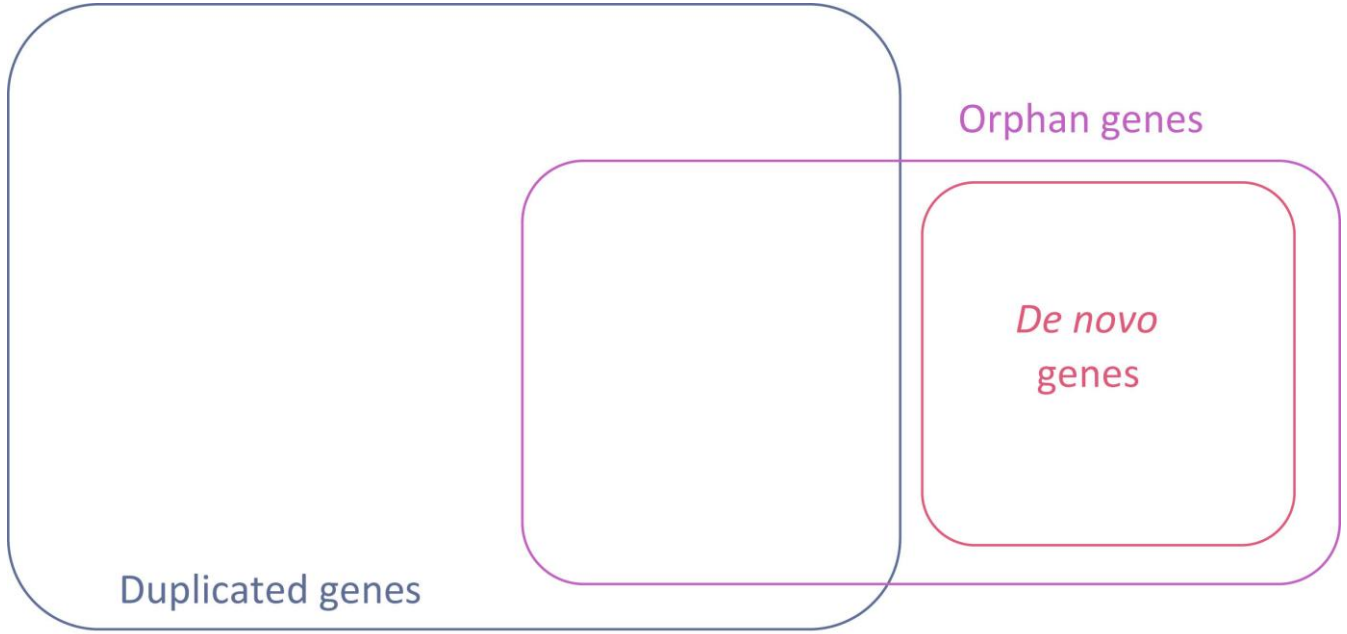
Springer Link

A Continuum of Evolving De Novo Genes Drives Protein-Coding Novelty in *Drosophila*

Heames, B., Schmitz, J. & Bornberg-Bauer, E., 2020

Introduction | orphan gene, de novo gene

- ✓ 特定の系統以外で相同性が見られない遺伝子を orphan gene と呼び、真核生物の遺伝子の最大 30% を構成することもある。
- ✓ そのうち、非遺伝子領域から新たに誕生した遺伝子を de novo gene と呼ぶ。



Rödelspergerl et al., 2019

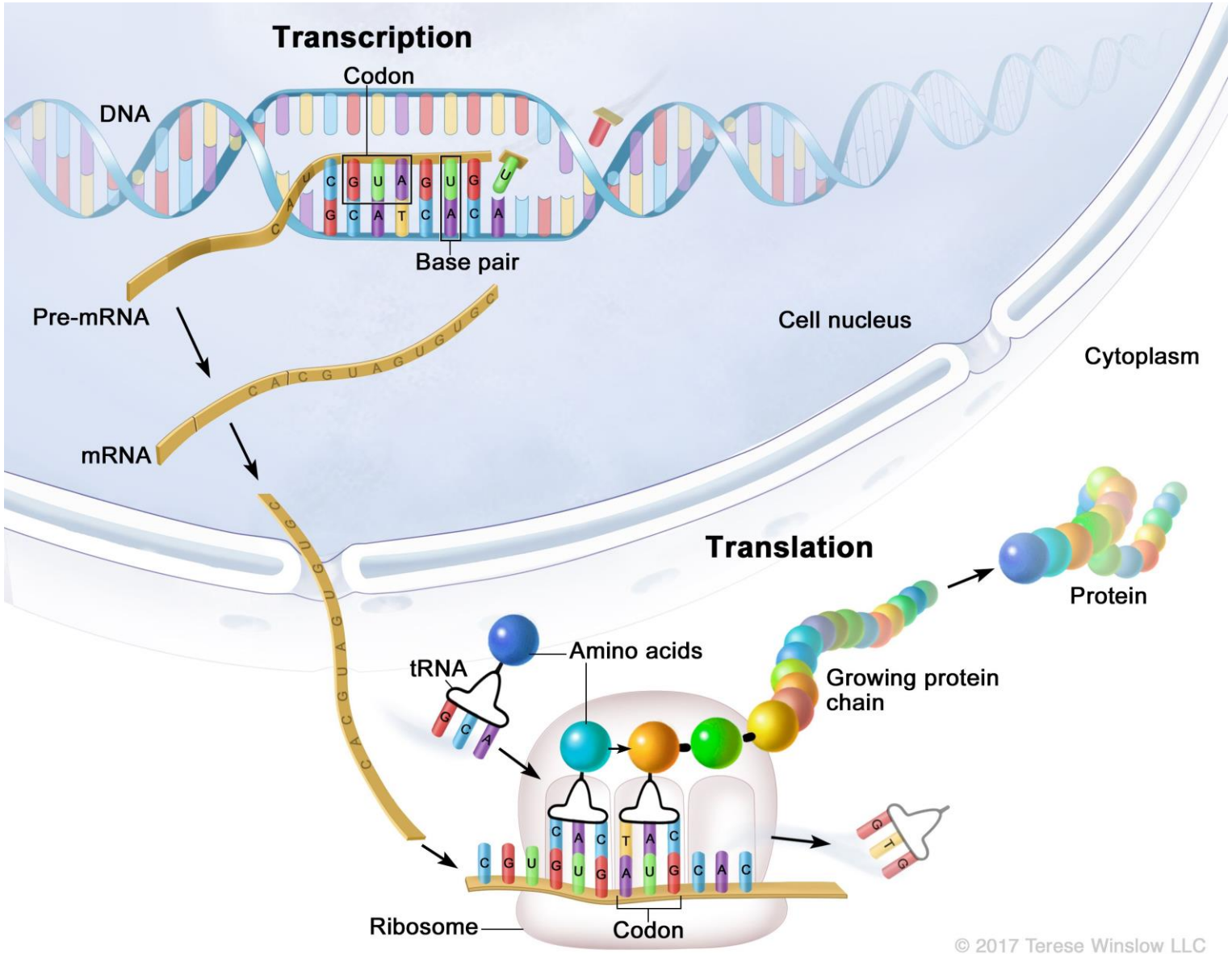
Introduction | de novo gene に関する 2 つの未解決テーマ

- ✓ de novo gene に関して、明らかになっていないテーマが大きく 2 つある。

1. 非遺伝子領域がタンパク質のコード能力をどのように獲得するか？
   - ➢ 転写、オープンリーディングフレーム（ORF）、リボソーム結合、翻訳を必要とする。

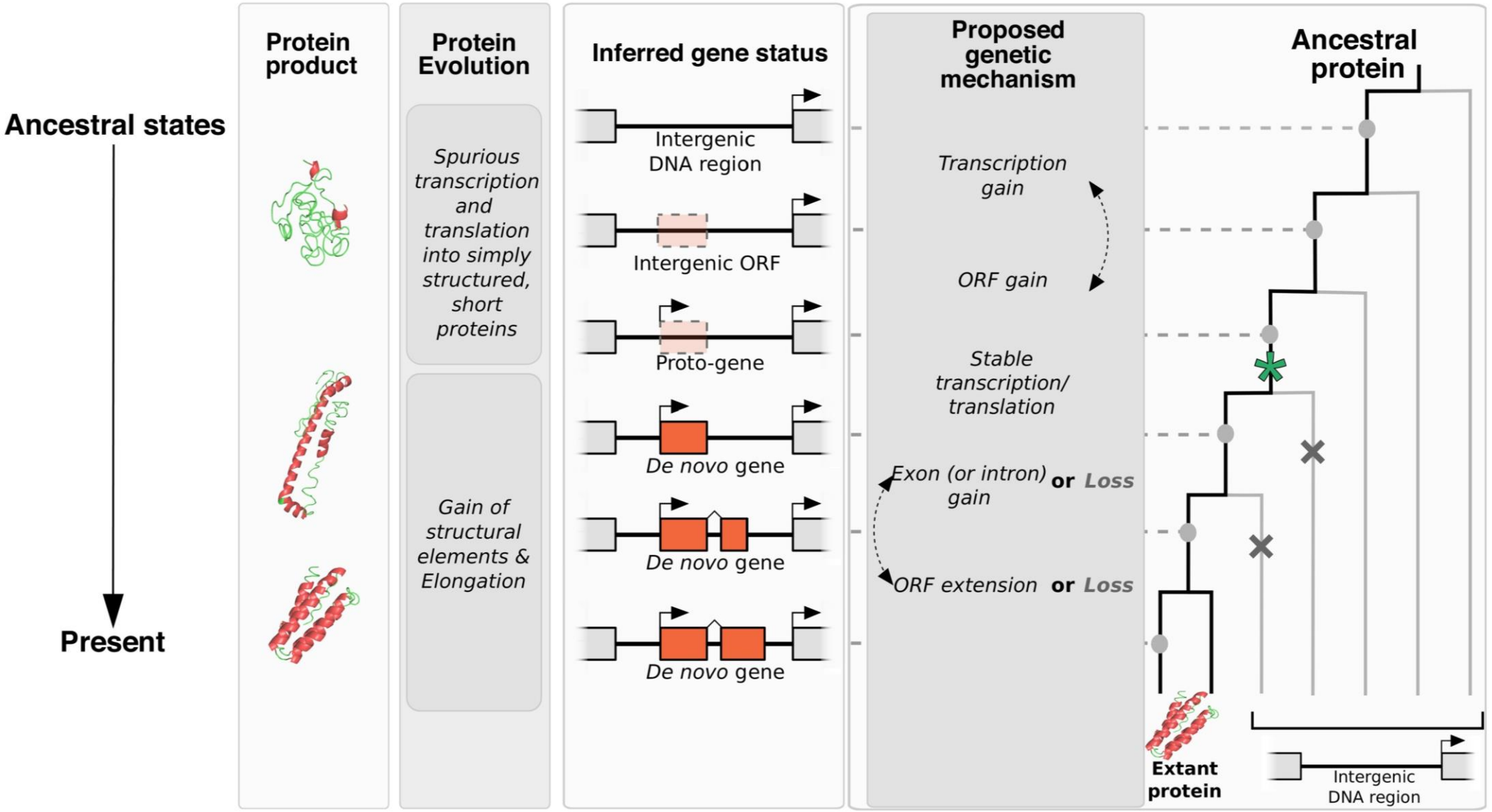
2. コードされるタンパク質の構造や機能はどのようになっているか？
   - ➢ 種にどのような影響を及ぼすのか。

Introduction | セントラルドグマ



© 2017 Terese Winslow LLC
U.S. Govt. has certain rights

Introduction | de novo gene 誕生メカニズム



Bornberg-Bauer et al., 2021

Purpose

- ✓ *Drosophila*（ショウジョウバエ）において、orphan gene に関する先行研究は多いが、de novo gene に関してはあまり研究されていない。

- ✓ 本研究における 3 つの目標

1. *Drosophila* 系統の orphan gene の起源を推定し de novo gene を同定する。

2. 同定した de novo gene の特性を明らかにする。

3. de novo gene がコードするタンパク質の進化の過程を推測する。

Material | ゲノム、プロテオームデータ

- ✓ *Drosophila* 12 種と外群 3 種のゲノム、プロテオーム、CDS、アノテーションデータを取得した。

Species	Release	Abbreviation
D. grimshawi	FlyBase r1.3	*D. gri*
D. mojavensis	FlyBase r1.4	*D. moj*
D. virilis	FlyBase r1.06	*D. vir*
D. willistoni	FlyBase r1.05	*D. wil*
D. persimilis	FlyBase r1.3	*D. per*
D. pseudoobscura	FlyBase r3.04	*D. pse*
D. sechellia	FlyBase r1.3	*D. sec*
D. simulans	FlyBase r2.02	*D. sim*
D. melanogaster	FlyBase r6.11	*D. mel*
D. erecta	FlyBase r1.05	*D. ere*
D. yakuba	FlyBase r1.05	*D. yak*
外群 **Anopheles gambiae*	GCA_000005575.1	*A. gam*
外群 **Lucilia cuprina*	GCA_001187945.1	*L. cup*
外群 **Ceratitis capitata*	GCF_000347755.1	*C. cap*

Material | 転写、翻訳に関するデータ

- ✓ Gene Expression Omnibus（GSE117217）より、 14,423 個の *D. melanogaster* の RNA-seq のメタアナリシスデータを取得した。
- ✓ 質量分析を用いたタンパク質同定に関する 2 つの先行研究（Brunner et al., 2007; Casas-Vila et al., 2017）と SmProt database より、 *D. melanogaster* の翻訳に関するデータを取得した。

Result | *Drosophila* 12 種の各遺伝子数

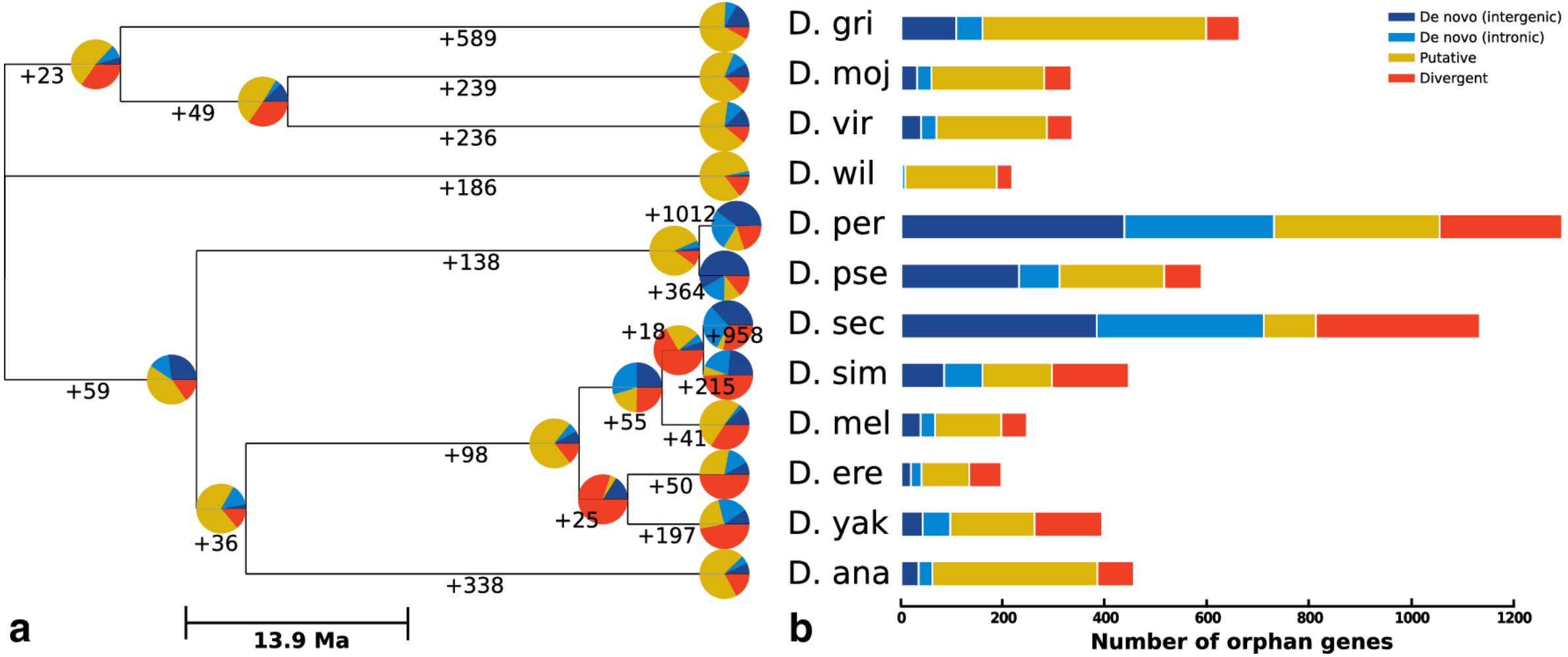
その他

外群の遺伝子と相同性あり

Species	Proteome	Orphans	De novo (intergenic)	De novo (intronic)	Putative	Divergent	Total de novo (all)	% De novo (all)
D. ana	14,365	455	34	27	323	71	61	13.4
D. yak	14,824	393	42	54	165	132	96	24.4
D. ere	13,605	196	19	21	93	63	40	20.4
D. mel	13,907	246	38	28	130	50	66	26.8
D. sim	14,179	445	84	75	136	150	159	35.7
D. sec	16,465	1133	383	327	102	321	710	62.7
D. pse	14,574	588	231	79	205	73	310	52.7
D. per	16,874	1294	437	293	324	240	730	56.4
D. wil	13,783	217	2	6	179	30	8	3.7
D. vir	13,620	335	39	30	216	50	69	20.6
D. moj	13,425	333	31	28	221	53	59	17.7
D. gri	14,982	662	108	51	438	65	159	24.0
Total	174,603	6297	1448	1019	2532	1298	2467	39.2

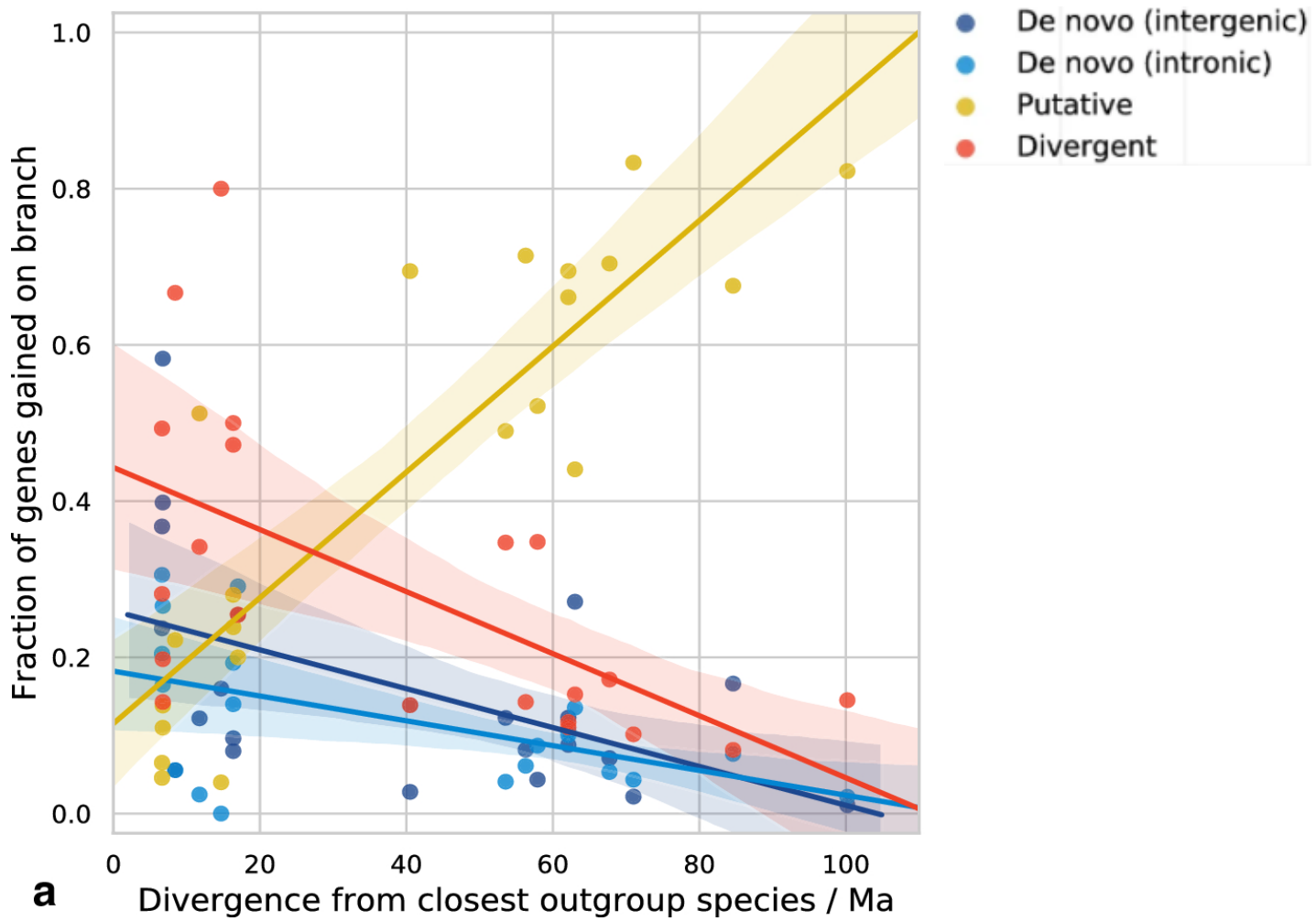
Result | *Drosophila* 12 種の各遺伝子数

- ✓ 種分化の際に遺伝子誕生率が高くなり、その後数百万年の間にこれらの遺伝子の大部分が徐々に失われる。



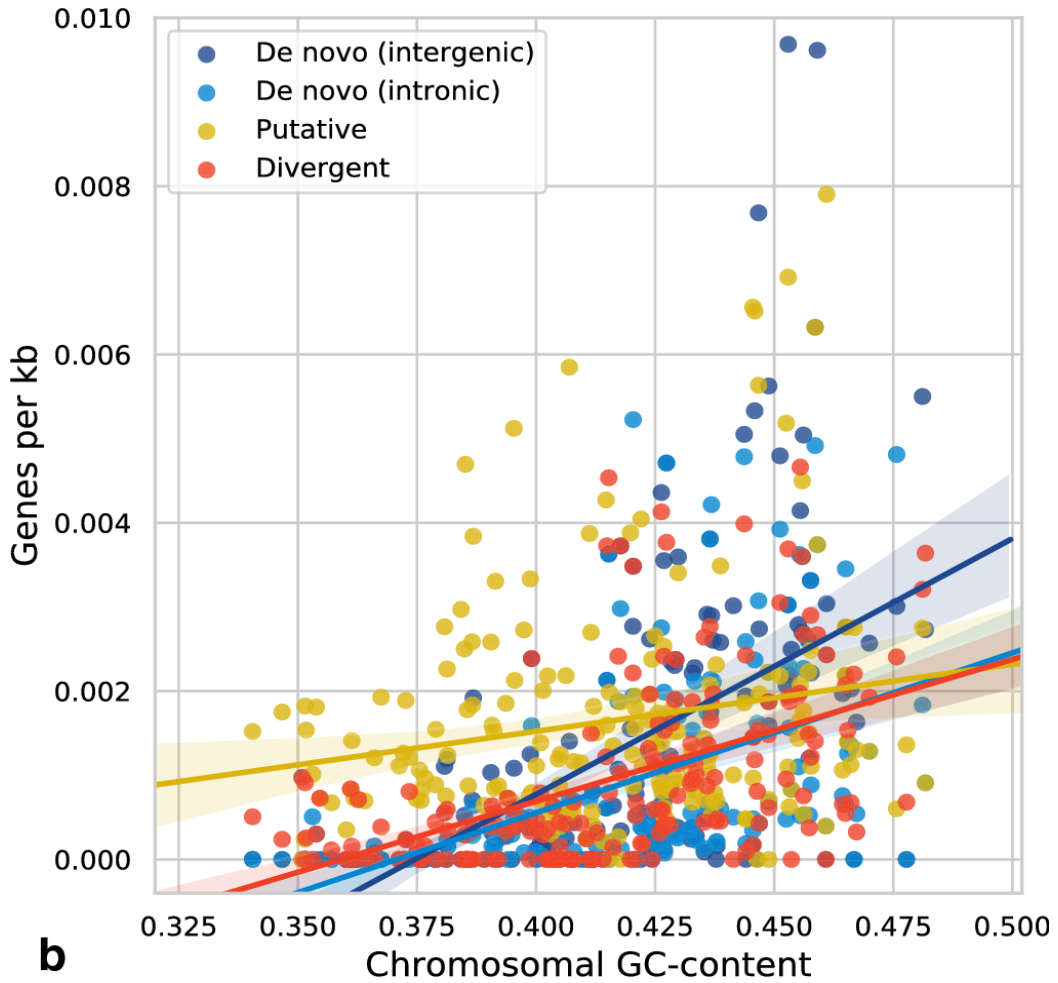
Result | ゲノムの分化と de novo gene

- ✓ 昆虫ゲノムは分化が早いことが知られており、putative に正の相関関係が見られることを踏まえると、同定した de novo gene は信頼できると考えられる。

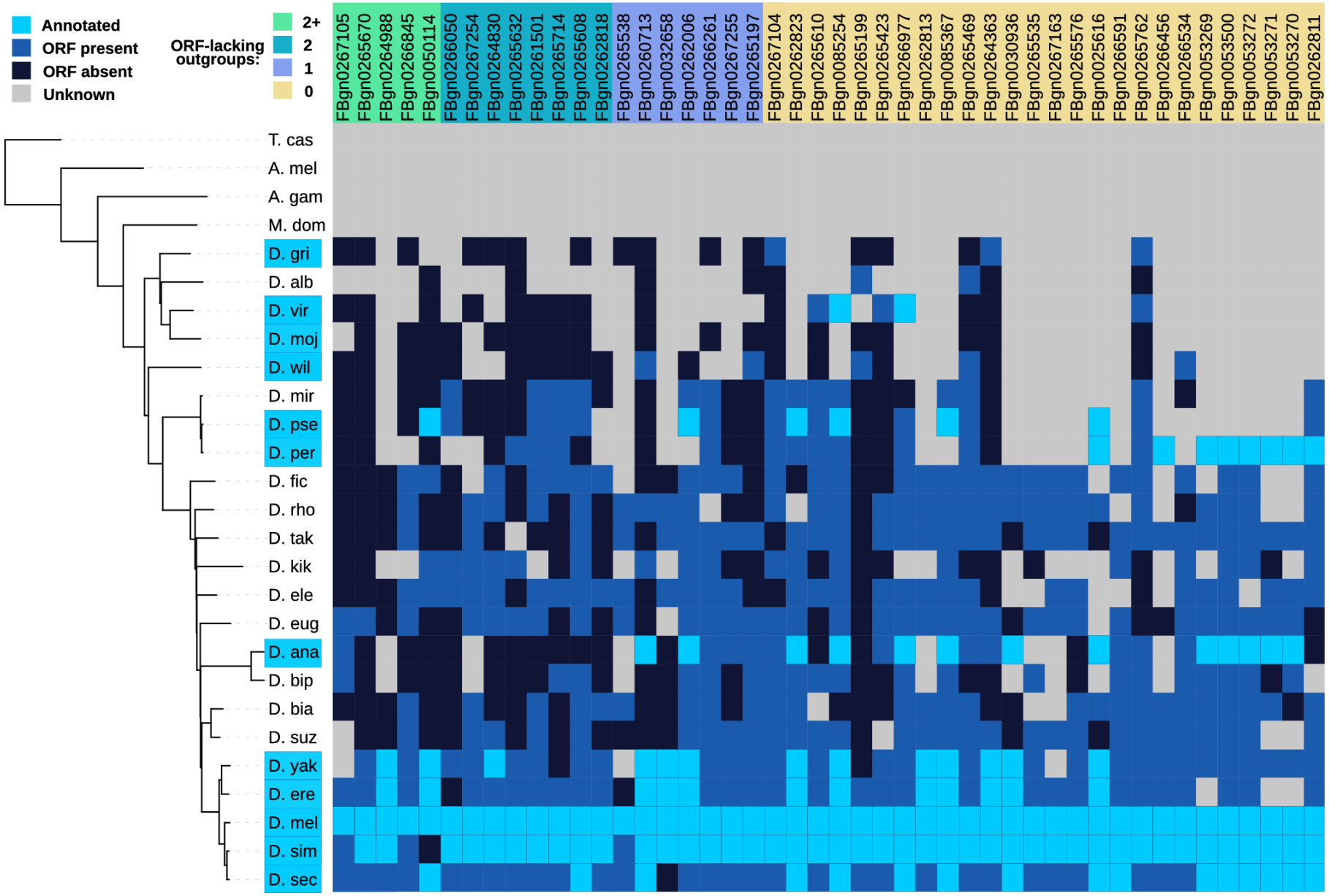


Result | GC 含量と de novo gene

- ✓ orphan gene 密度は染色体の GC 含量と正の相関関係があり、特に intergenic de novo gene は r = 0.56 と他に比べて強い相関を示した。

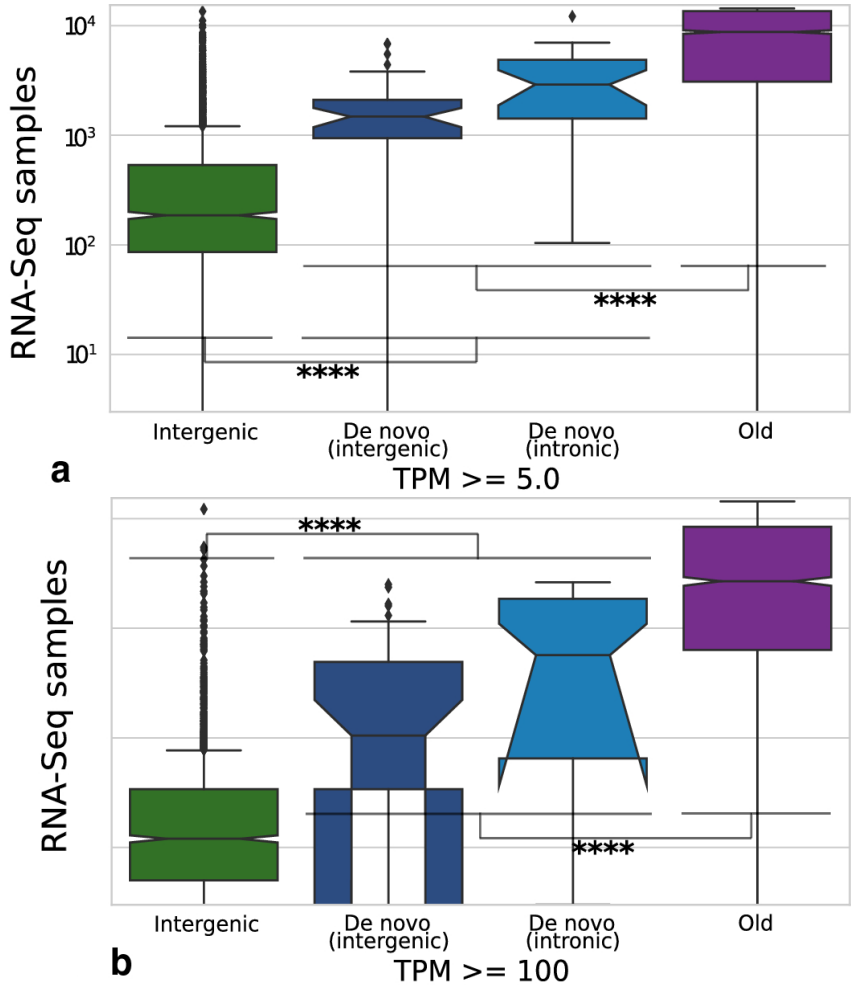


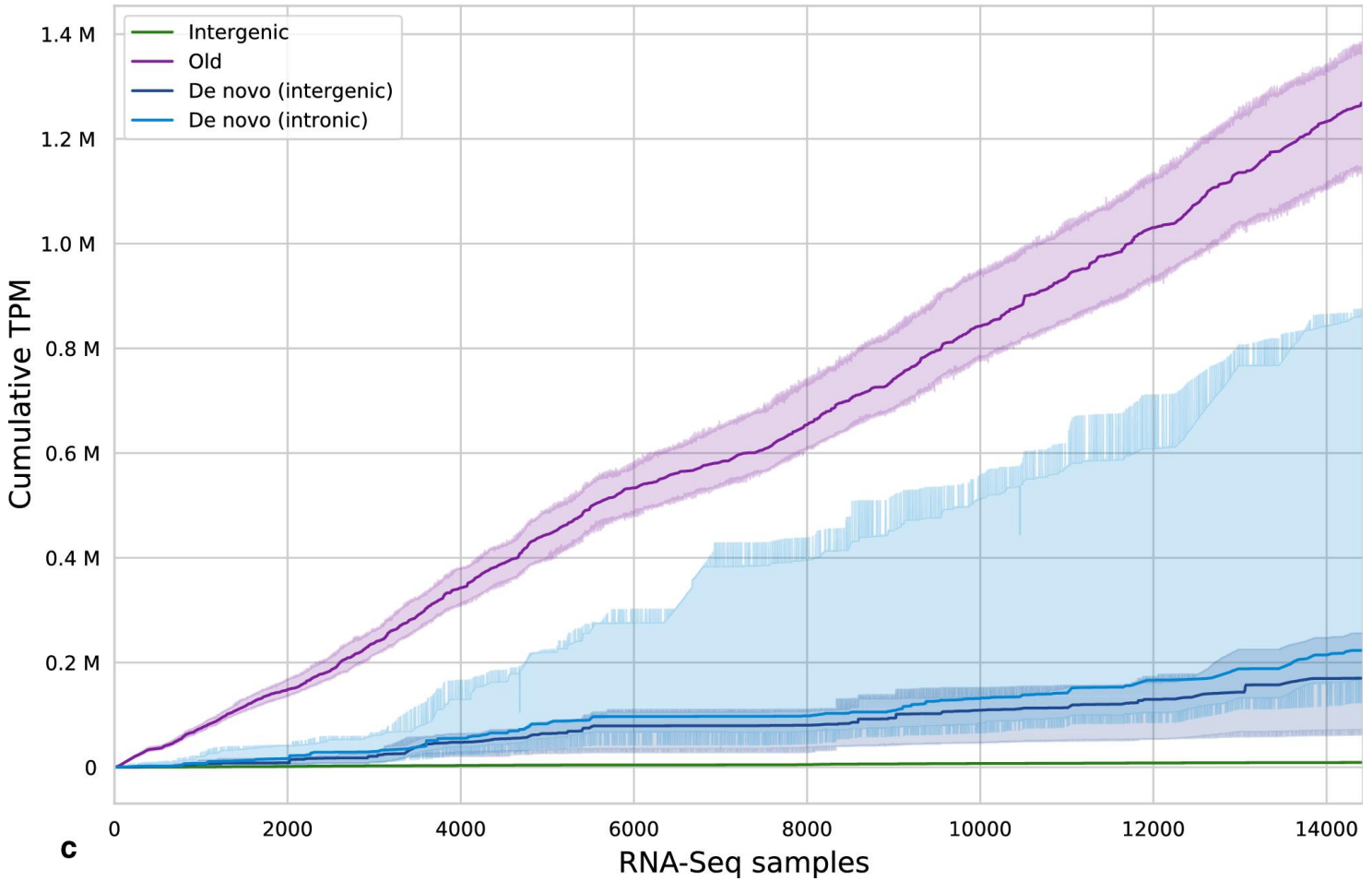
Result | ORF の保存と de novo gene



Result | de novo gene の発現量

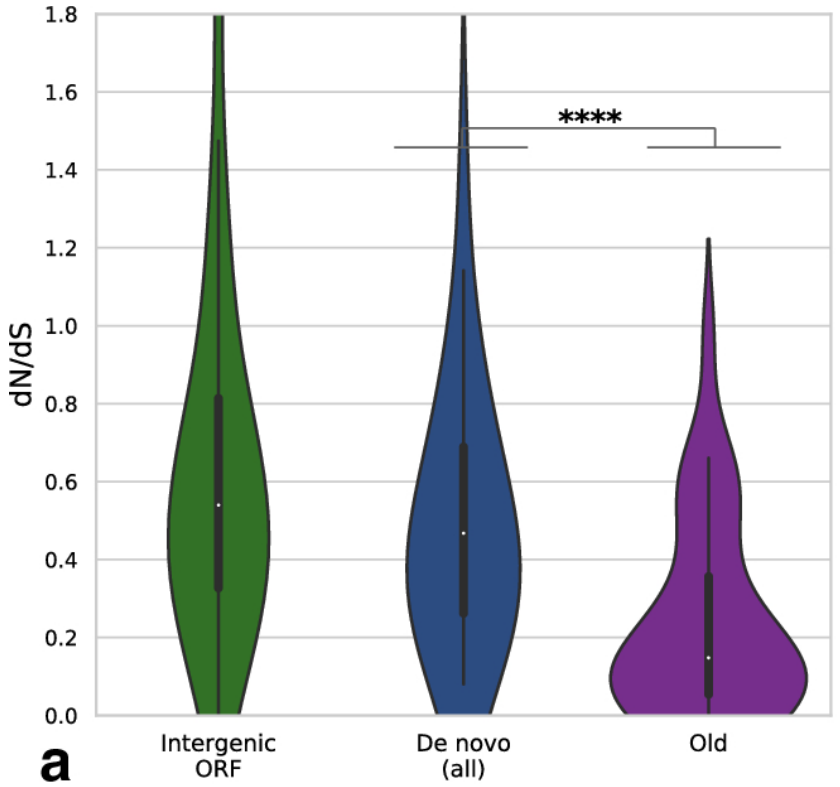
- ✓ de novo gene は遺伝子間領域より大きく、既存の遺伝子より小さく発現しており、最近誕生した結果を反映している。





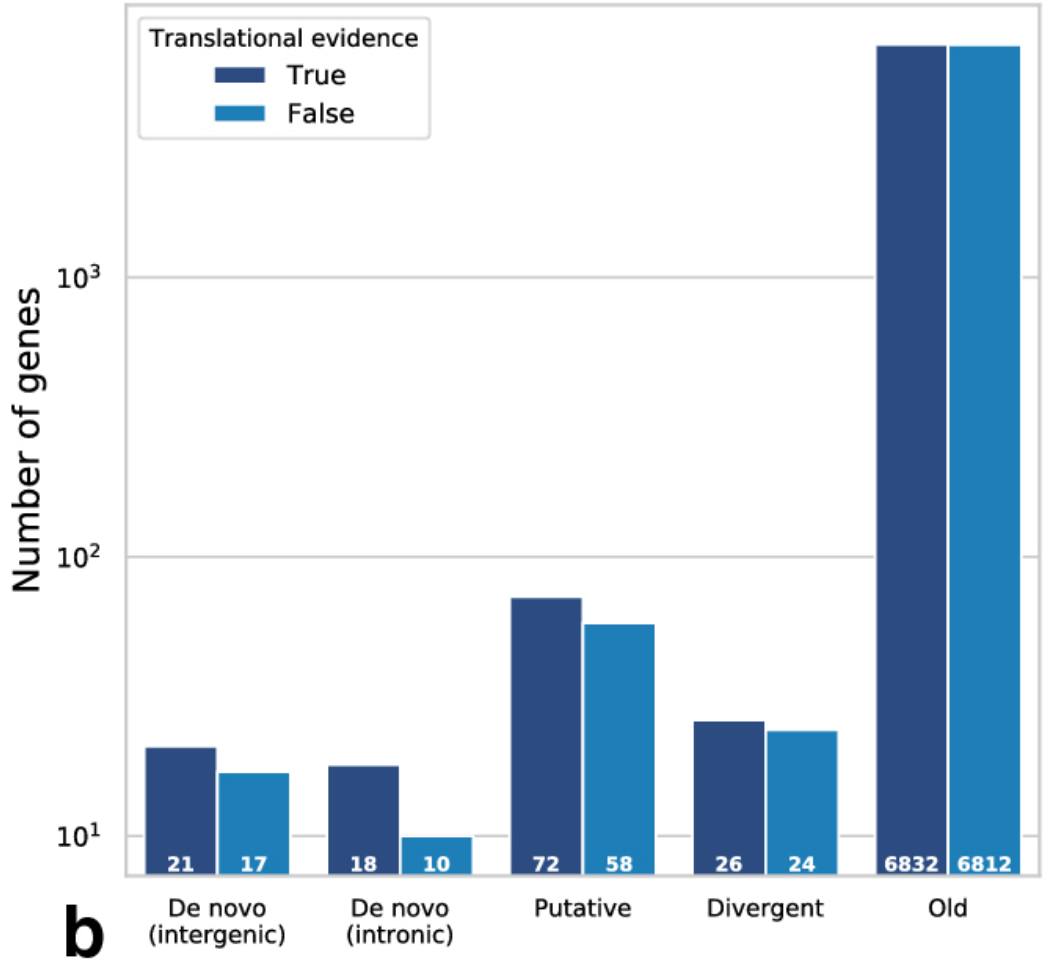
Result | 選択と de novo gene

- ✓ de novo gene は既存の遺伝子より弱い purifying selection がかかっている。
- ✓ 遺伝子間領域から de novo gene が誕生することを考えると、intergenic ORF にも弱い purifying selection がかかっていても矛盾しない。



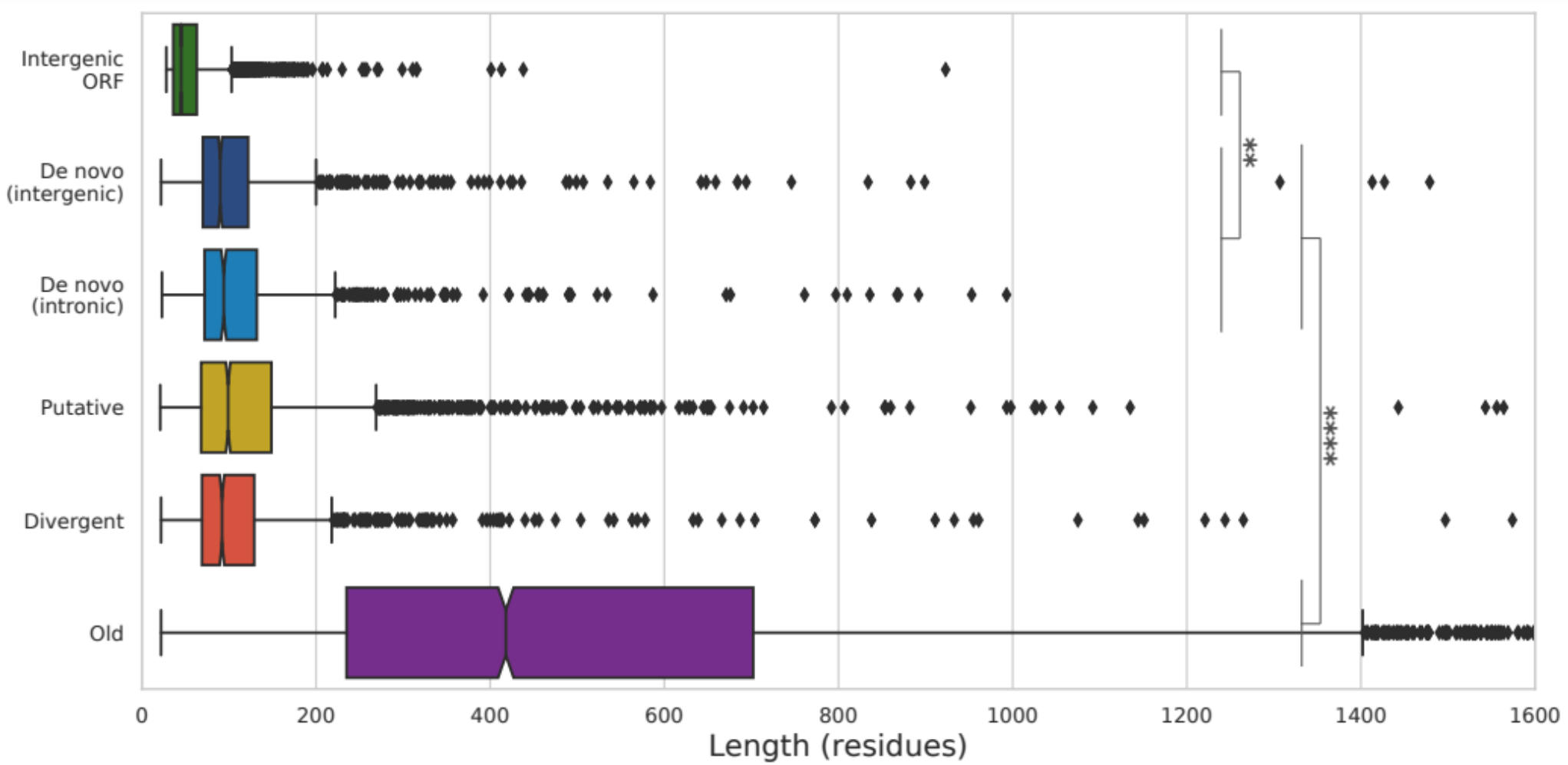
Result | de novo gene の翻訳能力

- ✓ de novo gene の 39 / 66 (59%) が少なくとも 1 つの翻訳の証拠を持っていることが分かった。



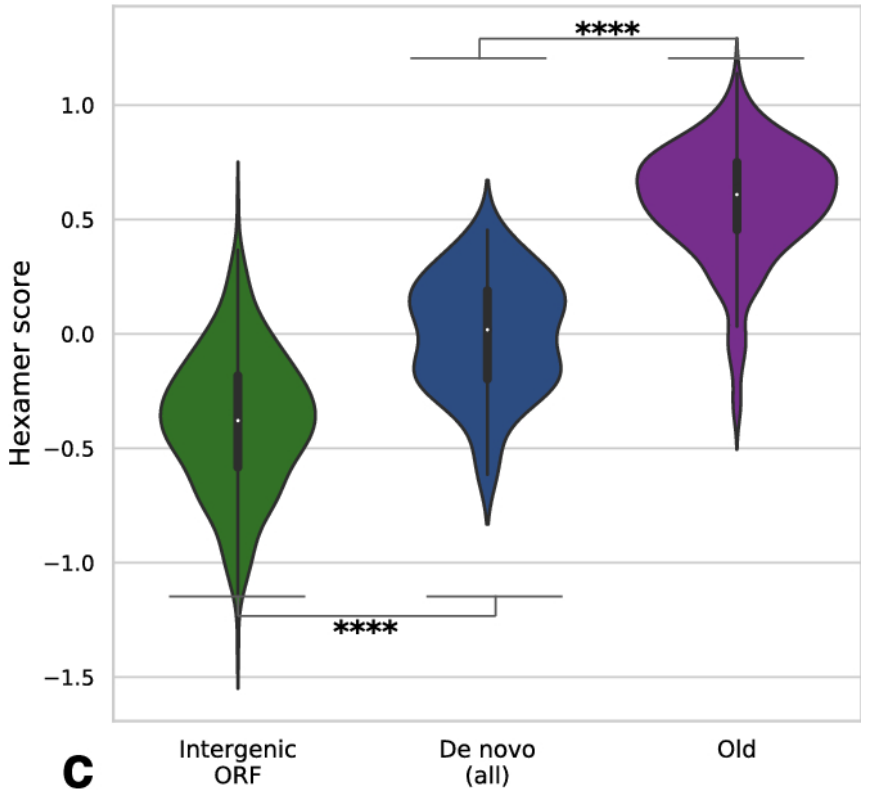
Result | de novo gene の配列特性

- ✓ de novo gene の長さは intergenic ORF と既存の遺伝子の中間を示すことから、徐々に長くなっていることを示唆している。



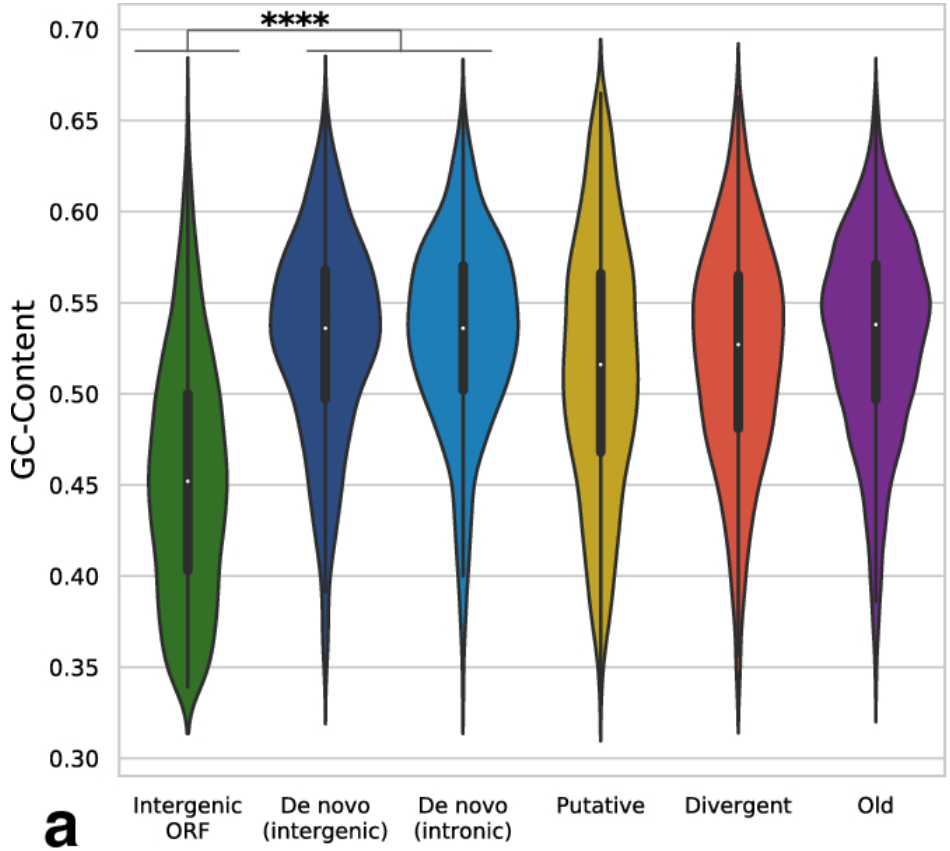
Result | de novo gene の配列特性

- ✓ Hexamer score とは特定の種における遺伝子の codon usage との類似性の尺度である。
- ✓ de novo gene の Hexamer score は intergenic ORF と既存の遺伝子の中間を示すことから、徐々に codon usage が最適化されていることを示唆している。



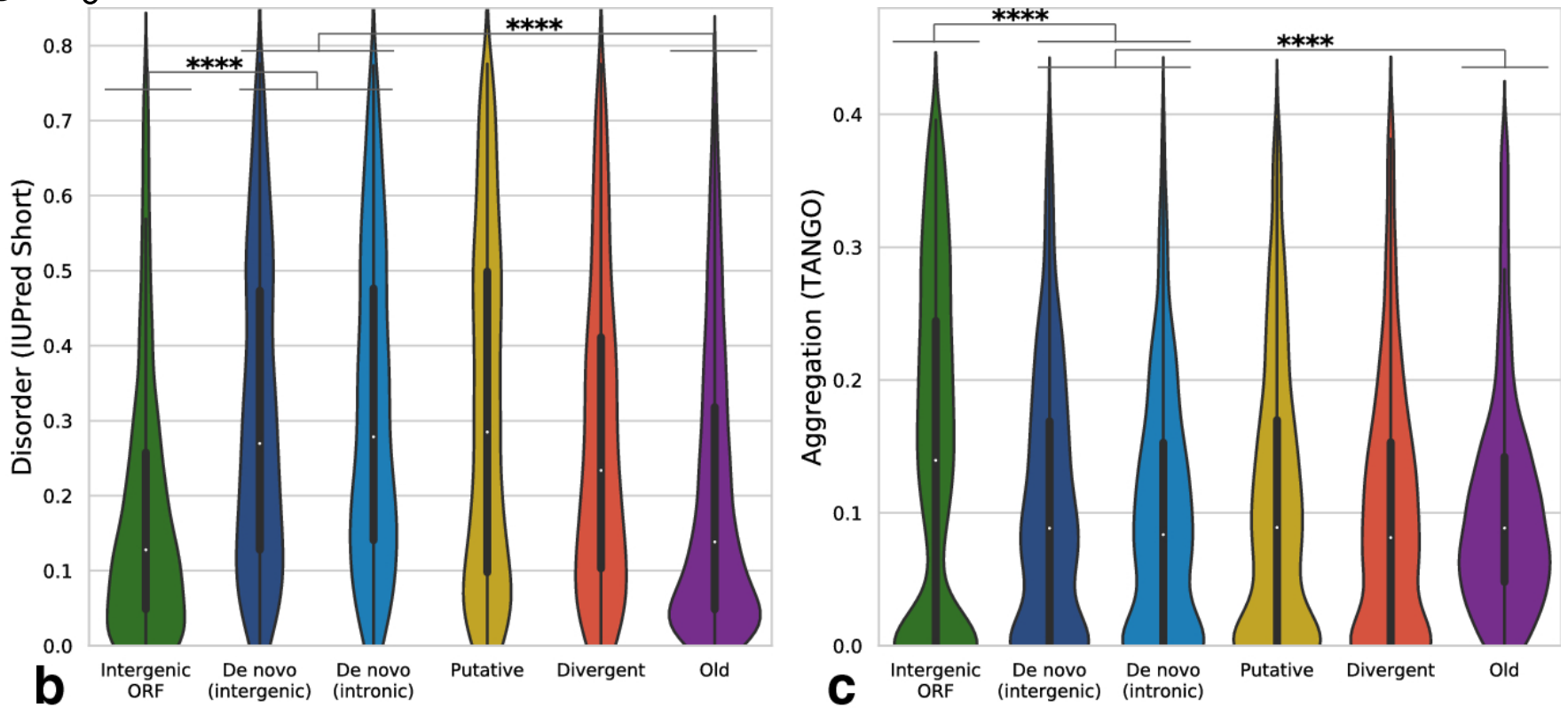
Result | de novo gene の配列特性

- ✓ de novo gene と既存の遺伝子の GC 含量は同程度であった。
- ✓ GC 含量が多い染色体領域からの偏った誕生が、この傾向に寄与している可能性がある。



Result | de novo gene がコードするタンパク質

- ✓ de novo gene は intergenic ORF と既存の遺伝子よりも disorder（天然状態で特定の構造を取らない）タンパク質をコードする。
- ✓ de novo gene は intergenic ORF よりも aggregation（凝集する）タンパク質をコードしない。



Conclusion

✓ 本研究における 3 つの目標

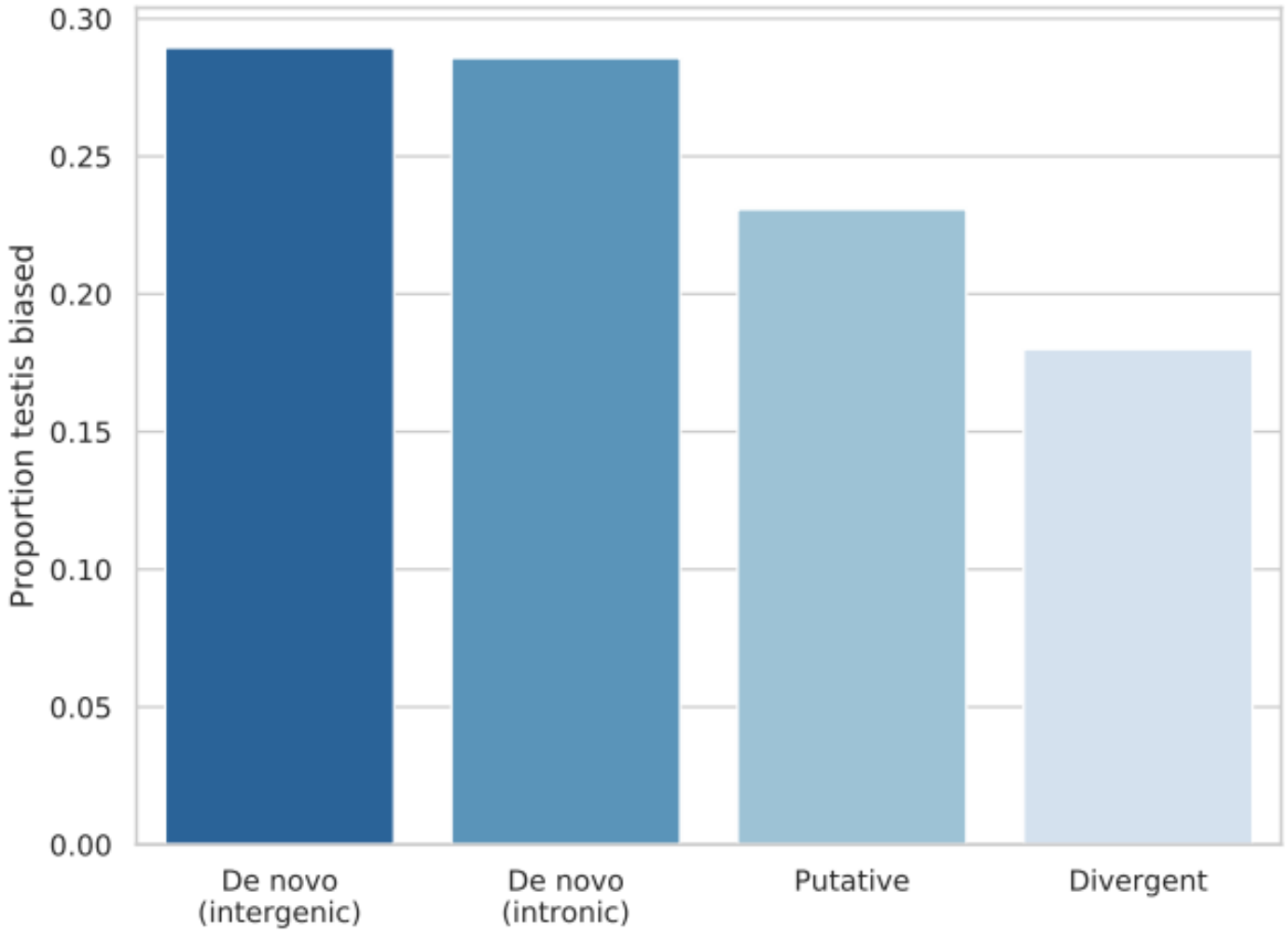
1. *Drosophila* 系統の orphan gene の起源を推定し de novo gene を同定する。
   - *Drosophila* 12 種の遺伝子を de novo (intergenic, intronic), putative, divergent に分類し、合計 2,467 個の de novo gene を同定した。

2. 同定した de novo gene の特性を明らかにする。
   - 長さや Hexamer score、発現量は intergenic ORF と既存の遺伝子の中間の値を示し、段階的に進化していることが示唆された。

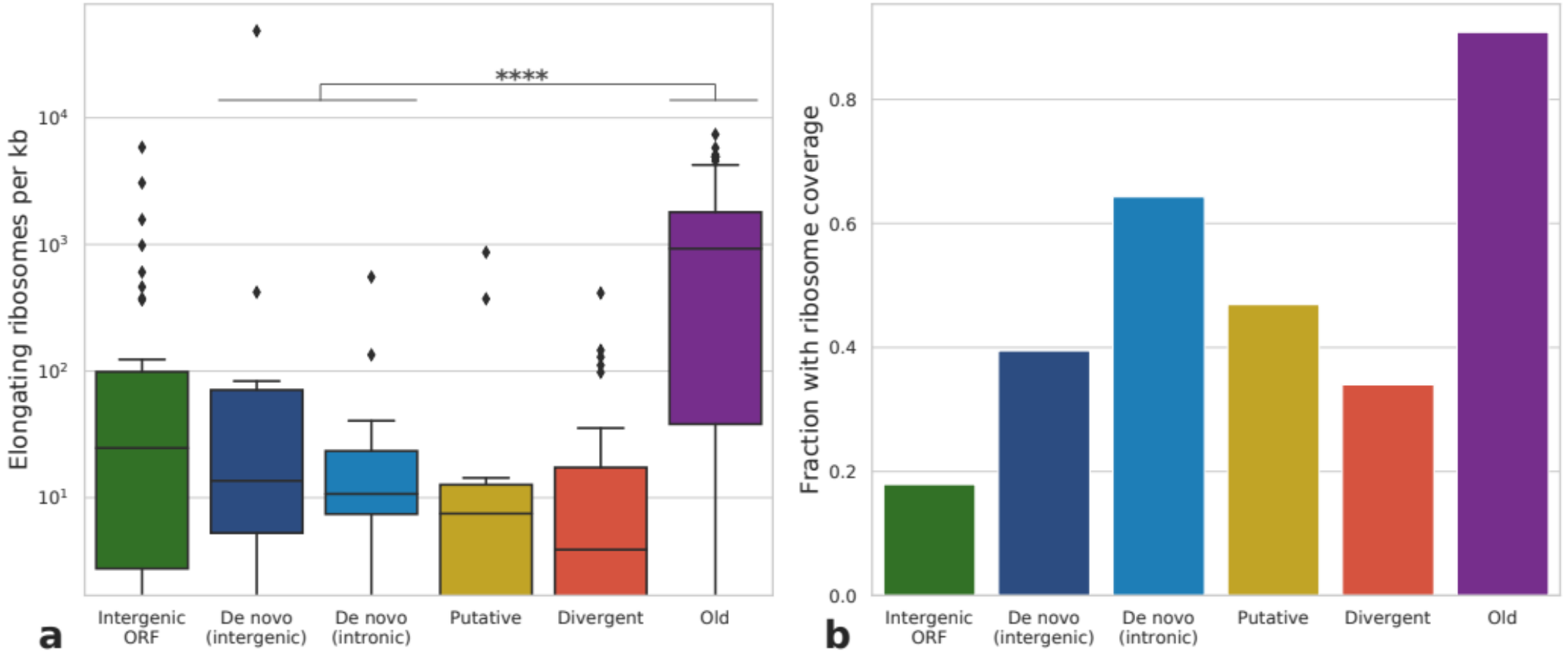
3. de novo gene がコードするタンパク質の進化の軌跡を推測する。
   - disorder を増加させ、aggregation を減少させるような弱い選択がかかっていることが示唆された。

Supplementary | 翻訳の証拠に関するベン図

Supplementary | 精巣における発現量



Supplementary | リボソームプロファイリング



Supplementary | TPM (Transcripts Per Million)

- ✓ RNA-Seq データから得られたリードカウントデータは、そのまま転写産物（遺伝子）発現量を表すわけではない。1 転写産物にマッピングされるリードの数は、サンプル中の総リード数（sequence depth）と転写産物の長さに影響される。サンプル中の総リード数が多いほど、1 転写産物あたりにマッピングされるリード数も多い。また、転写産物が長いほど、1 転写産物あたりにマッピングされるリード数も多い。そのため、RNA-Seq データから得られるリードカウントデータを転写産物発現量として利用するには、総リード数や転写産物長で補正する必要がある。

- ✓ TPM は、サンプル中に全転写産物が 100 万個存在するときに、各転写産物に何個あたりの転写産物が存在するのかを表す値である。