リード，(ii) a, b phasing により構築したハプロタイプ a と b のアレルに従って分類したリードの本数と総塩基数，アセンブル結果を示した．

例1) Sc type chr. XV: 270,132...293,126 bp (CBS2156)



例2) Sc type chr. VIII: 163,765…185,587 bp (CBS5832)

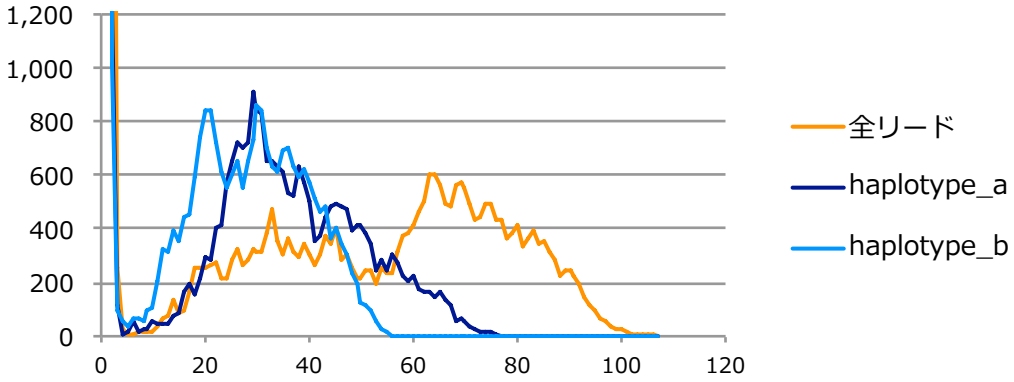


図 5-12 phasing 前後の 32-mer 頻度分布

例 1，2 について，phasing 前（全リード），phasing 後（haplotype_a, haplotype_b）におけるそれぞれのリードの集合の 32-mer 頻度分布.

例1) Sc type chr. XV: 270,132...293,126 bp (CBS2156)

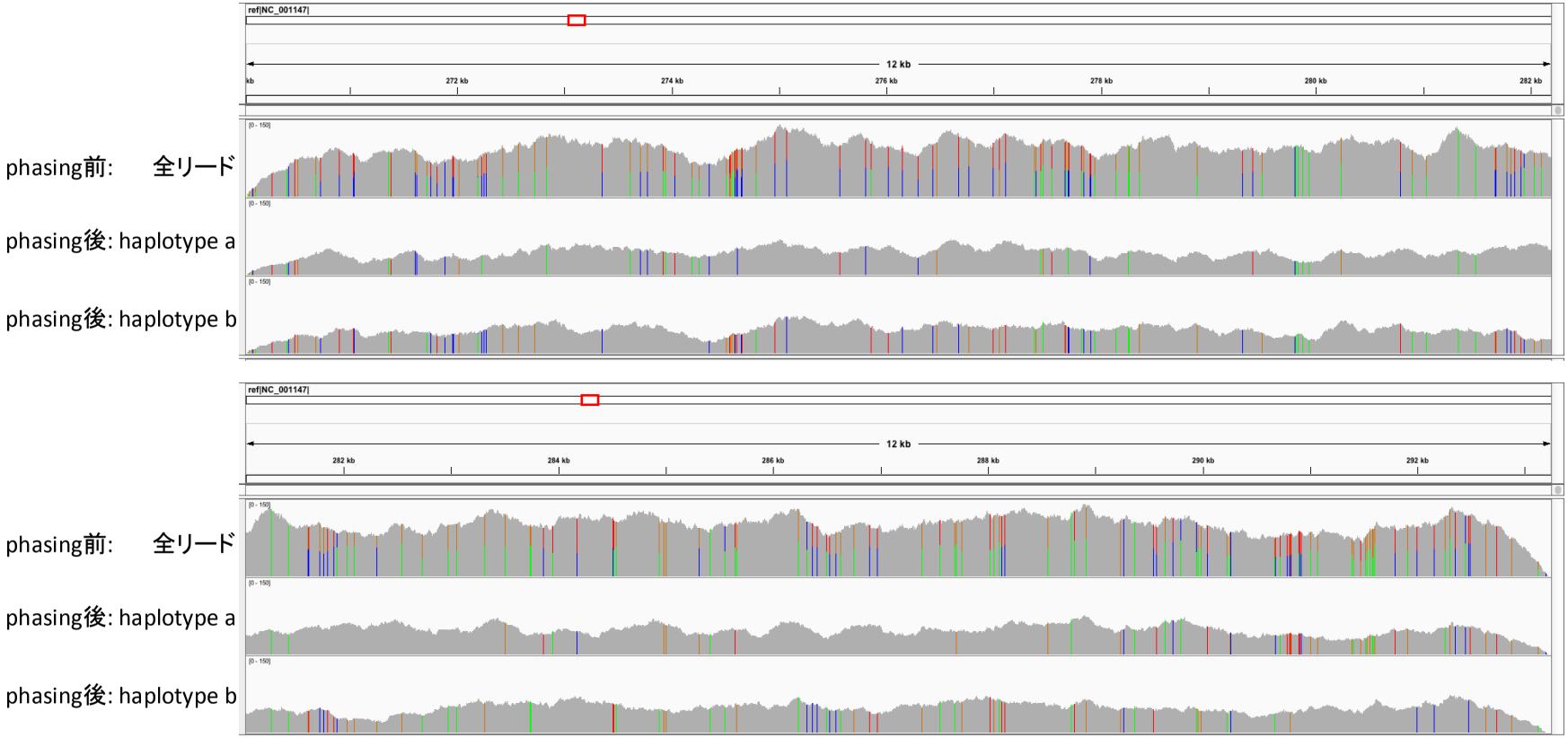


図 5-13 IGV による phasing 前後のリードの様子（例 1 CBS2156 株）

表 5-2 例 1 におけるマップされた全リードの集合（全リード），phasing の結果に従って分類したリードの集合（haplotype a, b）の *S. cerevisiae* S288c に対するマッピング結果

例2) Sc type chr. VIII: 163,765…185,587 bp (CBS5832)

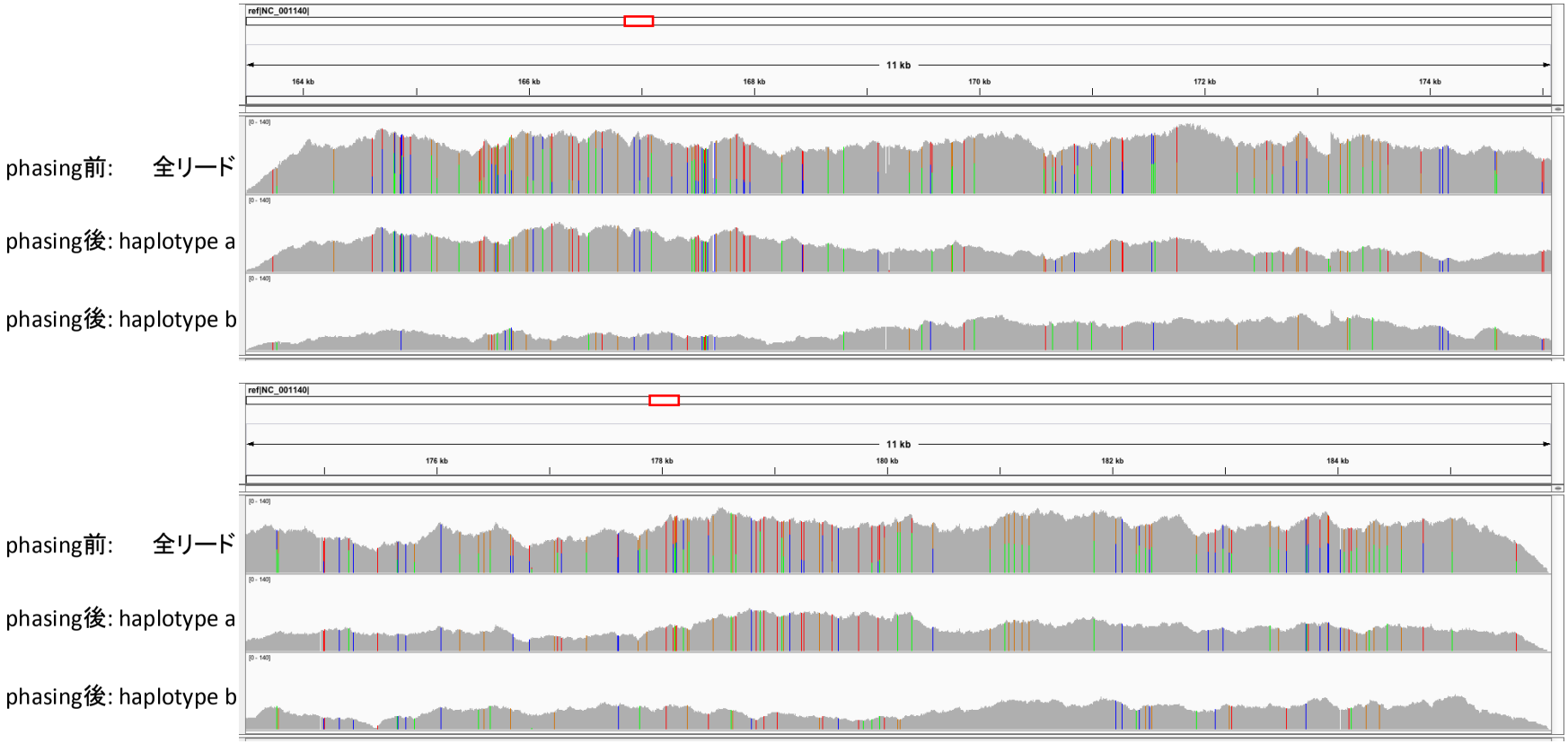


図 5-14 IGV による phasing 前後のリードの様子（例 2 CBS5832 株）

表 5-2 例 2 におけるマップされた全リードの集合（全リード）および phasing の結果に従って分類したリードの集合（haplotype a, b）の *S. cerevisiae* S288c に対するマッピング結果

続いて，W34/70 株において構築したハプロタイプのうち，Group 1 のコンセンサス配列との相同性が高いハプロタイプを W34/70_a，他方を W34/70_b と分類した時の Group 1 との相同性の分布を図 5-15 A に示す．ここでの相同性は，ブロック長に対する Group 1 のコンセンサス配列と各ハプロタイプの一致する塩基数の割合で算出している．ハプロタイプ W34/70_a と Group 1 のコンセンサス配列との相同性の分布では 99.9%にピークが存在し，ブロック全体で Group 1 に対して高い相同性があったが，ハプロタイプ W34/70_b との相同性は 99.5%ほどであり，W34/70 株のハプロタイプ間で Group 1 との相同性に違いが現れた（図 5-15 A）．この結果と比較するために，対象としているブロックの中の hetero SNV をランダムに分け（random phasing），同様に Group 1 との相同性を調べた結果を図 5-15 B に示した．分布は，100 回の施行の平均値をプロットし，エラーバーは標準偏差を示している．random phasing においても，相対的に相同性の高いハプロタイプを W34/70_a としているが，ハプロタイプ間に大きな違いが見られなかった．この結果から，phasing によって得られた W34/70 株のハプロタイプのうちの片方は Group 1 のハプロタイプと有意に相同性が高いことが明らかになった．さらに，*S. cerevisiae* S288c に対しても同様に phasing によって得られたハプロタイプを S288c 株との相対的な相同性に従って W34/70_a と W34/70_b に分類し，それぞれの相同性の分布を確認したが，S288c 株に対する W34/70 株のハプロタイプ間の相同性の違いは現れなかった（図 5-15 C）．この結果から *S. cerevisiae* S288c と一致するアレルは W34/70 株の片方のハプロタイプに連続して存在しないことが示唆された．W34/70 株以外の Group 2 に分類される 4 株についても同様に，ハプロタイプの片方において Group 1 のコンセンサス配列との高い相同性が確認された（図 5-17 左）．

Group 1 との相同性にしたがって分類したブロックごとのハプロタイプ（W34/70_a, W34/70_b）を統合した配列を W34/70 株のハプロタイプと仮定した時の，Group 1 のコンセンサス配列，W34/70 株のハプロタイプ（W34/70_a，W34/80_b），S288c 株の系統樹を図 5-16 に示した. Group 1 とハプロタイプ W34/70_a が系統的に近く, Group 1 に分類される株の Sc type のハプロタイプと Group 2 に分類される W34/70 株の Sc type のハプロイドは共通祖先に由来する可能性が示唆された．Group 2 に属する他 4 株の系統樹は，図 5-17 右に示した．

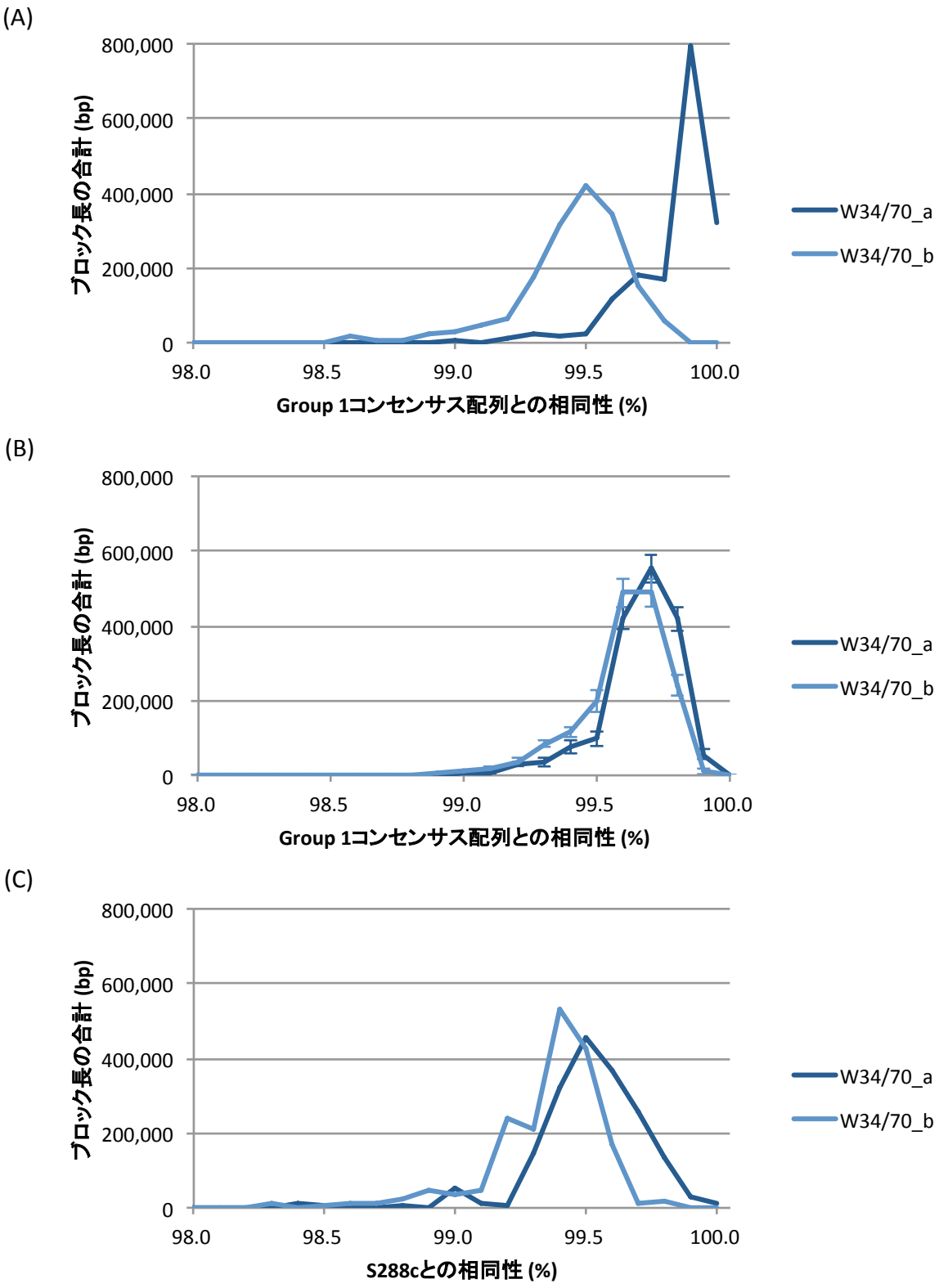


図 5-15 W34/70 株のハプロタイプ間の相同性の違い

(A) Group 1 のコンセンサス配列と W34/70 株の phasing した領域（W34/70_a, W34/70_b）との相同性．Group 1 との相同性が相対的に高い方を W34/70_a としている．(B) (A) と同じ領域に対して，hetero SNV のアレルをランダムに 2 つ

(W34/70_a, W34/70_b) に分けた時のそれぞれの相同性の分布．100 回の試行の平均値をプロットしている．(C) *S. cerevisiae* S288c と W34/70 株のハプロタイプとの相同性．S288c との相同性が相対的に高い方を W34/70_a としている．

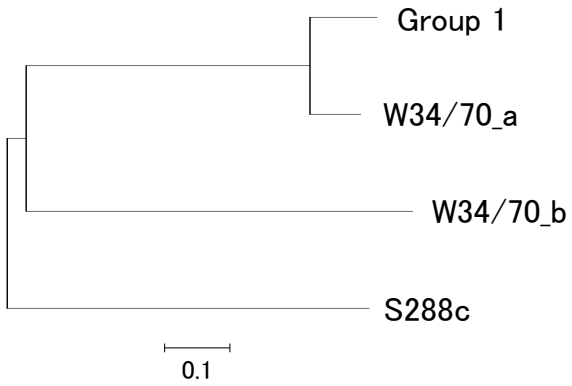


図 5-16 Group 1 と W34/70 株のハプロタイプとの系統樹

Group 1 のコンセンサス配列, W34/70 株のハプロタイプ（W34/70_a, W34/70_b）及び *S. cerevisiae* S288c における系統関係．

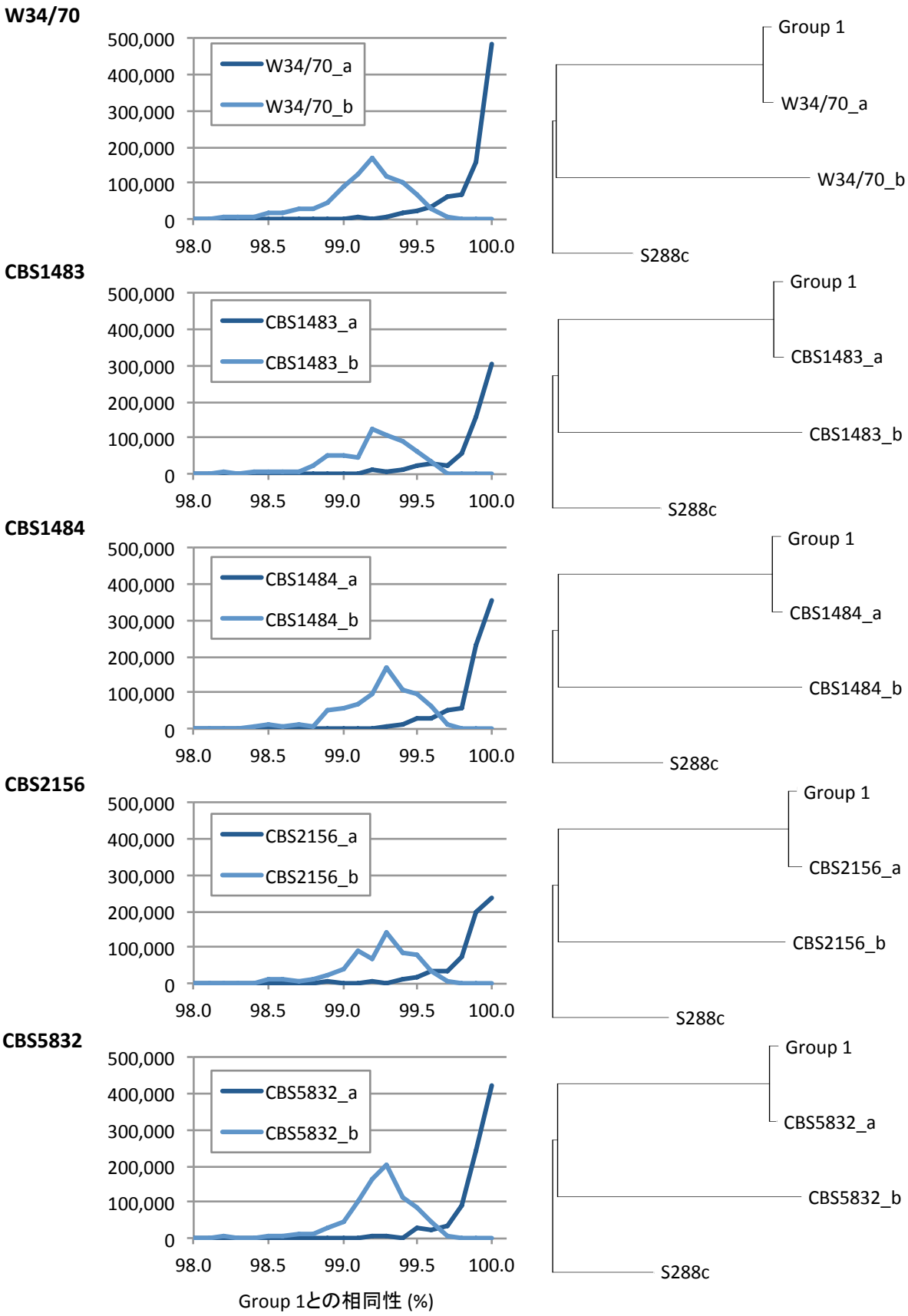


図 5-17 Group 1 と Group 2 のハプロタイプとの相同性と系統樹

左：Group 2 に属する 5 株について paired-end リードに基づいた phasing によ

り得られたハプロタイプと Group 1 のコンセンサス配列との相同性の分布図. 横軸は Group 1 のコンセンサス配列との相同性（%）, 縦軸はブロック長の合計（bp）を示している. ブロックごとに Group 1 のコンセンサス配列との相同性が相対的に高いハプロタイプを a，他方を b としている．右：Group 2 に属する 5 株から得られたハプロタイプと Group 1 のコンセンサス配列，*S. cerevisiae* S288c の系統樹.

最後に，phasing の評価を行った 2 つの領域について Group 1 に属する株と比較した結果を示す.

- 例 1 CBS2156 株（染色体 XV: 270,132…293,126 bp）

CBS2156 株のハプロタイプのアレルに従って Group 2 に属する他 4 株のリードを分類し，それぞれをアセンブルしたところ，CBS5832 株を除く 3 株（CBS1483 株，CBS1484 株，W34/70 株）ではハプロタイプ a と b がそれぞれ 1 本の配列として構築された．一方，CBS5832 株ではハプロタイプ b に対応するリードが回収できず，ハプロタイプ a のみ得られたことから，CBS5832 株についてはこの領域において LOH が起きたと考えらえる．Group 2 に属する 5 株から得られたハプロタイプ，Group 1 に属する株で染色体 XV を保存していた CBS1503 株，CBS1513 株，*S. cerevisiae* S288c の系統樹を図 5-18 に示した.

- 例 2 CBS5832 株（染色体 X: 163,765…185,587 bp）

例 1 と同様に CBS5832 株のアレルに従って, Group 2 に属する 4 株のハプロタイプを構築した. この領域においても Group 1 に属する CBS1503 株と CBS1513 株以外は染色体欠失により配列が得られなかった．Group 1 に属する 2 株（CBS1503, CBS1513），Group 2 に属する 5 株のハプロタイプ，S288c 間の配列に基づいた系統樹を図 5-19 に示した.

図 4-18 (A) において Group 1 と 2 は 2 つのクレードに分けられたが，Group 2 に属する 5 株についてハプロタイプの配列を構築し，系統樹を作成したところ，Group 1 に属する株が含まれるクレードに Group 2 のハプロタイプ a が位置することが確認された.

	株名	倍数性	ハプロタイプ	
Group 1	CBS1503	1	a	-
	CBS1513	1	a	-
	CBS1538	-	-	-
	CBS1174	-	-	-
	CBS2156	-	-	-
Group 2	W34/70	3	a	b
	CBS1483	3	a	b
	CBS1484	2	a	b
	CBS2156	2	a	b
	CBS5832	2	a	–

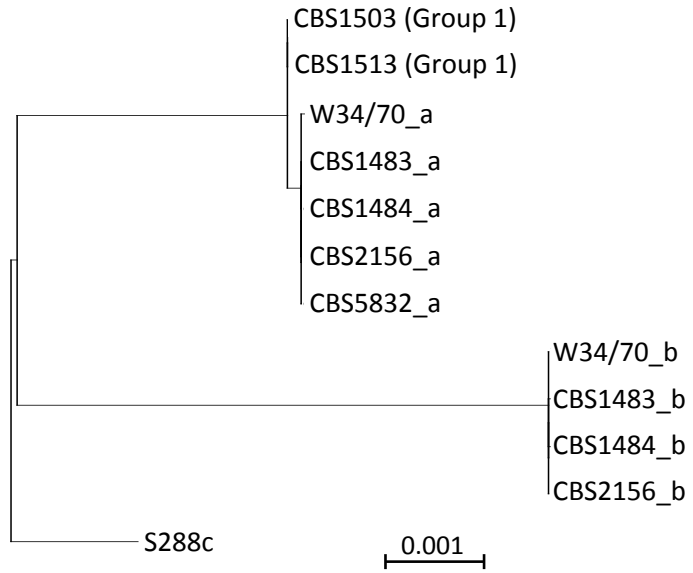


図 5-18 ハプロタイプに基づいた系統樹（例 1）

染色体 XV: 270,132…293,126 bp におけるハプロタイプに基づいた系統樹.

CBS2156 株で phasing された領域に対して Group 2 に属する他 4 株を加え，系統解析を行なった．左：対象領域の倍数性と各株で得られたハプロタイプ．Group 1 に属する株のハプロタイプを a としている．染色体欠失の場合は-で示した．CBS5832 株は 2 倍体であったが，ハプロタイプは a のみ得られた．右：対象領域における系統樹.

	株名	倍数性	ハプロタイプ	
Group 1	CBS1503	1	a	-
	CBS1513	1	a	-
	CBS1538	-	-	-
	CBS1174	-	-	-
	CBS2156	-	-	-
Group 2	W34/70	2	a	b
	CBS1483	4	a	b
	CBS1484	2	-	b
	CBS2156	2	a	b
	CBS5832	2	a	b

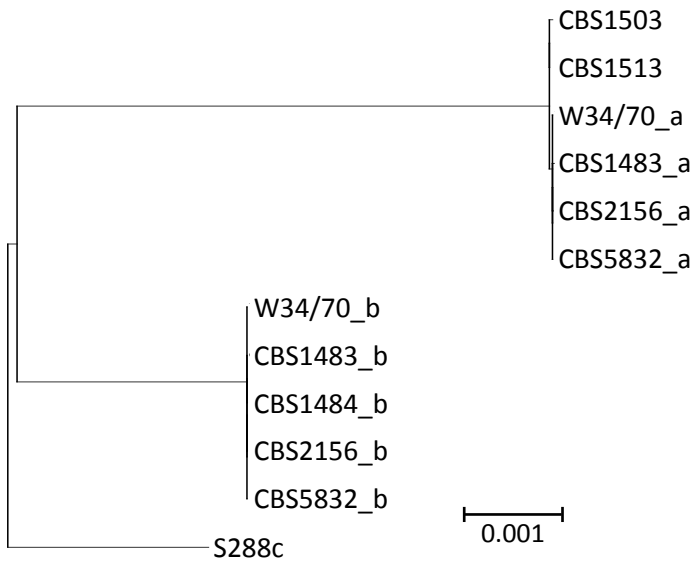


図 5-19 ハプロタイプに基づいた系統樹（例 2）

染色体 X: 163,765…185,587 bp におけるハプロタイプに基づいた系統樹.

CBS5832 株で phasing された領域について，Group 2 に属する他 4 株を加えて系統解析を行なった. 左 : 対象領域の倍数性と各株で得られたハプロタイプ. Group 1 に属する株のハプロタイプを a としている．染色体欠失の場合は-で示した.

CBS1484 株は 2 倍体であったが，ハプロタイプは b のみ得られた．右：対象領域における系統樹.

Group 1 – 2 間に存在する homo/hetero SNV の大半が「一致するサイト」であること注目し，Group 1 と 2 は共通祖先に由来するという仮説を立て，その仮説を検証するために，Group 2 に属する株の S type におけるハプロタイプ別のゲノムを構築し，Group 1 に属する株と比較した. Group 2 に属する 5 株からそれぞれ検出された hetero SNV のうち 79%以上を，連続する 5 つ以上の hetero SNV から構成されるハプロタイプのブロックに分類することに成功した．Group 2 において構築したハプロタイプのブロックと Group 1 の比較結果から，Group 2 に属する株のハプロタイプのうちの片方が，Group 1 に属する株と相同性が高い – つまり，Group 2 に属する株における hetero SNV のうち Group 1 と「一致するアレル」がハプロタイプ上において連続して存在している – ことが明らかになった.

5.4 仮説の検証（homo/homo サイト）

phasing により Sc type においてハプロタイプ別のゲノムが構築できた領域に関しては，Group 2 に属する株のハプロタイプの片方は Group 1 のハプロタイプと相同性が高く，Group 1 と 2 は共通祖先に由来することが示唆された．つまり，Group 1 – 2 間の SNV のうち 1/4 から 1/3 を占める homo/hetero SNV はハプロタイプを考慮したことで仮説を支持する結果が得られた．しかし，残りの Group 1- 2-間に存在する SNV – homo/homo SNV – は何に起因するものであるのか, について Group 2 で起きた LOH と合わせて考察する．*S. pastorianus* の Sc type の祖先種であるエールビール酵母 *S. cerevisiae* は高ヘテロ接合性であり（図 5-20 A），異種交配後の LOH によって， *S. pastorianus* Group 2 に属する株の Sc type において hetero SNV の不均一な分布が観察されると考えられる（図 5-20 B）．Group 1 – 2 は共通のハプロタイプに由来し，Group 2 に属する株において異種交配後のゲノム再構成時に LOH が起きた場合，Group 2 に属する株の Sc type の染色体は大きく 3 つの領域に分類される.

I. LOH が起きなかった領域, つまり Group 2 に属する株において祖先株エールビール酵母に由来する hetero SNV が観察される領域
II. LOH により，hetero SNV のうち Group 1 と異なるアレルが消失し，Group 1 と同じアレルが homo SNV として観察される領域
III. LOH により，hetero SNV のうち Group 1 と同じアレルが消失し，Group 1 と異なるアレルが homo SNV として観察される領域

この 3 つの領域における Group 1 – 2 間の SNV を考えると，まず，領域 I の場合，Group 1 – 2 間において homo/hetero SNV が観察される．この領域については，前の節において Group 2 に属する株のハプロタイプ別ゲノムの構築及び Group 1 との比較から Group 2 のハプロタイプの片方が Group 1 と相同性が高いことが示された．続いて，Group 2 に属する株において LOH が起きた領域について考察する．LOH により消失するハプロタイプはランダムに選択され，領域 II あるいは領域 III のような領域が観察されることになる．領域 II のように Group 1 と異なるアレルが消失した場合，Group 1 - 2 の間の SNV も消失する，つまり，Group 1 – 2 間に SNV が存在しない領域が観察される．一方，領域 III のように Group 1 と同じアレルが消失した領域においては，Group 1 – 2 間に homo SNV が観察される（図 5-20 B）．つまり，現在観察されている Group 1 – 2 における homo SNV は Group 2 の LOH により生じたものであり，系統的に由来が異なるために観察されている homo SNV ではないという可能性が示唆された．このように LOH によって Group 1- 2 間に homo SNV が生じた場合，染色体上に領域 I，II，III がモザイク状に観察されることが予想される．実際に，Group 2 に属する株の hetero SNV の染色体上の分布と Group 1 – 2 間の homo SNV の分布を重ね合わせると，図 5-21 のような分布が得られた．Group 1 – 2 間の homo SNV は染色体上に不均一に存在し，Group 1 – 2 間において変異が存在しない領域も確認された．この結果は，LOH により Group 1 – 2 間に homo SNV が生じたというモデルを支持する結果であり，Group 1 と 2 の Sc type のハプロタイプが共通祖先株に由来するという仮説に矛盾しない．

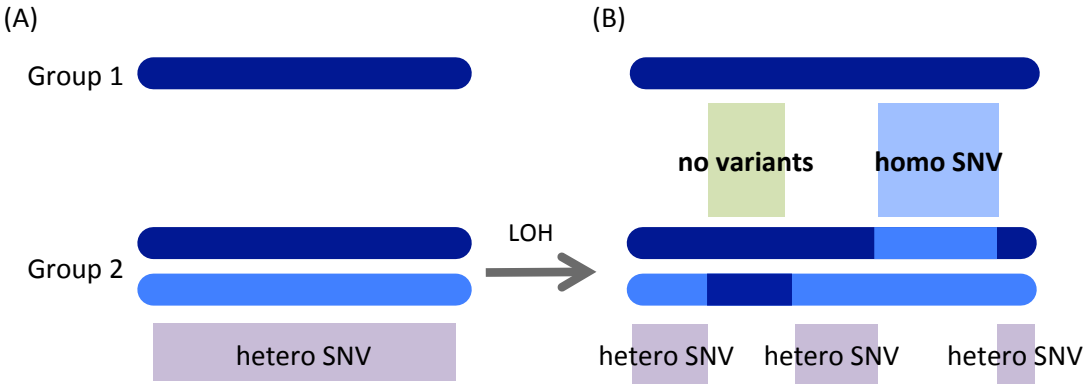
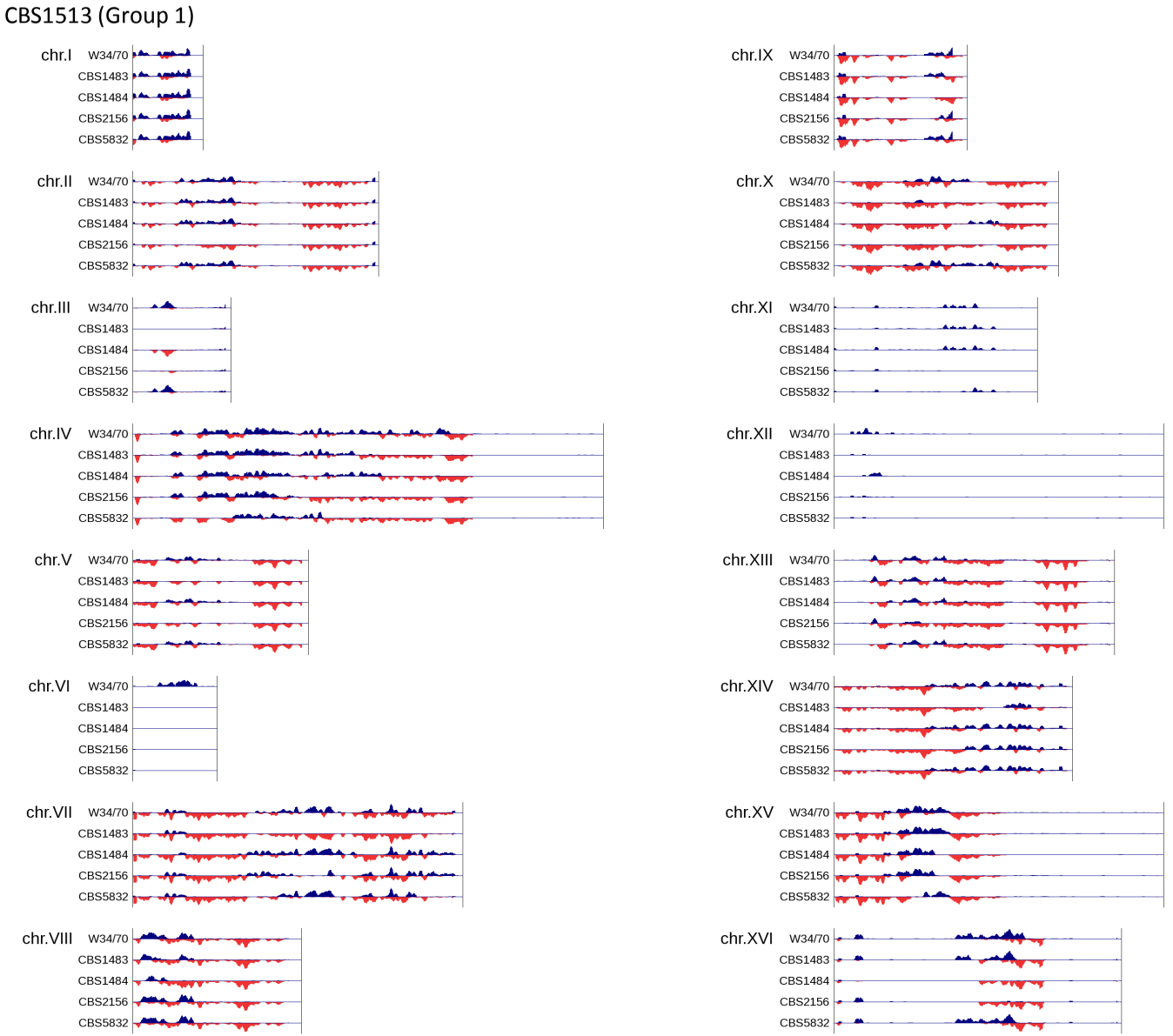
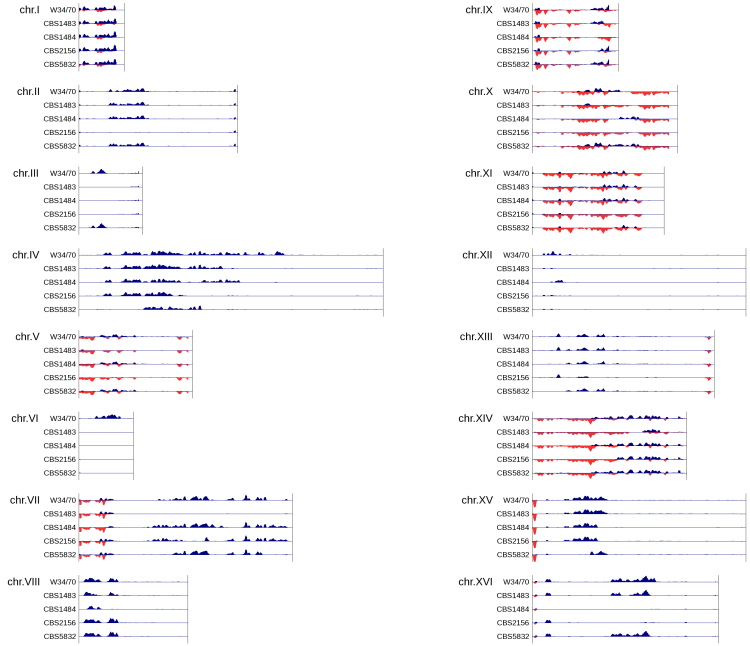


図 5-20 LOH による hetero SNV の不均一化と Group 間の変異

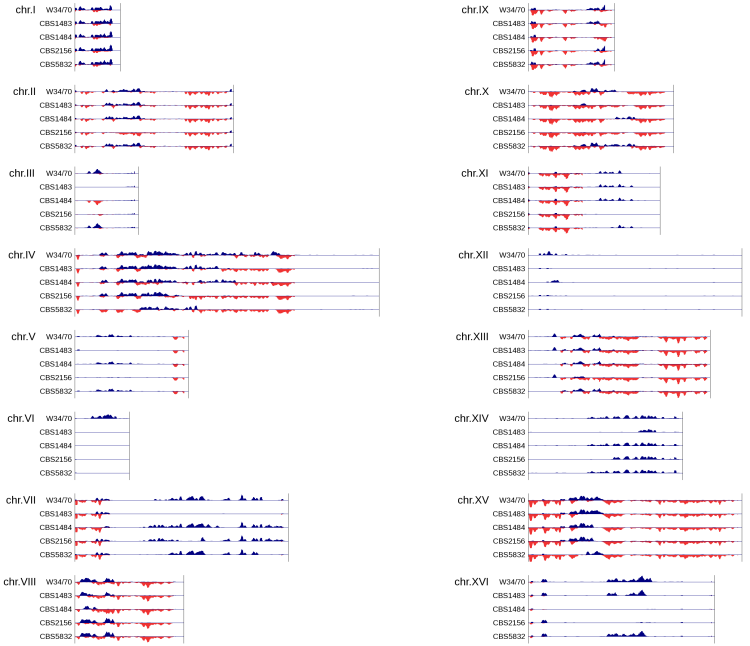
(A) Group 1 Sc type の染色体（1 倍体）を濃青, Group 2 Sc type の相同染色体（2 倍体）を濃青と水色で示した ．Group 間で共通のハプロタイプ（濃青）が存在し，Group 2 Sc type は高ヘテロ接合性である (B) Group 2 で LOH が起きた後の Group 1 と 2 の Sc type の染色体の様子．Group 2 において染色体全体に存在した hetero SNV が LOH により不均一化し，Group 1 と 2 の間で homo SNV が観察されるようになる．


CBS1513 (Group 1)
chr.I
chr.II
chr.III
chr.IV
chr.V
chr.VI
chr.VII
chr.VIII
chr.IX
chr.X
chr.XI
chr.XII
chr.XIII
chr.XIV
chr.XV
chr.XVI
W34/70
CBS1483
CBS1484
CBS2156
CBS5832

CBS1538 (Group 1)



CBS1503 (Group 1)



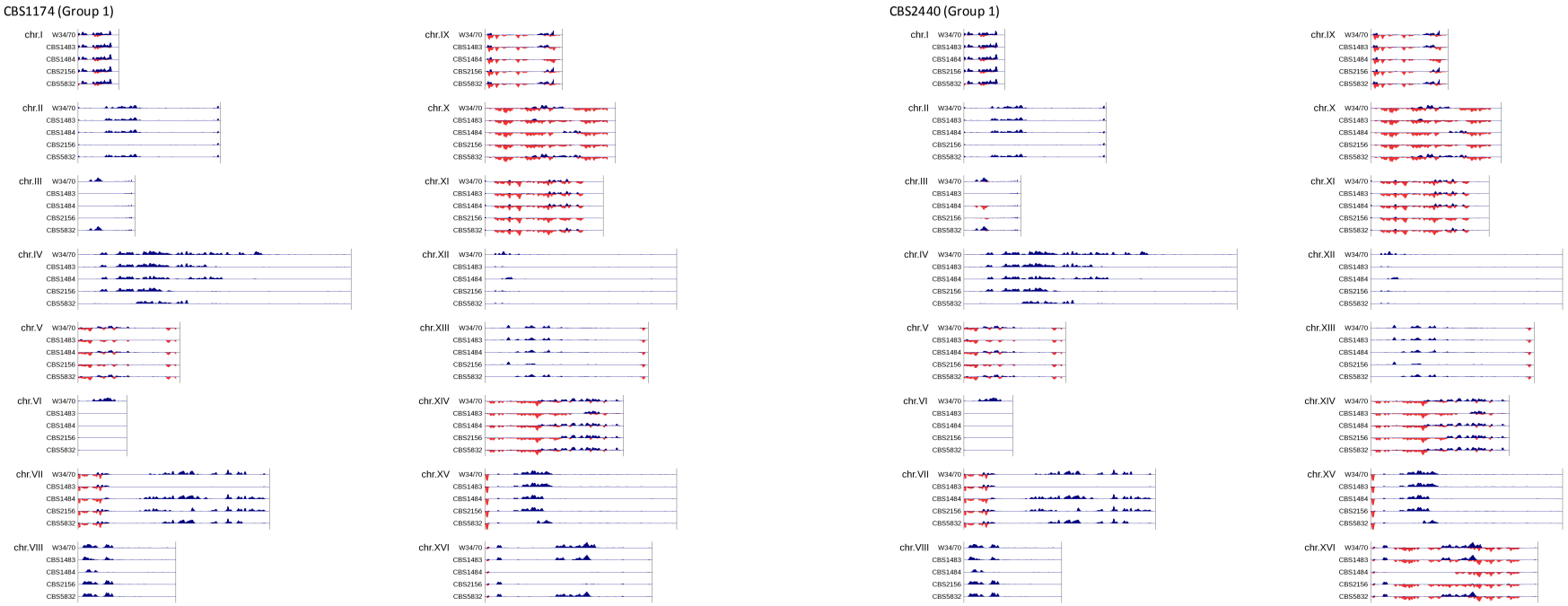


図 5-21 Group 2 に属する 5 株の hetero SNV と Group 1 – 2 間の homo SNV の分布

Group 2 に属する 5 株の hetero SNV の分布（濃青）と Group 1 - Group 2 各株間に存在する homo SNV の分布（赤）. step size 1kbp, window size 10 kbp の平均密度のヒストグラムであり，縦軸はそれぞれ 0 から 100 の範囲を示している.

5.5 考察

本章における解析の結果，Group 1 と Group 2 間で共通の異種交配が 1 回存在していることが強く示唆された．この結論は，両 Group に属する株間にブレイクポイントが一致する染色体間転座が存在すること，Se type において両 Group に属する株間の変異サイト数が少なく，系統的な差異が確認できなかったこととも矛盾しない．しかし, Sc type の倍数性は Group 1 に属する株では 1 倍体または染色体欠失であるのに対し，Group 2 に属する株においては 2 倍体または 3 倍体であり，Group 間で異なることが示された（図 4-10）．もし両 Group が共通祖先に由来した場合，なぜこのような倍数性に差が生じるのであろうか？以下，Sc type における倍数性の差異が生じた理由について考察する．

共通の異種交配に由来した場合に，倍数性の差異が生じた理由について 2 つの仮説が考えられる．Group 1 と 2 に共通な異種交配と染色体間転座の後に Group 1 の Sc type で染色体欠失が繰り返し起きたとする説（図 5-22 A）と Group 2 で異種交配が 2 回起きたとする説である（図 5-22 B）．前者の場合，祖先種の *S. cerevisiae* と *S. eubayanus* は共に 2 倍体であり，異種交配と染色体間転座の後に，現在の Group 1 に相当する株でのみ Sc type で染色体欠失が繰り返し起き, Group 2 の株は染色体間転座を繰り返しながらも，異質 2 倍体を維持してきたため倍数性に違いが生じたと考えられる（図 5-22 A）．この場合，Group 1 の株間の変異率が小さいことから，染色体欠失は Group 1 で共通に起きたと予想される．後者の場合は，1 倍体の *S. cerevisiae* と 2 倍体の *S. eubayanus* との異種交配と染色体間転座の後，Group 1 は染色体間転座を繰り返しながら現在の倍数性になり，Group 2 は別の *S. cerevisiae* と異種交配することで Sc type が 2 倍体となり，その後のゲノム再構成によって現在のような Group 1 と 2 の倍数性の違いが見られると考える（図 5-22 B）．どちらの説に関しても可能性は存在するため，得られたデータのみからでは，共通の異種交配の後，倍数性の違いがどのように生じたのかは明らかにすることができなかった．

また，染色体欠失が Sc type でのみ観察される理由については，Se type の染色体欠失が環境的な制約を受けている可能性が考えられる．*S. eubayanus* は耐寒性に優れており，*S. pastorianus* の耐寒能に大きく寄与していることが示唆されている（Libkind et al. 2011）．実際には，異種交配後に Se type においても染色体欠失は起きたが，そのような個体は低温のビール醸造環境下に適応できず，Sc type の染色体欠失を持つ株のみが生き残ることができ，結果として単離されると考えても矛盾はない．

Group 1 と 2 は地理的起源に基づいた分類であるとされており，前述の二つの説のどちらが正しいものであるのかについては，史実との照合が必要であると考えられる．ラガービール酵母の祖先種である *S. eubayanus* の由来に関しても諸説あったが，最終的に解析結果と歴史的背景からチベットに由来するという説が現在もっとも有力とな

っている．*S. eubayanus* がパタゴニアで初めて単離されたため，当初，ラガービール酵母の祖先種 *S. eubayanus* はパタゴニアに由来すると考えられていた（Libkind et al. 2011）．しかし，2014 年にアジアでも単離され，DNA 配列の比較結果から，チベットで単離された *S. eubayanus* が *S. pastorianus* の Se type に最も系統的に近いことが判明した（Bing et al. 2014）．歴史的背景においても，ラガービールは 15 世紀頃にはすでに作られ始めており，大西洋間の貿易は 16 世紀後半から始まったとされることから南アメリカから *S. eubayanus* がヨーロッパに持ち込まれたと考えるよりも，陸続きで昔から交易のあったアジアから持ち込まれたと考える方が妥当であると考えられている．このように，解析結果から得られた知見や仮説について，実験的手法や史実に基づいた検証がされることが望ましい．

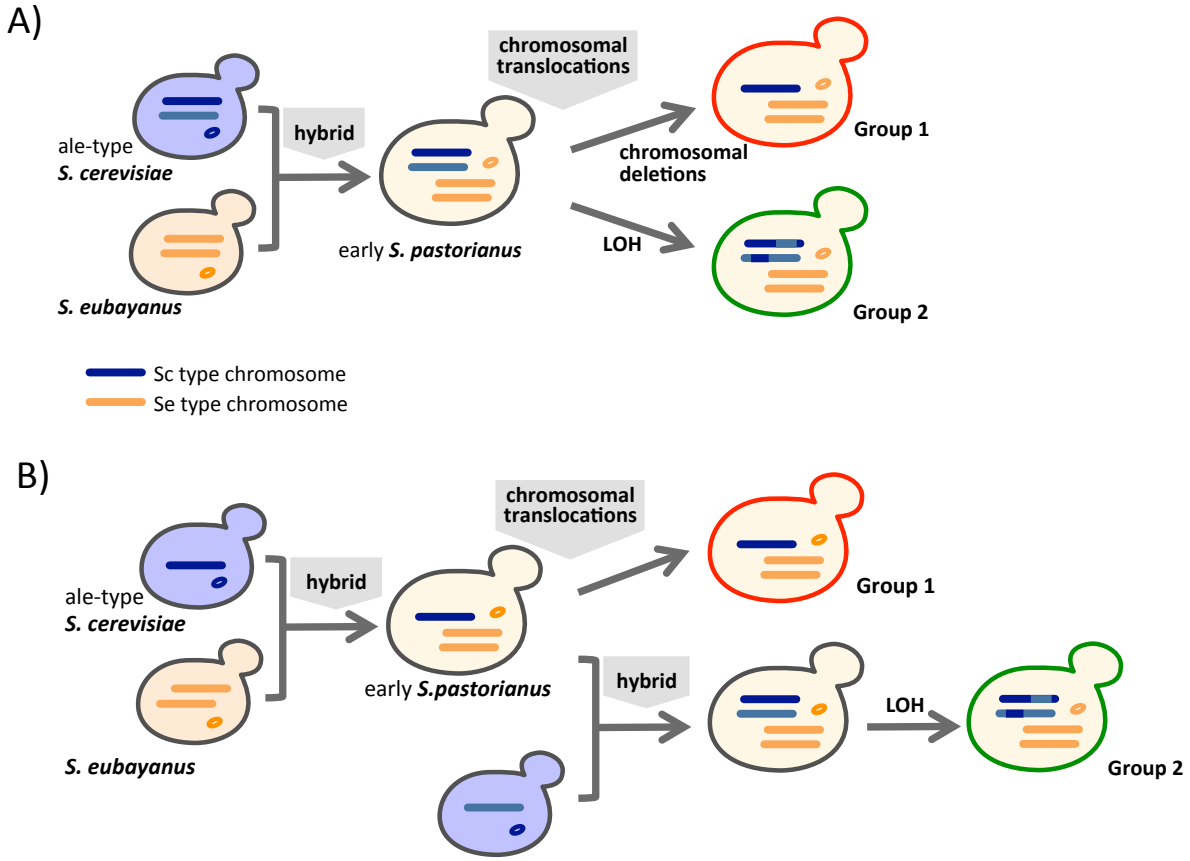


図 5-22 *S. pastorianus* の由来として考えられる 2 つの仮説の模式図

Group 1 と 2 の倍数性の違い，共通の染色体間転座の存在を踏まえた *S. pastorianus* の異種交配とその後のゲノム再構成の仮説．(A) 2 倍体の *S. cerevisiae*（エールビール酵母）と 2 倍体の *S. eubayanus* との異種交配と染色体間転座の後，Sc type での染色体間欠失と染色体間転座を繰り返した株が現在の Group 1 株，4 倍体を維持しながら Sc type での loss of heterozygosity (LOH) と染色体間転座を繰り返した株が現在の Group 2 であるという説．(B) 1 倍体の *S. cerevisiae* と 2 倍体の *S. eubayanus* との異種交配と染色体間転座の後，別の *S. cerevisiae* と異種交

配し，LOH と染色体間転座を繰り返した株が Group 2 であるという説.

5.6 まとめ

本章において，*S. pastorianus* の地理的起源に基づく分類と系統関係について，hetero SNV に着目し，そこから得られた仮説の検証を実施した．Group 2 に属する株の hetero SNV から，Sc type のハプロイドが共通祖先株に由来する可能性が示唆され，それを検証するために Group 2 に属する株において Sc type のハプロタイプ別ゲノムの構築を実施した．その結果，Group 2 のハプロタイプの片方は Group 1 のハプロタイプと有意に相同性が高いことが明らかになり，共通祖先に由来するという仮説を支持する結果が得られた．また，Group 1 と 2 間に存在する homo SNV は LOH に起因している可能性が示唆され，これらの結果から，最終的に，Group 1 と 2 の間には共通の異種交配が存在したと結論を出した．この結論はブレイクポイントが一致する染色体間転座の存在及び Se type の系統解析の結果とも一致し，これらの事象を矛盾なく説明できるものあると考えられる.

第6章　結論と展望

本研究は, ハイスループット DNA シークエンスデータを用いた全ゲノムレベルの解析により，地理的起源に基づく分類（Group 1, 2）の系統関係の解明を目的として実施された．具体的には，祖先種 *S. eubayanus* CBS12357 の染色体 16 本に相当するドラフトゲノムを構築し，祖先種 2 種（*S. cerevisiae* S288c*, S. eubayanus* CBS12357）のゲノムを基盤として *S. pastorianus* のゲノム解析を行なった．ミトコンドリア，倍数性，染色体間転座，系統解析の結果から地理的起源に基づく分類が系統的な由来の違いを意味するものであるか，そうでないのか，それぞれを支持する結果が得られ，これらの解析から最終的な結論を出すことはできなかった．しかし，Sc type における hetero SNV に着目したことで，Sc type のハプロタイプが同一の祖先株に由来するという可能性が示唆され，その検証を実施した．実際に，hetero SNV の連鎖関係を解くことによってハプロタイプ別の配列を構築し，ハプロタイプにおける配列間の比較を行った．その結果，*S. pastorianus* の地理的起源に基づいた分類の系統関係について，最終的に Group 1 と 2 の間で共通の異種交配が存在するという結論が導き出された．また，*S. pastorianus* Sc type において繰り返し loss of heterozygosity（LOH）が起きており，異種交配後のゲノム再構成は染色体欠失・重複や染色体間転座だけでないことが示唆された．

近年, さまざまなグループによるハイスループット DNA シークエンスデータを用いた *S. pastorianus* 株のゲノム解析が行われている中で（Walther et al. 2014; Baker et al. 2015; Van den Broek et al. 2015），本研究は，*S. pastorianus* Sc type における hetero SNV に着目した点とハプロタイプ別のゲノムに基づいた系統解析を実施した点で他の研究と大きく異なると言える．エールビール酵母 *S. cerevisiae* においては高ヘテロ接合性ゲノムであることが既に知られていたが（Borneman et al. 2011），ラガービール酵母 *S. pastorianus* の Group 2 に属する株において祖先株であるエールビール酵母に由来する hetero SNV が存在すること，また，異種交配後に LOH が起きたことが示唆されることは本研究で初めて示された成果である．また，ハプロタイプ別のゲノムの比較により，Group 1 に属する株のアレルと一致するアレルが Group 2 に属する株のハプロタイプ上に連続していることも明らかにした．これらの結果は地理的起源に基づく分類では分けられている両 Group が，系統的には共通祖先に由来することを示唆しており，地理的起源に基づく分類と系統関係について今まで相矛盾すると考えられていた様々な先行研究の結果を満足させることのできるモデルを示したことも大きな成果である．

今後，*S. pastorianus* の祖先種であるエールビール酵母 *S. cerevisiae* においてもハ

プロタイプ配列が構築され，*S. pastorianus* との比較が実現すれば，*S. pastorianus* の祖先種に近いエールビール酵母の特定，及び異種交配後に LOH が起きた領域の特定が可能になると考えられる．また，実験的に *S. cerevisiae* と *S. eubayanus* を異種交配させた株の配列データからも異種交配後のゲノム再構成についてより多くの知見が得られることが期待される．

2009 年にラガービール酵母のドラフトゲノムが完成し（Nakao et al. 2009），ハイスループット DNA シークエンサによる再解析も本研究以外でも多く行われているが，ラガービール酵母の形質として注目されている遺伝子の配列が構築できていないのが現状である．凝集に関わる *FLO* 遺伝子やマルトース発酵に関わる *MAL* 遺伝子などはテロメア領域に存在し，テロメア領域のリピート配列を克服できず配列構築が困難な状況である．さらに *MAL* 遺伝子についてはゲノム中に複数コピー存在することが知られているが，それぞれの配列の相同性が高いため，配列構築が困難で，配列が部分的に構築できた場合も染色体上のどこに位置するものであるかを特定することができない場合が多い．

このような問題点を解決するにあたって，今後，一分子シークエンス技術が大きな助けになると考えられる．Pacific Biosciences（PacBio）社の DNA シークエンサから得られる配列長は 20 kbp 程度あり，リピート配列を含むテロメア領域やゲノム中に複数コピー存在する配列の構築においても，ロングリードを用いることでそれぞれの配列を構築することは可能であると考えらえる．また PacBio リードによりハプロタイプの構築が実現すれば，より広範囲での phasing が可能になると考えられる．今回 paired-end リードで構築したハプロタイプ配列の最大長は，PacBio リード 1 本でカバーすることが可能である．しかし，エラー率 15%のリードを基にハプロタイプの構築をすることは容易なことではなく，十分なシークエンスカバレッジと正確なアセンブルが要求されると考えられ，技術的に解決すべき問題は多いと考えられる．

参考文献

Akao T, Yashiro I, Hosoyama A. 2011. Whole-genome sequencing of sake yeast Saccharomyces cerevisiae Kyokai no. 7. *DNA Res* **18**: 423–34.

Alkan C, Coe BP, Eichler EE. 2011. Genome structural variation discovery and genotyping. *Nat Rev Genet* **12**: 363–376.

Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25**: 3389–3402.

Babrzadeh F, Jalili R, Wang C, Shokralla S, Pierce S, Robinson-Mosher A, Nyren P, Shafer RW, Basso LC, de Amorim H V, et al. 2012. Whole-genome sequencing of the efficient industrial fuel-ethanol fermentative Saccharomyces cerevisiae strain CAT-1. *Mol Genet Genomics* **287**: 485–494.

Baker E, Wang B, Bellora N, Peris D, Hulfachor AB, Koshalek JA, Adams M, Libkind D, Hittinger CT. 2015. The genome sequence of Saccharomyces eubayanus and the domestication of lager-brewing yeasts. *Mol Biol Evol* **32**: 2818–2831.

Bennett RJ, Forche A, Berman J, Heitman J, Carter DA, Dyer PS, Kwon-chung KJ, Fraser JA, Tamara L, Roemer T, et al. 2014. Rapid Mechanisms for Generating Genome Diversity: Whole Ploidy Shifts, Aneuploidy, and Loss of Heterozygosity. *Cold Spring Harb Perspect Med* **4**: a019604.

Bing J, Han P-J, Liu W-Q, Wang Q-M, Bai F-Y. 2014. Evidence for a Far East Asian origin of lager beer yeast. *Curr Biol* **24**: R380–R381.

Borneman AR, Desany BA, Riches D, Affourtit JP, Forgan AH, Pretorius IS, Egholm M, Chambers PJ. 2012. The genome sequence of the wine yeast VIN7 reveals an allotriploid hybrid genome with Saccharomyces cerevisiae and Saccharomyces kudriavzevii origins. *FEMS Yeast Res* **12**: 88–96.

Borneman AR, Desany BA, Riches D, Affourtit JP, Forgan AH, Pretorius IS, Egholm M, Chambers PJ. 2011. Whole-genome comparison reveals novel genetic elements that characterize the genome of industrial strains of Saccharomyces cerevisiae. *PLoS Genet* **7**: e1001287.

Bradbury JE, Richards KD, Niederer HA, Lee SA, Rod Dunbar P, Gardner RC. 2006. A homozygous diploid subset of commercial wine yeast strains. *Antonie Van Leeuwenhoek* **89**: 27–37.

Cherry JM, Adler C, Ball C, Chervitz SA, Dwight SS, Hester ET, Jia Y, Juvik G, Roe T, Schroeder M, et al. 1998. SGD: Saccharomyces Genome Database. *Nucleic Acids Res* **26**: 73–79.

Corran HS. 1975. *A History of Brewing*. David & Charles, London.

Curtin CD, Borneman AR, Chambers PJ, Pretorius IS. 2012. De-novo assembly and analysis of the heterozygous triploid genome of the wine spoilage yeast Dekkera bruxellensis AWRI1499. *PLoS One* **7**: e33840.

Dunn B, Sherlock G. 2008. Reconstruction of the genome origins and evolution of the hybrid lager yeast Saccharomyces pastorianus. *Genome Res* **18**: 1610–23.

Durbin RM, Altshuler DL, Durbin RM, Abecasis GR, Bentley DR, Chakravarti A, Clark AG, Collins FS, De La Vega FM, Donnelly P, et al. 2010. A map of human genome variation from population-scale sequencing. *Nature* **467**: 1061–1073.

Forche A, Abbey D, Pisithkul T. 2011. Stress Alters Rates and Types of Loss of Heterozygosity in Candida albicans. *MBio*.

Foury F, Roganti T, Lecrenier N, Purnelle B. 1998. The complete sequence of the mitochondrial genome of Saccharomyces cerevisiae. *FEBS Lett* **440**: 325–331.

Gibson BR, Storgårds E, Krogerus K, Vidgren V. 2013. Comparative physiology and fermentation performance of Saaz and Frohberg lager yeast strains and the parental species Saccharomyces eubayanus. *Yeast* **30**: 255–266.

Glenn TC. 2011. Field guide to next-generation DNA sequencers. *Mol Ecol Resour* **11**: 759–769.

Goffeau A, Barrell BG, Bussey H, Davis RW, Dujon B, Feldmann H, Galibert F, Hoheisel JD, Jacq C, Johnston M, et al. 1996. Life with 6000 Genes. *Science (80-)* **274**: 546–567.

Guindon S, Lethiec F, Duroux P, Gascuel O. 2005. PHYML Online - A web server for fast maximum likelihood-based phylogenetic inference. *Nucleic Acids Res* **33**: 557–559.

Hebly M, Brickwedde A, Bolat I, Driessen MRM, de Hulster E a. F, van den Broek M, Pronk JT, Geertman J-M, Daran J-M, Daran-Lapujade P. 2015. S. cerevisiae x S. eubayanus interspecific hybrid, the best of both worlds and beyond. *FEMS Yeast Res* **15**: fov005–fov005.

Hewitt SK, Donaldson IJ, Lovell SC, Delneri D. 2014. Sequencing and characterisation of rearrangements in three S. pastorianus strains reveals the presence of chimeric genes and gives evidence of breakpoint reuse. *PLoS One* **9**.

Kajitani R, Toshimoto K, Noguchi H, Toyoda A, Ogura Y, Okuno M, Yabana M, Harada M, Nagayasu E, Maruyama H, et al. 2014. Efficient de novo assembly of highly heterozygous genomes from whole-genome shotgun short reads. *Genome Res* **24**: 1384–1395.

Kent WJ. 2002. BLAT---The BLAST-Like Alignment Tool. *Genome Res* **12**: 656–664.

Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R, Jones SJ, Marra MA. 2009. Circos: an Information Aesthetic for Comparative Genomics. *Genome Res* **19**: 1639–1645.

Lamour AKH, Mudge J, Gobena D, Hurtado-Gonzales OP, Bharti K, Donahoo RS, Finley S, Huitema E, Hulvey J, Platt D, et al. 2012. Genome sequencing and mapping reveal loss of heterozygosity as a mechanism for rapid adaptation in the vegetable pathogen Phytophthora capsici. *Mol Plant Microbe Interact* **25**: 1350–60.

Legras JL, Merdinoglu D, Cornuet JM, Karst F. 2007. Bread, beer and wine: Saccharomyces cerevisiae diversity reflects human history. *Mol Ecol* **16**: 2091–2102.

Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**: 1754–1760.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**: 2078–2079.

Li H, Homer N. 2010. A survey of sequence alignment algorithms for next-generation sequencing. *Brief Bioinform* **11**: 473–483.

Li K-B. 2003. ClustalW-MPI: ClustalW analysis using distributed and parallel computing. *Bioinformatics* **19**: 1585–1586.

Libkind D, Hittinger CT, Valério E. 2011. Microbe domestication and the identification of the wild genetic stock of lager-brewing yeast. *Proc Natl Acad Sci U S A* **108**: 14539–14544.

Liti G, Carter DM, Moses AM, Warringer J, Parts L, James SA, Davey RP, Roberts IN, Burt A, Koufopanou V, et al. 2009. Population genomics of domestic and wild yeasts. *Nature* **458**: 337–41.

Liti G, Peruffo A, James S a., Roberts IN, Louis EJ. 2005. Inferences of evolutionary relationships from a population survey of LTR-retrotransposons and telomeric-associated sequences in the Saccharomyces sensu stricto complex. *Yeast* **22**: 177–192.

Louis VL, Despons L, Friedrich A, Martin T, Durrens P, Casaregola S, Neuveglise C, Fairhead C, Marck C, Cruz J a., et al. 2012. Pichia sorbitophila, an Interspecies Yeast Hybrid, Reveals Early Steps of Genome Resolution After Polyploidization. *G3 Genes, Genomes, Genet* **2**: 299–311.

Mallet J. 2005. Hybridization as an invasion of the genome. *Ecol Evol* **20**: 229–237.

McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, et al. 2010. The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* **20**: 1297–1303.

Metzker ML. 2010. Sequencing technologies - the next generation. *Nat Rev Genet* **11**: 31–46.

Miller JR, Koren S, Sutton G. 2011. Assembly Algorithms for Next-Generation Sequecing Data. *NIH Public Access* **95**: 315–327.

Mira NP, Münsterkötter M, Dias-Valada F, Santos J, Palma M, Roque FC, Guerreiro JF, Rodrigues F, Sousa MJ, Leão C, et al. 2014. The genome sequence of the highly acetic acid-tolerant zygosaccharomyces bailii-derived interspecies hybrid strain ISA1307, isolated from a sparkling wine plant. *DNA Res* **21**: 299–313.

Monerawela C, James TC, Wolfe KH, Bond U. 2015. Loss of lager specific genes and subtelomeric regions define two different Saccharomyces cerevisiae lineages for Saccharomyces pastorianus Group I and II strains. *FEMS Yeast Res* **15**: fou008–fou008.

Nakao Y, Kanamori T, Itoh T, Kodama Y, Rainieri S, Nakamura N, Shimonaga T, Hattori M, Ashikari T. 2009. Genome sequence of the lager brewing yeast, an interspecies hybrid. *DNA Res* **16**: 115–29.

Nguyen H-V, Legras J-L, Neuvéglise C, Gaillardin C. 2011. Deciphering the hybridisation history leading to the Lager lineage based on the mosaic genomes of Saccharomyces bayanus strains NBRC1948 and CBS380. *PLoS One* **6**: e25821.

Nijhoffpublishers M. 1987. Three newly delimited species of Saccharomyces sensu stricto. *Antonie Van Leeuwenhoek* **84**: 77–84.

Oliver SG, van der Aart QJM, Agostoni-Carbone ML, Aigle M, Alberghina L, Alexandraki D, Antoine G, Anwar R, Ballesta JPG, Benit P, et al. 1992. The complete DNA sequence of yeast chromosome III. *Nature* **357**: 38–46.

P L. 1909. *Mikroskopische Betriebskontrolle in den Garungswerben*. 6th edn.

Peris D, Lopes C a, Belloch C, Querol A, Barrio E. 2012. Comparative genomics among Saccharomyces cerevisiae × Saccharomyces kudriavzevii natural hybrid strains isolated from wine and beer reveals different origins. *BMC Genomics* **13**: 407.

Querol A, Bond U. 2009. The complex and dynamic genomes of industrial yeasts: MINIREVIEW. *FEMS Microbiol Lett* **293**: 1–10.

Rainieri S, Kodama Y, Kaneko Y, Mikata K, Nakao Y, Ashikari T. 2006. Pure and mixed genetic lines of Saccharomyces bayanus and Saccharomyces pastorianus and their contribution to the lager brewing strain genome. *Appl Environ Microbiol* **72**: 3968–74.

Rainieri S, Kodama Y, Nakao Y, Pulvirenti A, Giudici P. 2008. The inheritance of mtDNA in lager brewing strains. *FEMS Yeast Res* **8**: 586–96.

Scannell DR, Zill O a., Rokas A, Payen C, Dunham MJ, Eisen MB, Rine J, Johnston M, Hittinger CT, Andrews BJ. 2011. The Awesome Power of Yeast Evolutionary Genetics: New Genome Sequences and Strain Resources for the Saccharomyces sensu stricto Genus. *G3 Genes, Genomes, Genet* **1**: 11–25.

Sipiczki M. 2008. Interspecies hybridization and recombination in Saccharomyces wine yeasts. *FEMS Yeast Res* **8**: 996–1007.

Tadami H, Shikata-Miyoshi M, Ogata T. 2014. Aneuploidy, copy number variation and

unique chromosomal structures in bottom-fermenting yeast revealed by array-CGH. *J Inst Brew* **120**: 27–37.

Van den Broek M, Bolat I, Nijkamp JF, Ramos E, Luttik M a. H, Koopman F, Geertman JM, de Ridder D, Pronk JT, Daran JM. 2015. Chromosomal copy number variation in Saccharomyces pastorianus evidence for extensive genome dynamics in industrial lager brewing strains. *Appl Environ Microbiol* **81**: 6253–6267.

Walther A, Hesselbart A, Wendland J. 2014. Genome sequence of Saccharomyces carlsbergensis, the world's first pure culture lager yeast. *G3 Genes, Genomes, Genet* **4**: 783–93.

Wendland J. 2014. Lager Yeast Comes of Age. *Eukaryot Cell* **13**: 1256–1265.

謝辞

本研究を進めるにあたり，ラガービール酵母の DNA サンプルを提供して頂いた児玉 由紀子博士（サントリーグローバルイノベーションセンター）に感謝申し上げます．酵母の培養，シークエンスを行なっていただきました東京工業大学 流水 利恵技術員に御礼申し上げます．

博士課程在学中は博士課程教育リーディングプログラム 情報生命博士教育院より経済的，教育的なご支援をいただいたことにつきまして御礼申し上げます．

有意義な討論を頂いた同研究室の方々，卒業生の方々に感謝いたします．