

成熟したショートリードのさらなる有効活用の可能性 ～フィデリティを高めて自然科学の難題に挑む～



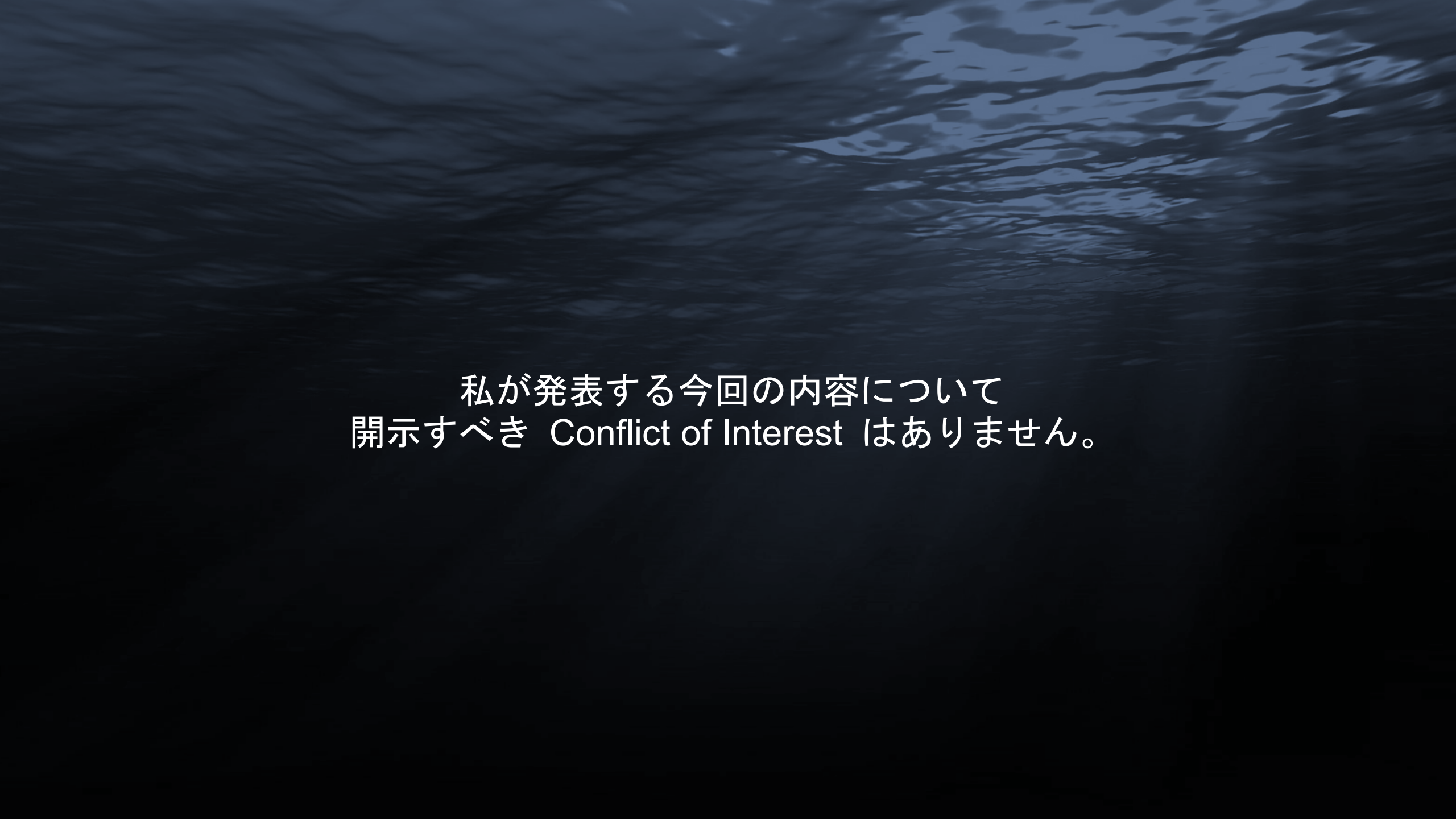
理化学研究所
生命機能科学研究センター
分子配列比較解析チーム
@phyloinfokobe



リーダー
工樂 樹洋 Shigehiro Kuraku



兼 関西学院大院理工学研究科 / 神戸大院理学研究科 客員准教授 / 甲南大学院自然科学 非常勤講師

The background of the slide is a close-up photograph of water with fine, dark ripples. The lighting is somewhat dim, giving it a moody, blue-grey appearance.

私が発表する今回の内容について
開示すべき Conflict of Interest はありません。

自己紹介

2005 京都大にて博士号（分子進化学）

-2007 理研神戸でJRA～RA～PD
（進化発生学）

2007-12 ドイツ南部で大学教員
欧米式ゲノムプロジェクトや
コアラボ運営 との接点

2012- 理研神戸
独自研究＋技術支援＋教育普及

NGSについての最近の雑感

“ポスト「現場の会」時代”に入り、俯瞰した情報が得にくくなっている!?

技術は普及したが、チョイスが安定化・固定化した結果、
多様な経験を持つ人はさほど増えてはいないのでは？

持続可能なコアラボとは？ cf. 第6期基本計画「研究基盤の充実による研究力向上」

日本流 ‘Expectation management’の難しさ

コアラボ人材としてのタレント育成パス

どこで読むのがよいのか？

受託企業 or アカデミア？

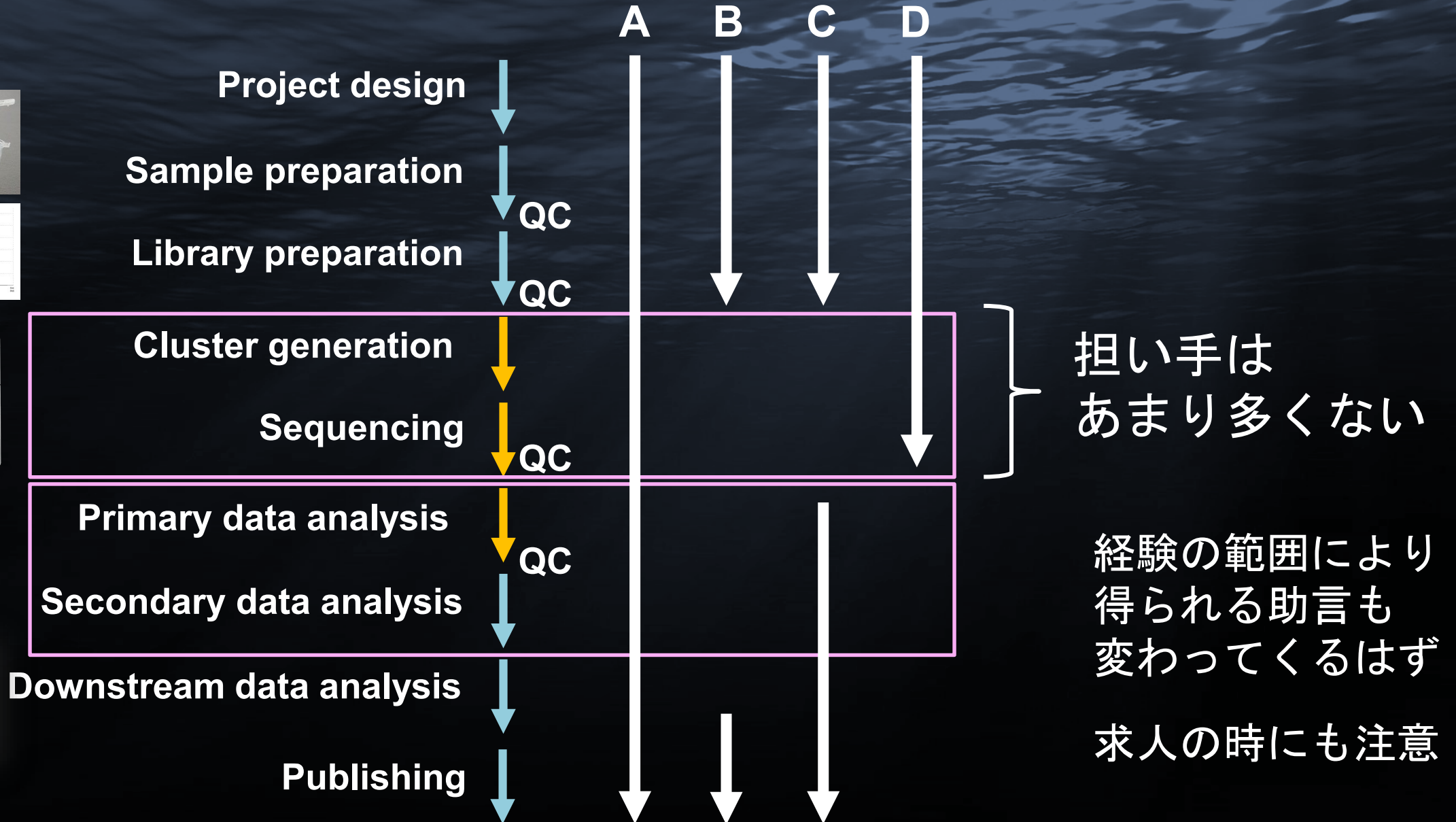
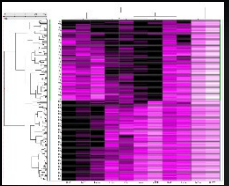
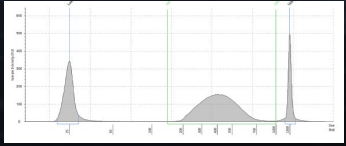
日本 or 海外？

アカデミアの価値はどこにあるか？

核酸以外の試料の柔軟な受入（シングルセル、クロマチン、標識実験）

技術の科学的評価 cf. 日本でHi-Cが広まらないのはなぜか？

「NGSの経験がある」人には何種類かいる



技術的な実績（～2018）

さらなる汎用化をめざして

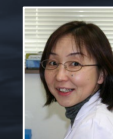
効率的な*de novo*トランスクリプトームデータ取得法の提案

ライブラリインサート長の調節&アセンブリ結果の評価

Hara et al., 2015 BMC Genomics

Tanegashima et al., 2018 Sci Data

原雄一郎 種子島千春



長距離（>20 Kbp）メイトペアリード取得の効率化・節約

iMate プロトコルのリリース

Tatsumi et al., 2015 Biotechniques

辰見香織



ゲノムアセンブリ結果の評価法の検討

ウェブサーバ gVolante（ジーボランチ）の公開

Nishimura et al., Bioinformatics 2017

Nishimura et al., Methods Mol. Biol. 2019

西村理



原雄一郎



ChIP-seqのためのChIPアッセイ条件の最適化

多様な細胞種・分子種を想定して

Kadota et al., 2017 Sci. Rep.

実験医学 2016年10月号 Vol.34 No.16.

門田満隆



技術的な実績（2018～）

さらなる汎用化をめざして

染色体レベルのゲノム情報構築法の検討

Hi-Cデータ取得の効率化とHi-C scaffoldingプログラムの試用

Kadota et al., 2020 GigaScience

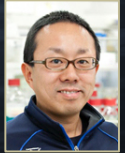
西村理



門田満隆



三浦 尚



平谷伊智朗

10x Chromiumを使用したシングルセルRNA-seq

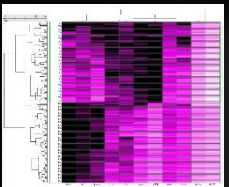
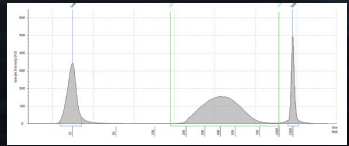
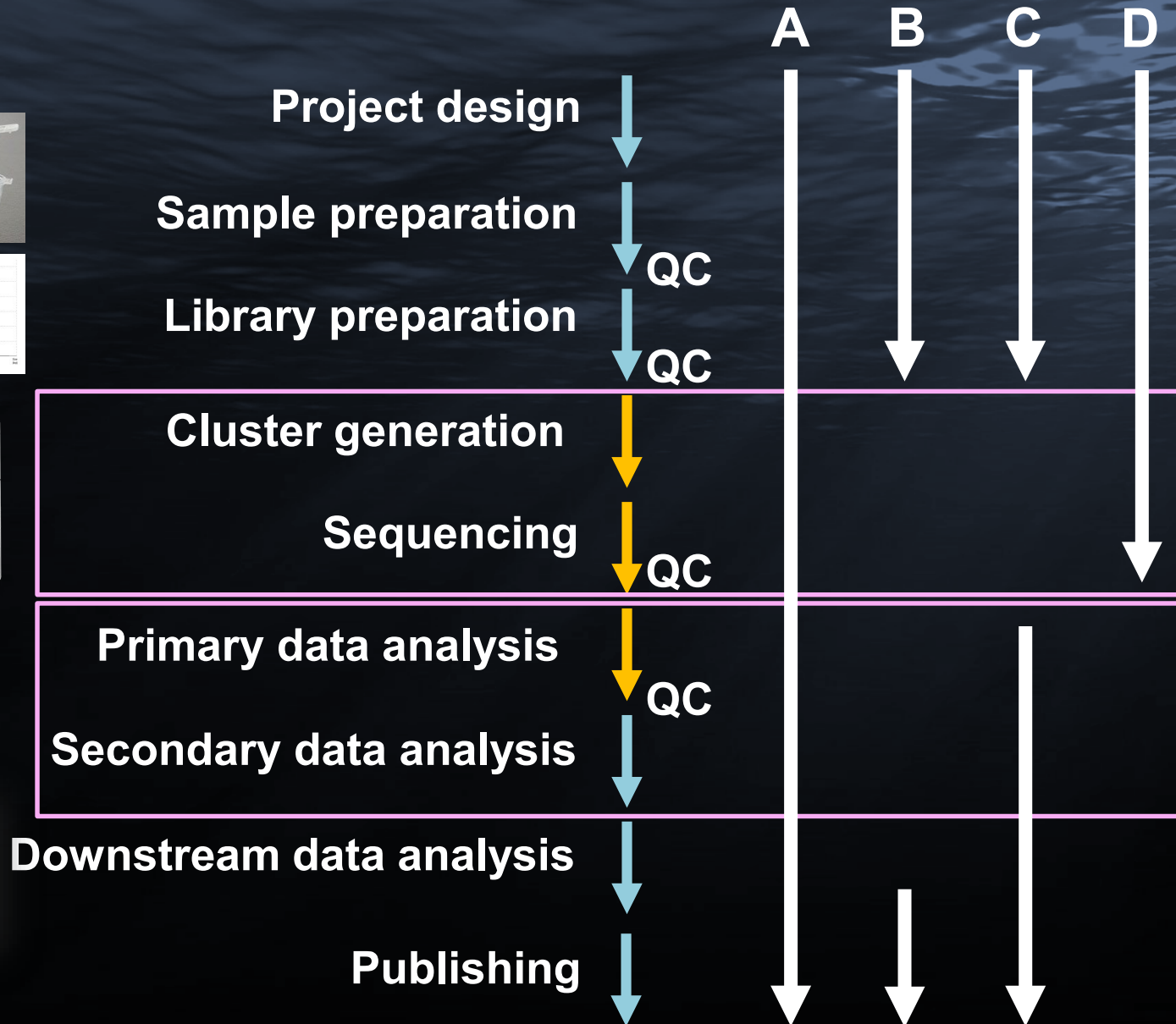
3'UTRに重点を置いた遺伝子モデル作成法の検討 (cellranger mkgtf/mkref)

標識塩基S4Uを利用したRNA代謝プロファイリング (SLAM-seq)

市販キットの試用および汎用品の使用による節約化の検討

種別ごとの最適化 と 種別を越えた最適化

「NGSの経験がある」人には何種類かいる



シークエンサを自ら動かす立場から

個々のライブラリの評価

サイズ分布と定量方法

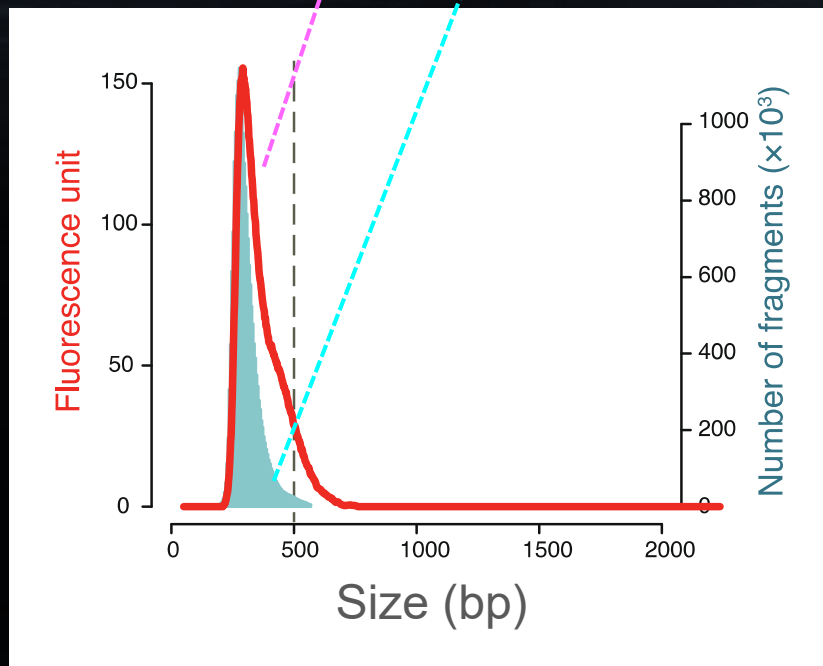
ライブラリの混ぜ方

ラン結果から何を学ぶか？

ライブラリ中のすべての分子が 均等に読まれるわけではない

アジレントBioanalyzerによる
ライブラリのインサート長分布

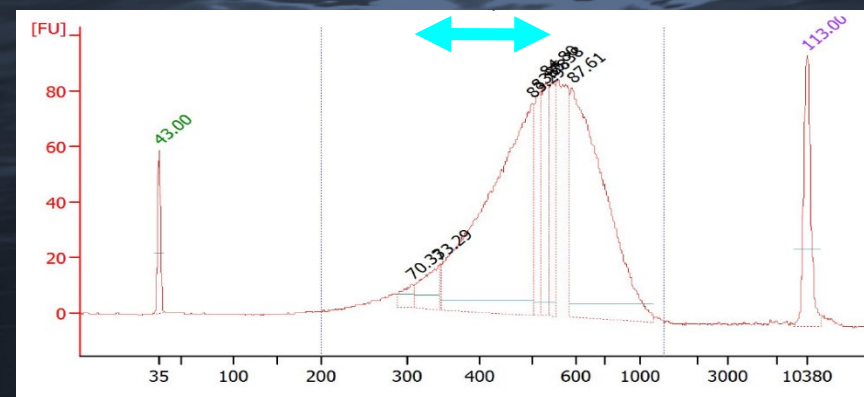
マッピングの結果判明した
実際に読まれたインサート長分布



Hara et al., 2015. *BMC Genomics*, 16:977

これを知ったうえで ライブラリのサイズ分布を操る

1) 断片化



2) PCR 増幅

必要量得るためのサイクル数

ex) 1000ng totalRNA → 6

10ng totalRNA → 13

3) サイズセレクション

磁気ビーズで特定の分画を除く

シークエンサを自ら動かす立場から

個々のライブラリの評価

サイズ分布と定量方法

ライブラリの混ぜ方

インデックスの配列類似性、塩基組成、カラーバランスを考慮

ラン結果から何を学ぶか？

インデックス配列の組合せについて - より確かなリード取得のために

参照：イルミナ公式資料 'Index Adapters Pooling Guide'

配列類似性

似た配列のインデックスを同時使用すると
ライブラリ毎のリードの振分け精度が落ちる

蛍光のカラーバランス

サイクル毎のカラーバランスが偏って
いると、クラスターの位置決めが
正しく行われない

塩基のバランス

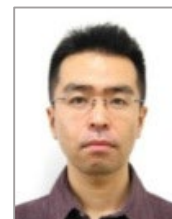
サイクル毎の塩基バランスが悪いと、
ベースコールの精度が下がる

調製済ライブラリ群を受け付ける際
にも事前に確認をしてもらう

オリジナルツール iLike & iMaster

<https://transcriptome.riken.jp/iLike/>

西村理





References

A507	TAAGTT
A508	TAGACC
A701	ATCACG
A702	ACAGTG
A703	CAGATC

Query (Queries)

A507	TAAGTT
A508	TAGACC
A701	ATCACG
A702	ACAGTG
A703	CAGATC

↓

Results:

A508 has similarity with the following
A703, 2 mismatch(es)



iMaster

A tool to check the balance of the index sequence

Lane 1									
Position	1	2	3	4	5	6	7	8	9
Red	100%	50%	50%	50%	25%	50%	50%	50%	100%
Green	0%	50%	50%	50%	75%	50%	50%	50%	0%
A	50%	25%	25%	50%	0%	0%	0%	25%	100%
T	0%	25%	50%	25%	50%	25%	0%	50%	0%
G	0%	25%	0%	25%	25%	25%	50%	0%	0%
C	50%	25%	25%	0%	25%	50%	50%	25%	0%

Lane 1		
Name	Index Sequence	Conc/Ratio
index5	CATAGCCTA	1
index8	ATAGTGGCA	1
index3	CCTATCCTA	1
index1	AGCTCTGAA	1

シークエンサを自ら動かす立場から

個々のライブラリの評価

サイズ分布と定量方法

ライブラリの混ぜ方

インデックスの類似性、塩基組成、カラーバランスを考慮

ラン結果から何を学ぶか？（研究の目的に向かう前に）

期待通りのリード数が得られているか？

ライブラリの種類に特有の傾向はないか？

生物種の特徴がどう表れているか？

外注時でも所属機関のコアラボに助言を求めてみては？

ショートリードと決めたら、次は？

どこで読むか？

急ぐなら、in-houseでHiSeq1500 あるいは MiSeq
待てるなら、外注でHiSeqX か NovaSeq

リード長

SR80 (HiSeq1500) – RNA-seq for DE解析, ChIP-seq
PE127 (HiSeq1500) – *de novo* RNA-seq, Hi-C
PE150 (HiSeqX) – 全ゲノム

ライブラリ調製

リード長に合わせたインサート長に
必要量は？ PCRサイクル数は最少に

NovaSeq 6000 SPフローセルをPE250で試用

(参照：イルミナカスタマーレビュー)

同じライブラリをHiSeq1500 (Rapid v2) でもシーケンスし比較

シーケンス直後に確認

想定(800M)を超えるリードを取得(885M)

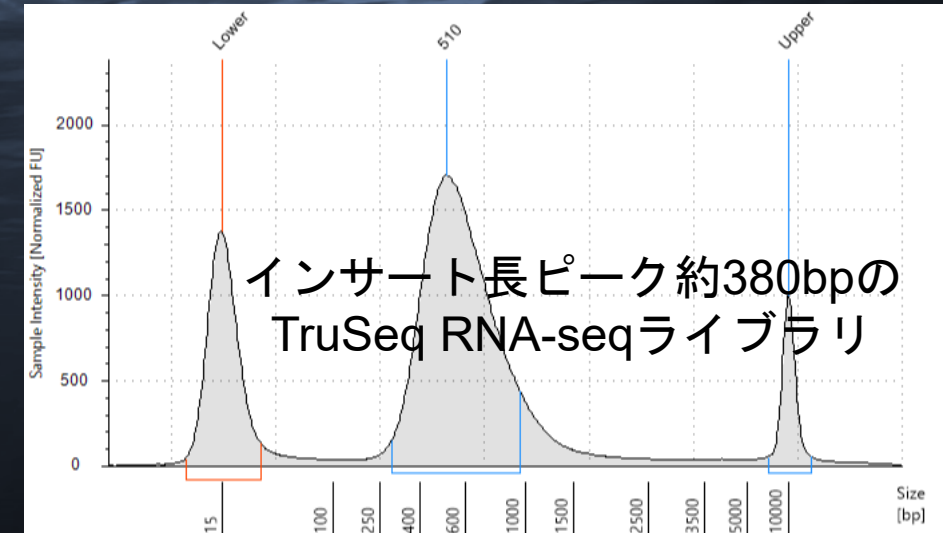
Qスコアの差はベースコール法の違いに起因

さらに下流での比較

取得リード中のGC含量やインサート長には大差なし

脊椎動物複数種で良好なトランスクリプトームアセンブリを達成

→ NovaSeq 6000 SPフローセルでより安価にPE250リードを取得可能



技術的な実績（2018～）

さらなる汎用化をめざして

染色体レベルのゲノム情報構築法の検討

Hi-Cデータ取得の効率化とHi-C scaffoldingプログラムの試用

Kadota et al., 2020 GigaScience

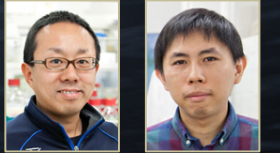
西村理



門田満隆



三浦 尚



平谷伊智朗

10x Chromiumを使用したシングルセルRNA-seq

3'UTRに重点を置いた遺伝子モデル作成法の検討 (cellranger mkgtf/mkref)

標識塩基S4Uを利用したRNA代謝プロファイリング (SLAM-seq)

市販キットの試用および汎用品の使用による節約化の検討

ゲノムシーケンス：どのレベルを目指すか？（あくまでも仮の分け方）

「着手」グレード

「非染色体」グレード

「染色体」グレード

「VGP」グレード

‘3.4.2.QV40’ phased, gapless で♂♀



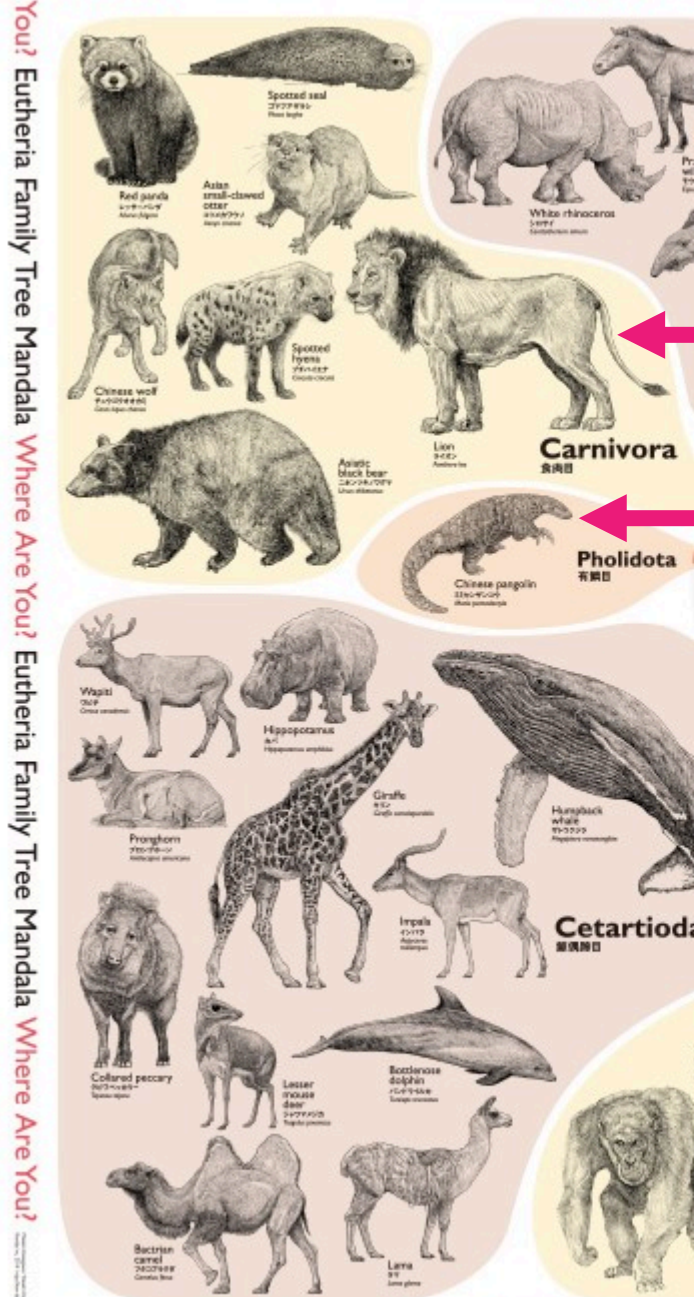
Joana Damas,  Graham M. Hughes
Marco Corbo,  Michael Hiller, 
Diane P. Genereux,  Ross Swofford
Kerstin Lindblad-Toh,  Emma C. T. 

doi: <https://doi.org/10.1101/2020.04.10.090000>

This article is a preprint and has not been

系統樹マンダラ【真獣類編】
(キウイラボ 発行)

Where Are You? Eutheria Family Tree Mandala Where Are



Eutheria Family Tree Mandala **Where Are You?** Eutheria

VERY HIGH

[illegible]

HIGH

[illegible]

ゲノムシーケンス：どのレベルを目指すか？（あくまでも仮の分け方）

「着手」グレード

「非染色体」グレード

「染色体」グレード

「VGP」グレード

‘3.4.2.QV40’ phased, gapless で♂♀



繋いだ結果をどう評価するか？

gVolante <https://gvolante.riken.jp/>

N50長

短い配列を含めるか否かで数値が大きく増減することもあり、大きければよい、というわけではない（ヒトなら約171Mbp）

単一コピーオーソログのカバー率

参照遺伝子セットの適性 (cf. 脊椎動物向けCVG)

遺伝子探索パイプラインの感度 (CEGMA > BUSCO)

BUSCO v2～v3～v4：基準が変わってよいのか？

ある程度以上になると繋がりの指標としては不十分

染色体グレードアセンブリの評価

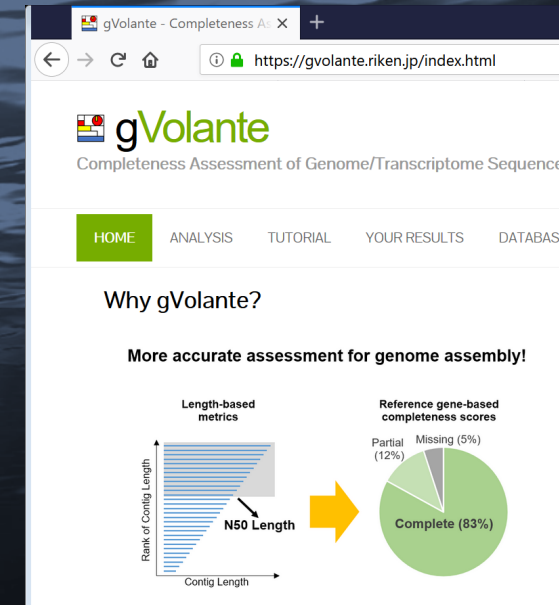
核型や連鎖地図など配列以外の情報を参照する必要

gap数や総塩基長に占める>10Mbp配列長の割合なども

西村理

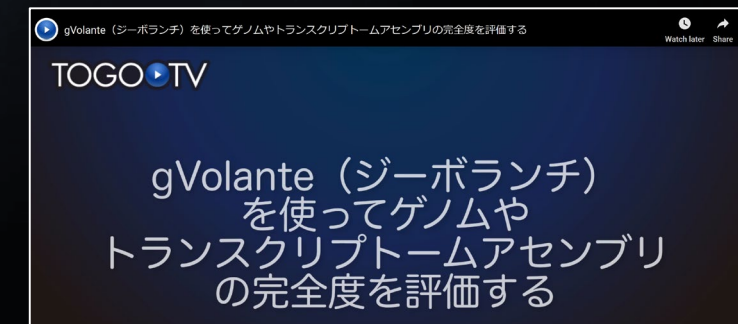


原雄一郎



Nishimura et al., Bioinformatics 2017

Nishimura et al., Methods Mol. Biol. 2019



統合TVからチュートリアルムービー

ゲノム「解読」の成果を伝える表現に注意

プレスリリースを読んでいて感じたことはありませんか？

そもそも「解読」？→「写経」しただけ

「N50長がX百倍に」→大きければよいというものではない

「遺伝子がX万個」→プログラムのかけ方で大きく変わる

「高精度の(high-quality)」より
「染色体規模の(chromosome-scale)」のほうが上？

発信者の認識とセンスが問われるところ

ゲノムシーケンス：どのレベルを目指すか？ (あくまでも仮の分け方)

「着手」グレード

メイトペア
ロングリード

N50長が<100kb

BUSCO/CEGMAスコア(Complete)が<80%

生物学的大発見ナシに
単一種で出版するのは
かなり困難

「非染色体」グレード

Hi-C
連鎖地図

BUSCO/CEGMAスコア(Complete)が80-100%

最長の配列も染色体長にはまだ到底届かない

「染色体」グレード

染色体規模の配列の総塩基長が全体の90%以上

染色体の数(n)と長さをほぼ再現 (核型情報が必要)

× 殆どの場合
ゼロから計画
しなおし

「VGP」グレード

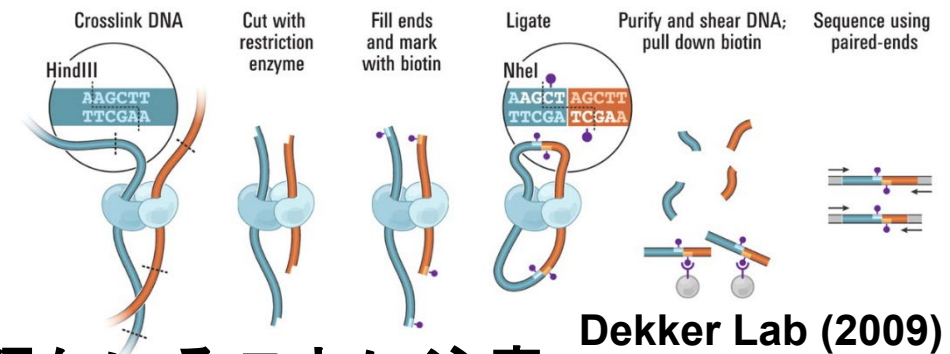
‘3.4.2.QV40’ phased, gapless で♂♀

Hi-C : 3Dゲノム解析からゲノムスキャフォールディングへ

ライゲーション部位を挟むペアエンドリード取得

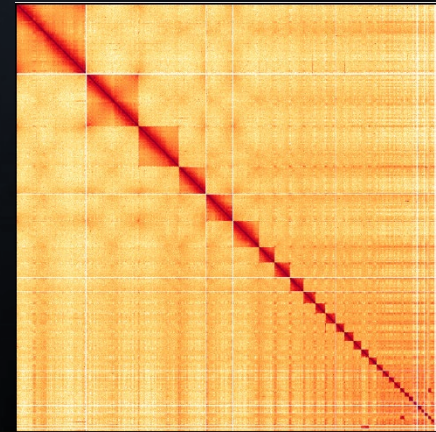
使用する試料はDNAではなくクロマチン（細胞・組織）

ここでもまた、Hi-Cの「経験がある」人には何種類かいることに注意



DNA上で近い領域間ほどクロマチン上のコンタクトが多いことを利用しスキャフォールディング

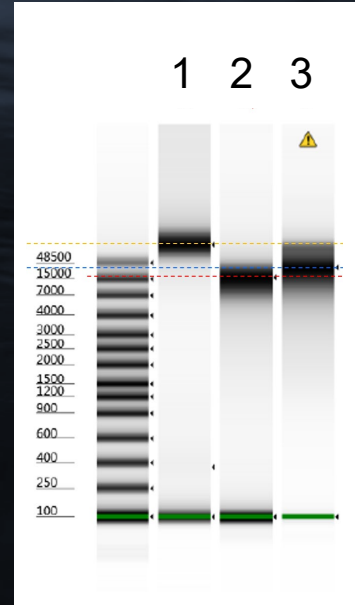
3Dゲノム解析とスキャフォールディングの間では「理想のライブラリ」は基本的に同じ



Hi-Cスキャフォールディングではギャップ(N)は埋まらず、むしろギャップ数は増える
不適なライブラリであることを知らずに大量シーケンスしないように

Hi-Cライブラリの評価：読む前に問題に気づけるか？

Hi-C → QC1 → ライブラリプレップ → QC2 → シーケンス



1. Before HindIII digestion

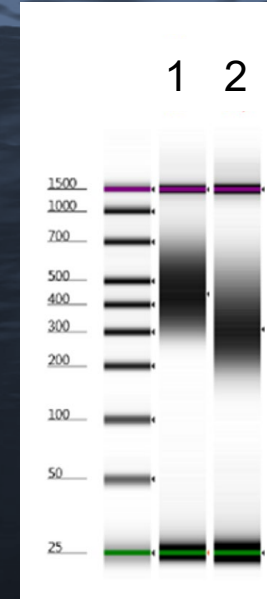
NNNAAGCTTNNN
NNNTTCGAANNN

2. After HindIII digestion

NNNAAGCTTNNN
NNNTTCGAANNN

3. Hi-C DNA after fill-in and ligation

NNNAAGCTAGCTTNNN
NNNTTCGATCGAANN



1. Before NheI digestion

NNNAAGCTAGCTTNNN
NNNTTCGATCGAANN

2. After NheI digestion

NNNAAGCTAGCTTNNN
NNNTTCGATCGAANN

QCパスしたもののみ大量シーケンス

Hi-Cスキャフォolding最適化のための条件検討

Kadota et al., *GigaScience* 2020

Softshell turtle
Pelodiscus sinensis

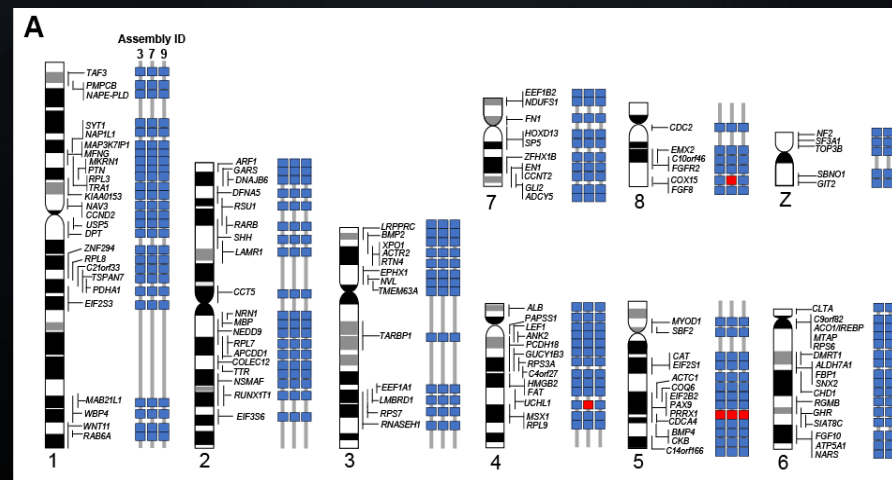
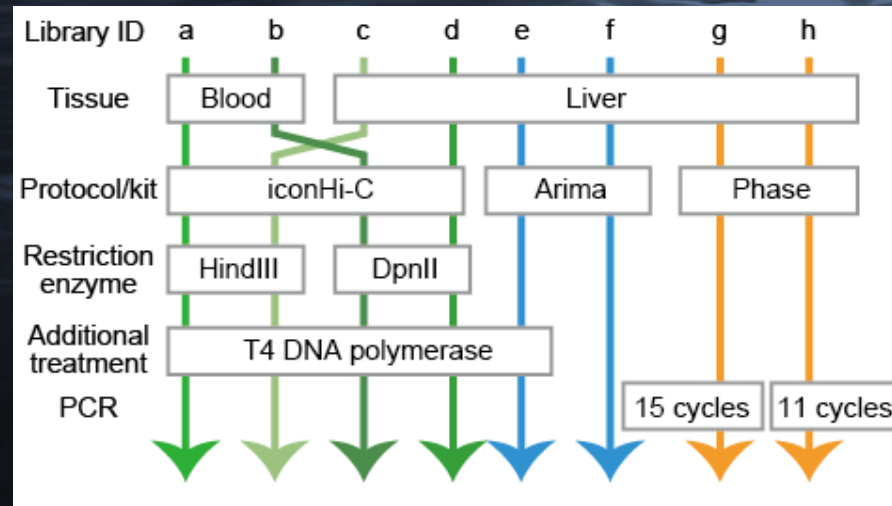
情報の揃った種を選び、WetおよびDryのさまざまな条件を検討

検討項目

- ・使用する組織
- ・固定時間
- ・制限酵素の選択
- ・反応時間
- 制限酵素
ライゲーション
- ・PCRサイクル数

- ・ プログラムの選択
- ・ 入力配列最短長
- ・ 使用Hi-Cライブラリ
- ・ 入力Hi-Cリード数
- ・ Misjoin 修正ラウンド数

22通りの間で比較



→ 検討結果に基づき
iconHi-C (アイコニック)
プロトコルを提案

Hi-Cスキャフォールディング考：多数種での実績から

Kadota et al., GigaScience 2020

ライブラリ調製

同じ生物種でも組織のチョイスにより、結果に大きな差

細胞数を確保できる**単一**の組織を選ぶ

有償キット間でも、仕様やパフォーマンスに大きな差

我々のiconHi-Cプロトコルは、有償キットほど迅速性に重きを置いていない

スキャフォールディング

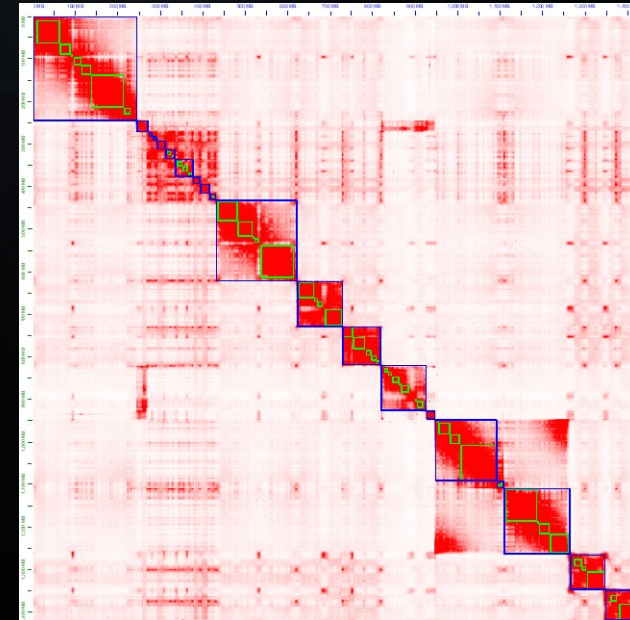
基になるゲノムアセンブリの出来も勝敗を分ける

配列長の下限の設定(3d-dnaの既定値は15,000bp)も結果を大きく左右

スキャフォールディング結果を評価するための**核型**情報が必要

あまり知られていないが、

Juicebox上での'review'（マニュアルでの修正）で、核型との類似性などを総合的に判断し、ファイナライズ



まとめ

「オミクス」におけるNGSで最も大事ななのは、
偏りが少なく多様性の高いライブラリを調製すること
そのうえで、どれだけ速く、安くリード取得できるか

染色体グレードへのゲノムアセンブリの改善
については適宜ご相談ください

装置の配備だけでなく、人材の育成と
ノウハウの蓄積や発信をも視野に入れた技術基盤のデザインを

特定の拠点に「丸投げ」するより、「セカンドオピニオン」も経て、
できるだけ多くの現場に技術が広がるように