Abstract #959F

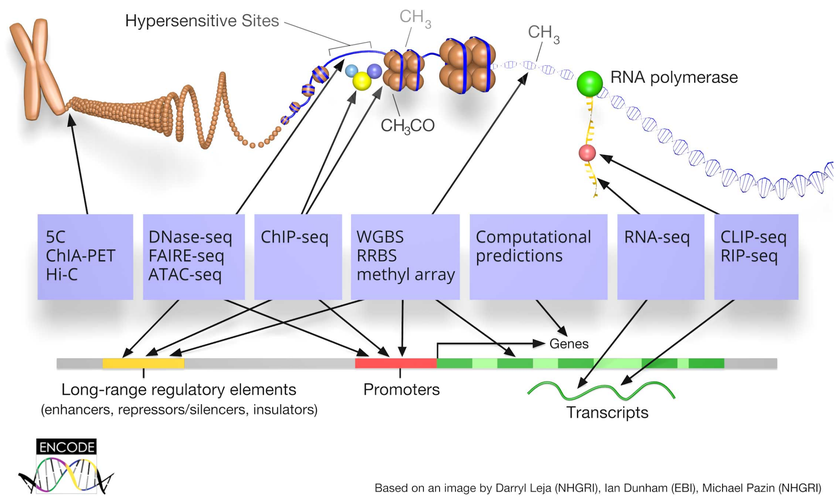
Identifying genomic sites with highly predictable DNA accessibility

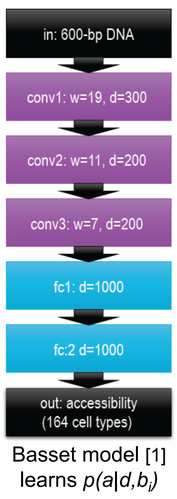
AUTHORS

Kamil Wnuk[1], Jeremi Sudol[1], Shahrooz Rabizadeh[1], Christopher Szeto[2], Charles Vaske[2]. [1]NantOmics LLC., Culver City, CA; [2]NantOmics LLC., Santa Cruz, CA

BACKGROUND

- DNA accessibility, chromatin regulation, and genome methylation are all key drivers of transcriptional events promoting tumor growth





Basset model [1] learns $p(a|d,b_i)$

- DNA-based prediction tasks have all been cell/tissue type specific
  - Basset [1] predicts accessibility given DNA & discrete tissue type: $p(a|d,b_i)$
- Cell or tissue type can be predicted from RNA-seq [3]: $p(b|r)$ can be learned
- ENCODE [2] has local structure w.r.t. tissue type in RNA-seq and DNase-seq data

METHODS

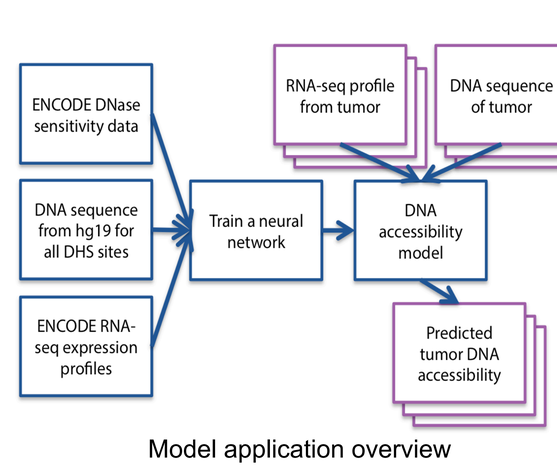
- Factorizing convolutional layers improved the baseline model on Basset dataset

Test metric on Basset dataset	ROC AUC	PR AUC
Basset model	0.895	0.561
Ours (no RNA-seq)	0.910	0.605

- Let the neural net implicitly handle tissue type: add RNA-seq signature as input
  - we learn $p(a|d,r) = \Sigma_i^B p(a|d,b_i)p(b_i|r)$
- L1000 genes from Library of Integrated Network-based Cellular Signatures (LINCS)
- New data: 74 unique tissue types from ENCODE; matched RNA-seq, DNase-seq

	Unique tissues	DNase-seq samples	RNA-seq samples
Train	66	198	281
Valid	10	11	12
Test	8	11	11

Dataset Partitions



Model application overview

Our model [4] learns $p(a|d,r)$

RESULTS

Figure 1. Promoter and promoter flank (P&F) accessibility is highly predictable

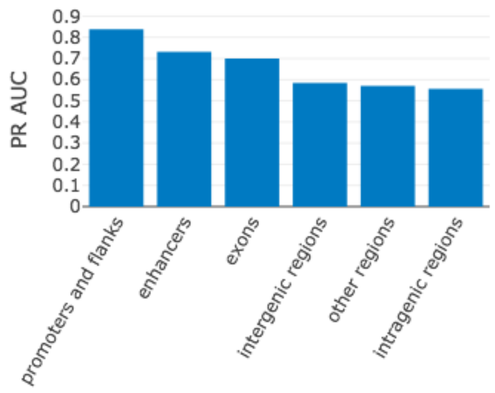


Table 1. P&F predictability is retained even when holding out similar tissue types

Dataset Partition	Held out tissue types		Held out samples with tissue type overlap	
Test metric	ROC AUC	PR AUC	ROC AUC	PR AUC
Over all sites	0.897	0.621	0.913	0.725
Promoter & Flank	0.876	0.839	0.914	0.911

Figure 2. Tissue-type affects accessibility prediction accuracy

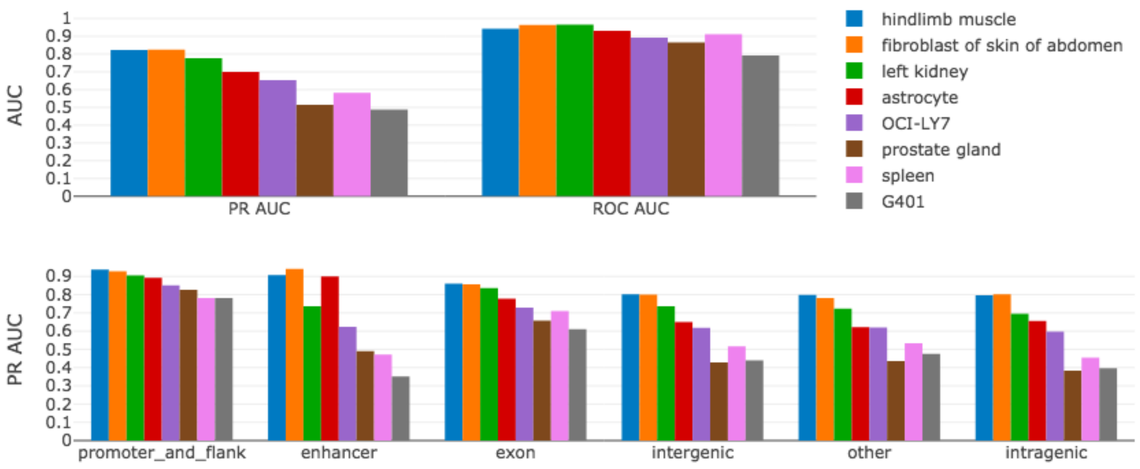
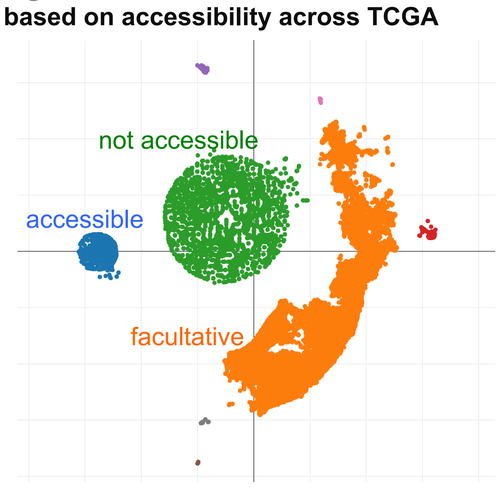


Table 2. Prediction performance is less correlated with test sample similarity to training data at P&F sites than when evaluated over all potentially accessible sites

	Pearson corr.	Pearson p-value	Spearman rho	Spearman p-value
Overall	-0.7472	1.77e-05	-0.7080	7.52e-05
P&F	-0.6795	1.87e-04	-0.5417	5.16e-03

PR AUC vs. min RNA-seq dist. to training

Figure 3. P&F sites form distinct clusters based on accessibility across TCGA



Site cluster property	Number of TCGA P&F sites
Constitutively not accessible	40,823
Constitutively accessible	10,878
Facultative	52,891

Figure 4. INDEL mutations have more impact on site accessibility than SNP mutations

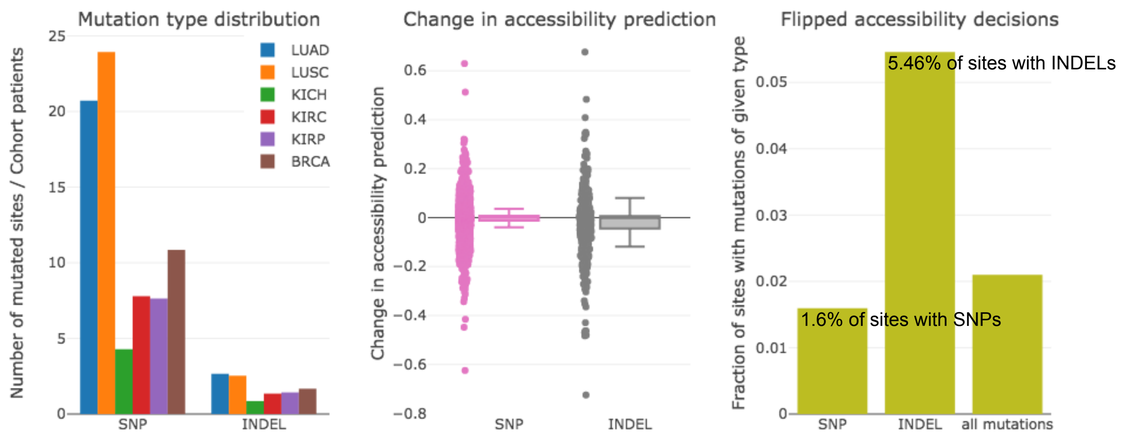
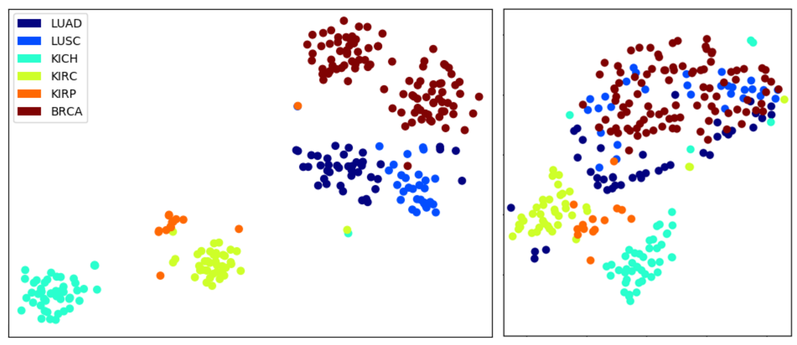


Figure 5. TCGA samples cluster differently when t-SNE is based on predicted accessibility rather than RNA-seq



RNA-seq predicted accessibility

CONCLUSIONS

- **Adding an RNA-seq signature as input** allows the model to figure out how tissue type/state affect DNA-sequence-based prediction tasks
  - No need to train one model per type or multitask outputs
  - Applies to new types not seen in training
- At **promoter and promoter flank regions** of the genome it is **possible to predict DNA accessibility** to much higher precision than any prior results
- Performance is independent of whether sites overlap with L1000 genes
- **Some tissues are more challenging**, but not purely due to distance from training
  - Most difficult test tissue, G401, is most different from training examples
  - However, astrocytes are more different than prostate or spleen
- **INDEL mutations cause more accessibility predictions to flip** than SNPs
- Clustering cohorts based on accessibility gives distinct assignment from RNA-seq
  - Distinct differentially expressed pathways
- Offers a **distinct perspective from analysis of RNA-seq** alone

REFERENCES

1. D. R. Kelley, J. Snoek, J. L. Rinn, "Basset: Learning the regulatory code of the accessible genome with deep convolutional neural networks," Genome Research, 2016, 26(7)
2. ENCODE Project Consortium, "An integrated encyclopedia of DNA elements in the human genome," Nature, 2012, 489(7414)
3. A. Conesa, et al. "A survey of best practices for RNA-seq data analysis," Genome Biology, 2013, 17(1)
4. https://www.biorxiv.org/content/early/2017/12/05/229385

Corresponding author: kwnuk@nantomics.com