研究基盤EXPO2025 @ 金沢大学 2024.1.27
「研究データ基盤構築とそのエコシステム化を考えるシンポジウム」

# オープンサイエンス時代のゲノム人類学

## - データ相互利活用に向けた課題と展望 -

金沢大学医薬保健研究域付属サピエンス進化医学研究センター
石谷 孔司

考古学

・古人骨/動物骨
・植物遺物
・古微生物/ ウイルス etc.

**生命情報科学**
**Bioinformatics**

・AI技術
・機械学習
・数理モデル
・データ品質管理
・解析環境構築　etc.

進化学

・系統/起源探索
・遺伝的多様性
・集団動態 etc.

医学

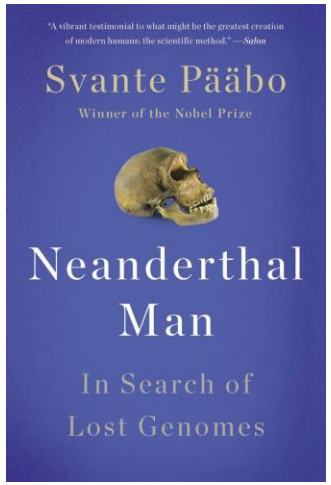
・過去の感染症
・疾患の起源
・疾患リスク etc.

# Novel prize 2022

**“for his discoveries concerning the genomes of extinct hominins and human evolution”**

Prof. Svante Pääbo
(Max Planck Institute for Evolutionary Anthropology,　OIST)

- 絶滅したネアンデルタール人のゲノム解読
- これまで知られていなかったヒト科「デニソワ」を発見
- アフリカから移住してきたホモ・サピエンスに、絶滅した人類から遺伝子が受け継がれていることを突き止めた



# ゲノム人類学とパレオゲノミクス

**ゲノム人類学** | ゲノムから人類の進化や起源に迫る

×

**パレオゲノミクス** | 過去の生物のゲノム情報を読み解く

↳ 進化や形質の多様性にある**missing link**を直接明らかにする

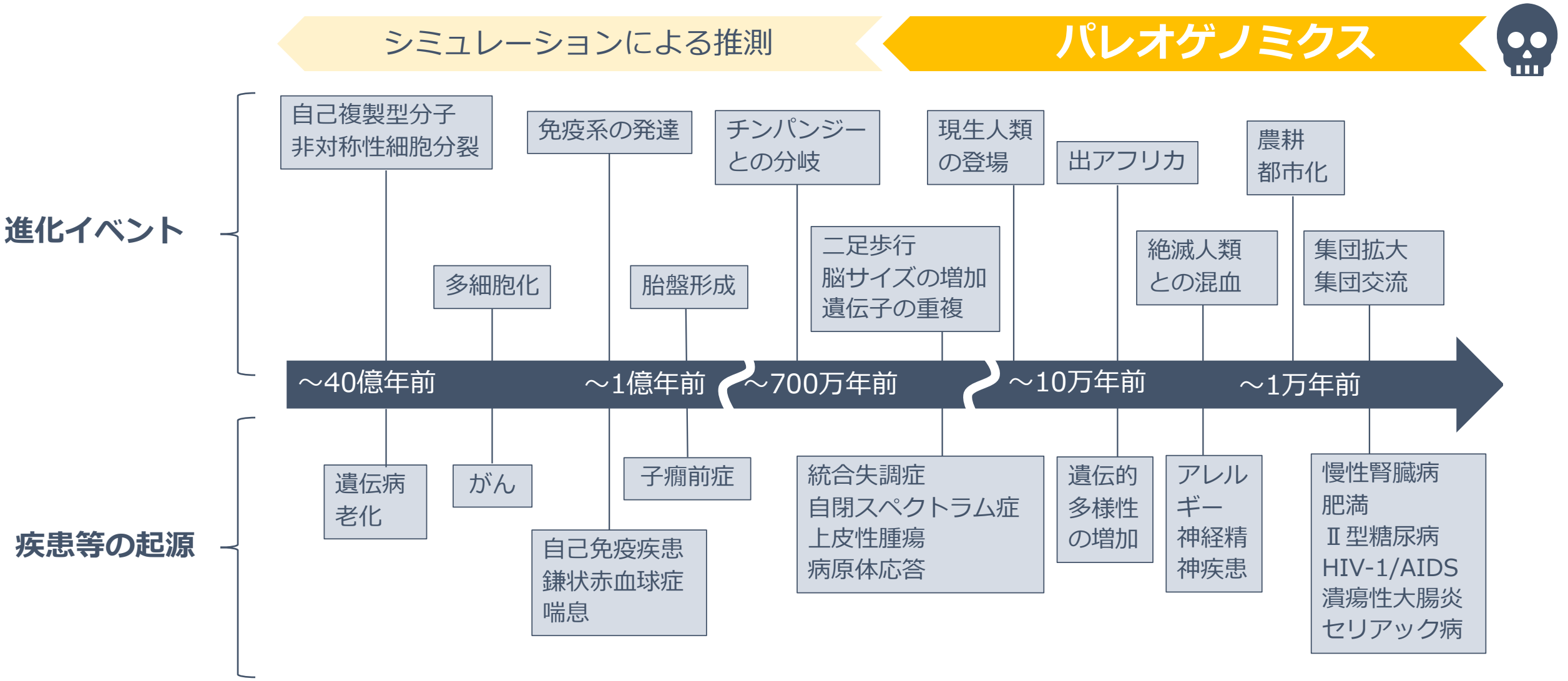
nature
https://doi.org/10.1038/s41586-020-2818-3
Accelerated Article Preview

**The major genetic risk factor for severe COVID-19 is inherited from Neanderthals**

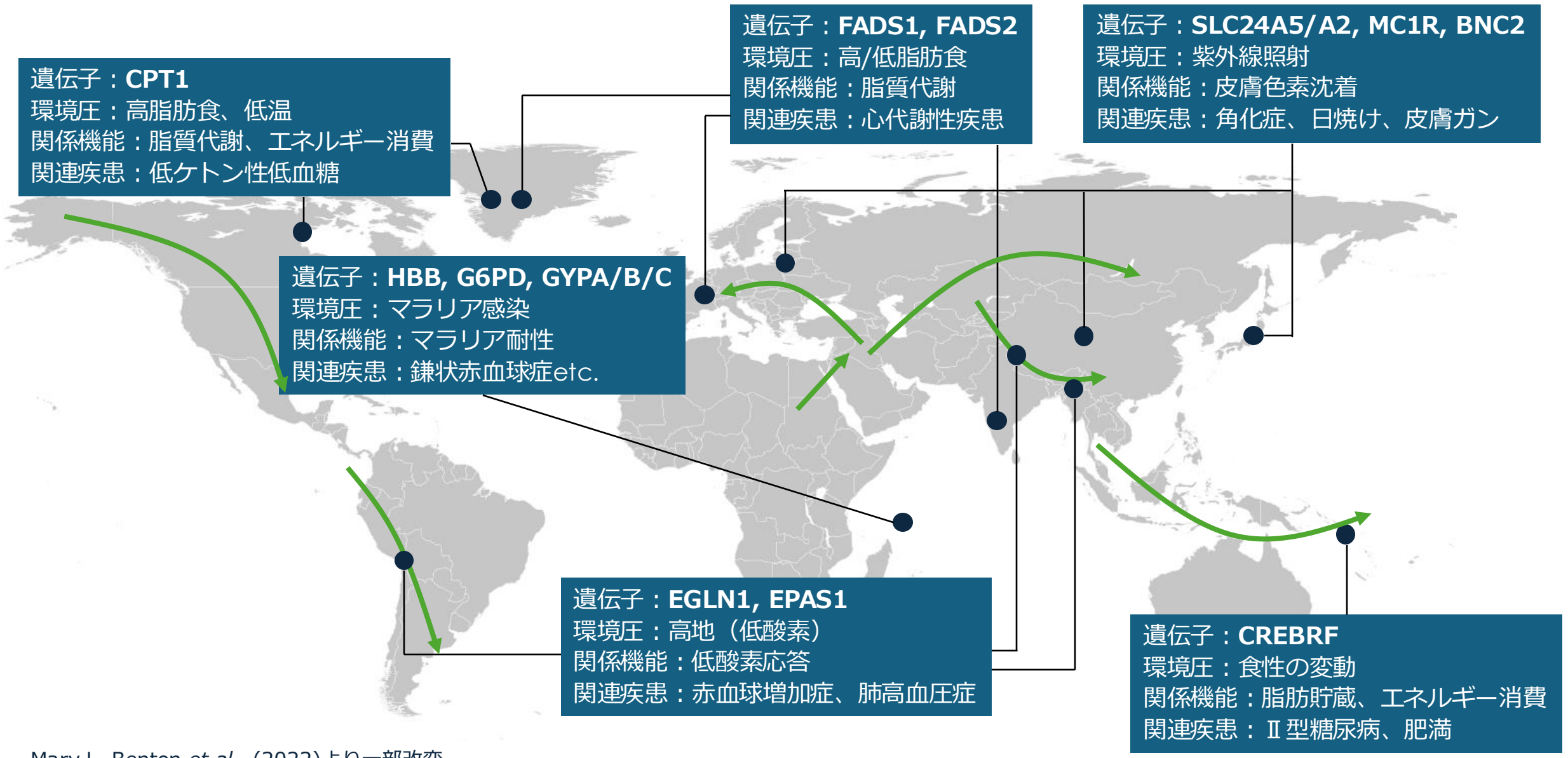
最近の研究事例：
COVID-19の重症化においてネアンデルタール人のバリアントが最大3倍のリスクをもたらすことが報告される

# パレオゲノミクスの有用性

## 過去に生じた特異的な変異や進化等、当時の状態を”直接”観察

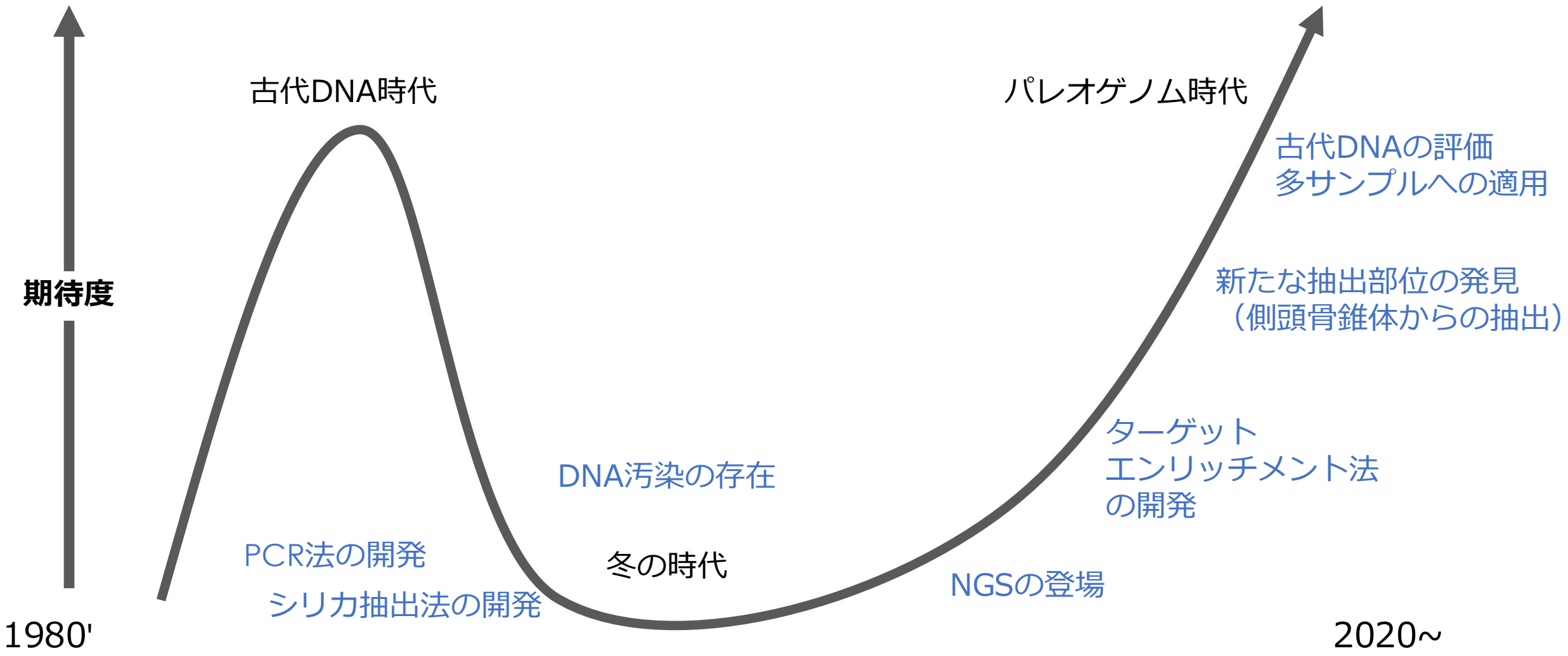


# ホモ・サピエンスの拡散と環境適応




Mary L. Benton *et al.*, (2022)より一部改変

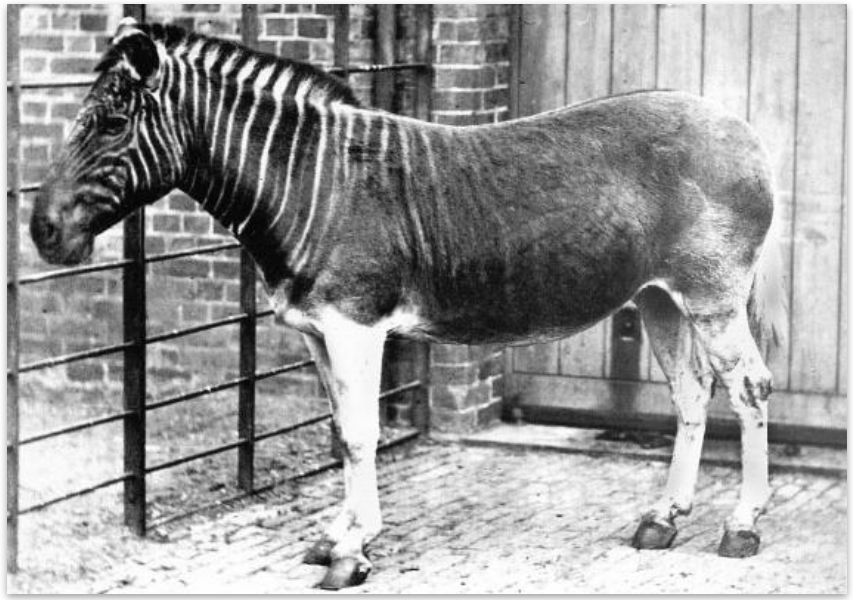
# 古代ＤＮＡ研究のハイプサイクル




| 黎明期 | 発展期 | 幻滅期 | 第二次発展期（普及期） |
|---|---|---|---|

# 古代DNA研究のはじまり

クアッガのDNA分析



- 博物館標本の筋組織を使用
- 現代試料の1%程のDNA量しか得られなかった
- Mountain ZebraのmtDNAの229bpと比較した結果.

  共通祖先が300-400万年前であると推定された（当時）

Unidentified reading frame 1

*Quagga* C CCA ATC CTG CTC GCC GTA GCA TTC CTC ACA CTA GTT GAA CGA AAA GTC TTA GGC TAC ATA CAA CTT CGT AAA GGA CCC AAC ATC GTA GGC CCC TAT GGC CTA CTA CAA CCC ATT AC

*Zebra* . . . . . T . . . . G . . . . . T . . . . . C . . . . . G*.

Cytochrome oxidase I

*Quagga* A GGA GGA TTC GTT CAC TGA TTC CCT CTA TTC TCA GGA TAC ACA CTC AAC CAA ACC TGA GCA AAA ATT CAC TTT ACA ATT ATA TTC GTA GGG GTC AAC ATA ATT TTC TTC CCA

*Zebra* G . . . . . T . . . . . G . . . . . C . . . . . A . . . . T . . . . C* . . . . . .

比較解析されたmtDNAの部分領域　Higuchi *et al*., *Nature* (1984) Fig.1 より

# 古代DNA研究のはじまり

2,400年前のエジプトミイラ（軟組織）のDNA分析 (Pääbo S. 1985 *Nature*)

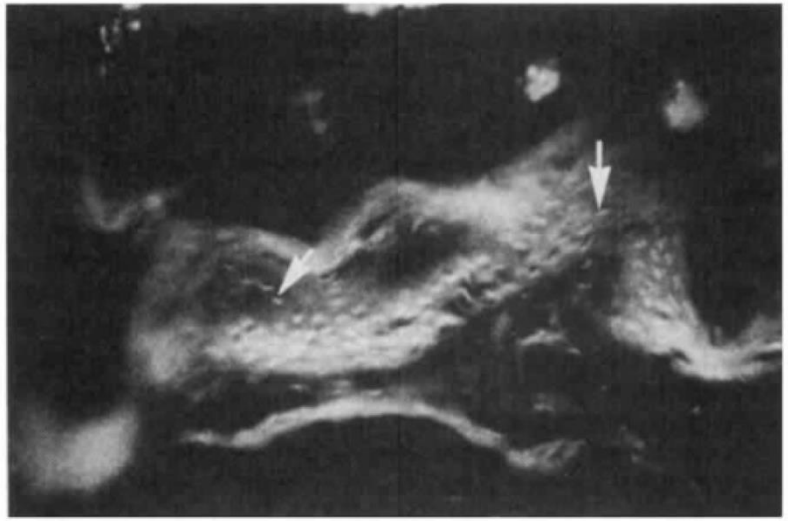


Fig.1 ミイラの組織片

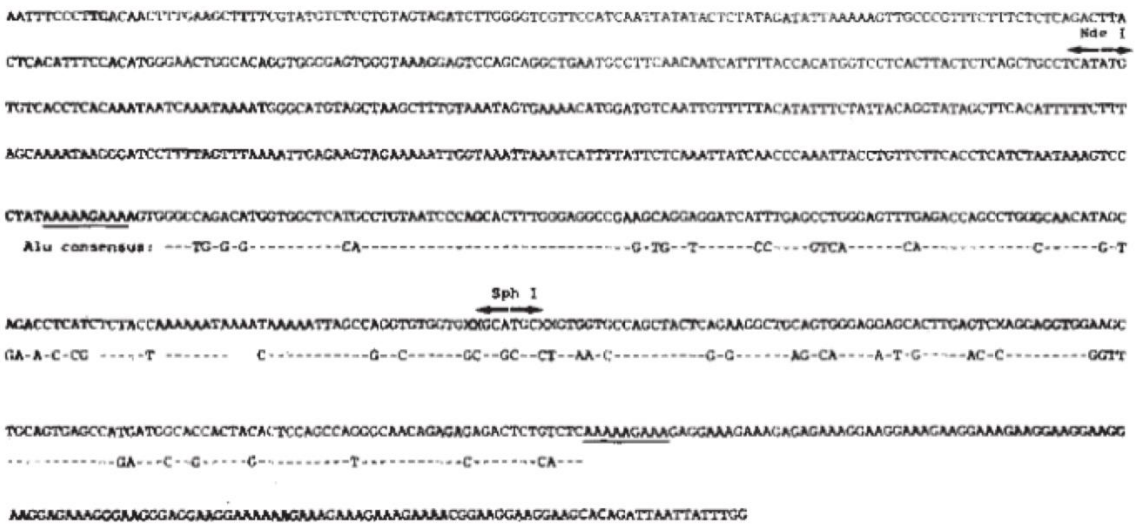



Fig.2 比較解析されたAlu配列

・死後数千年のヒト軟組織からのDNA抽出及びクローニング

・２種類のヒトAlu配列を検出

**この後、PCR法の活用により**

**古代DNA研究は普及期へ**

# 恐竜はヒトの近縁種？

**8,000万年前の恐竜に対する古代DNA分析** (Hedges and Schweitzer, 1995 *Science*)

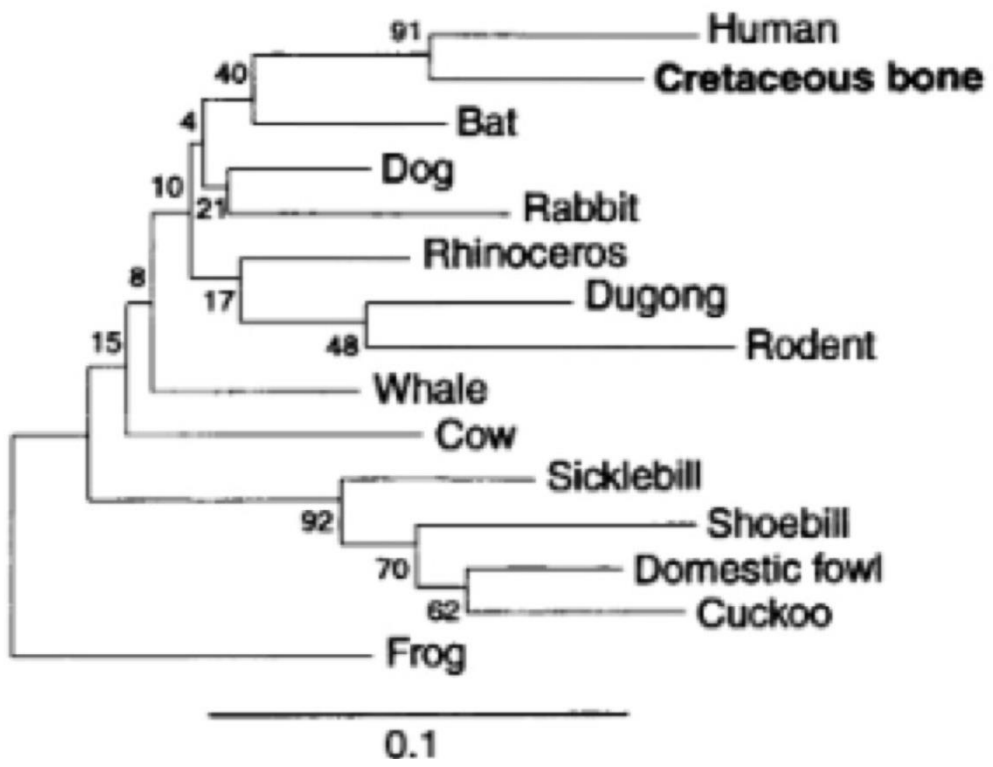



図. mtDNA Cytbの分子系統樹

8,000万年前の**恐竜**の骨を使ってDNAを抽出、

得られたDNA配列を用いて系統解析した結果

**ヒトと最も近縁** であった

# DNAの半減期

**考古遺物に基づくDNAの崩壊予測** (Allentoft *et al*., 2012 )

| temperature | *k* per site per year | half-life (years), 30 bp | half-life (years), 100 bp | half-life (years), 500 bp | average length at 10 kyr | time (years) until average length = 1 bp |
|---|---|---|---|---|---|---|
| 25°C | $4.5 \times 10^{-5}$ | 500 | 150 | 30 | 2 bp | 22 000 |
| 15°C | $7.6 \times 10^{-6}$ | 3000 | 900 | 180 | 13 bp | 131 000 |
| 5°C | $1.1 \times 10^{-6}$ | 20 000 | 6000 | 1200 | 88 bp | 882 000 |
| −5°C | $1.5 \times 10^{-7}$ | 158 000 | 47 000 | 9500 | 683 bp | 6 830 000 |

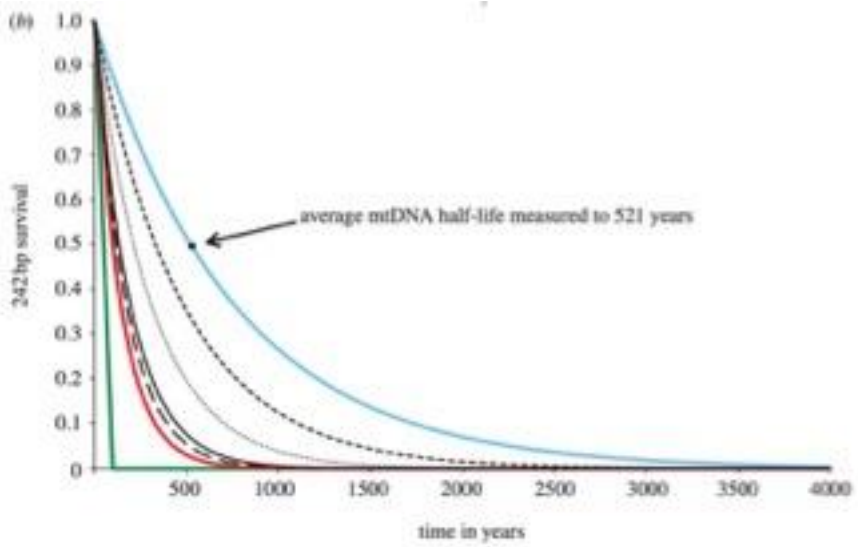



図. 経過時間におけるDNA配列の崩壊予測

DNA配列の崩壊率から (半減期を521年とした場合)

最も理想的な条件でも、地上には680万年前以上の

DNA配列 ( > 1bp) は存在しないとされている

※温度や湿度等の条件に大きく左右される

# ＤＮＡ汚染

**8,000万年前の恐竜に対する古代DNA分析** (Hedges and Schweitzer, 1995 *Science*)

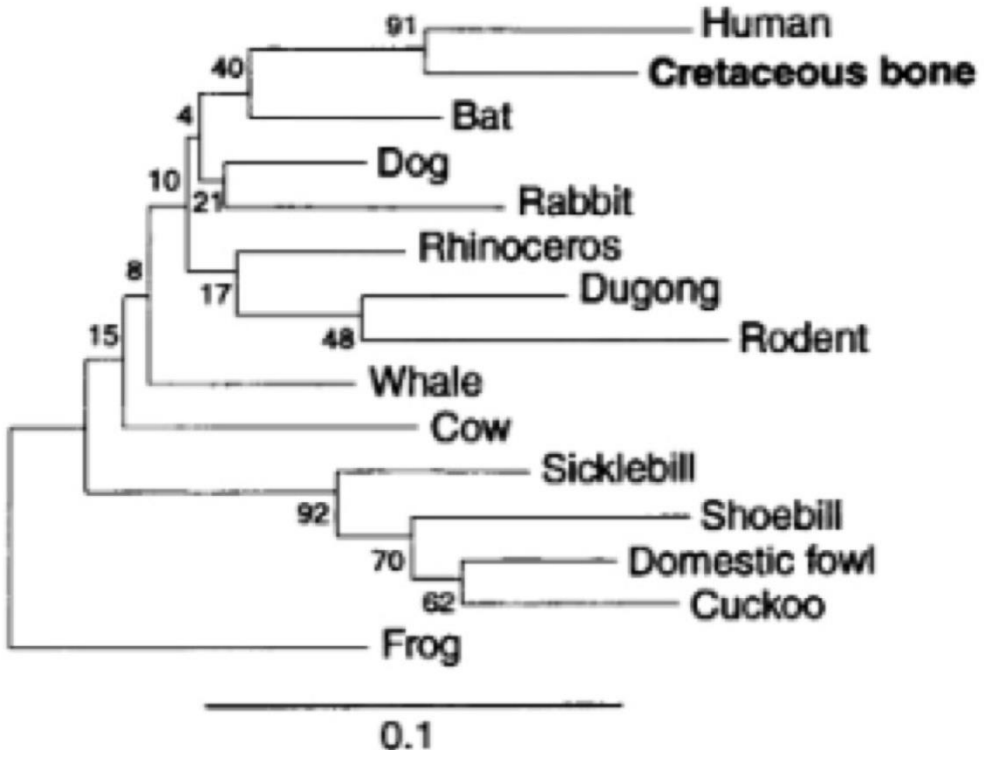



図. mtDNA Cytbの分子系統樹

8,000万年前の**恐竜**の骨を使ってDNAを抽出、

得られたDNA配列を用いて系統解析した結果

**ヒトと最も近縁** であった

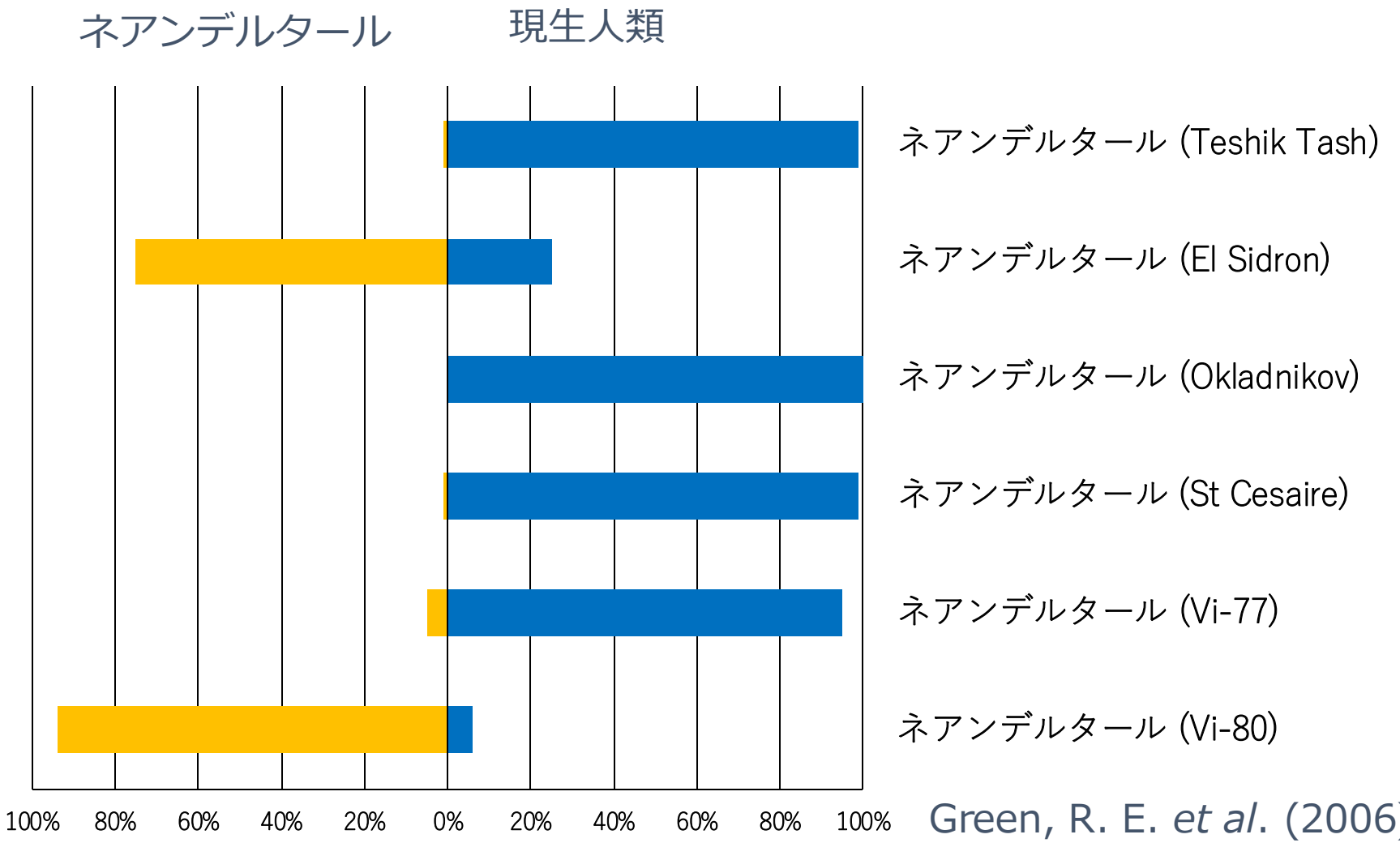
実は、、、

**ヒトDNAの混入**

**（コンタミネーション）**

# ＤＮＡ汚染

## ネアンデルタール人骨試料における現代人DNAの混入例




Green, R. E. *et al*. (2006)のデータを再解析

# ＤＮＡ汚染

## バクテリア・真菌類DNAの混入

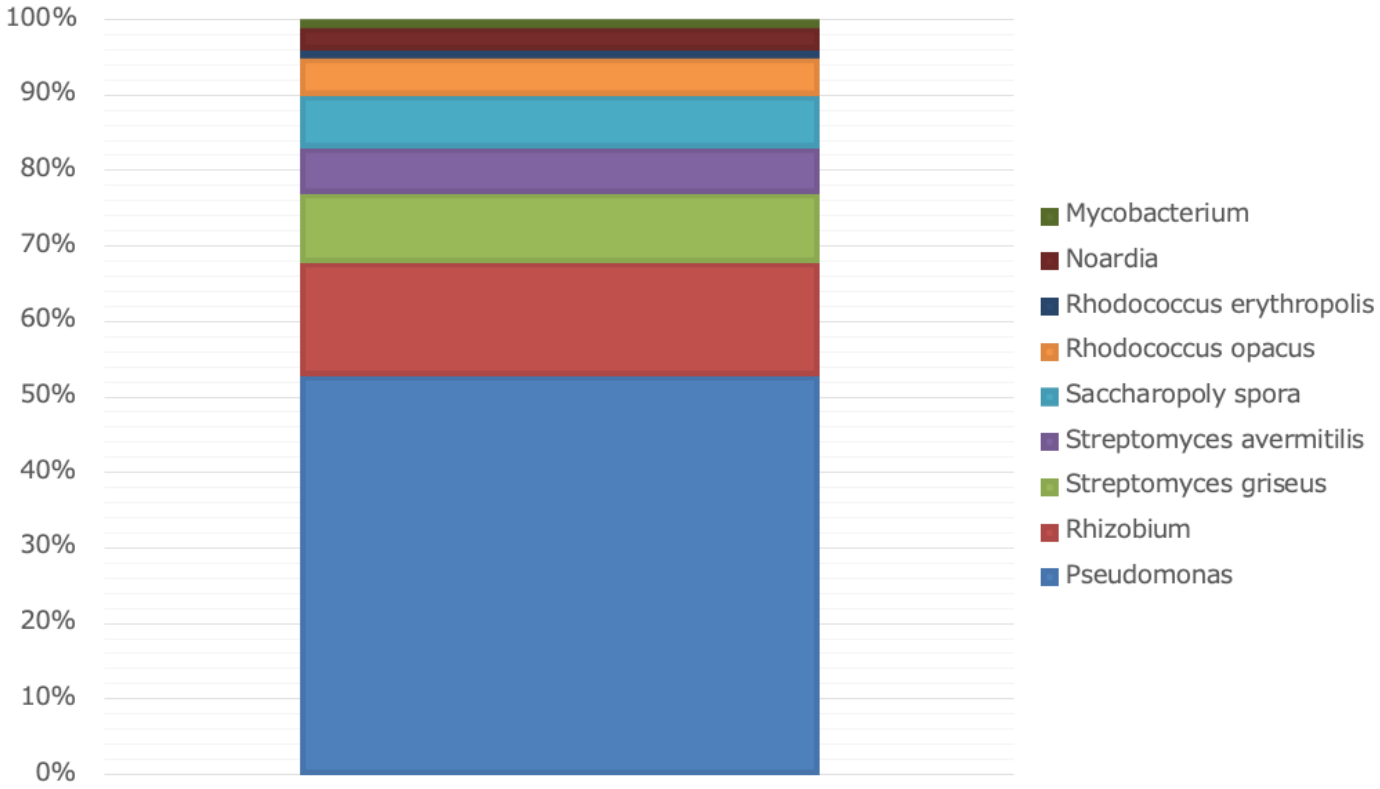



図. 古人骨抽出ＤＮＡに含まれる生物種の組成

バクテリア・真菌類由来のＤＮＡが
9割以上を占めることも

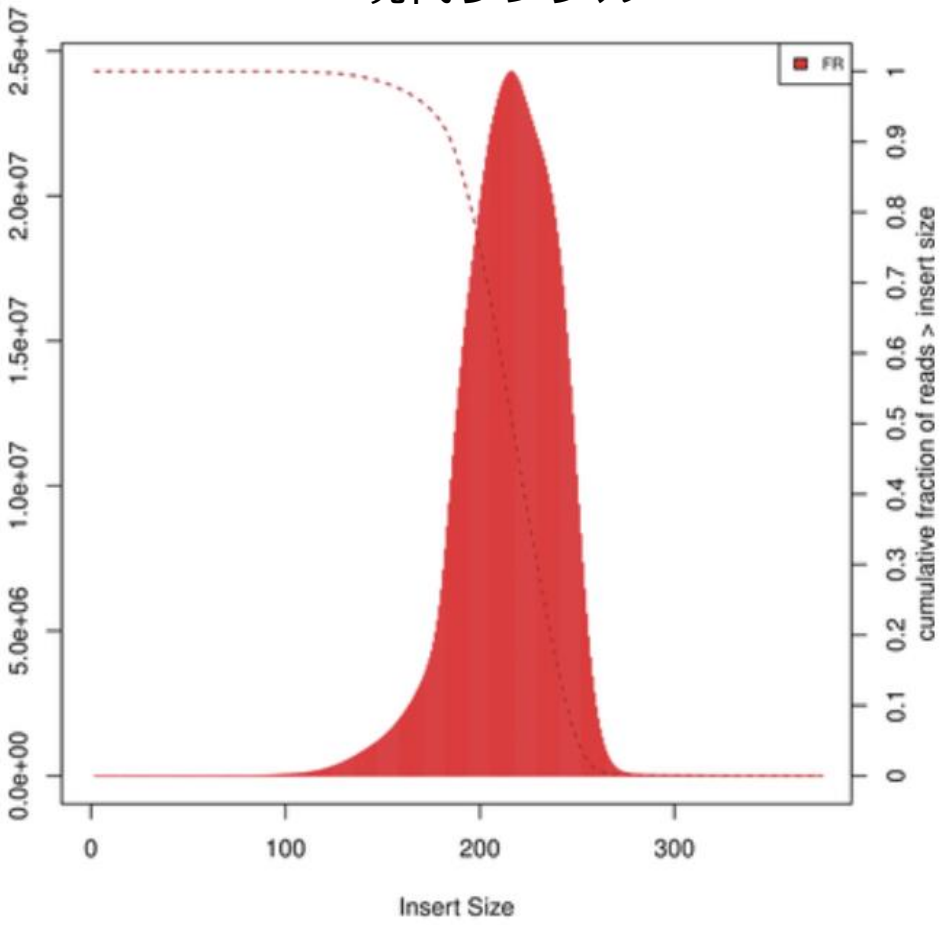
現代DNA試料の**数～10倍以上のデータを取得**する必要がある（データ & 計算量の負荷が高い）

# Ancient DNA Analysis

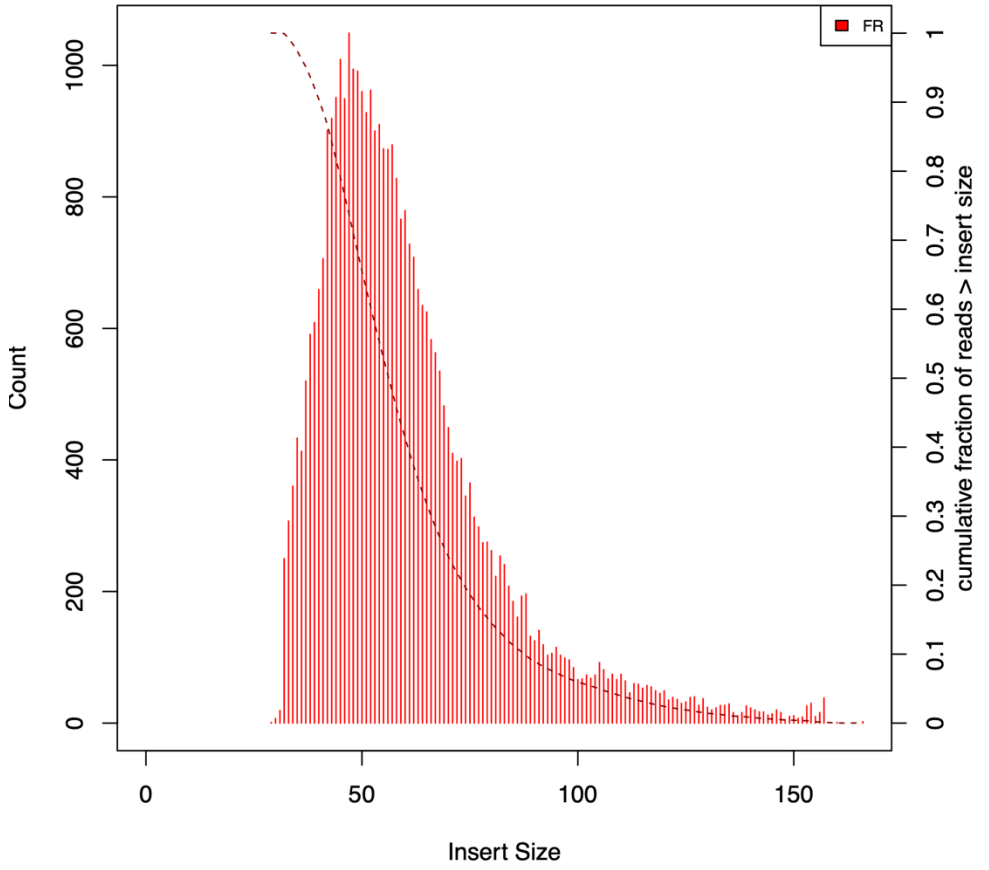
# DNA Fragmentation

```
$picard CollectInsertSizeMetrics INPUT=input.bam O=input_insert.txt H=input_insert.pdf
```

現代サンプル




古代サンプル



# 脱アミノ化

```
Consensus   TACATATTATGCTTGGCCTTACATGAGGACCTACATTTTGAAAGTTTATCTCAAGTGTATAG
Clone 1     ................T....T.....................................
Clone 2     ....................................................................
Clone 3     ....................................................................
Clone 4     ....................................................................
Clone 5     ................T....T.....................................
Clone 6     ................T....T.....................................
Clone 7     ....................G.....................................
Clone 8     ................T....T.....................................
                            ↑    ↑
```

図1. 2万6千年前のクマのDNA増幅産物

M. Hofreiter, D. Serre, H. N. Poinar, M. Kuch, S. Pääbo, ANCIENT DNA. *Nature Reviews Genetics* **2**, 353–359 (2001).

シトシンの脱アミノ化により、炭素4位のアミノ基が酸素原子に変わり、ウラシルと変化する。DNAシーケンサーは、ウラシルとチミンを区別できないため、データ上はチミンとして観察される。

脱アミノ化を検出できれば、古DNAの証拠として使える

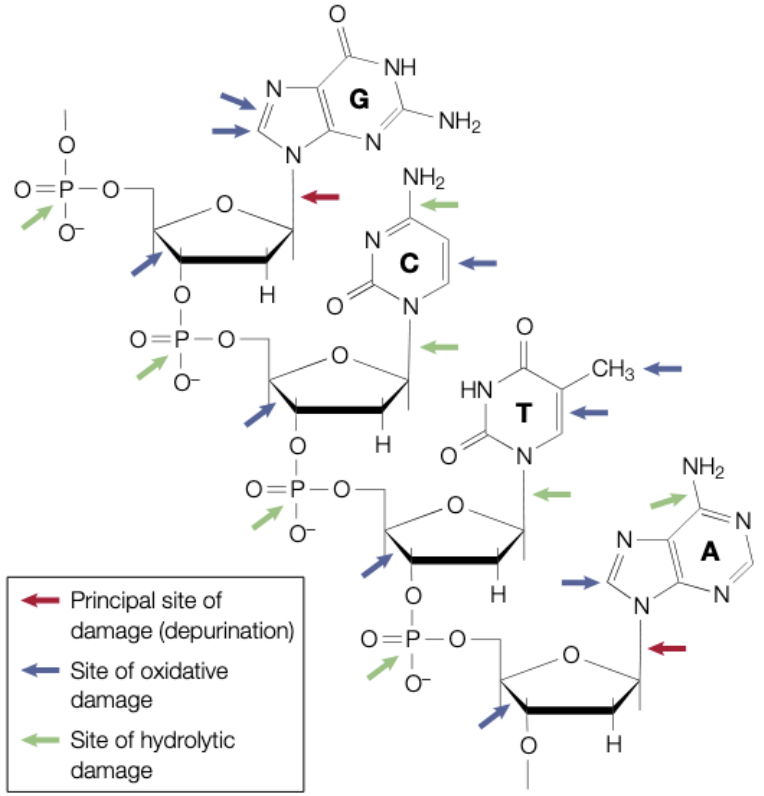



Figure 1 | **DNA damage shown or likely to affect ancient DNA.** A short segment of one strand of the DNA double helix is shown with the four common bases. Principal sites of damage are indicated by red arrows. Sites susceptible to hydrolytic attack are indicated by green arrows and those prone to oxidative damage by blue arrows. G, guanine; C, cytosine; T, thymine; A, adenine. (Modified with permission from REF. 4 © (1993) Macmillan Magazines Ltd.)

# Misincorporation pattern

## Patterns of damage in genomic DNA sequences from a Neandertal

**Adrian W. Briggs*†, Udo Stenzel*, Philip L. F. Johnson‡, Richard E. Green*, Janet Kelso*, Kay Prüfer*, Matthias Meyer*, Johannes Krause*, Michael T. Ronan§, Michael Lachmann*, and Svante Pääbo*†**

*Max Planck Institute for Evolutionary Anthropology, Deutscher Platz 6, D-04103 Leipzig, Germany; ‡Biophysics Graduate Group, University of California, Berkeley, CA 94720; and §454 Life Sciences, Branford, CT 06405

Contributed by Svante Pääbo, May 25, 2007 (sent for review April 25, 2007)

**Table 1. Maximum likelihood estimates (MLE) for four features of Neandertal DNA sequences**

| Parameter | MLE | 95% C.I. |
|---|---|---|
| Deamination, double-stranded DNA ($\hat{\delta}$) | 0.0097 | (0.0087, 0.011) |
| Deamination, single-stranded DNA ($\hat{\delta}_{ss}$) | 0.68 | (0.65, 0.71) |
| Nick frequency per base ($\hat{\nu}$) | 0.024 | (0.017, 0.036) |
| Length of single-stranded overhangs ($\hat{\lambda}$) | 0.36 | (0.35, 0.38) |

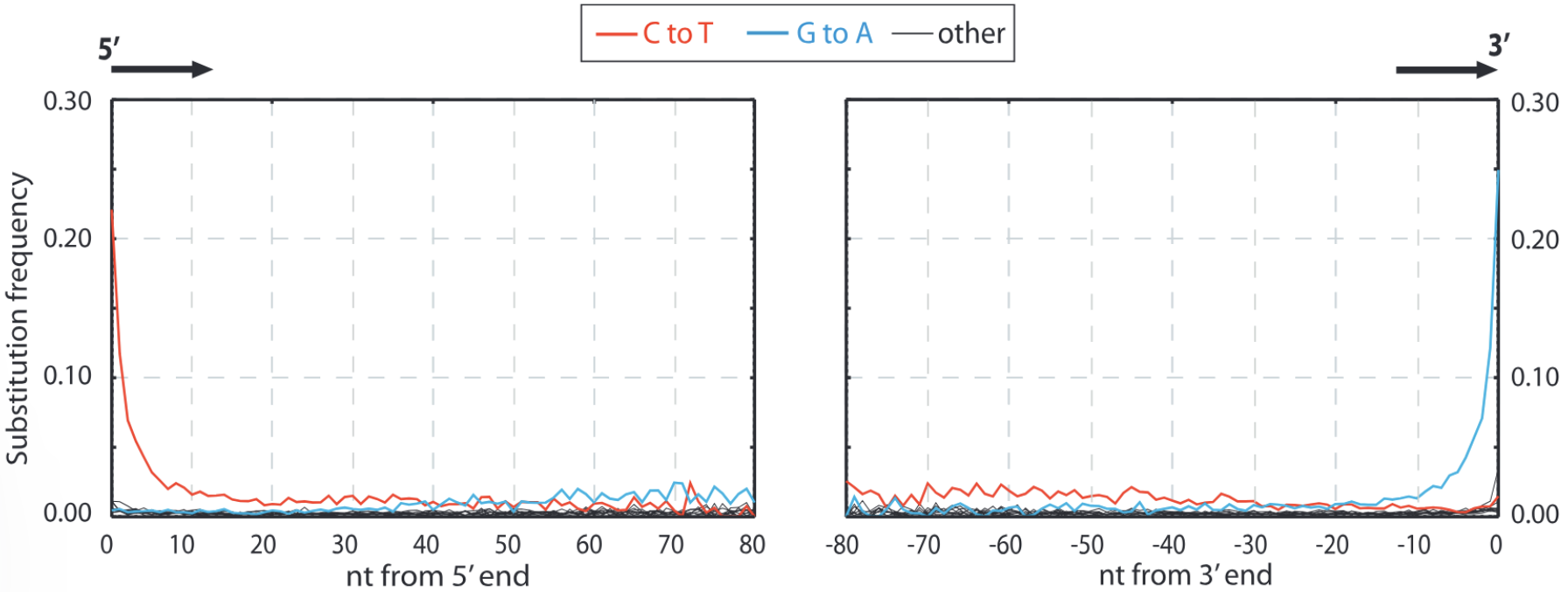
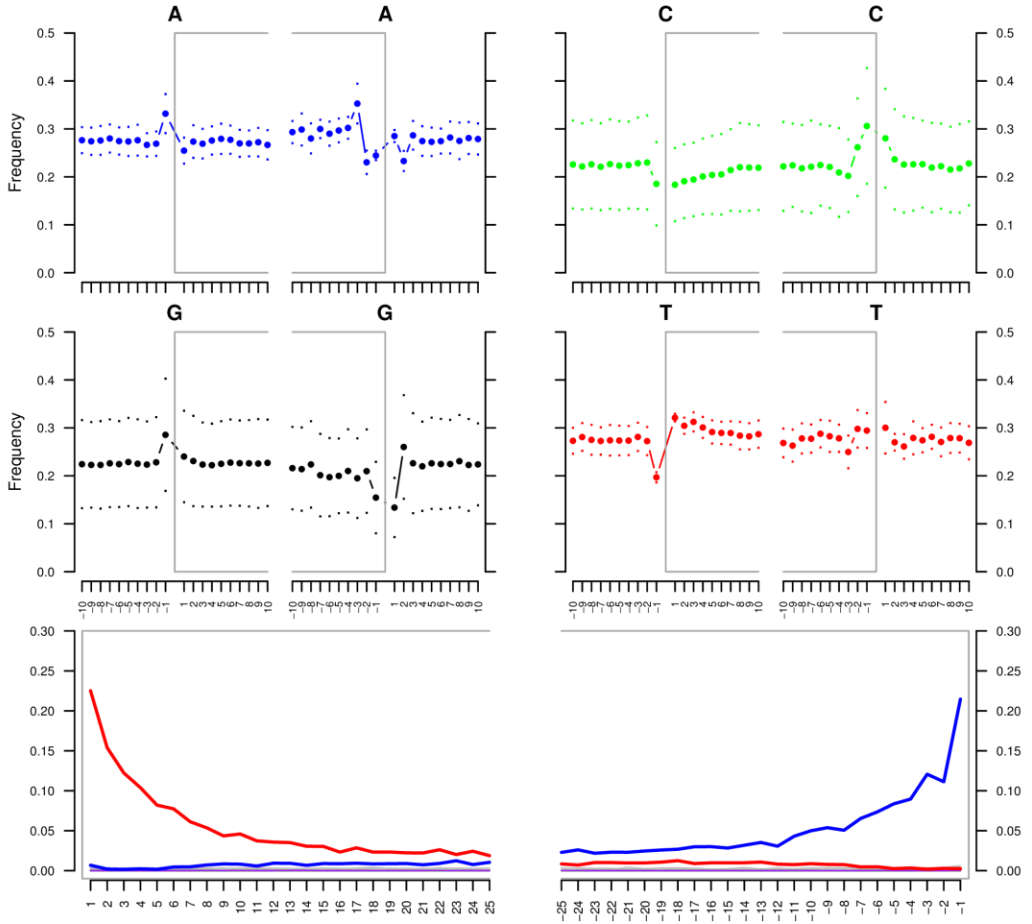
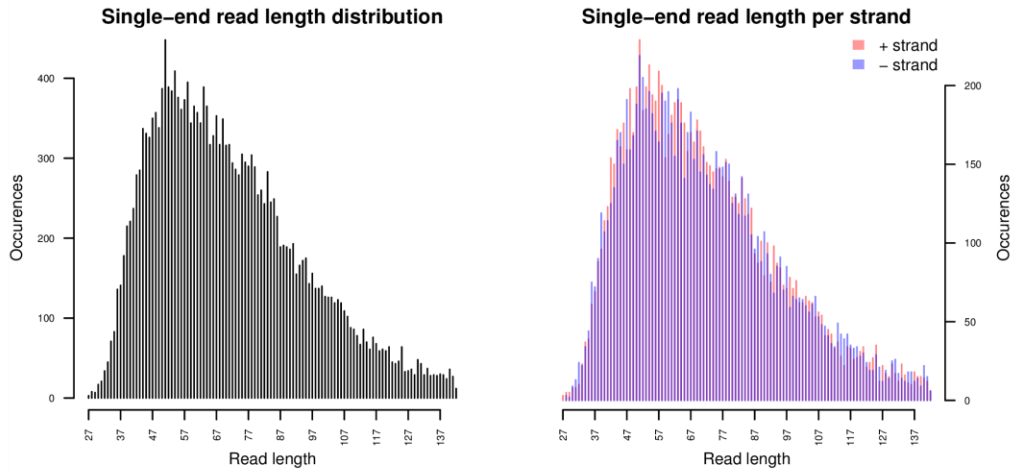


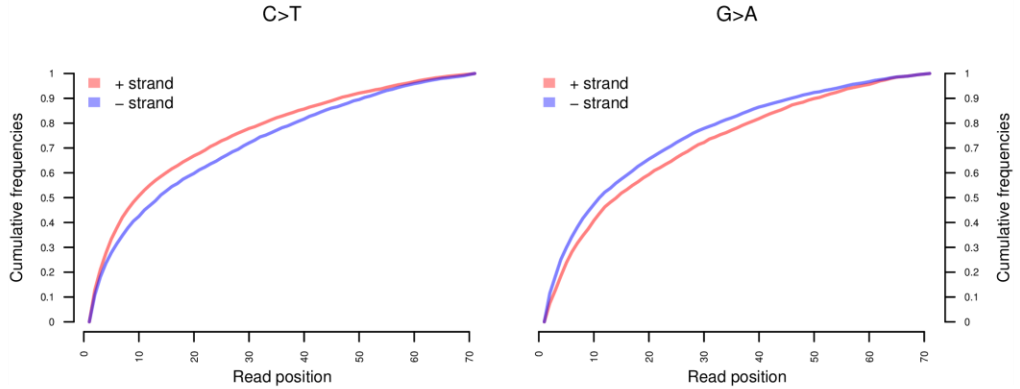

**Fig. 3.** Misincorporation patterns in Neandertal DNA sequences. The frequencies of the 12 possible mismatches are plotted as a function of distance from 5′- and 3′-ends. At each position, the substitution frequency, e.g., C-T, is calculated as the proportion of human reference sequence positions carrying C where the 454 sequence is T. The 10 5′- and 10 3′-most nucleotides were removed from the 3′- and 5′-graphs, respectively.

# MapDamage

$mapDamage -i input.bam -r ucsc.hg19.fasta



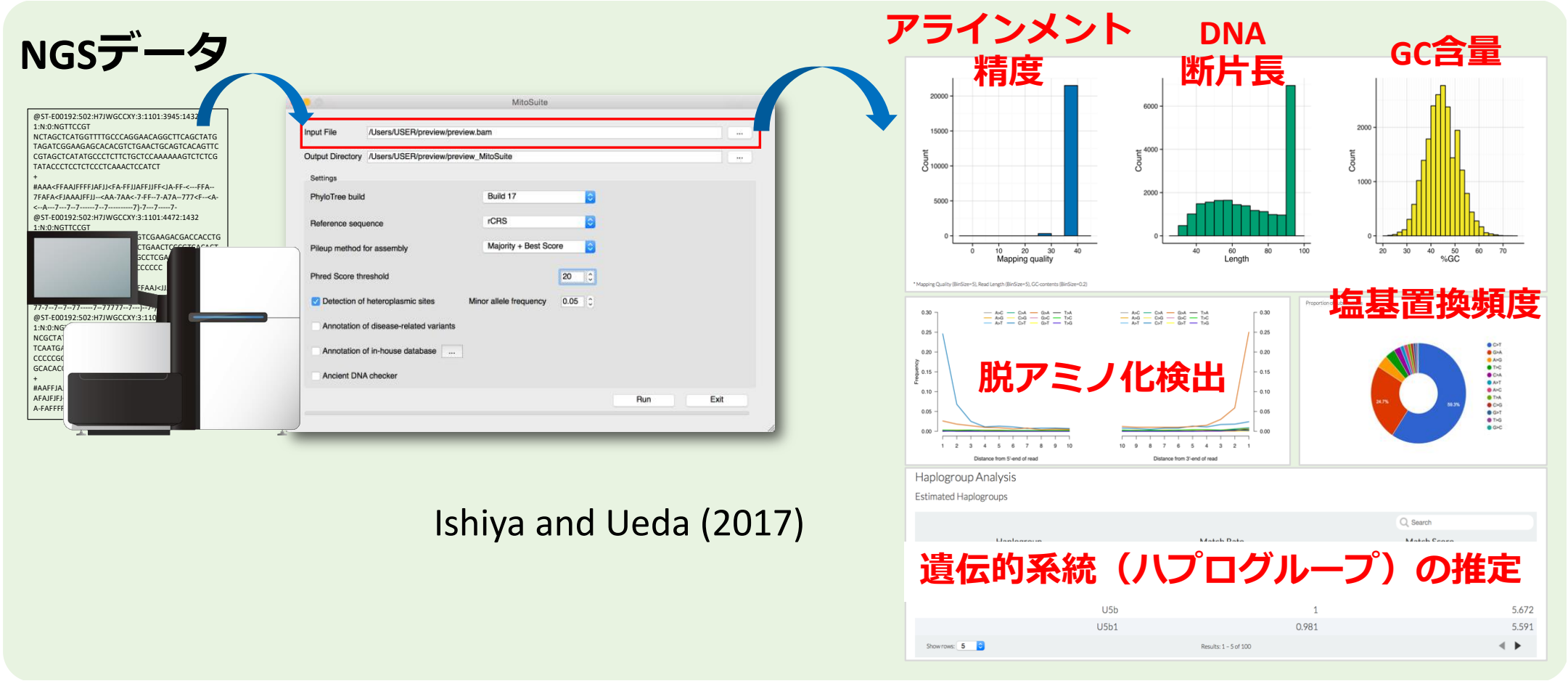




A. Ginolhac, M. Rasmussen, M. T. P. Gilbert, E. Willerslev, L. Orlando, mapDamage: testing for damage patterns in ancient DNA sequences. *Bioinformatics* **27**, 2153–2155 (2011).

# 古代人ゲノムデータの品質管理解析

GUIで品質管理・脱アミノ化検出・変異解析等を一括で行う統合解析ソフト

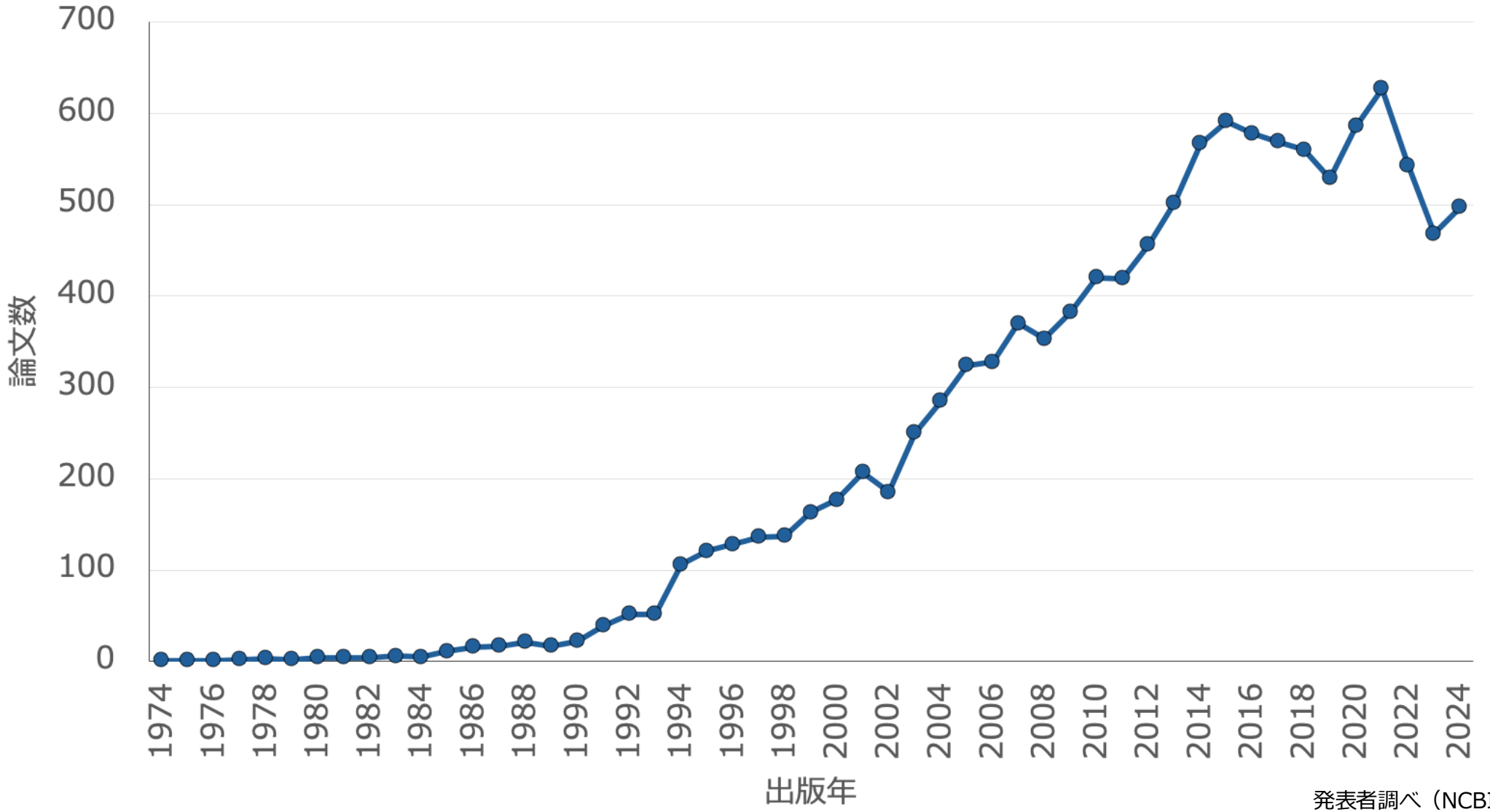



Ishiya and Ueda (2017)

K. Ishiya, S. Ueda, MitoSuite: a graphical tool for human mitochondrial genome profiling in massive parallel sequencing. *PeerJ* **5**, e3406 (2017).

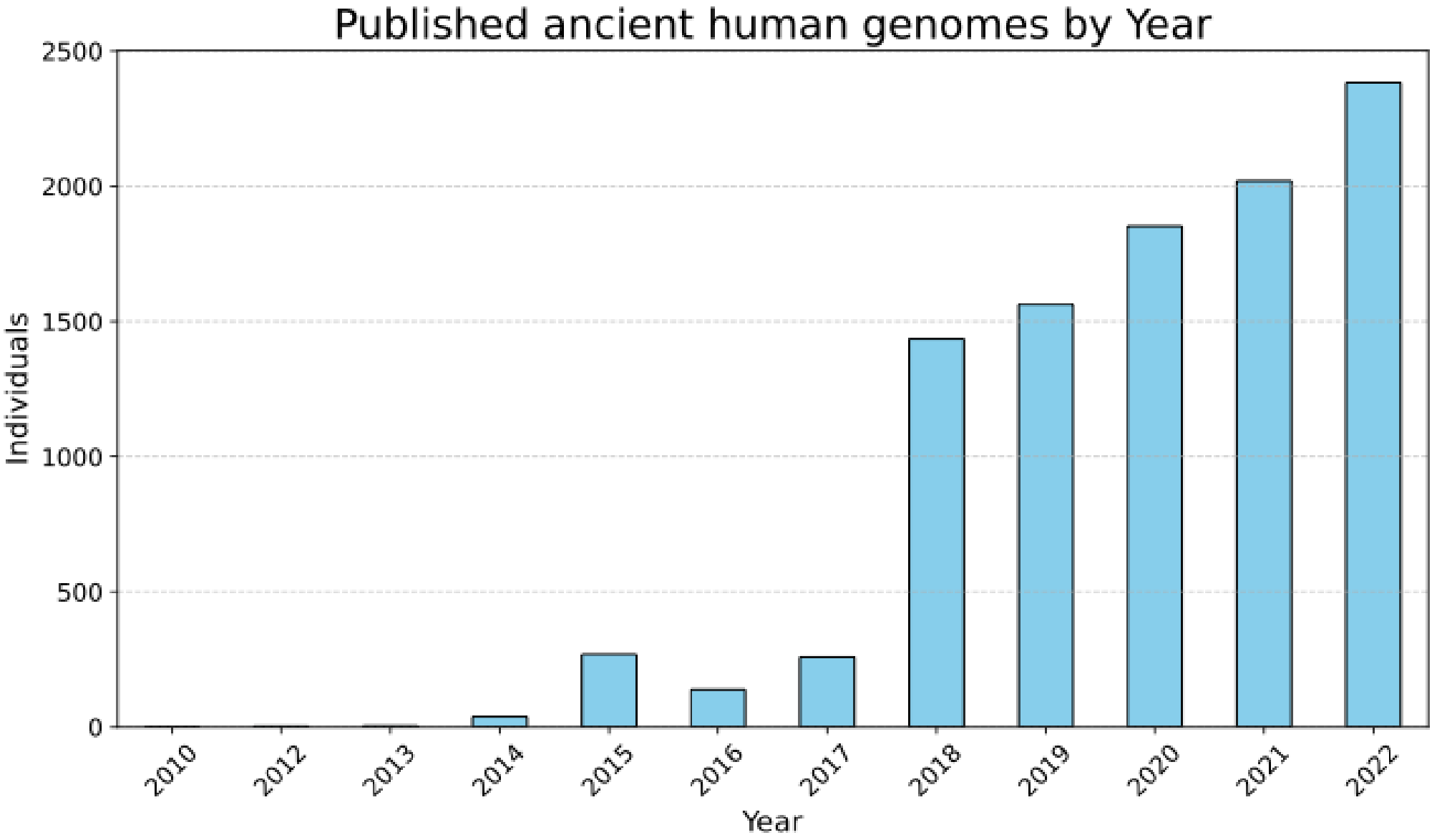
# パレオゲノミクス研究の発展

パレオゲノミクス(古ＤＮＡ）関連研究の報告例




発表者調べ（NCBI PubMed 2024年末時点）

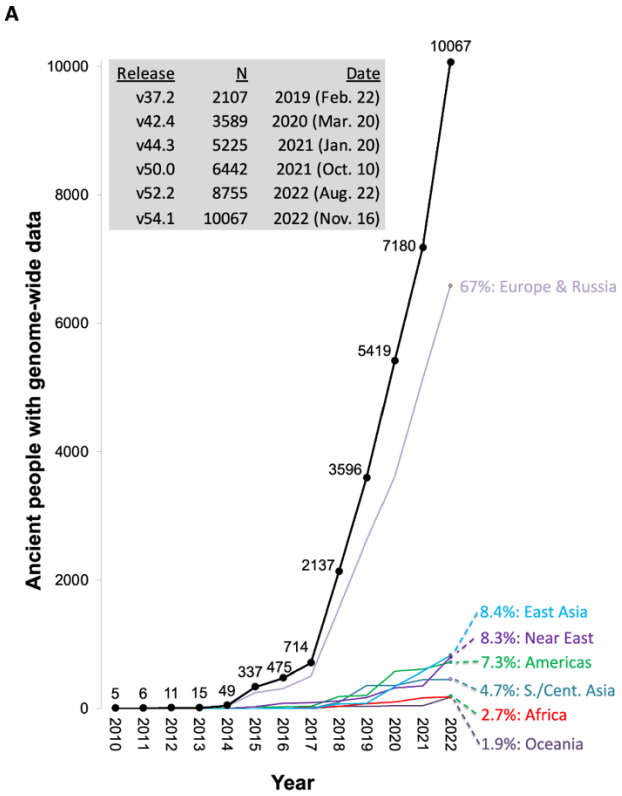
# 古代人ゲノムの公開数の推移




発表者調べ（古代人ゲノム 9,990個体を対象）

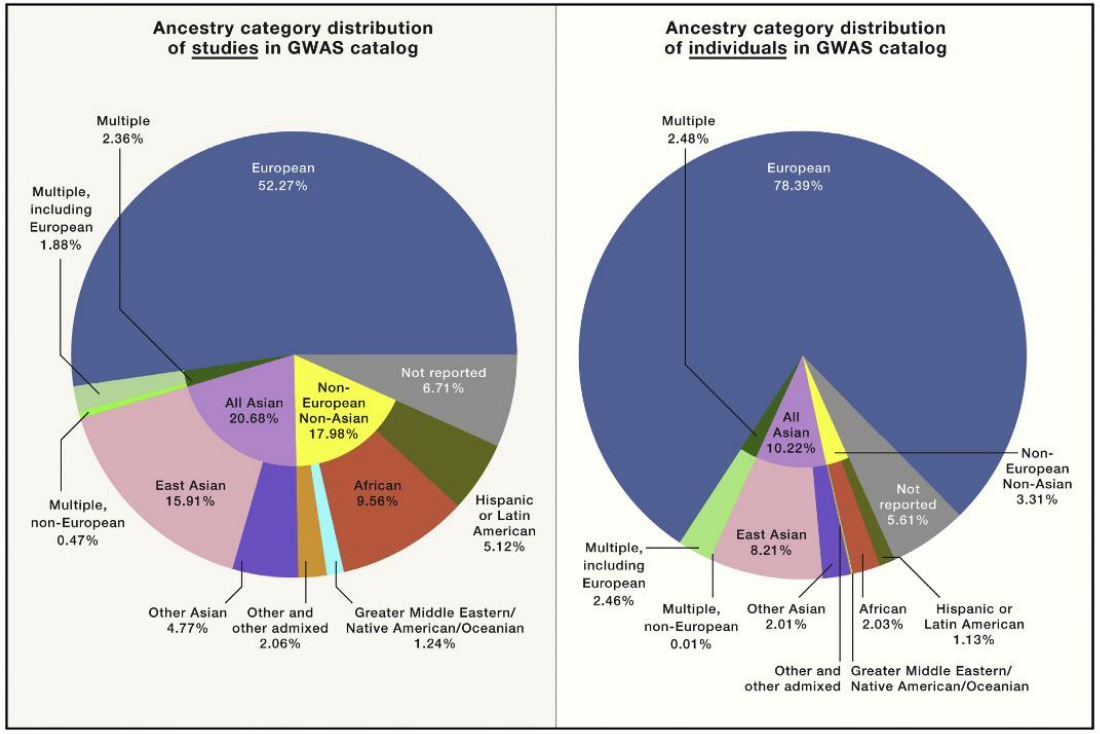
# ゲノム人類学におけるオープンサイエンス化と傾向

## 約6割以上が欧米の古代人ゲノムデータが占める




S. Mallick, A. Micco, M. Mah, H. Ringbauer, I. Lazaridis, I. Olalde, N. Patterson, D. Reich, The Allen Ancient DNA Resource (AADR) a curated compendium of ancient human genomes. Sci. Data 11, 182 (2024).

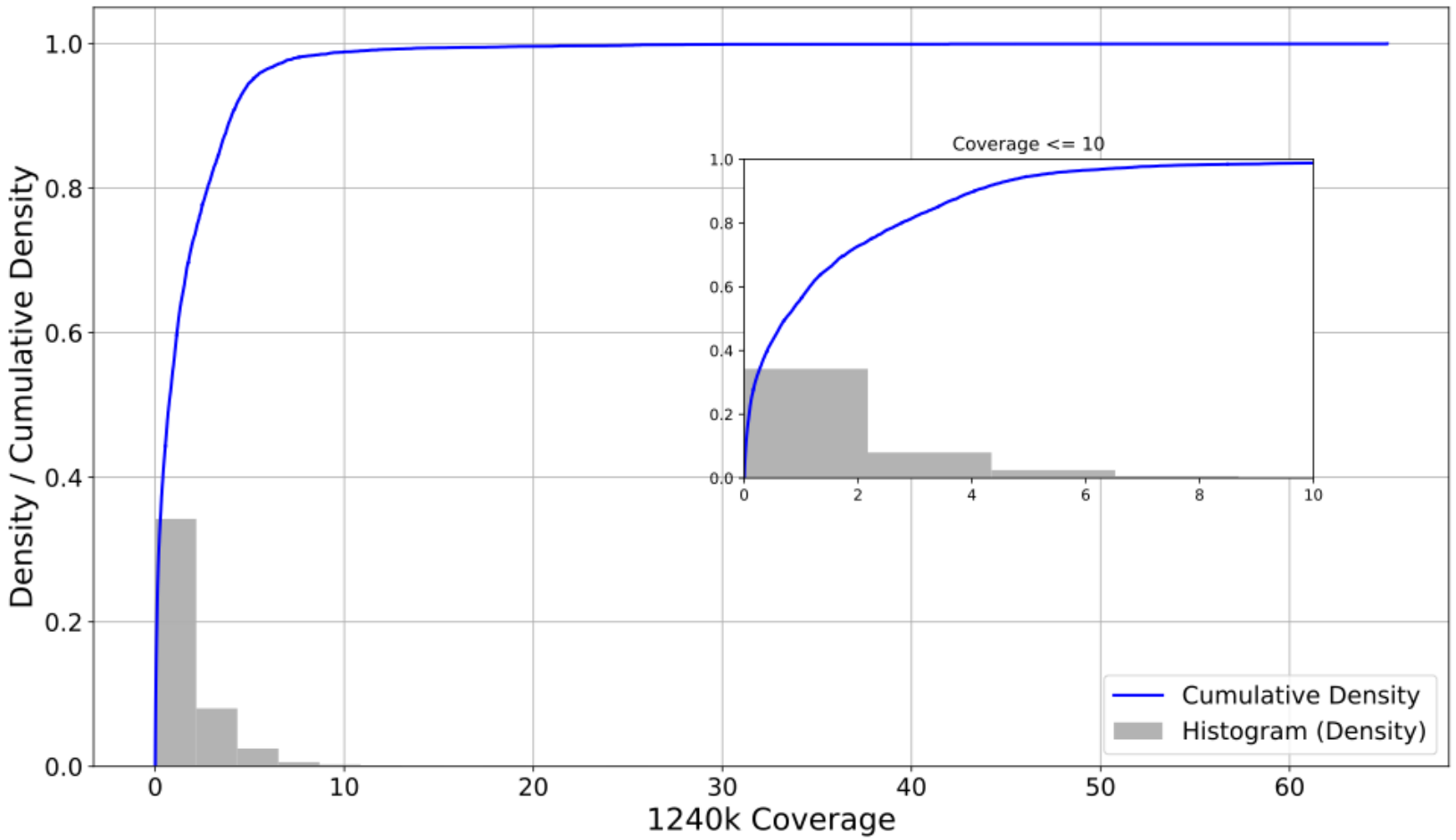
## 現代人ゲノムデータでも同様の傾向




G. Sirugo, S. M. Williams, S. A. Tishkoff, The Missing Diversity in Human Genetic Studies. Cell 177, 26–31 (2019).

**データベースに格納される集団に不均衡が生じている**

# 古代人ゲノムデータにおける課題




再現性や
信頼性の担保
が困難

発表者調べ（古代人ゲノム 9,990個体を対象）

**公開済みの古代人ゲノムの半数以上が1X以下（データ品質の改善が必要）**

# 古DNA解析の効率化

## 古ゲノム解析に必要な作業の一例

1. FASTQの品質チェック（例：リード配列が正確に読めているか？データに異常はないか？コンタミはないか？）
2. アダプター配列のトリミング（例：シーケンスのために付加したアダプター配列の除去）
3. FASTQのフィルタリング（例：低品質の塩基はエラーが含まれるので除外）
4. 参照ゲノムへのアラインメント（例：短いリード配列やエラーを許容できるか？）
5. アラインメント結果の確認（例：何%のリード配列がアラインメントされたか？ペア関係が壊れていないか？）
6. アラインメント結果のフィルタリング（例：誤ってアラインメントされたリードを除去）
7. 重複リードの除去（例：PCRやクラスター形成時に生じた複製による影響の除外）
8. 古DNAのダメージ評価（例：古DNAに特徴的な脱アミノや脱プリンが見られるか？）
9. 脱アミノ化の補正（例：脱アミノ化が生じたリード配列末端付近を除去）
10. ペア関係の修復（例：末端除去によって影響のあるペア関係を修復）
11. インサートサイズの計算（例：DNA断片化と断片長の確認）
12. 塩基配列の品質スコアの再校正（例：ハードウェア、試薬の品質、ランごとのバイアス修正）
13. 変異のコール（例：どの位置に変異が含まれているか？）
14. 変異のフィルタリング（例：誤って検出された変異の除去）
15. 変異のアノテーション（例：既知の変異と対応させる、新規変異の検出等）
16. 変異情報の統合化（例：既知集団ゲノムとのデータ統合）

… + 集団遺伝学解析

**多数のツールの導入（インストール）、各ツールごとのアルゴリズムや条件設定への理解が必須**
**大規模なシーケンスを行う場合、より大きな計算設備や時間も多くかかってくる**

# 生命科学研究におけるWFMの利用

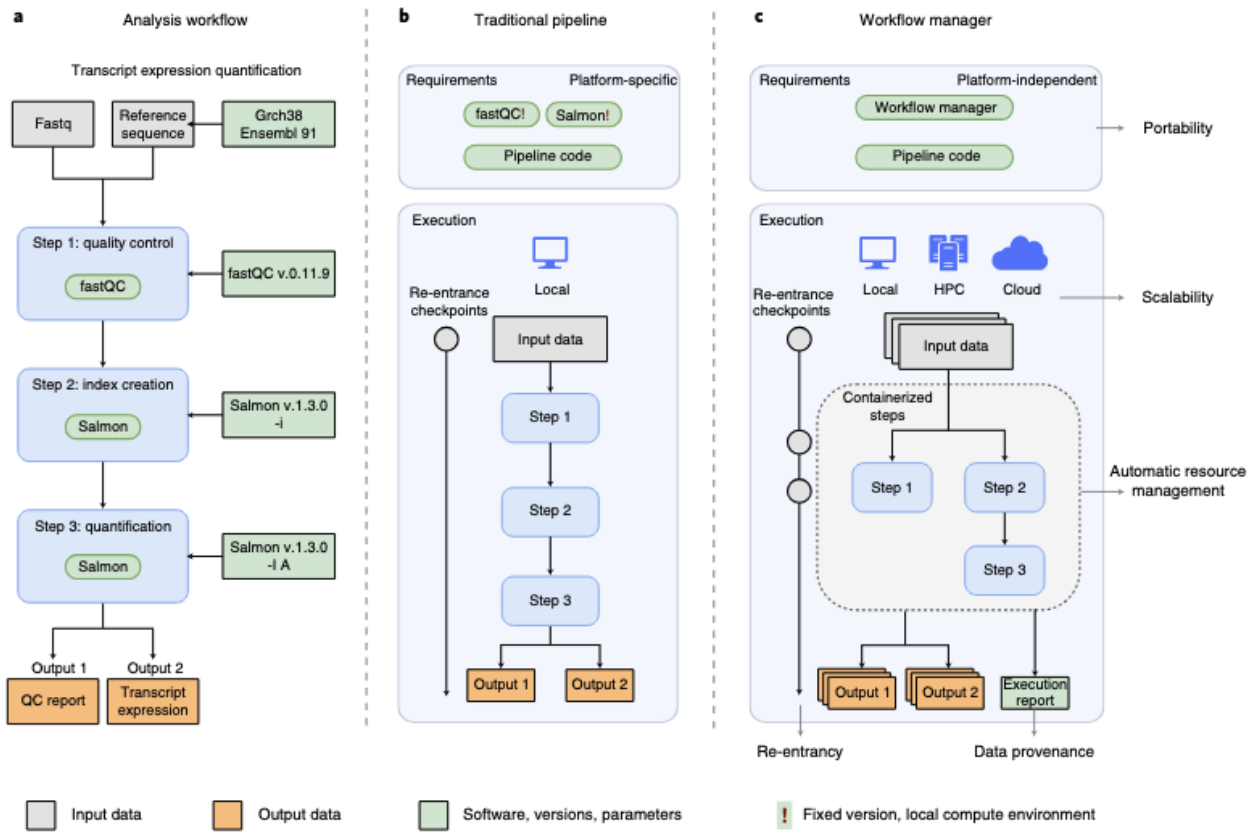
nature methods
PERSPECTIVE
https://doi.org/10.1038/s41592-021-01254-9
Check for updates

## Reproducible, scalable, and shareable analysis pipelines with bioinformatics workflow managers

Laura Wratten[1], Andreas Wilm[2] and Jonathan Göke[1]✉

**The rapid growth of high-throughput technologies has transformed biomedical research. With the increasing amount and complexity of data, scalability and reproducibility have become essential not just for experiments, but also for computational analysis. However, transforming data into information involves running a large number of tools, optimizing parameters, and integrating dynamically changing reference data. Workflow managers were developed in response to such challenges. They simplify pipeline development, optimize resource usage, handle software installation and versions, and run on different compute platforms, enabling workflow portability and sharing. In this Perspective, we highlight key features of workflow managers, compare commonly used approaches for bioinformatics workflows, and provide a guide for computational and noncomputational users. We outline community-curated pipeline initiatives that enable novice and experienced users to perform complex, best-practice analyses without having to manually assemble workflows. In sum, we illustrate how workflow managers contribute to making computational analysis in biomedical research shareable, scalable, and reproducible.**

L. Wratten, A. Wilm, J. Göke, Reproducible, scalable, and shareable analysis pipelines with bioinformatics workflow managers. Nat Methods 18, 1161–1168 (2021).




**ＷＦＭは解析のスケーラビリティや再現性を担保する上で重要**

# 生命科学研究におけるWFMの利用

| Tool | Class | Ease of use[a] | Expressiveness[b] | Portability[c] | Scalability[d] | Learning resources[e] | Pipeline initiatives[f] |
|---|---|---|---|---|---|---|---|
| Galaxy | Graphical | ●●● | ●○○ | ●●● | ●●● | ●●● | ●●○ |
| KNIME | Graphical | ●●● | ●○○ | ○○○ | ●●◐ | ●●● | ●●○ |
| Nextflow | DSL | ●●○ | ●●● | ●●● | ●●● | ●●● | ●●● |
| Snakemake | DSL | ●●○ | ●●● | ●●◐ | ●●● | ●●○ | ●●● |
| GenPipes | DSL | ●●○ | ●●● | ●●○ | ●●○ | ●●○ | ●●○ |
| bPipe | DSL | ●●○ | ●●● | ●●○ | ●●◐ | ●●○ | ●○○ |
| Pachyderm | DSL | ●●○ | ●●● | ●○○ | ●●○ | ●●● | ○○○ |
| SciPipe | Library | ●●○ | ●●● | ○○○ | ○○○ | ●●○ | ○○○ |
| Luigi | Library | ●●○ | ●●● | ●○○ | ●●◐ | ●●○ | ○○○ |
| Cromwell + WDL | Execution + workflow specification | ●○○ | ●●○ | ●●● | ●●◐ | ●●○ | ●●○ |
| cwltool + CWL | Execution + workflow specification | ●○○ | ●●○ | ●●◐ | ○○○ | ●●● | ●●○ |
| Toil + CWL/WDL/Python | Execution + workflow specification | ●○○ | ●●● | ●◐○ | ●●● | ●●○ | ●●○ |

GUIで操作性に優れる

スケーラビリティに優れており、高速

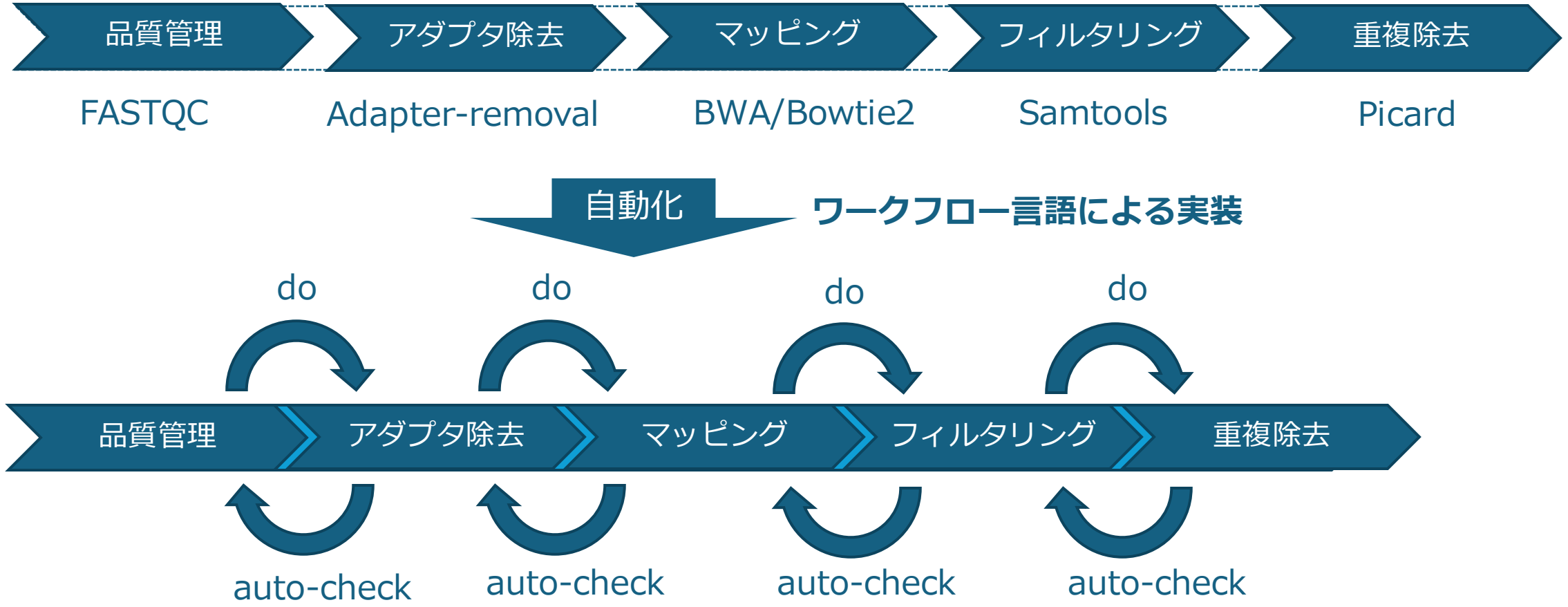
パレオゲノミクス解析に向けに開発・運用中

L. Wratten, A. Wilm, J. Göke, Reproducible, scalable, and shareable analysis pipelines with bioinformatics workflow managers. Nat Methods 18, 1161–1168 (2021).

# ワークフロー言語による自動化

## 自動化（Automation）

例）ゲノムマッピング　　従来は複数ソフトウェアを独立に実行、目視での確認、パラメータ設定、バージョン管理等が必要




**処理・対応を自動化 → 解析時間の短縮、ヒューマンエラーや解析エラーの防止**

# ワークフロー言語による自動化

## 自動化（Automation）

例）ワークフロー言語による自動化プログラムの実装例

```
SAMPLES = ["A", "B"]


rule all:
    input:
        "plots/quals.svg"


rule bwa_map:
    input:
        "data/genome.fa",
        "data/samples/{sample}.fastq"
    output:
        "mapped_reads/{sample}.bam"
    shell:
        "bwa mem {input} | samtools view -Sb - > {output}"


rule samtools_sort:
    input:
        "mapped_reads/{sample}.bam"
    output:
        "sorted_reads/{sample}.bam"
    shell:
        "samtools sort -T sorted_reads/{wildcards.sample} "
        "-O bam {input} > {output}"
```

初期設定

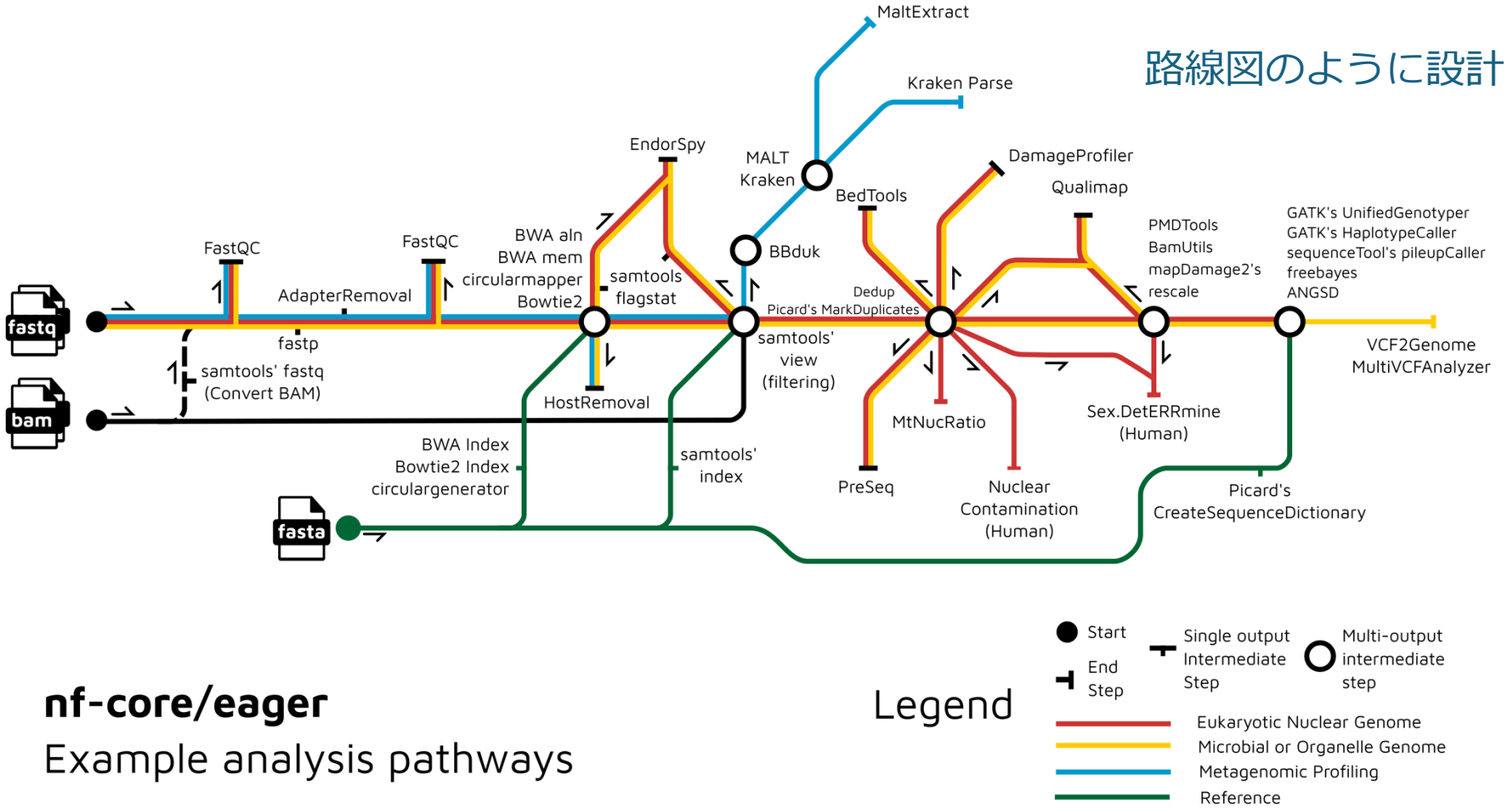
解析①（入力①→処理①→出力①）

解析②（入力①→処理②→出力②）

# パレオゲノミクス解析の効率化

## 設計（Design） & 最適化（Optimization）

路線図のように設計&処理効率の最適化を行う




**nf-core/eager**

Example analysis pathways




J. A. F. Yates, T. C. Lamnidis, M. Borry, A. A. Valtueña, Z. Fagernäs, S. Clayton, M. U. Garcia, J. Neukamm, A. Peltzer, Reproducible, portable, and efficient ancient genome reconstruction with nf-core/eager. *PeerJ* **9**, e10947 (2021).

# 他分野データ連携・応用

## 血液凝固亢進症やニコチン依存症関連変異はネアンデルタール人との交雑でもたらされた？

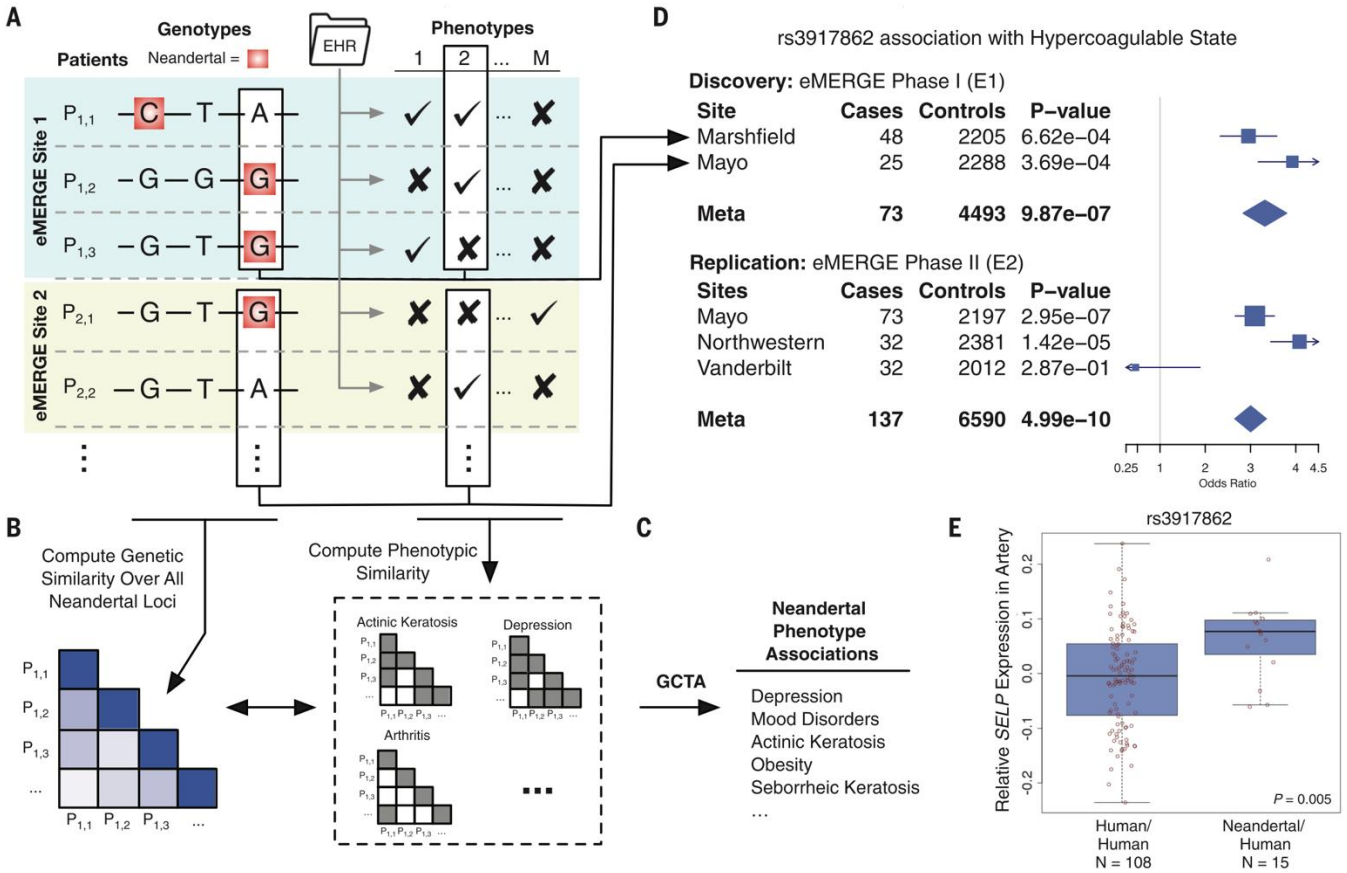
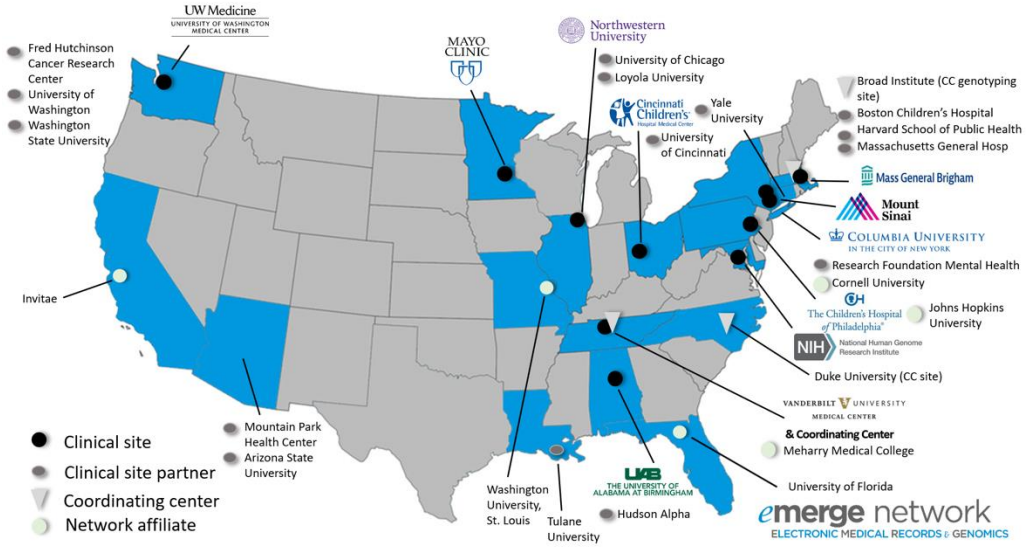



図1より引用

全米各地で得られたEHR+ヒトゲノム（eMERGE プロジェクト)と連携



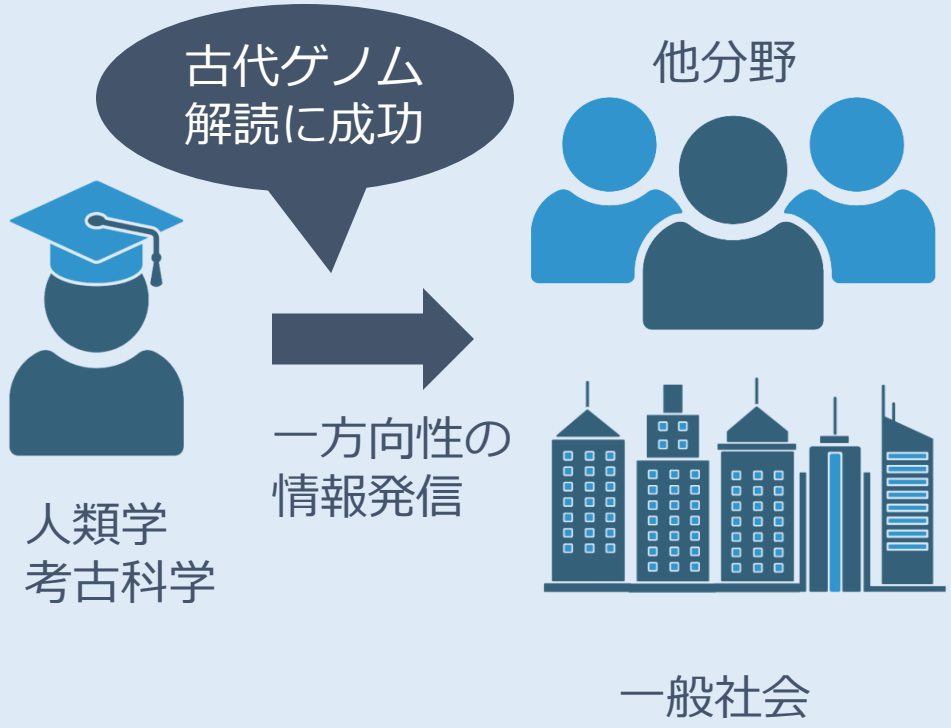

Simonti *et al.*, (2016)

**現代人の電子カルテ・全ゲノムデータとの連携により疾患リスクアリルを探索**

# ゲノム人類学 × パレオゲノミクス

## 従来までのアプローチ

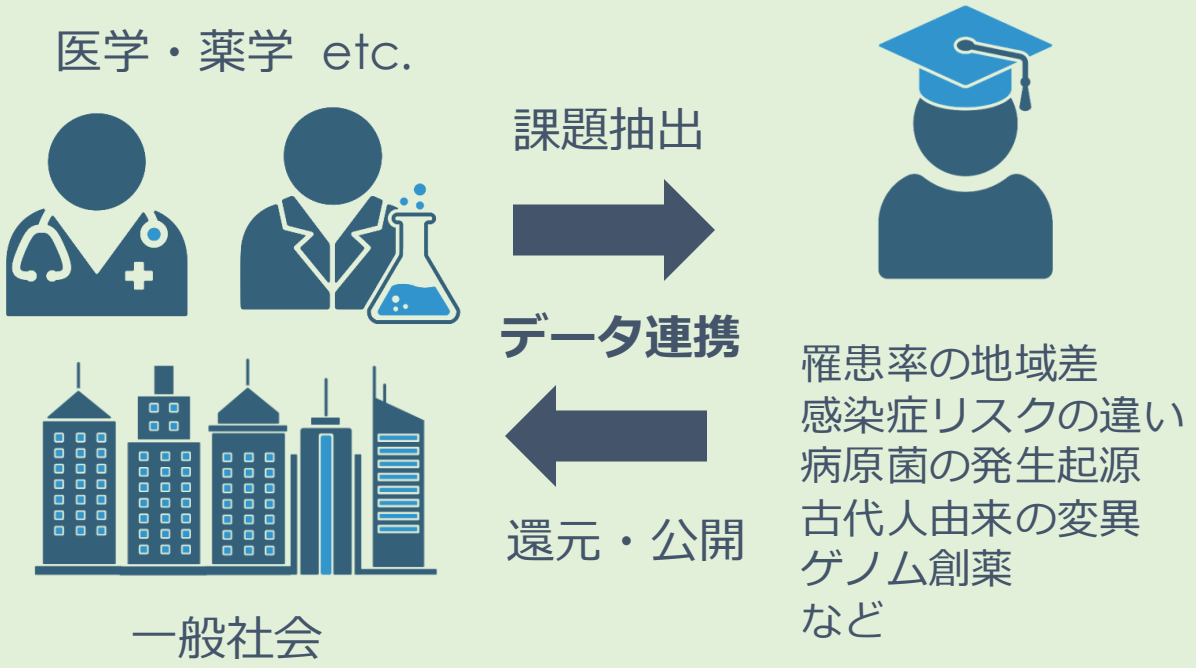
### ヒトの進化史・集団史を知る




新規性はあるが、融合研究や連携を想定してない

## 今後のアプローチ

### 病気・形質・薬効における
### 進化的背景やその影響を知る




**共創的な課題解決と総合知の創出**

# 謝辞


## 共同研究者

東京大学 植田 信太郎 先生
國學院大學　谷口 康浩 先生
人類学研究機構 松下 孝幸 先生
人類学研究機構 松下 真実 先生
総合研究大学院大学 王 瀝 先生
東京大学 近藤 修 先生
総合研究大学院大学 五条堀 淳 先生
東邦大学 水野 文月 先生
農研機構 熊谷 真彦 先生
金沢大学 覚張 隆史 先生

## 研究機関・研究室

サピエンス進化医学研究センターの皆さま
國學院大學考古学研究室の皆さま
東邦大学法医学教室の皆さま

## 外部研究資金

科研費
KAKENHI

生物考古情報の復元に向けたパレオメタゲノミクス技術の開発
日本学術振興会: 科学研究費助成事業 学術変革領域研究(A)

半定住狩猟採集民の社会組織と葬制：骨考古学先端技術との連携による先史社会の復元
日本学術振興会: 科学研究費助成事業 基盤研究(S)

パレオゲノミクス解析プラットフォーム開発とその応用
日本学術振興会: 科学研究費助成事業 学術変革領域研究(A)

シン・パレオゲノミクスが創る博物館資料群活用の新展開
日本学術振興会: 科学研究費助成事業 基盤研究(A)

古ゲノム分析による日本列島の穀物利用史の解明
日本学術振興会: 科学研究費助成事業 基盤研究(C)