

遺伝子データの世界から 博物館データの世界へ アクセスする

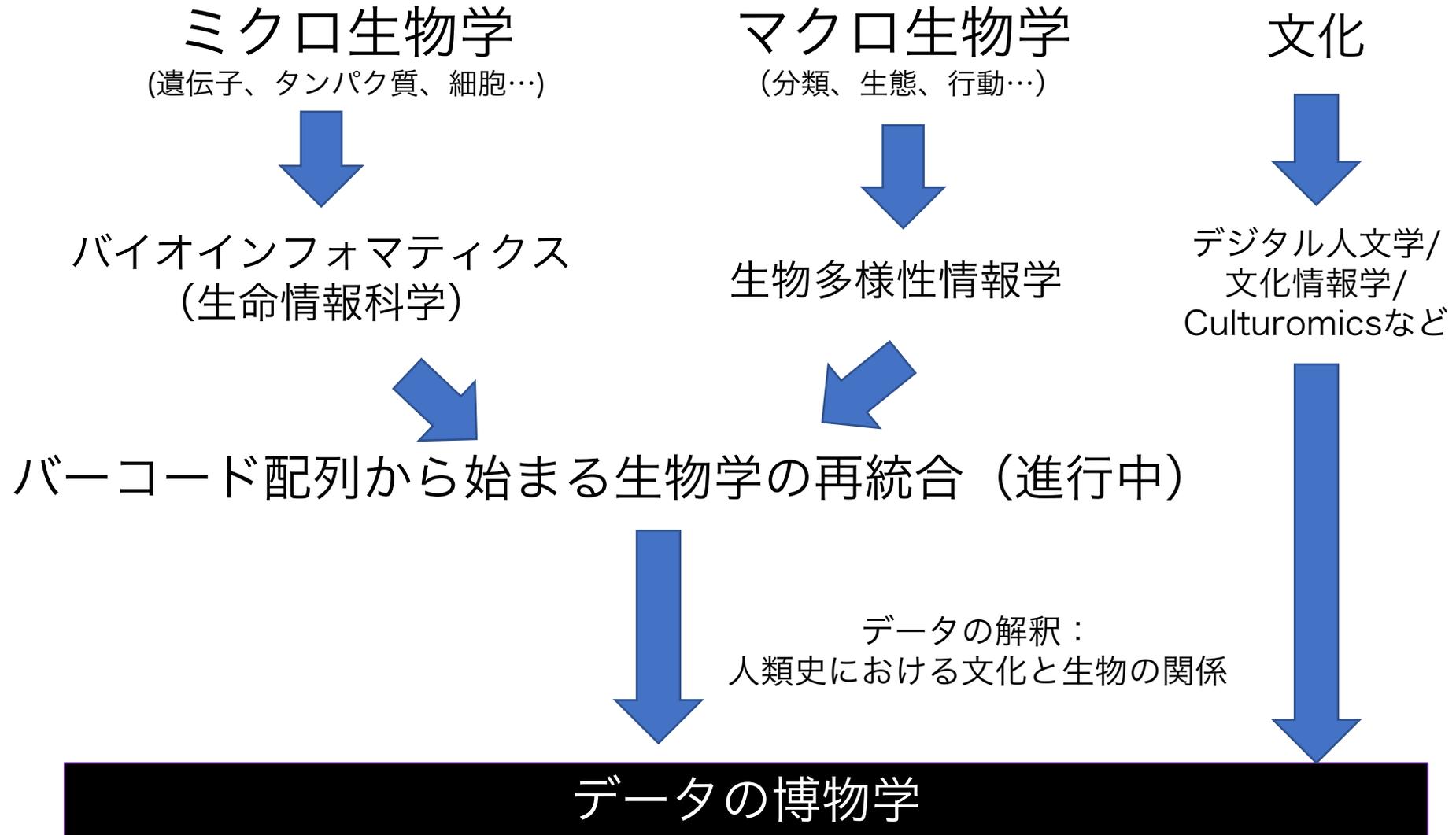
遺伝子⇔博物館標本が検索・利用できるようになると嬉しい

松前ひろみ（東海大学）

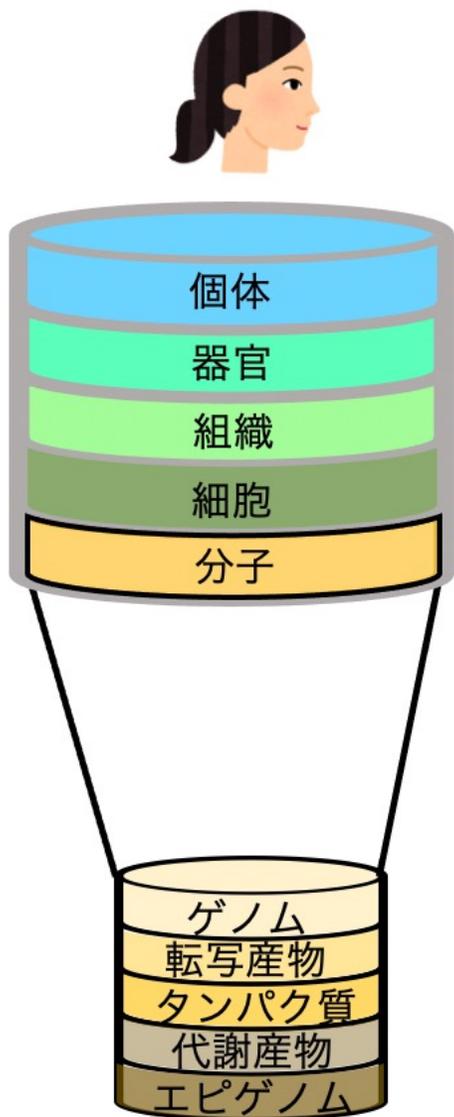
matsumae.hiromi.g@tokai.ac.jp

第39回 自然史標本データ研究会（S-Net研究会）オンライン
2022/06/25

学術分野を超えてデータの重なりを見る

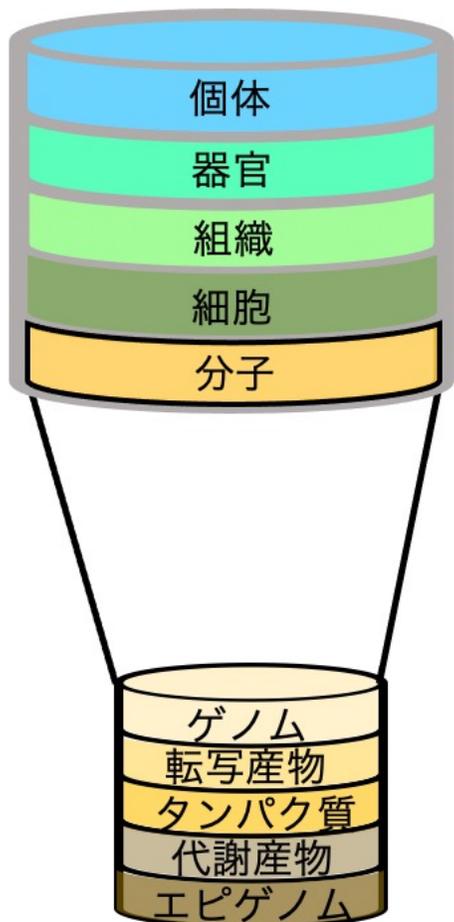


バイオインフォマティクス： 分子の世界のデータサイエンス



対象	ome (総体)	Omics (総体の解析法)
gene (遺伝子)	genome (ゲノム)	Genomics (ゲノミクス)
transcript (転写産物)	transcriptome (トランスクリプトーム)	Transcriptomics (トランスクリプトミクス)
Protein (タンパク質)	Proteome (プロテオーム)	Proteomics (プロテオミクス)

バイオインフォマティクス： 分子の世界のデータサイエンス



museomics
 ミュゼオミクス
 (museum + omics)

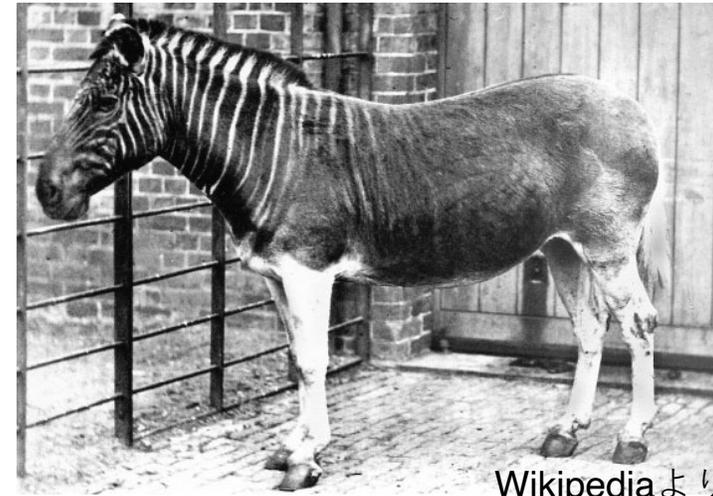
博物館標本からゲノムを復元する研究

- DNAバーコードを意味することもあるが、どちらかというとな全ゲノム情報
- 転写産物は分解しやすく乾燥標本に残らない
- タンパク質は技術的難易度が高い・網羅性が低い

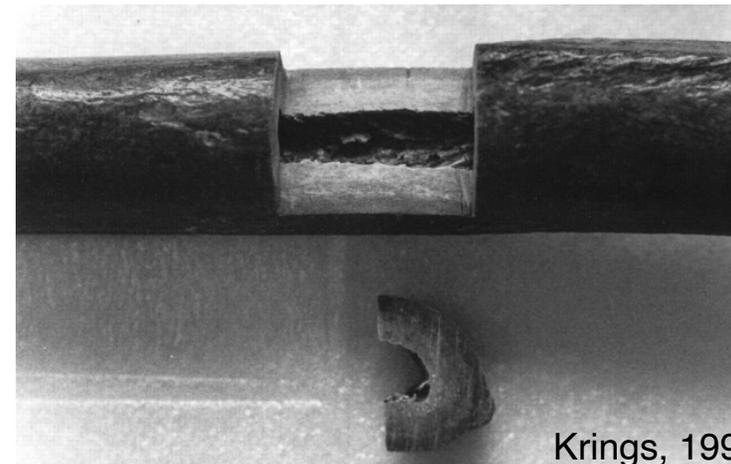
対象	ome (総体)	Omics (総体の解析法)
gene (遺伝子)	genome (ゲノム)	Genomics (ゲノミクス)
transcript (転写産物)	transcriptome (トランスクリプトーム)	Transcriptomics (トランスクリプトミクス)
Protein (タンパク質)	Proteome (プロテオーム)	Proteomics (プロテオミクス)

Museomicsを支える 古代ゲノム学(Paleogenomics, ancient DNA)

- 古人骨・化石などからDNAを抽出
 - ゲノム科学の技術発展と共に発達した分野
- ① Sanger法・PCR法の出現により発展
- 絶滅動物クアッガのDNA解析 (Higuchi et al. 1984, Science)
 - 絶滅人類・ネアンデルタール人**タイプ標本**・上腕骨のDNA分析 (Krings, et al, 1997, Cell)
 - 塩基配列のコンタミネーション・変異・エラーの識別が困難
 - 少ない分子量の検出の限界

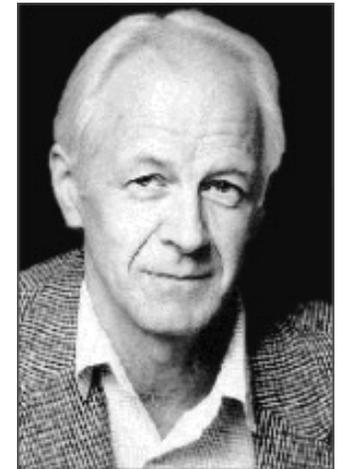


Wikipediaより



Krings, 1997, Cell

Alan Wilson



Svante Paabo

② 次世代シーケンサー (Next Generation Sequencer; NGS) 2008~

- PCR法よりも大量データが取れる→統計的に変異・エラー・コンタミネーションかを検証できる
- 断片化した古代ゲノム向き。

short read型 (70-300塩基対)

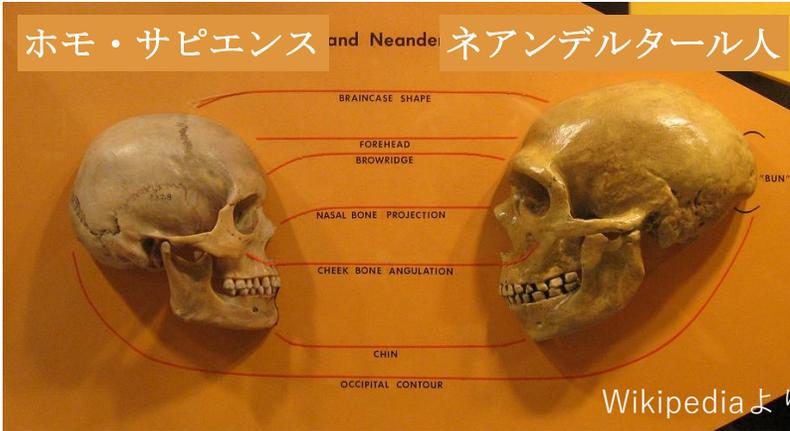
短いが1塩基当たりのエラー率が低い

NGSで6-7万年前の骨片から新しい絶滅人類の系統を発見



スヴァンテ・ペーボ

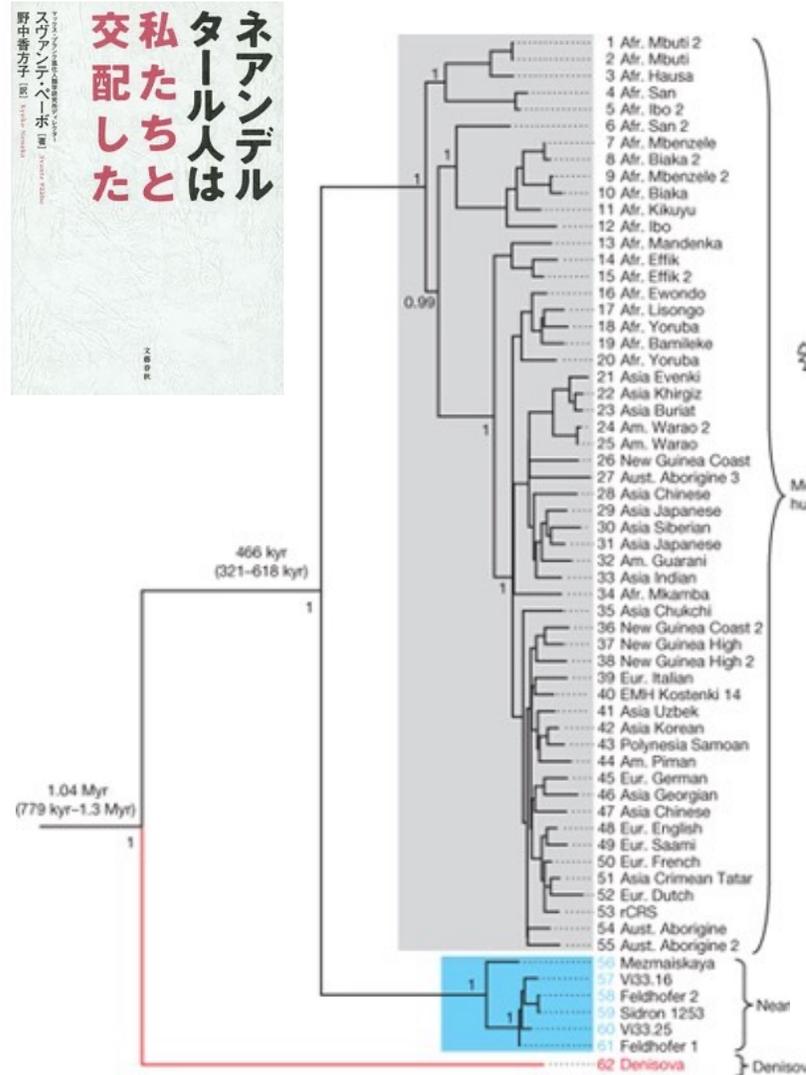
Krause, et al, Nature 2010



Wikipediaより

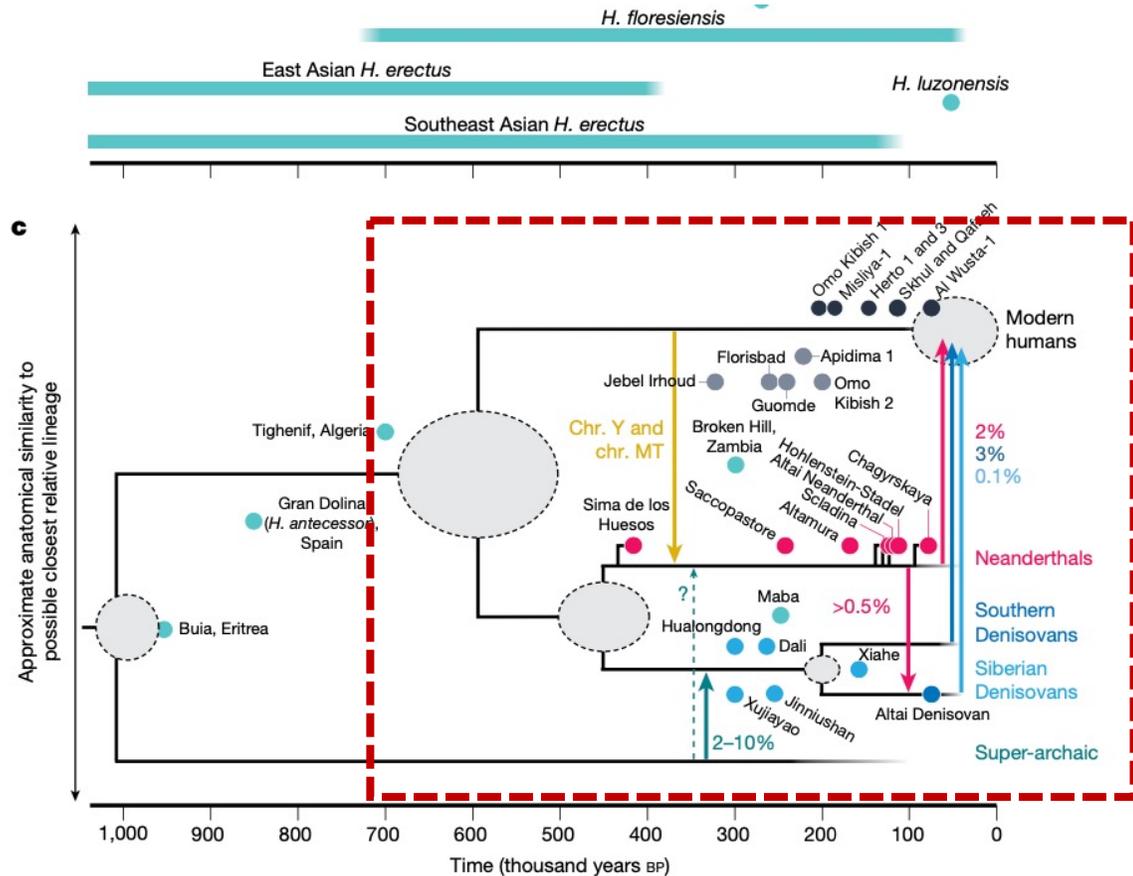


Wikipediaより



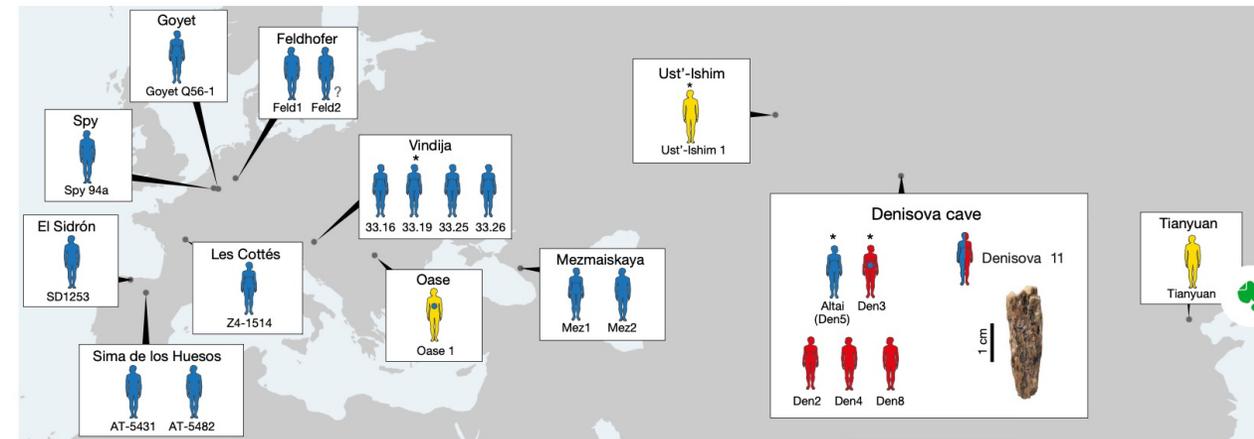
全身骨格がなくても、科学的に未知の情報が得られる可能性を示唆

ホモ属の混血関係



(Bergström et al, Nature 2021)

現生人類の一部は、絶滅人類であるネアンデルタール人・デニソワ人と3者間で混血していた



(Slon et al, Nature 2018)

博物館標本のゲノム解析の限界

ホモ属における進化（遺伝的祖先の解明）

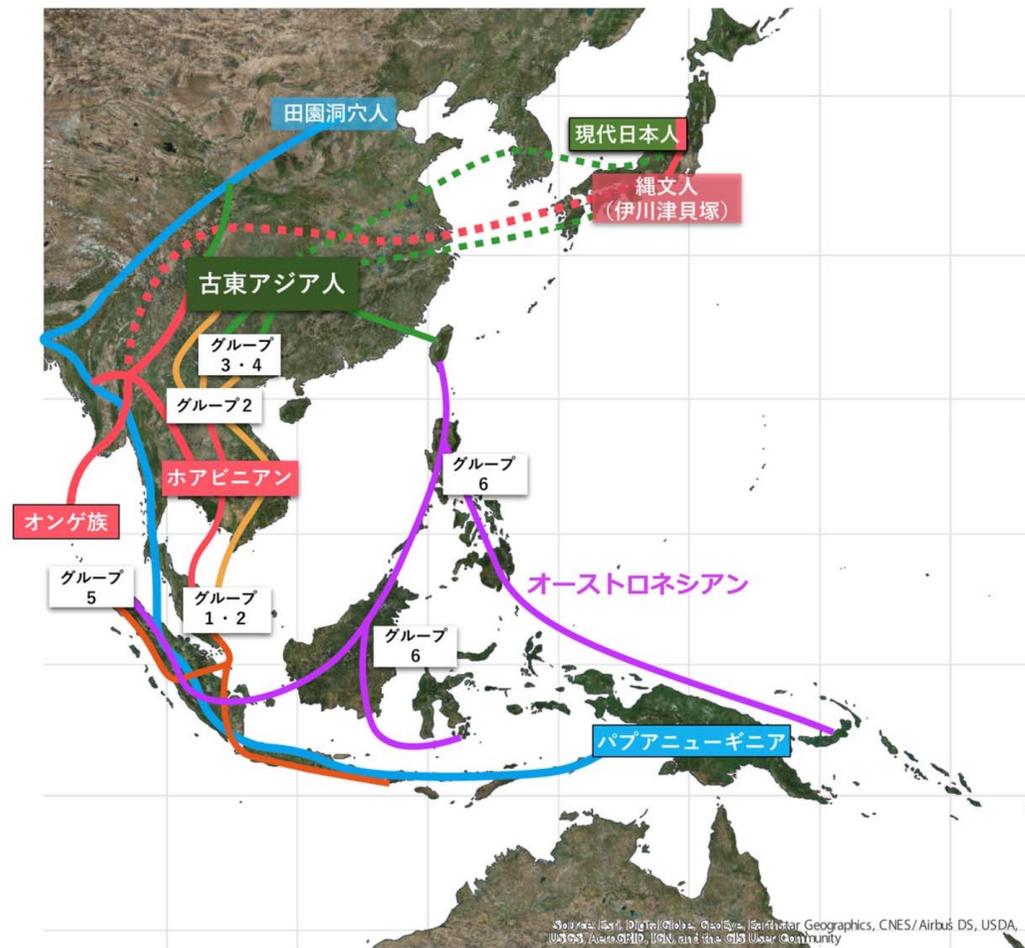
- 分かる → 系統
- 現在の技術・データでは分からない → 生物学的なヒトらしさを決定づける解剖・生理・行動・認知・文化面などの違い (Bergström et al, Nature 2021)

⇒ 文化をヒト特有の形質と捉えて解析する意味
文化進化 (cultural evolution)

ゲノムから読み解く東ユーラシア沿岸部の人類史

縄文人のゲノムも解読できる時代に

Paleogenomics



文化の歴史を測る代表的な方法：言語学

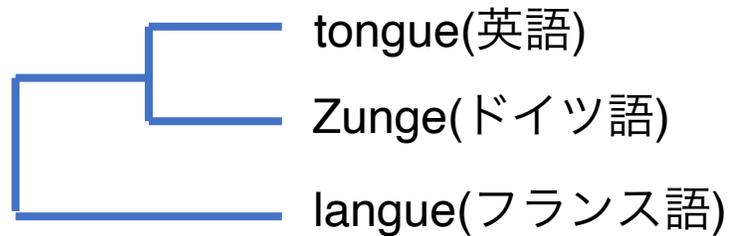
従来法 = 基礎語彙を使った言語比較法

- 語彙のデータセットが充実（標準データ、標準的な解析法）
- 基礎語彙 = 外来語になりにくい語彙のセット（身体のパーツなど）
- 語族 = 基礎語彙を共有する言語のグループ

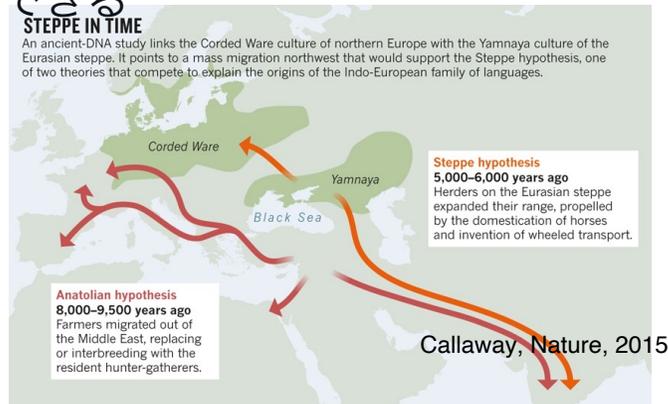
ヨーロッパ

→ 言語族レベルの言語多様性が低い

例) インド・ヨーロッパ語族



インド・ヨーロッパ語族の話者の遺伝的起源を古代ゲノムで解析し、言語と遺伝の歴史を比較できる



文化の歴史を測る代表的な方法：言語学

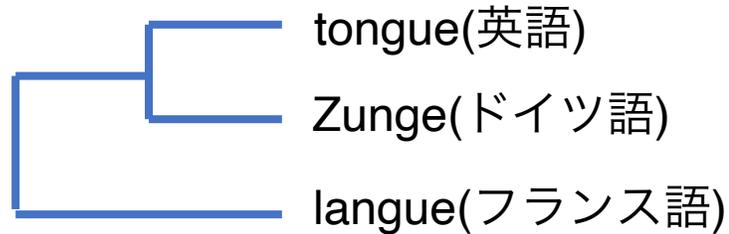
従来法 = 基礎語彙を使った言語比較法

- 語彙のデータセットが充実（標準データ、標準的な解析法）
- 基礎語彙 = 外来語になりにくい語彙のセット（身体のパーツなど）
- 語族 = 基礎語彙を共有する言語のグループ

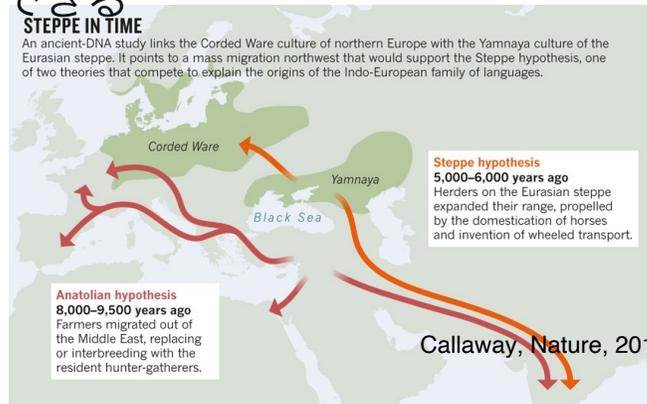
ヨーロッパ

→ 言語族レベルの言語多様性が低い

例) インド・ヨーロッパ語族



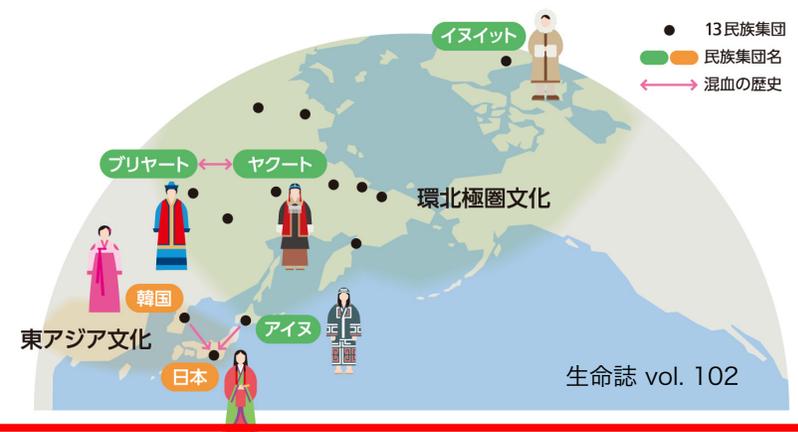
インド・ヨーロッパ語族の話者の遺伝的起源を古代ゲノムで解析し、言語と遺伝の歴史を比較できる



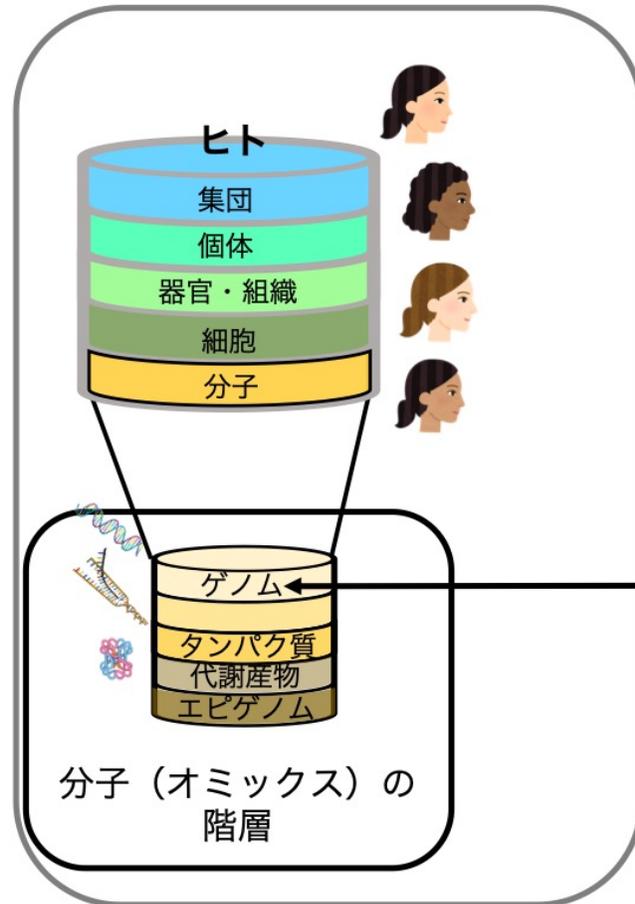
東ユーラシア沿岸部

→ 言語族レベルの言語多様性が高い

- shita (標準語/日本語族)
- ?? — aw(北海道アイヌ語/アイヌ語族)
- hyo(韓国語/韓国語族)



ゲノムは過去まで遡れるようになったが、 文化の計測・遡りには限界



生物と文化を同じようなデータ形式に落とし込んでデータ解析が行える研究対象は限られており、学際的な解釈に留まる研究が中心

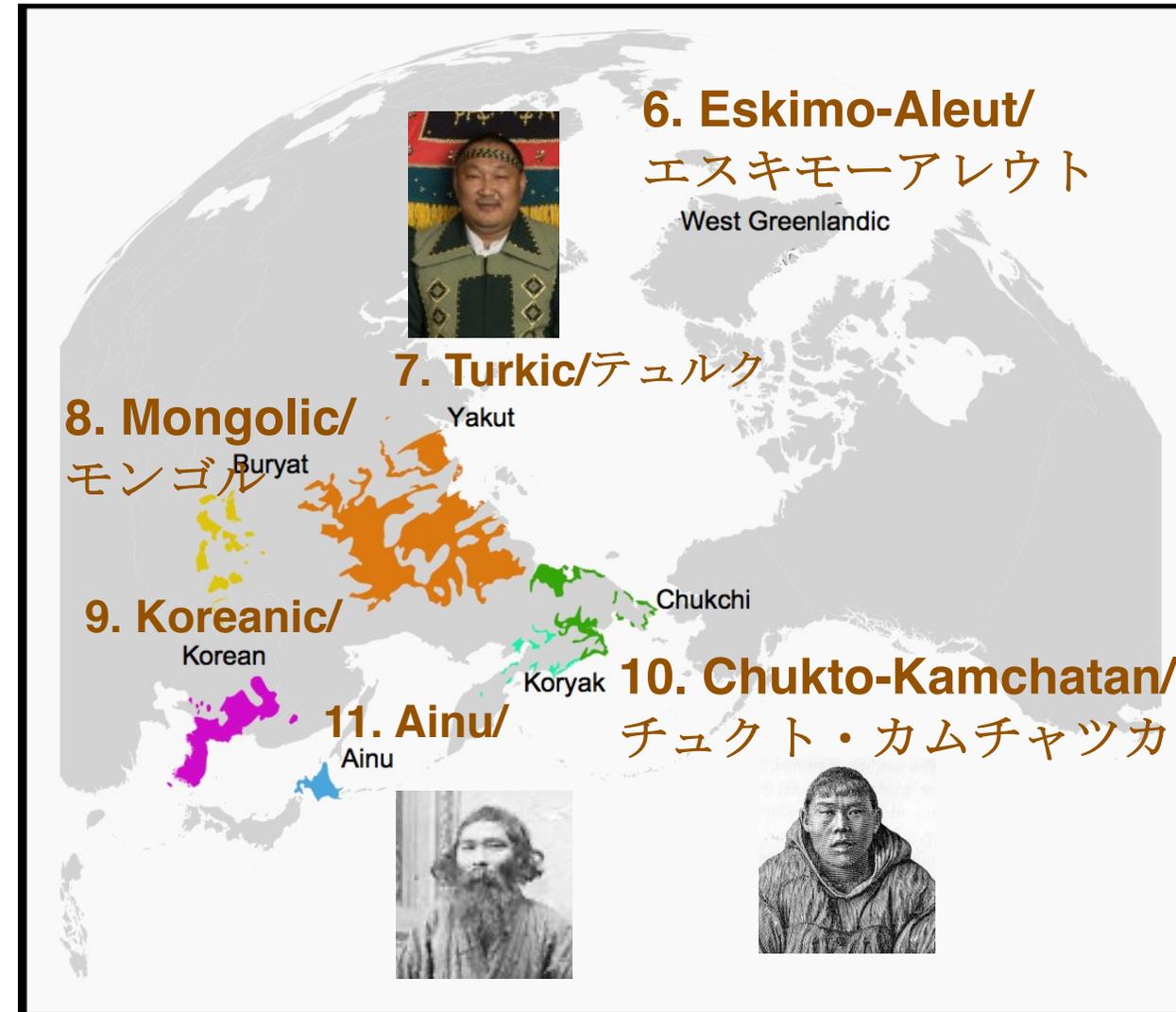
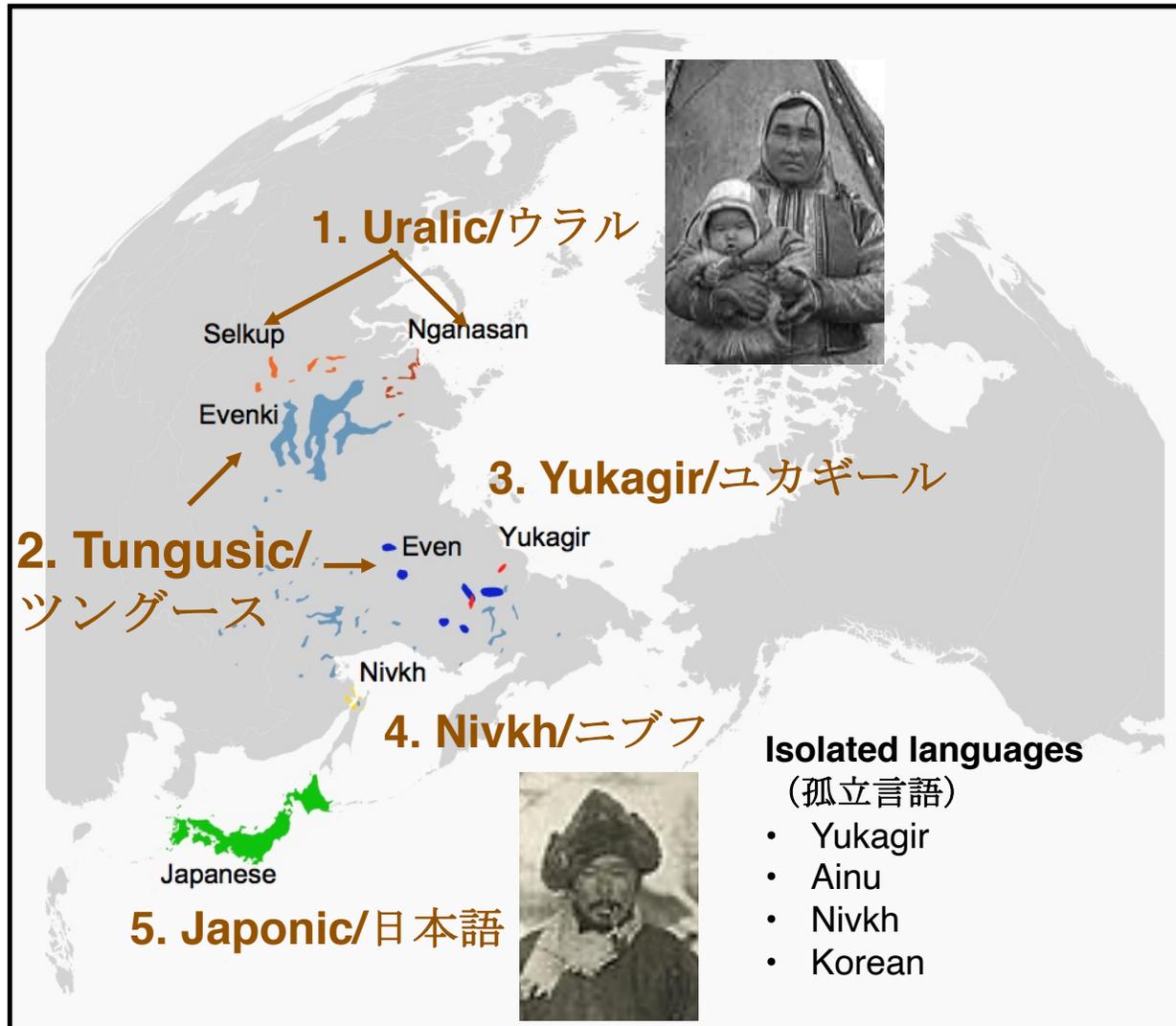
**人類文化の類似性と
ゲノム進化史の関連**

?



東ユーラシア沿岸部近傍の11言語族

⇔ヨーロッパは3言語族(インド・ヨーロッパ語族、ウラル語族、バスク語族)

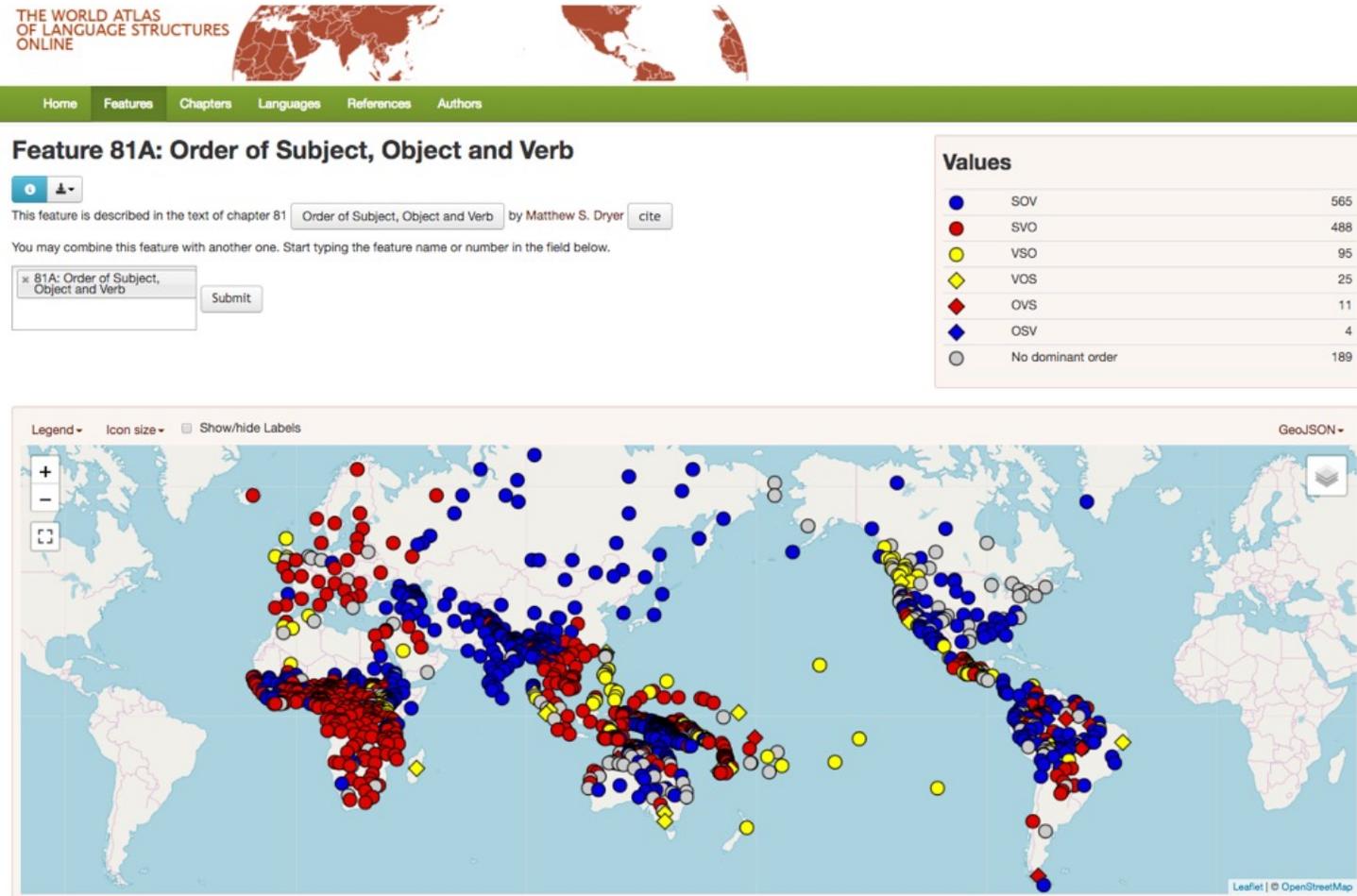


語彙以外の文化マーカを探る 文法の要素例：語順

英語: SVO

日本語: SOV

語順のバリエーションは7つ
⇔DNA (ATGC)は4つ

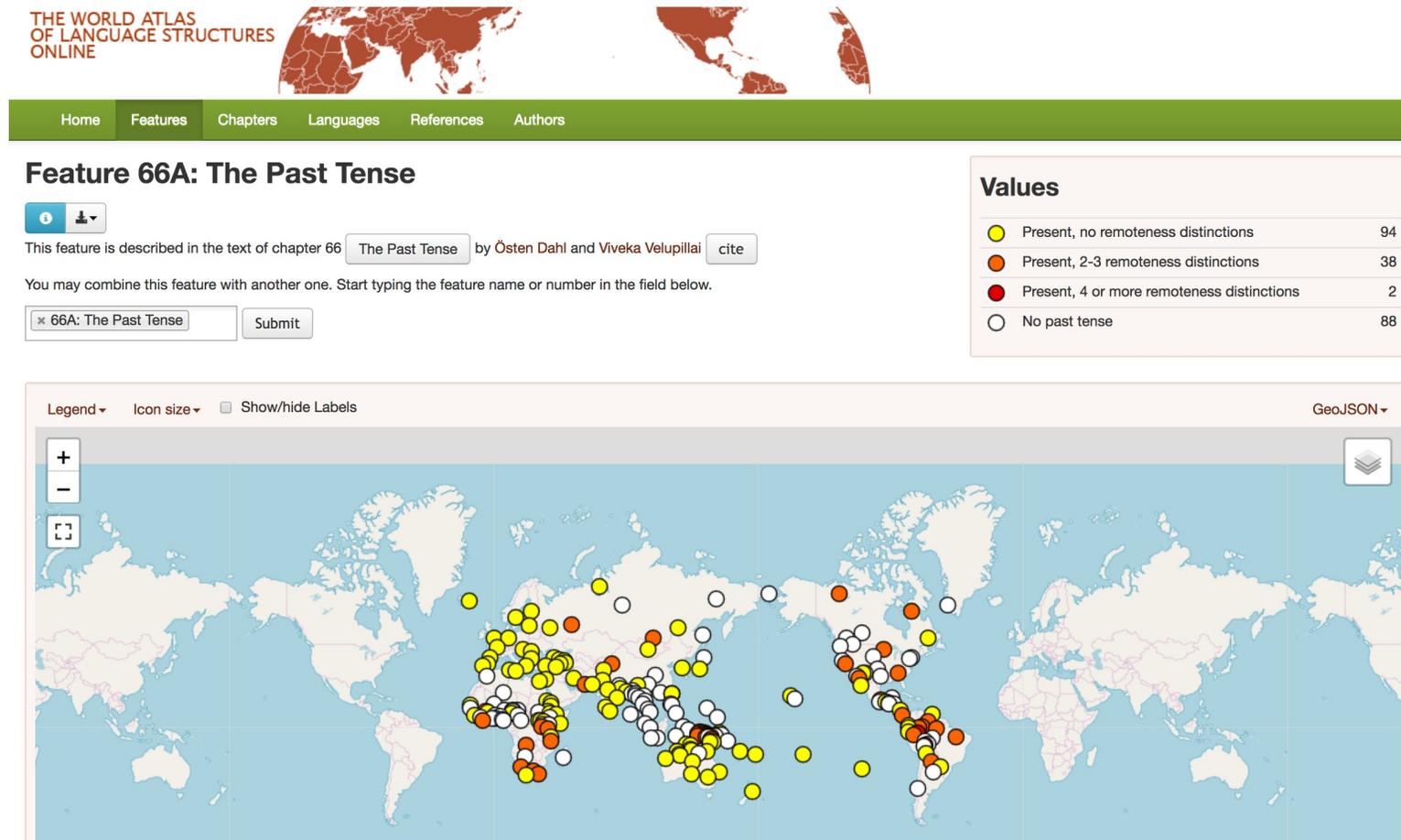


語彙以外の文化マーカ―を探る

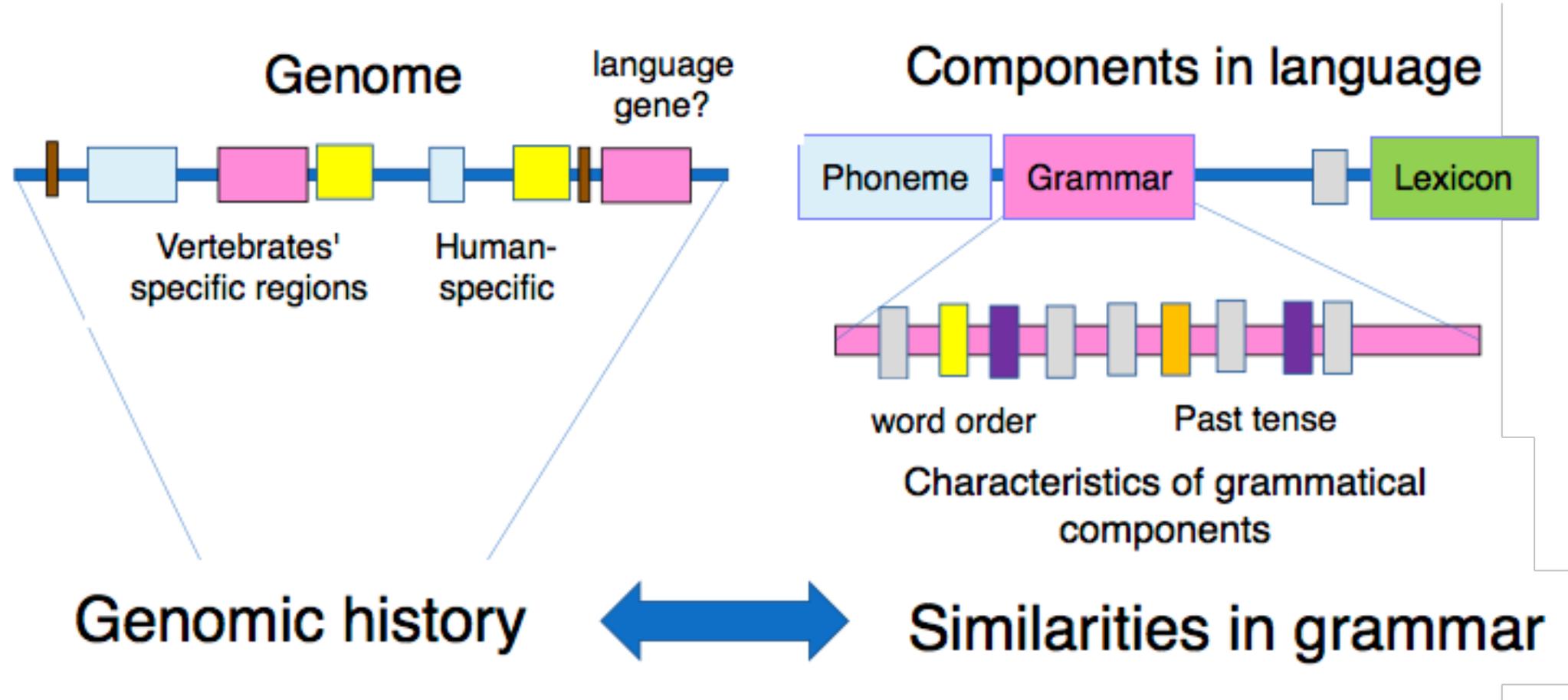
文法の要素例：過去形

- 動詞に過去形がある（黄色）
- 過去形を使わない（白）
- 過去形に、近い過去と遠い過去がある（オレンジ）

過去形の取り方で4パターン



ゲノム解析のアナロジーで、
文化の類似性（not 系統関係）を解析してみる



北東アジアの音楽のつながり

イヌイット(Eskimo-Aleut語族)
の喉歌

<https://www.youtube.com/watch?v=qnGM0BIA95I>

サハリンアイヌ(アイヌ語族)
の喉歌

<https://www.youtube.com/watch?v=maaJlJq7Gow>

必ず女性が向き合って歌う

こうした文化の類似性は過去の民族間の関係を示しているのか？

Music (songs) can be quantified like genetic data (SNPs)

CantoCore + CantoMetrics Dataset

41 Musical factors of songs (rhythms, pitch, etc.)

population

each song
(individual)



Patrick Savage Steven Brown

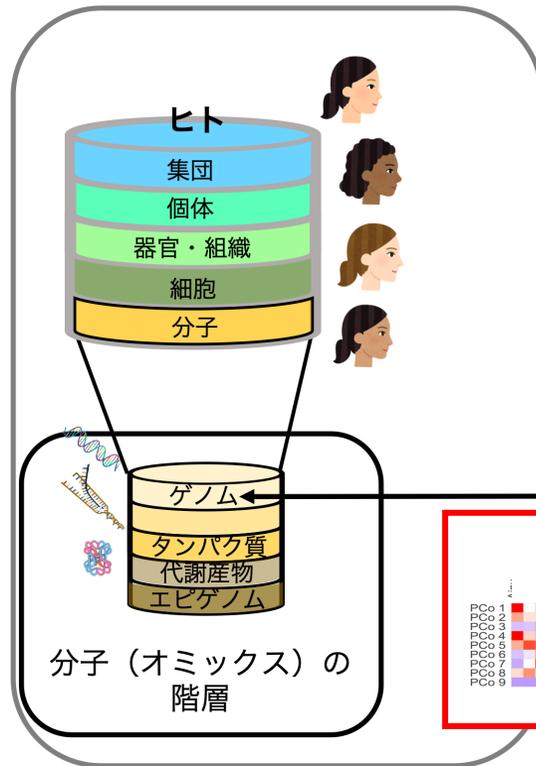
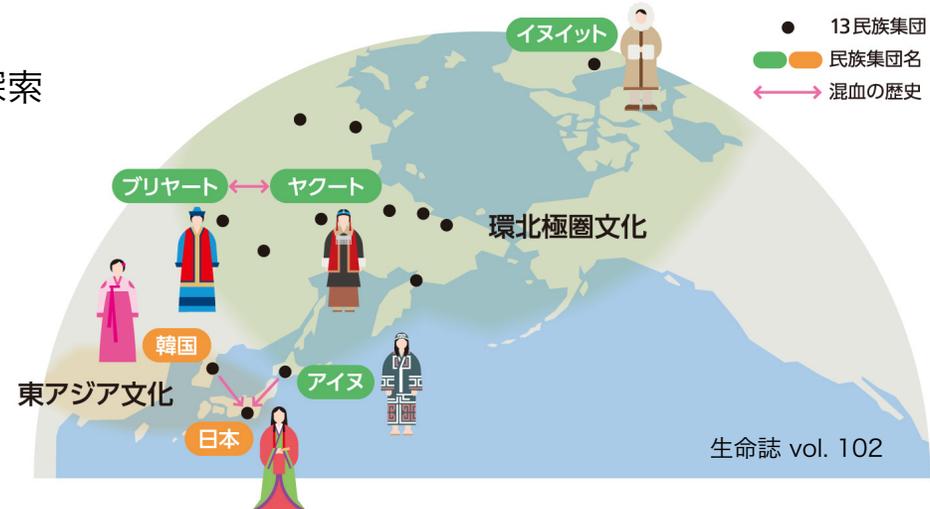
	A	B	C	D	P	T	W	X	Y	Z	AA	AB	AC	AD	AE	AF
1	Song N.	Population name	1a. Beat (a-met)	8a. Discrete pit	CM30. Tremolo	35. Rasp (extre	CantoCore ID 1	((CantoCore#)	2. No of beats	3. Beat sub-div	4. No. of sub-b	5. Syncopation	6. Motivic redu	7. Durational vi	8. Tonality	9. Mode
2	1	Ainu (Hokkaido)	0	1	0	1	1528	a	n/a	n/a	n/a	n/a	i	iii	ea	a
3	2	Ainu (Hokkaido)	0	1	1	0	1693	a	n/a	n/a	n/a	n/a	i	i	ad	n/a
4	3	Ainu (Hokkaido)	1	1	1	1	1406	d	a	c	a	iii	iii	ii	e	e
5	4	Ainu (Hokkaido)	1	1	1	1	1542	d	a	c	a	iii	iii	i	ea	a
6	5	Ainu (Hokkaido)	1	1	1	1	1401	d	a	c	a	iii	iii	i	e	d
7	6	Ainu (Hokkaido)	1	1	1	1	1563	d	a	c	a	ii	iii	ii	e	d
8	7	Ainu (Hokkaido)	1	1	0	1	1347	d	a	c	a	ii	iii	i	ae	a
9	8	Ainu (Hokkaido)	1	1	0	1	1458	d	a	c	b	i	iii	i	e	a
10	9	Ainu (Hokkaido)	1	1	1	1	1331	d	a	c	a	ii	iii	ii	e	d
11	0	Ainu (Hokkaido)	1	1	0	1	1414	d	a	c	a	iii	iii	i	ed	d
12	1	Ainu (Hokkaido)	1	1	0	0	1395	d	a	c	a	ii	iii	i	eac	a
13	2	Ainu (Hokkaido)	1	1	1	1	1567	d	a	c	a	i	iii	i	ea	a
14	3	Ainu (Hokkaido)	1	1	0	1	1343	d	a	c	a	i	iii	i	e	a
15	4	Ainu (Hokkaido)	1	1	0	1	1517	d	a	c	a	i	iii	i	ae	e
16	5	Ainu (Hokkaido)	1	1	0	1	1427	d	b	c	a	iii	iii	i	e	a
17	6	Ainu (Hokkaido)	1	1	0	0	1702	d	a	c	a	i	iii	i	be	d
18	7	Ainu (Hokkaido)	1	1	1	1	1481	d	a	c	b	i	iii	ii	e	d
19	8	Ainu (Hokkaido)	1	1	1	1	1727	d	a	c	b	i	iii	i	ed	e
20	9	Ainu (Hokkaido)	1	1	1	1	1682	d	a	c	a	iii	iii	ii	e	a
21	0	Ainu (Hokkaido)	1	1	1	1	1338	d	a	c	a	iii	iii	i	ea	d
22	1	Ainu (Hokkaido)	1	1	1	1	1560	d	a	c	b	ii	iii	i	ea	a
23	2	Ainu (Hokkaido)	1	1	1	1	1326	d	a	c	b	ii	iii	i	e	a
24	3	Ainu (Hokkaido)	1	1	0	1	1350	ab	n/a	n/a	n/a	n/a	i	iii	ea	d
25	4	Ainu (Hokkaido)	1	1	0	1	1515	d	a	c	a	i	iii	i	ea	a
26	5	Ainu (Hokkaido)	1	1	0	1	1575	d	a	c	a	iii	iii	i	ea	a
27	6	Ainu (Hokkaido)	1	1	1	1	1723	d	a	c	b	ii	iii	ii	ea	d
28	7	Ainu (Hokkaido)	0	1	0	1	1666	a	n/a	n/a	n/a	n/a	ii	ii	e	e
29	8	Ainu (Hokkaido)	1	1	0	1	1320	b	n/a	n/a	n/a	n/a	ii	ii	e	d
30	9	Ainu (Hokkaido)	1	1	0	1	1459	d	a	c	b	ii	iii	i	ea	a
31	0	Ainu (Hokkaido)	1	1	0	1	1465	d	a	c	a	i	iii	ii	e	d
32	1	Ainu (Sakhalin)	1	1	0	1	2233	d	a	c	a	ii	iii	i	ea	a
33	2	Ainu (Sakhalin)	1	1	0	0	2234	d	a	c	a	ii	iii	ii	ea	d
34	3	Ainu (Sakhalin)	1	1	0	0	2261	ab	n/a	n/a	n/a	n/a	ii	iii	e	d
35	4	Ainu (Sakhalin)	0	0	0	0	2235	a	n/a	n/a	n/a	n/a	i	ii	a	n/a

(Romax 1976; Savage et al. 2012; Savage et al. 2015)

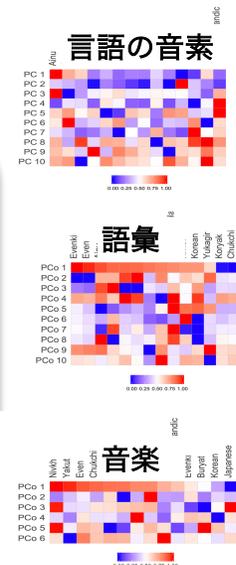
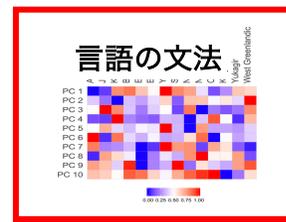
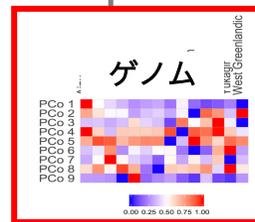
人類学研究から発見した生物データと文化データの融合点

北東アジアは言語多様性の高い地域

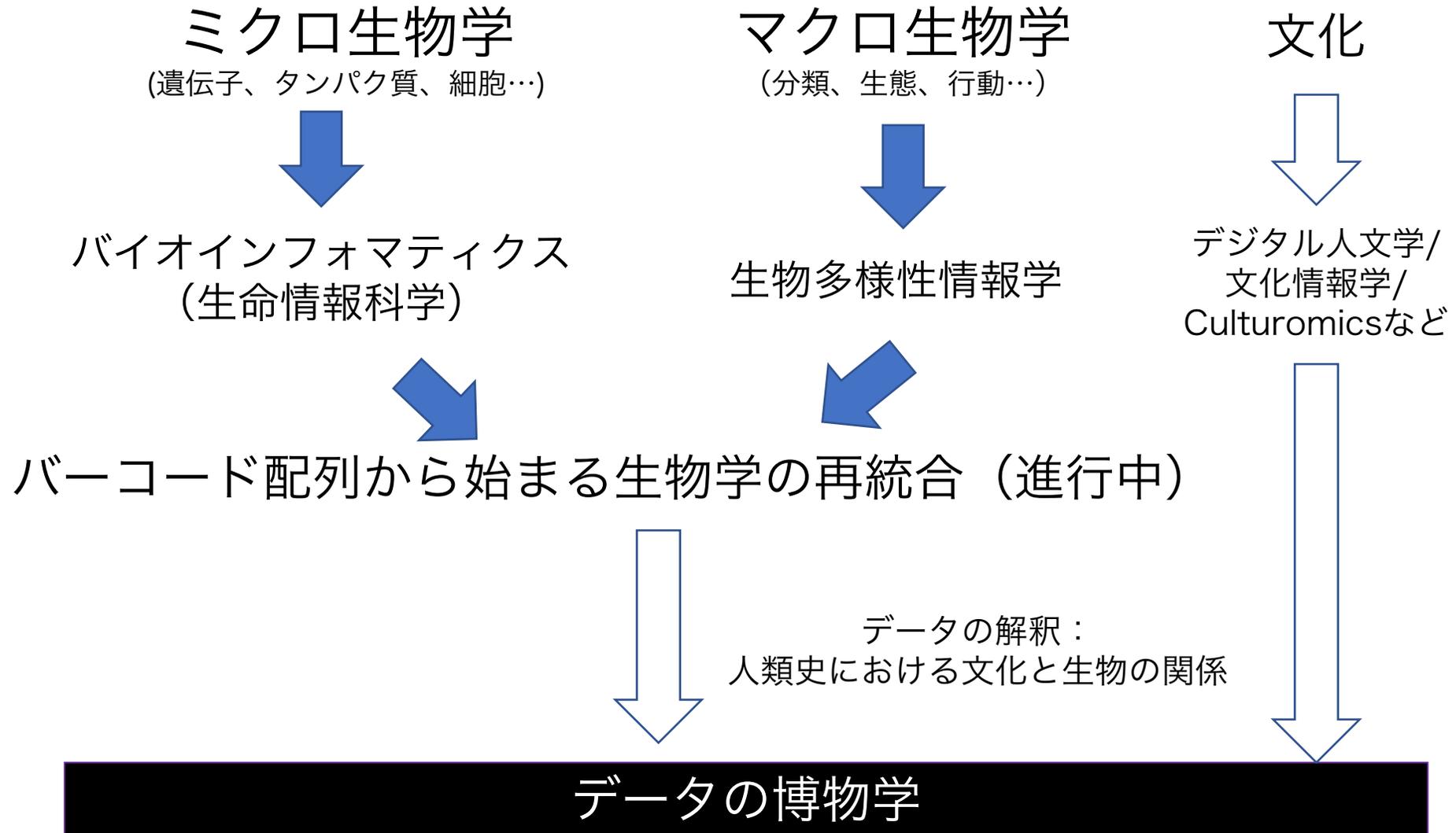
- どんな文化的マーカーならば歴史を反映するのか？
- 文化をコード化し、ゲノム史と相関する文化的マーカーを探索



言語 (文法) の類似性とゲノム史の関連を初めて発見

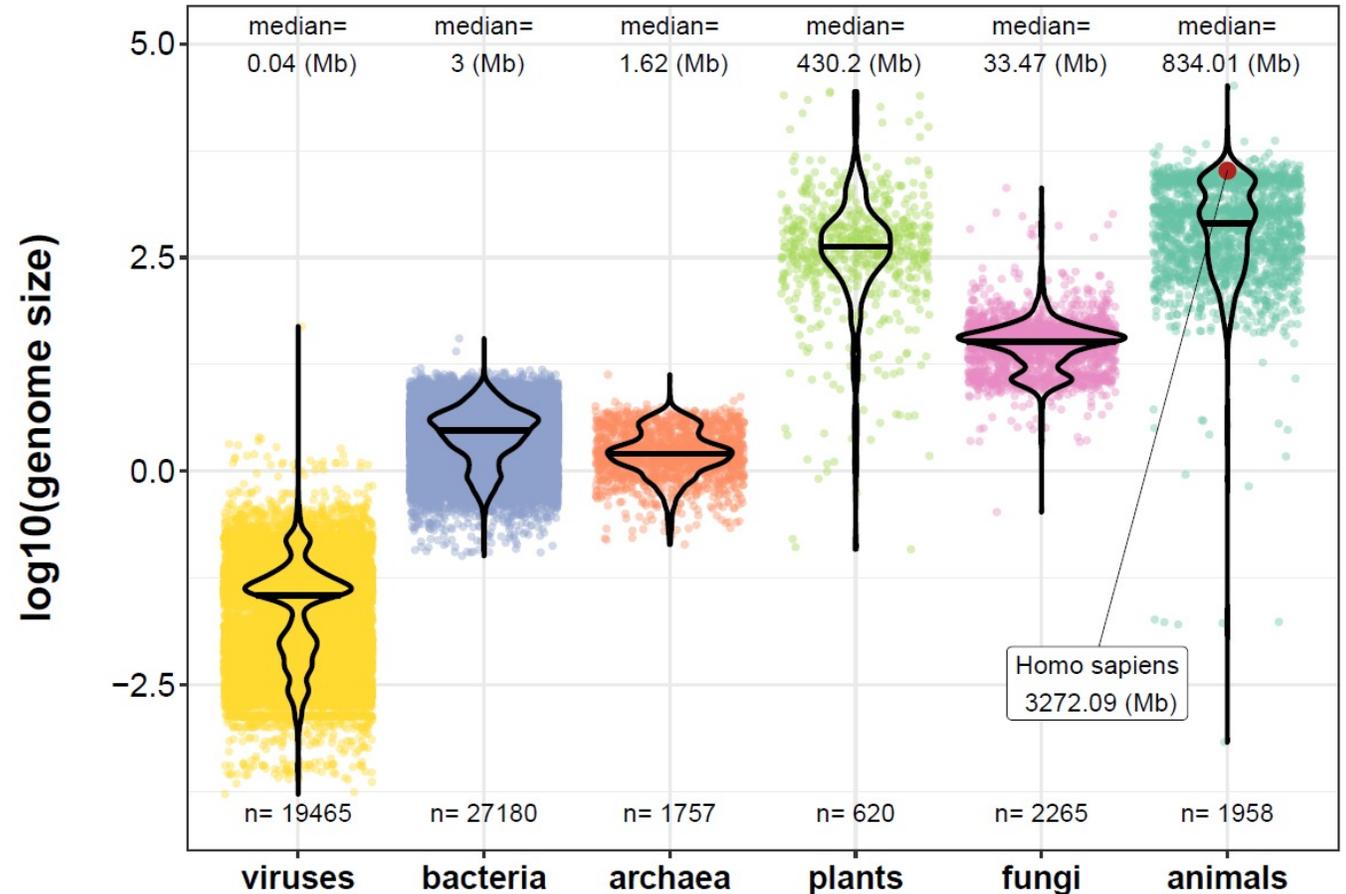


学術分野を超えてデータの重なりを作る



DNAバーコーディングと全ゲノム解析では読み取れる情報の内容が異なる

- バーコード領域：数百塩基
- ゲノムサイズ：生物種によって非常に異なる
 - 特に植物は倍数化で大きくなりやすく、ゲノム解析できていないものも多い
 - 針葉樹（数十～数百億塩基対）



NCBIに登録されているゲノムデータ

Oxford University Scientific Society, 2020.

<https://www.facebook.com/OUSciSoc/photos/a.320619098037316/2804993669599834/?type=3>

take home message

- バーコード配列は”同種っぽいクラスター”を集めるには便利
- ヒトでもそれ以外でも、**形態的・行動的特徴**など個体レベルの特徴は、**ゲノムで説明しきれていない**
 - 例) 交尾器の違いを説明できない
- **分類学の知見や視点が活きるのはこれから**
 - 特に、**全ゲノムなどのオミックス解析時代** (ポジショントーク)

Revisiting museomics : 博物館の全てのデータの博物学

-ome: 総体

拡張版
museome
museomics

(ゲノムに限定しない) 博物館由来の全データの解析

博物館標本xDNAx生物多様性データ

- DNA以外の分子情報
- 生物x文化

(松前他、バイオインフォマティクス学会ニュースレター、2018; 松前・神保他、査読中)