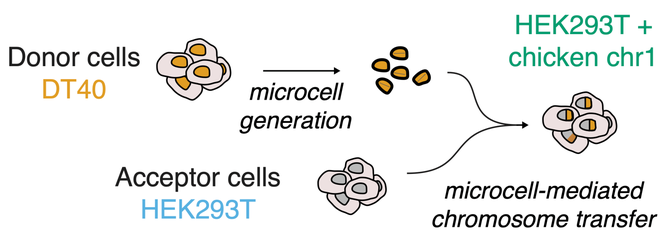
Predicting DNA function when grafted across species

Sergio Garcia Busto[1], Jacob Hepkema[1], Pierre Murat[1], Leopold Parts[1,2]

[1]Wellcome Sanger Institute, Wellcome Genome Campus, Hinxton, Cambridgeshire, CB10 1SA
[2]ExpressionEdits, Haverhill Research Park, Haverhill, CB9 7LR

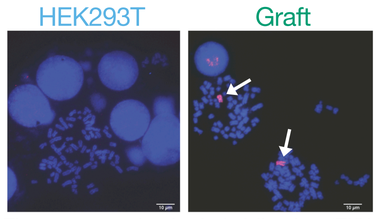
① The chicken graft: a foreign chromosome

We have **human HEK293T cells** (hg38) **with a chromosome from a chicken cell line, DT40** (gg6). This was generated via microcell-mediated chromosome transfer by our collaborators* at the Laboratory of Molecular Biology (LMB). The foreign chromosome (*gg6_1*) is **197 Mb long**, and has been characterised extensively in the Parts lab.



*Thanks to Gianluca Petris and Guillaume Guilbaud, LMB.

The transferred DNA can be seen with a *DT40-specific* probe



Fluorescence in situ hybridisation (FISH); red, DT40 gDNA probes.

② Why should we try to predict graft data?

Deep learning (DL) models trained on genomics attempt to learn a sequence function or DNA '*grammar*'. But the chicken graft has **two** distinct characteristics:

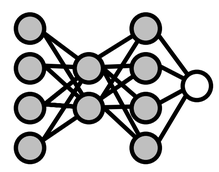
Sequence divergence through 319 million years

Foreign cellular context

By evaluating how DL models predict graft data when exposed to different sequences, we can **assess the extent to which grafted grammar is learnable**

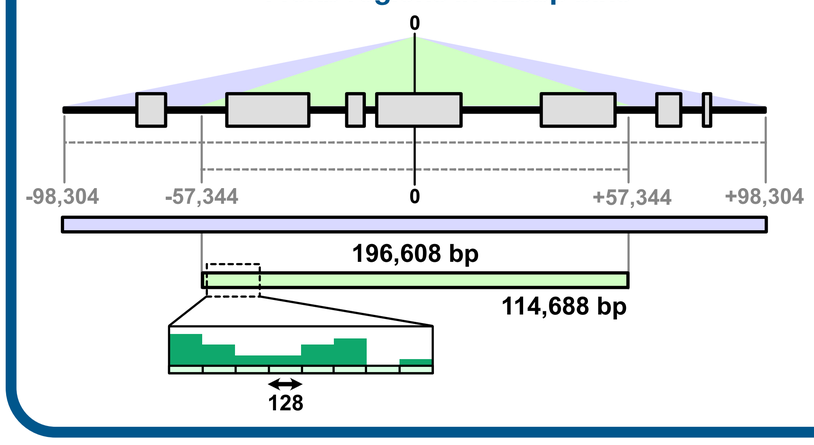
*This poster focuses on **H3K9me3**, a histone mark associated with **heterochromatin.***

Aims:
(i) Evaluate the ability of DL models to **learn regulatory grammar of the grafted chromosome** using H3K9me3
(ii) Interpret results by fine-tuning different models and **comparing predictions to data from native and grafted contexts**

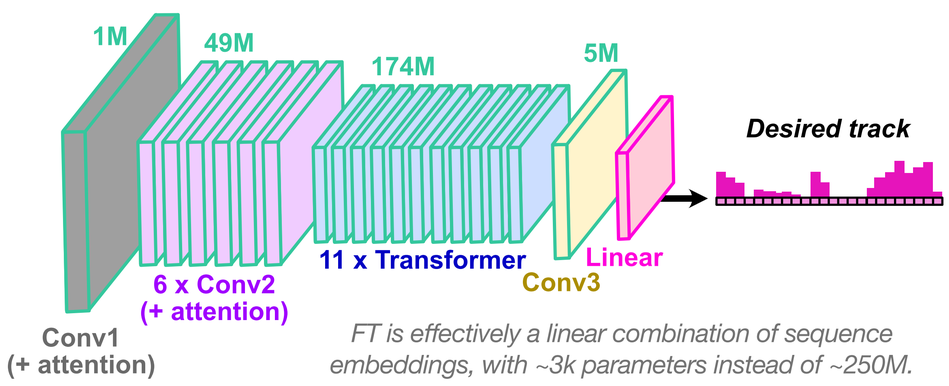


③ Enformer: a Large Language Model (LLM) that predicts genomic tracks

Enformer is **trained on human and mouse genomic tracks, has a context window of ~200kb and predicts the central ~115kb regions in 128bp bins**

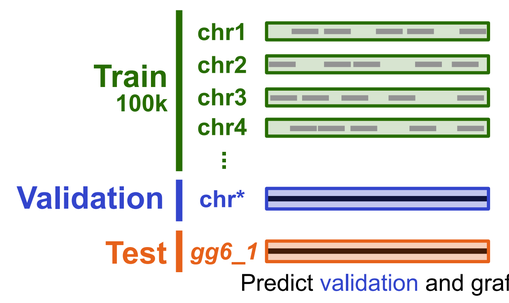


Enformer *fine-tuning* (FT): model **weights are frozen**, and a **custom linear layer is added and trained.** Learned sequence features (embeddings) are used to learn desired grammar.



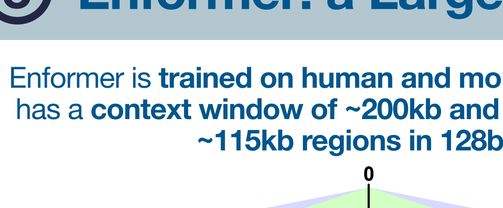
FT is effectively a linear combination of sequence embeddings, with ~3k parameters instead of ~250M.

For **training**, 100,000 sequences of ~200kbp are randomly sampled, and an **entire chromosome is held out** to evaluate FT efficacy



No cross-validation, but validation loss is monitored. Trained for 500 epochs.*

④ Enformer can learn native grammar after fine-tuning, but cannot learn grammar from graft



Poor prediction

Pretrained Enformer

Can it predict a graft mark (H3K9me3)?

Yes — No: *Default prediction is **poor***

Fine-tune on human part of graft

Can it predict an *unseen human chromosome*?

No — Yes

*Learned **human grammar***

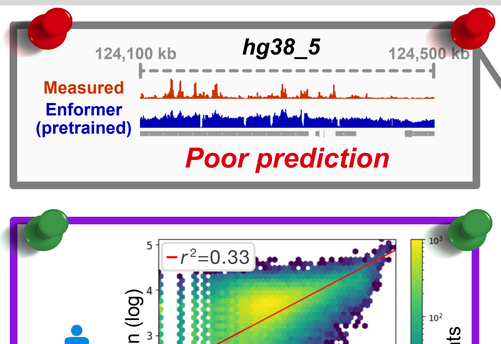
Can it predict a chicken chromosome?

Yes — No

Can it predict the grafted chromosome?

No — Yes

Good correlation** in human and chicken **validation**, **learned both native grammars



Poor correlation:** learning both native grammars **is not enough to predict graft

Fine-tune on human and chicken

Can it predict *unseen chicken and human*?

No — Yes: ***Ensure grafted chromosome can be predicted in native context***

Can it predict the grafted chromosome?

No: ***Enformer cannot learn graft grammar from native grammars***

Yes: *Learnable native grammars sufficient to learn graft grammar*

Fine-tune on chicken data (DT40)

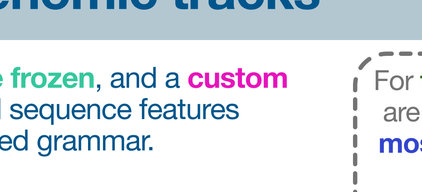
Can it predict an *unseen chicken*?

No — Yes: ***Ensure grafted chromosome can be predicted in native context***

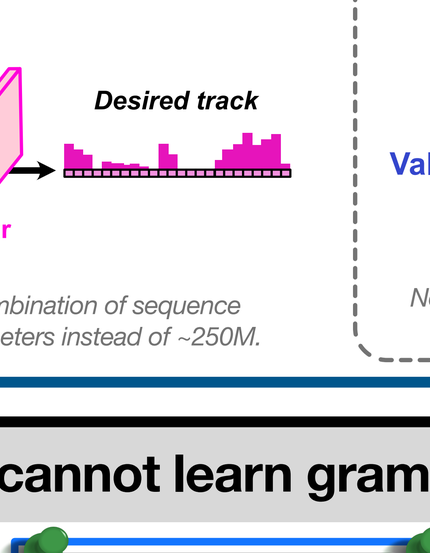
Can it predict the grafted chromosome?

No: ***Grammar different to learnable native chicken or human grammars***

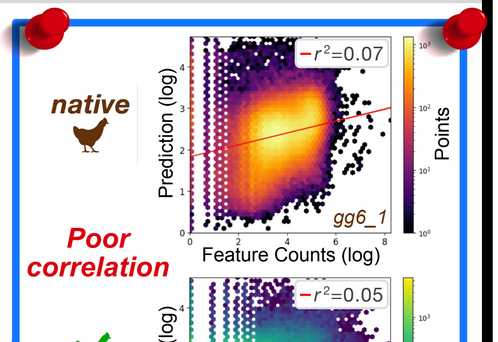
Yes: *Grammar similar to learnable native chicken grammar*



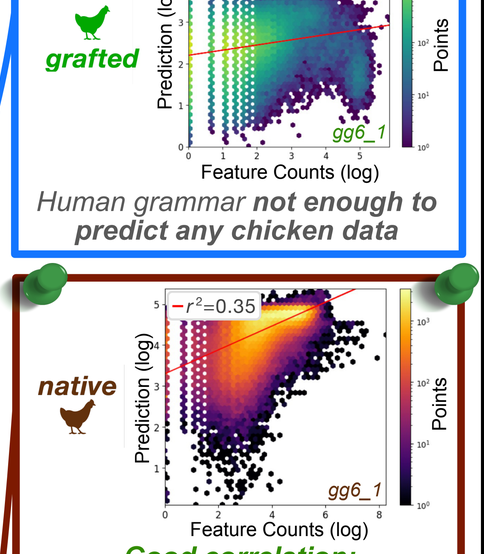
***Good correlation** (validation)*



Poor correlation



*Human grammar **not enough to predict any chicken data***



***Good correlation:** learned chicken grammar*

Poor correlation:** chicken grammar **not enough to predict graft data

r^2 Held out ~1kb bins Log-scale	Pretrained	Human FT	Chicken FT	Chicken + human FT	+ control
hg38_12	0.04	0.50	0.09	0.33	-
gg6_5	0.05	0.10	0.40	0.35	-
gg6_1	0.06	0.07	0.35	0.28	-
gg6_1	0.01	0.05	0.07	0.06	0.15

⑤ Conclusions and future directions

- ★ **After task-specific training, Enformer can learn *native* regulatory grammar**
- ★ **Enformer cannot learn regulatory rules of the grafted chromosome, suggesting they are neither chicken nor human-like**
- ★ Apply workflow to other histone marks and data types
- ★ Integrate biological information to inspect model mispredictions

References

Avsec, Ž., Agarwal, V., Visentin, D. *et al.* Effective gene expression prediction from sequence by integrating long-range interactions. *Nat Methods* **18**, 1196–1203 (2021).