Introduction to Karas & Impact of Genomic Context on Variant Calling

Arkarachai Fungtammasan (Chai)
Steve Osazuwa
Naina Thangaraj
Jason Chin

DNAnexus®
The Global Network For Genomics™

Purpose

- Introduction to Keras and how to use it for deep learning in genomics
- Use case on studying the impact of genomic context on variant calling

About myself

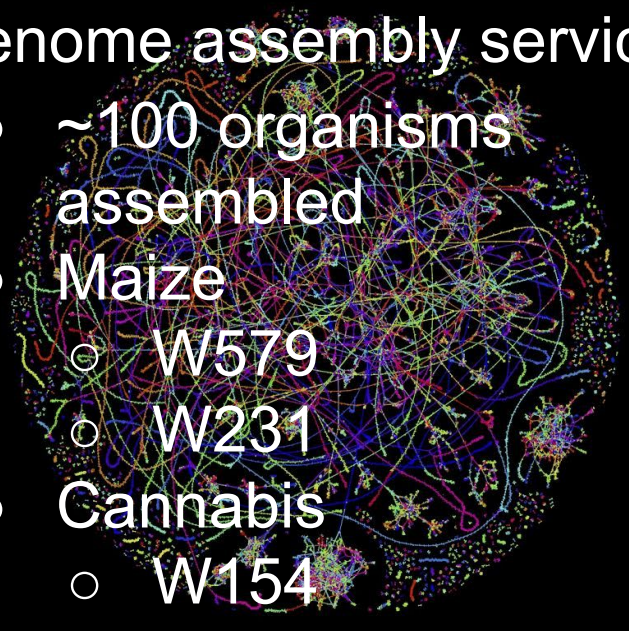
About myself

Booth 227

DNAnexus

Genome assembly service

- ~100 organisms assembled
- Maize
  - W579
  - W231
- Cannabis
  - W154



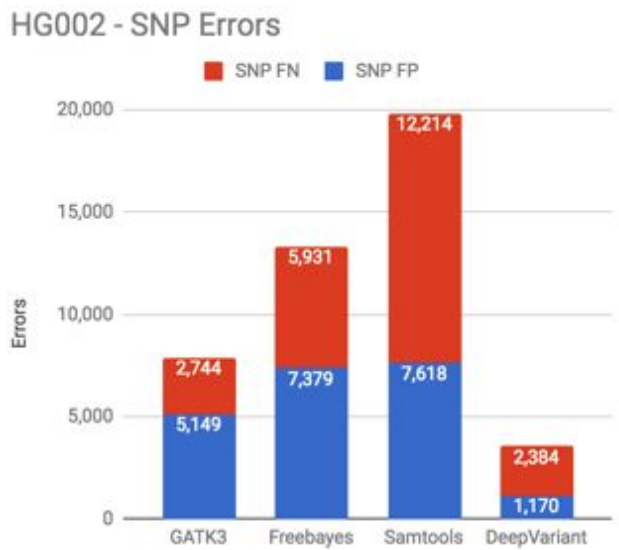
Science frontier program

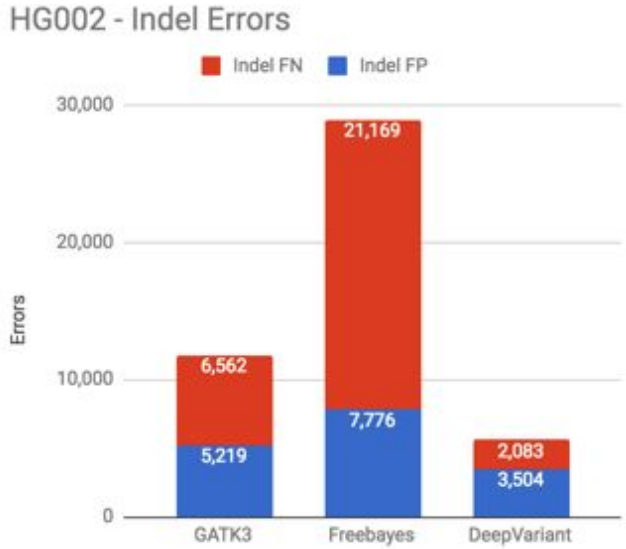
- 10% of time to work on cutting edge science problems and contribute back to the scientific community

- Results frequently published in bllog posts https://blog.dnanexus.com, preprints, or presented as a conference talk

Science frontier program

EVALUATING DEEPVARIANT: A NEW DEEP LEARNING VARIANT CALLER FROM THE GOOGLE BRAIN TEAM





Outlines

- What is Keras?
- How to get started?
- Basic Keras code structure
- Real world application: Impact of genomic context on variant calling

Outlines

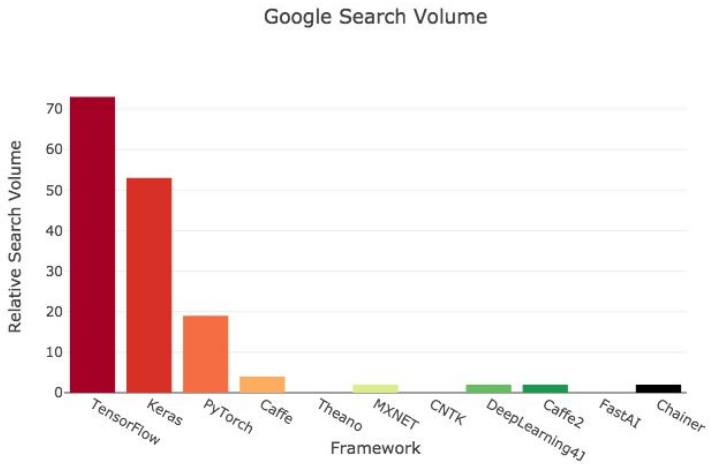
- What is Keras?
- How to get started?
- Basic Keras code structure
- Real world application: Impact of genomic context on variant calling

Keras (care-ras)

- High level Python library/API for deep learning run on top of TensorFlow, CNTK, and Theano
- Designed with user friendliness, modularity, and extensibility principles
- Support common stuffs:
  - Model: CNN, RNN, combination
  - Hardware: CPU, GPU, TPU, Spark

Why Keras ?

- Easy to learn and code. “Keras is an API designed for human beings, not machines.”
- Popularity



How to get started?

- https://keras.io
  - Install backend
  - pip install keras
- https://www.tensorflow.org/tutorials/
  - Install tensorflow
  - from tensorflow import keras
- Kaggle online kernel

Keras code structure

1 Model

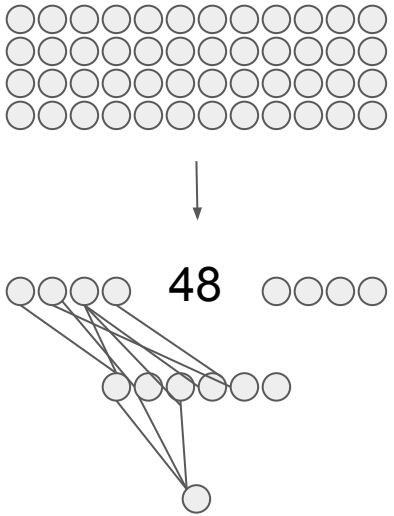
```
def keras_dna_model():

    model = Sequential()

    model.add(Flatten(input_shape=(4, 12)))

    model.add(Dense(6, activation='relu'))

    model.add(Dense(1, activation='sigmoid'))

    return model
```

Picture from: flicker.com

Keras code structure

1 Model

```
def keras_dna_model():

    model = Sequential()

    model.add(Flatten(input_shape=(4, 12)))

    model.add(Dense(6, activation='relu'))

    model.add(Dense(1, activation='sigmoid'))

    return model
```



Keras code structure

1 Model

```
def keras_dna_model(input_shape):

    X_input = Input(input_shape)

    X = Flatten()(X_input)

    X = Dense(6, activation='relu', name='n1')(X)

    X = Dense(1, activation='sigmoid', name='n2')(X)

    model = Model(inputs=X_input, outputs=X, name='keras_dna_model')

    return model
```

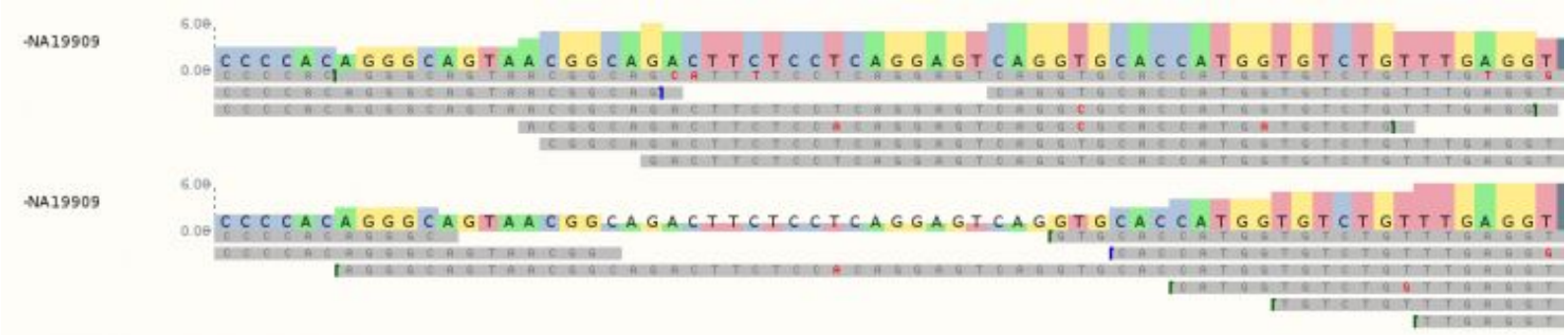
Picture from: flicker.com

Keras code structure

2 Run model

```
model = keras_dna_model(input_dimension)

model.compile(optimizer='adam', loss='binary_crossentropy', metrics=['accuracy'])

model.fit(X_train, Y_train, epochs=2, batch_size=100)



model.predict(X_test)

model.evaluate(X_test, Y_test, batch_size=20168, verbose=1)
```

Impact of genomic context in variant calling



Picture from: EMBL

Impact of genomic context in variant calling

Right flanking

NTTTTTTTTTTTT concordant with ground truth = 0.71

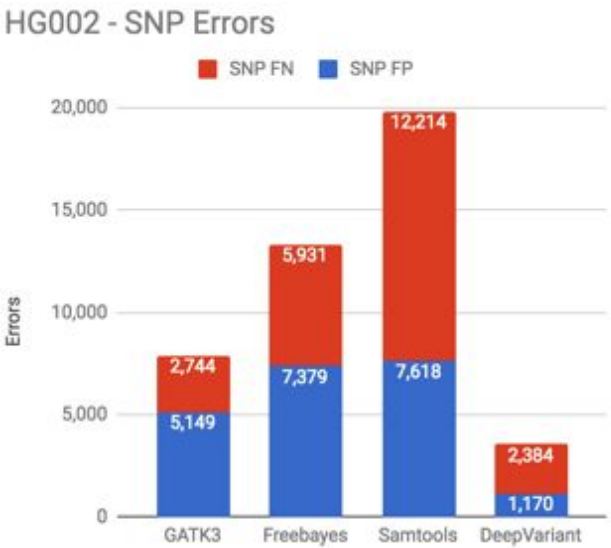
NTTTTTTTTTTTC concordant with ground truth = 0.92

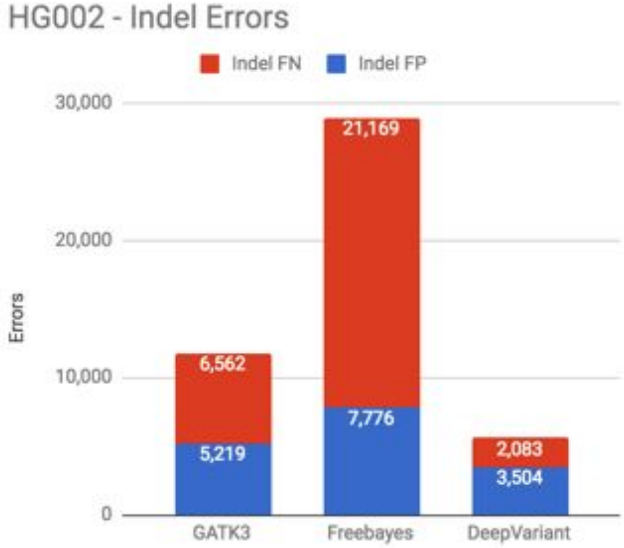
*calculation based on high-confidence regions only

Problem formulation

Feature		Response
ATTCGACGGGG	→	Correct variant (1)
AAAATCCTAAAA	→	Incorrect variant (0)

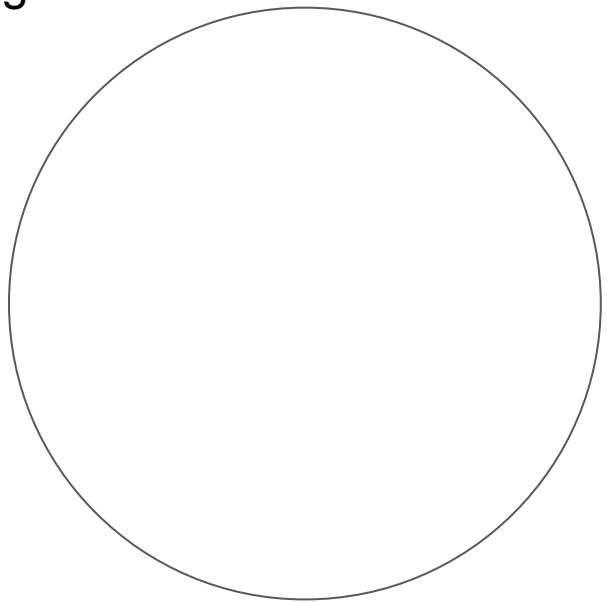
Data


HG002 - SNP Errors
SNP FN
SNP FP
20,000
15,000
10,000
5,000
0
Errors
2,744
5,149
5,931
7,379
12,214
7,618
2,384
1,170
GATK3
Freebayes
Samtools
DeepVariant


HG002 - Indel Errors
Indel FN
Indel FP
30,000
20,000
10,000
0
Errors
6,562
5,219
21,169
7,776
2,083
3,504
GATK3
Freebayes
DeepVariant

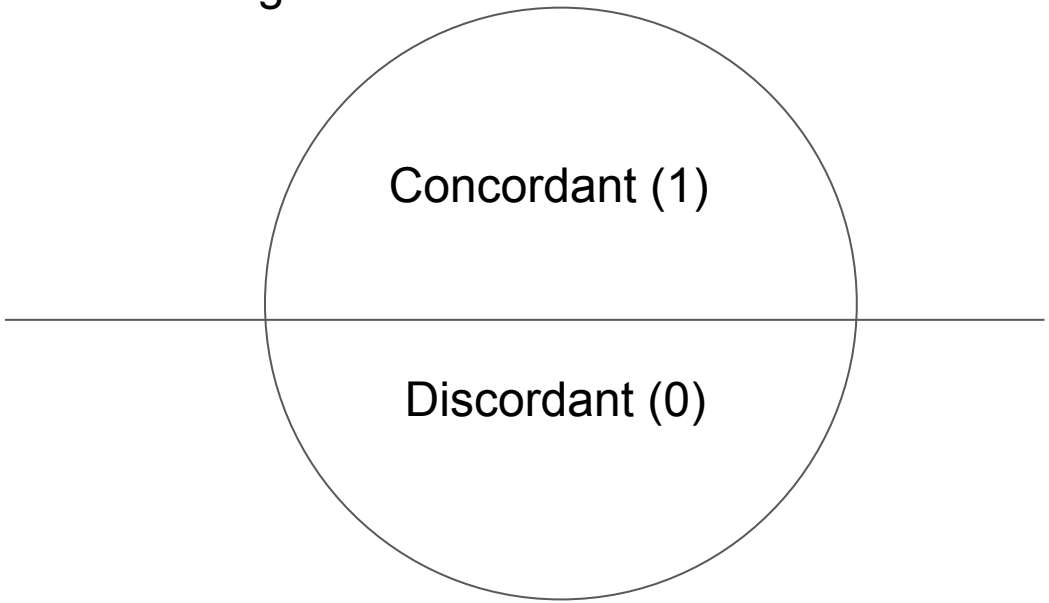
Data

Indel DeepVariant calling
on HG002



Data

Indel DeepVariant calling
on HG002



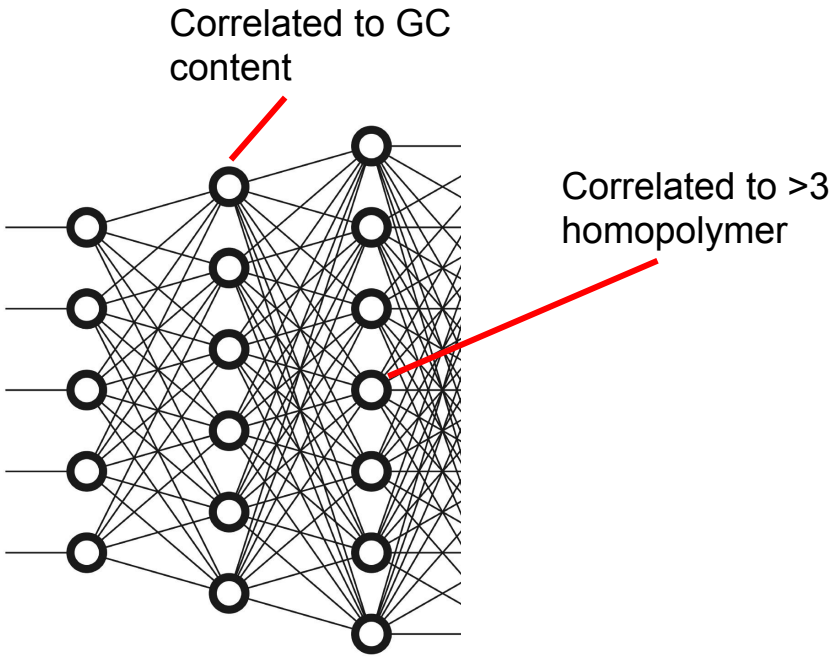
GIAB HG002
Variant truth set

Why deep learning?

	feature1	feature2	response
	GC%	>3 homopolymer	
1	0.2	1	1
2	0.6	1	0
3	0.4	0	0
4	0.2	1	1

Why deep learning?

```
A T C A G C A T

1 0 0 1 0 0 1 0
0 0 1 0 0 1 0 0
0 0 0 0 1 0 0 0
0 1 0 0 0 0 0 1
```



Notebook demo

Caveats

- Unoptimized model
- Size of context
- DeepVariant probably learn a lot about genomic context by itself
- Summarize of attribute in one matrix

Next Step

- Production scale of parameter experiment using Papermill https://papermill.readthedocs.io/en/latest/

Thanks to

Jason Chin, Steve Osazuwa, Brett Hannigan, Naina Thangaraj

Jason Williams and Nirav Merchant

DNAnexus®
The Global Network For Genomics™

Experiment and collaboration

https://github.com/Arkarachai/pag2019_demo_keras_for_genomics

chai@dnanexus.com

DNAnexus®
The Global Network For Genomics™

Besides the caveats, how is the model perform?

- Bayes error

  sum(max(concordant_i,discordant_i))/total_case = 74%

Data

Indel DeepVariant calling

GIAB
Variant truth set