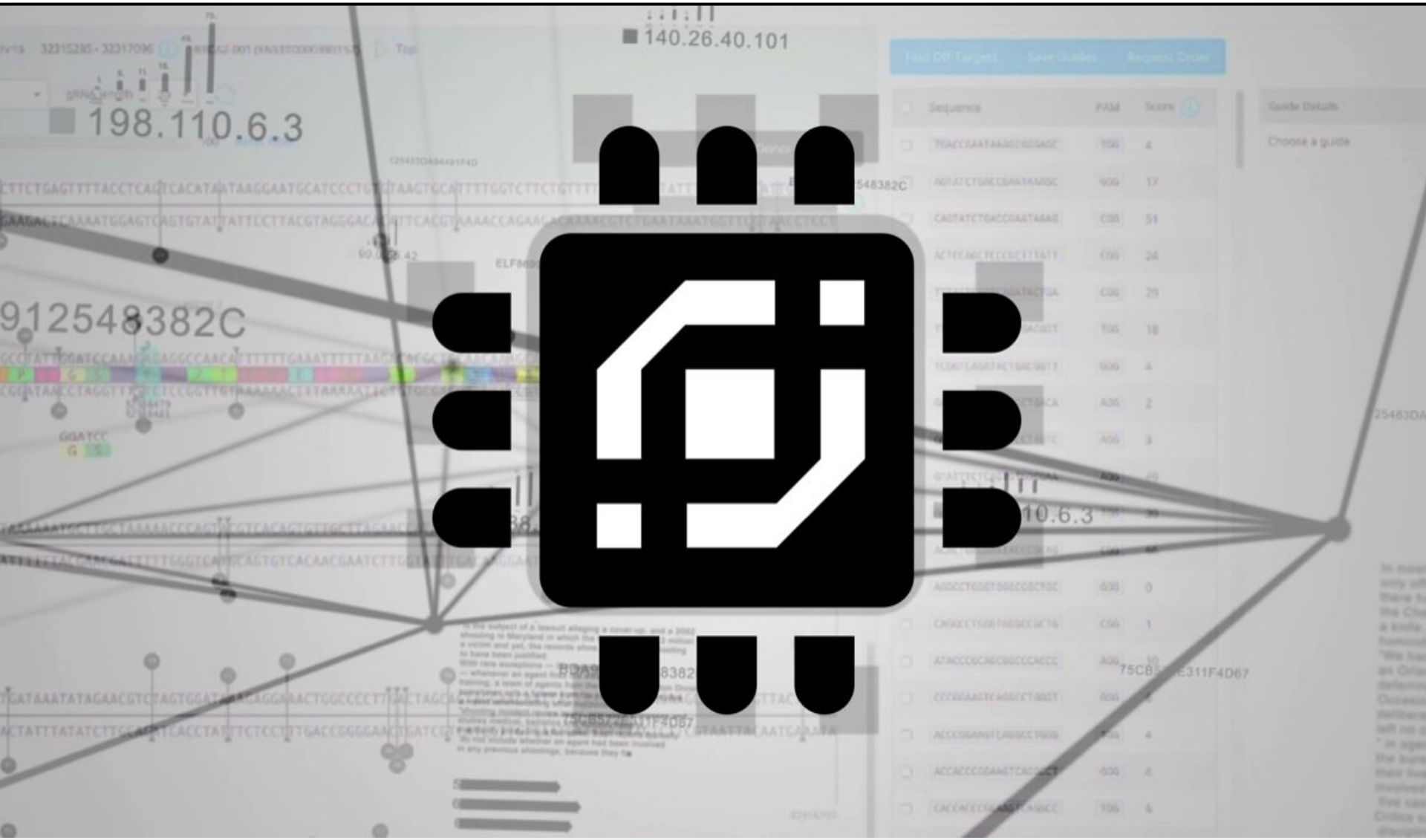
DESKTOP
GENETICS

REPROGRAMMING THE HUMAN GENOME WITH PYTHON

RILEY DOYLE, CEO AND TECHNICAL LEAD
MARK DUNNE, DATA SCIENTIST

AI-POWERED GENOME EDITING

DESKGEN IMPROVES THE SAFETY AND EFFECTIVENESS OF CRISPR



WHO ARE WE?

INTERDISCIPLINARY TEAM BASED IN LONDON

Riley Doyle

CEO & Technical Lead
@doyle.riley
github.com/rodoyle

Background in Biochemical Engineering

Python since 2008

Mark Dunne

Data Scientist

github.com/MarkDunne

Background in Computer Science

Python since 2012

GET A COPY OF THESE SLIDES

SLIDES, FUTURE MEETUPS, CRISPR RESOURCES, JOB OPPORTUNITIES

Send an empty email to

PyCon@deskgen.com

AGENDA

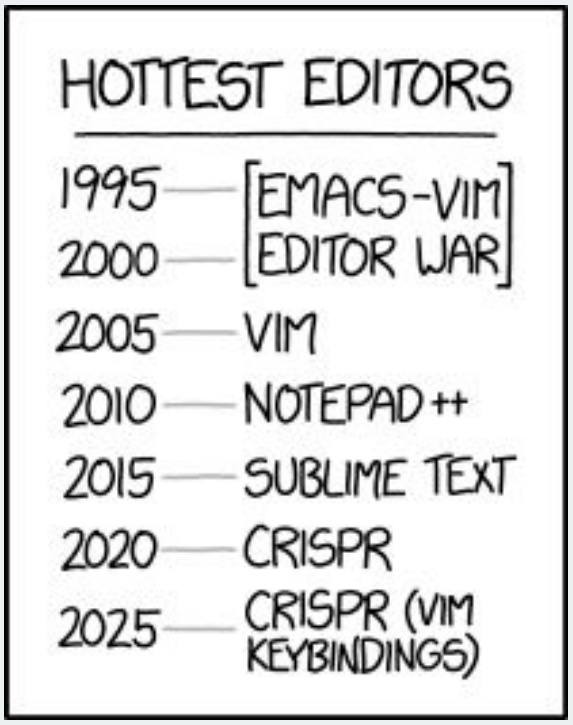
1. Brief intro to CRISPR
2. Applying machine learning to DNA
3. Our CRISPR design process
4. The path forward

DESKTOP
GENETICS

1. BRIEF INTRO TO CRISPR

FROM XKCD ...

HOW DO YOU "PROGRAM BIOLOGY?"



Biology	Python
Cell	Computer
DNA	*.py source files
Genome	All Source files
RNAs	binaries
Proteins	Objects
CRISPR	Sed (ie. s/'ATG'//g+)

BIGGEST BIOTECH BREAKTHROUGH OF THE CENTURY

GLOBAL COVERAGE ACROSS SCIENCE AND TECH MEDIA

GENE EDITING SAVES GIRL DYING FROM LEUKAEMIA IN WORLD FIRST

5 November 2015 NewScientist

HIV GENES HAVE BEEN CUT OUT OF LIVE ANIMALS USING CRISPR

15 May 2016 TIME

CHINA USED CRISPR TO FIGHT CANCER IN A REAL, LIVE HUMAN

18 November 2016 WIRED

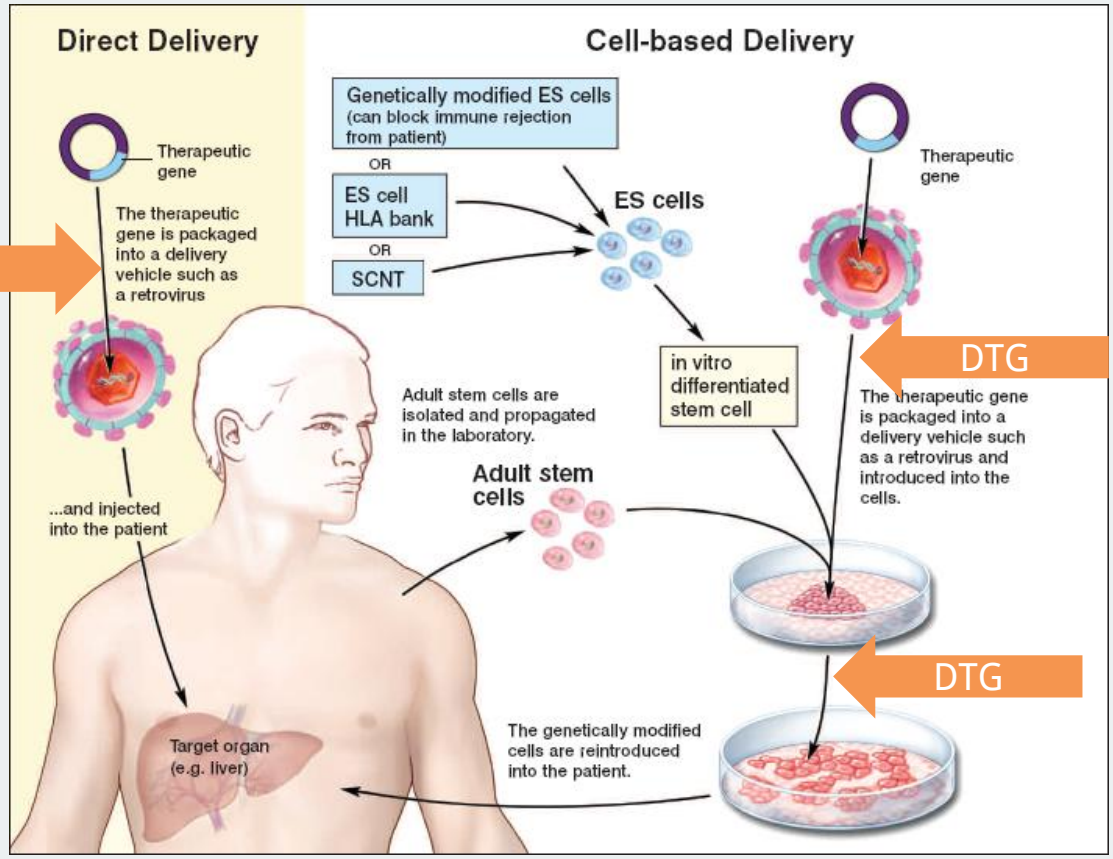
CRISPR: GENE EDITING IS JUST THE BEGINNING

07 March 2016 NATURE

CELL & GENE THERAPY TACKLES DISEASES

CRISPR IS USED TO TREAT PATIENTS AND DISCOVER CURES

DTG

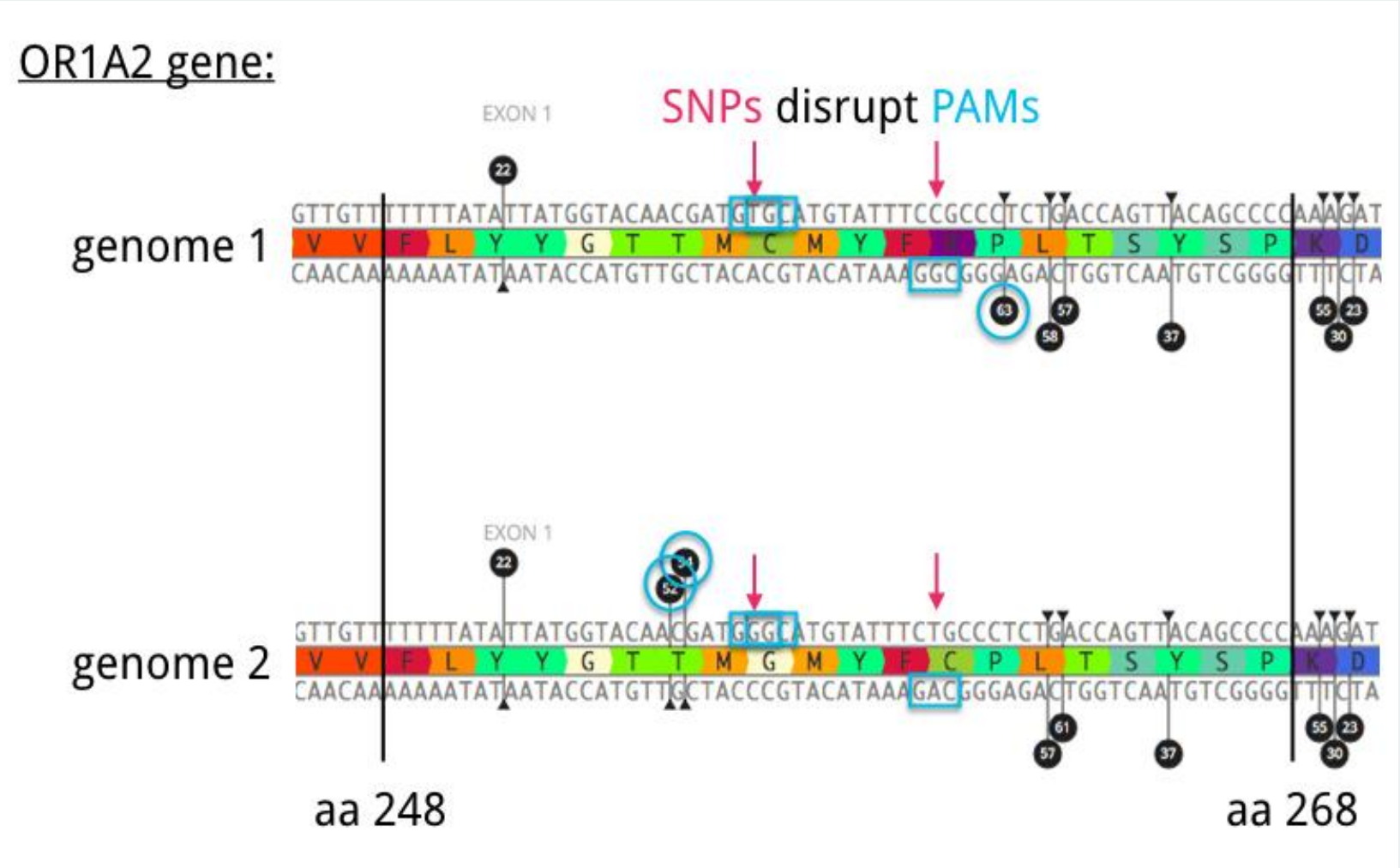


DTG

DTG

REAL GENOMES HAVE MUTATIONS

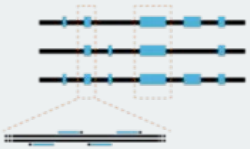
EVERYONE HAS 4 to 5 MILLION VARIANTS IN THEIR GENOME



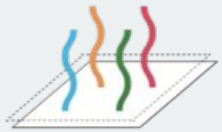
GENOME EDITING PROCESS

AI REQUIRED TO AUTOMATE DECISION MAKING THROUGHOUT THE PROCESS

CHARACTERIZE



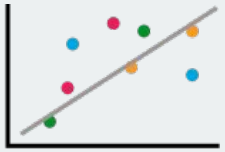
DESIGN



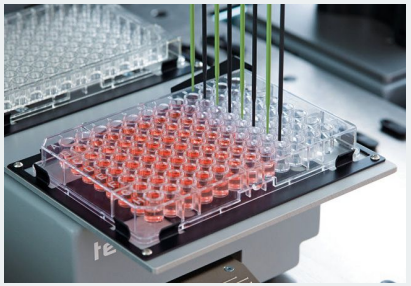
MANUFACTURE

SCREEN

SEQUENCE

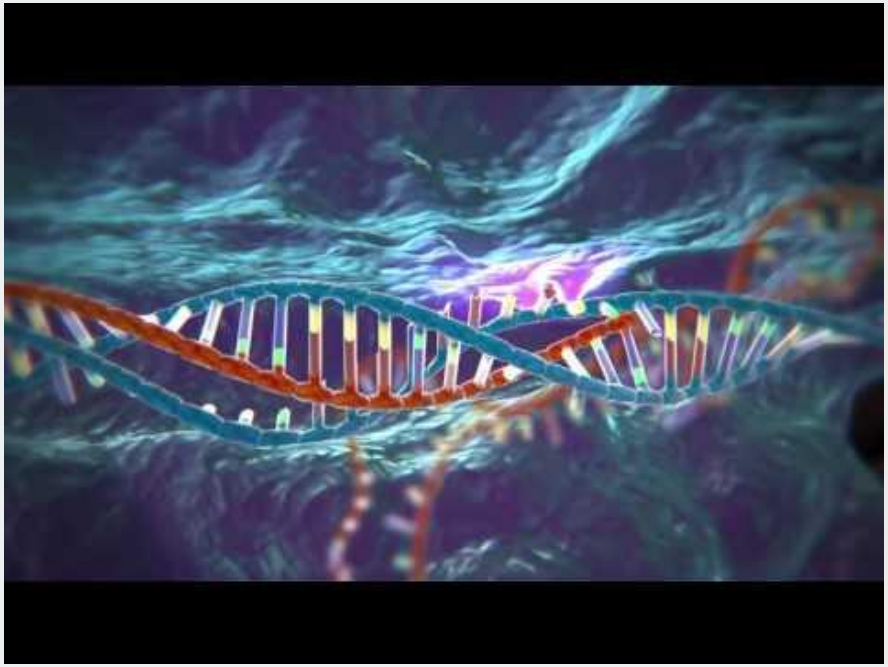


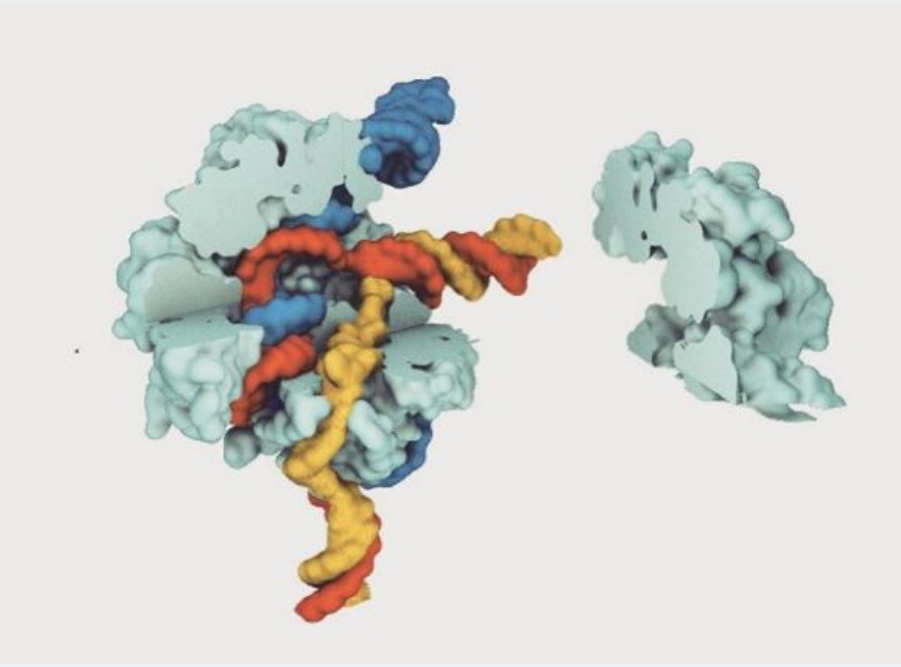
ANALYZE



CRISPR AT A GLANCE

MOLECULAR INTERACTIONS AND MODELS





Email PyCon@deskgen.com
for video and 3D molecule

CRISPR OVERVIEW

PROGRAMMABLE TWO COMPONENT SYSTEM

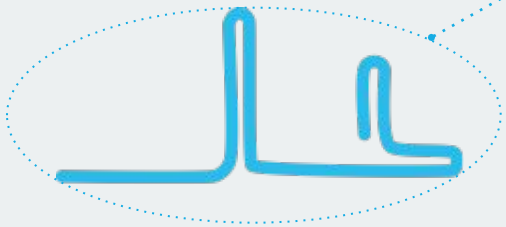
CAS9
NUCLEASE

RNA
COMPONENT

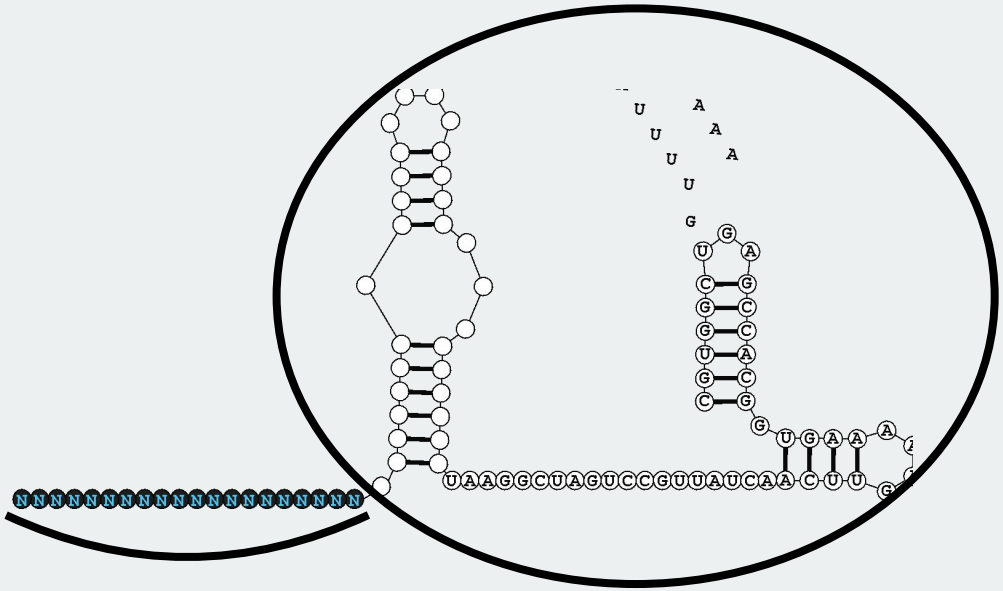
CRISPR OVERVIEW

PROGRAMMABLE TWO COMPONENT SYSTEM

CAS9
NUCLEASE



RNA
COMPONENT



VARIABLE
20 BP GUIDE RNA

(**sgRNA**)

CONSTANT
REGION

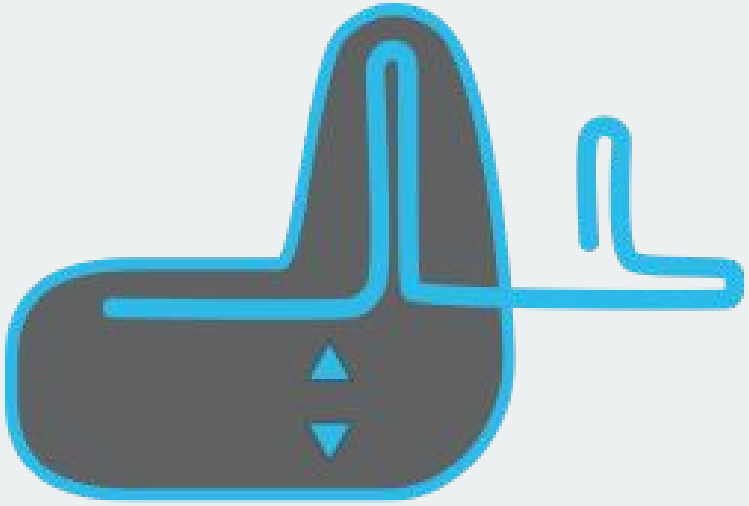
(**tracrRNA**)

CRISPR OVERVIEW

PROGRAMMABLE TWO COMPONENT SYSTEM

CAS9
NUCLEASE

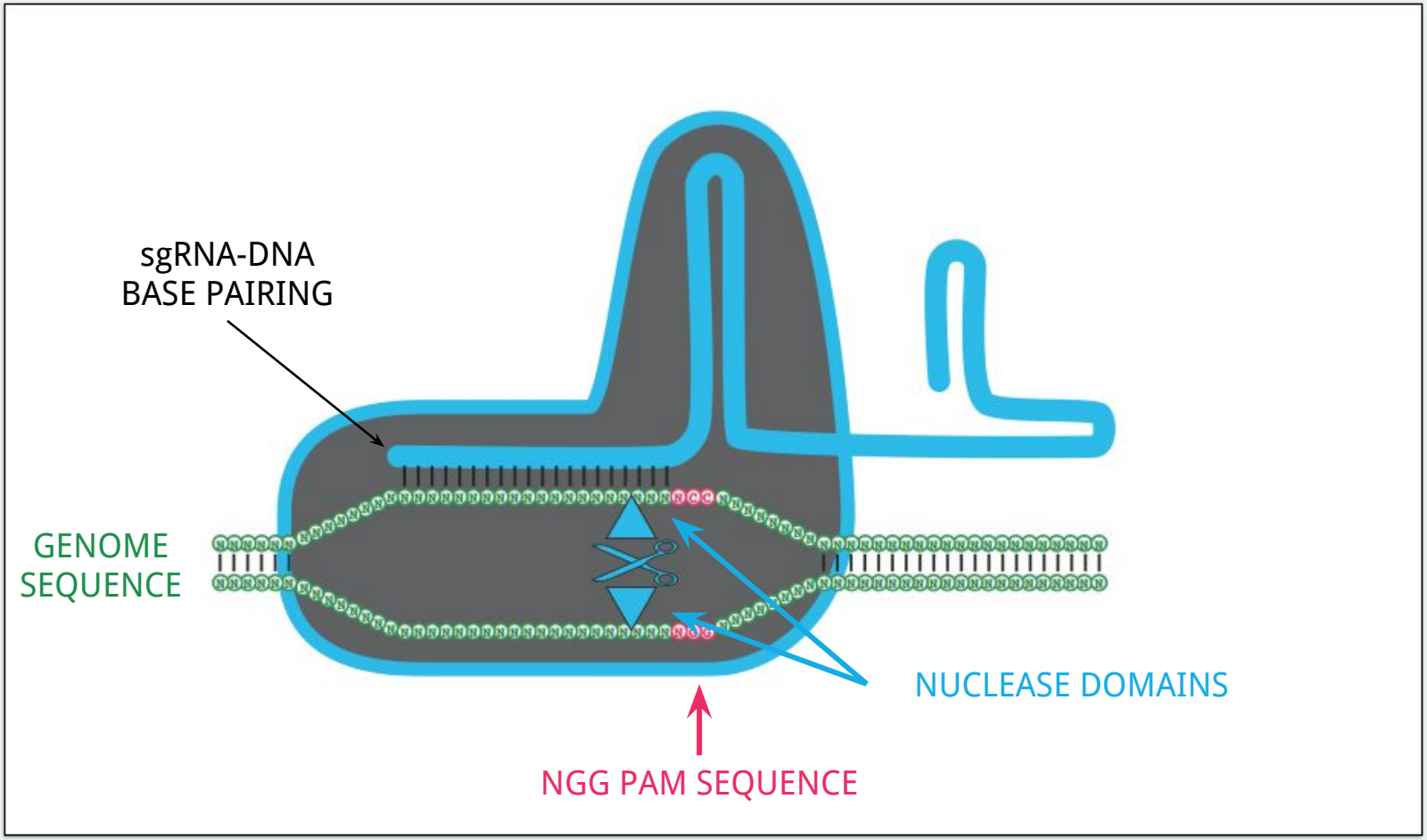
RNA
COMPONENT



ACTIVE
RNA-GUIDED CAS9
COMPLEX

CRISPR OVERVIEW

CUT + REPAIR = GENOME EDITING

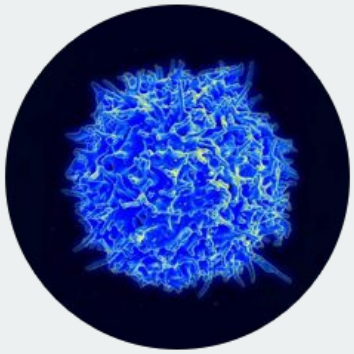


WHY EDIT GENOMES?

RESEARCH AND DEVELOPMENT → CLINICAL CURES

- Degenerative blindness
- Custom cancer models

- Humanization of heart valves
- Swine fever resistance



- HIV eradicated *in vitro*
- Immuno-oncology

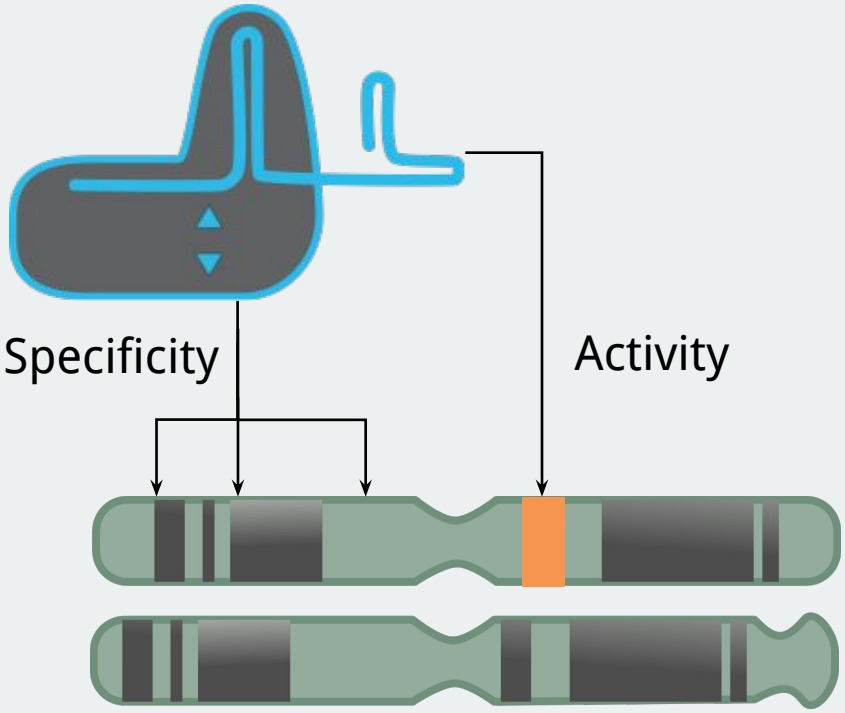
- Clinical trials cured cancer
- Clinical trials cured HIV

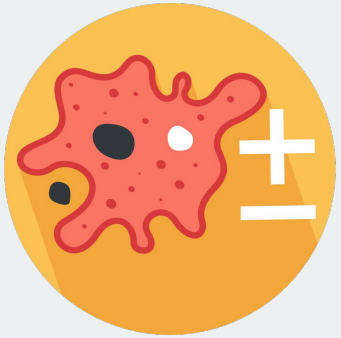
DESKTOP
GENETICS

2. APPLYING MACHINE LEARNING TO DNA

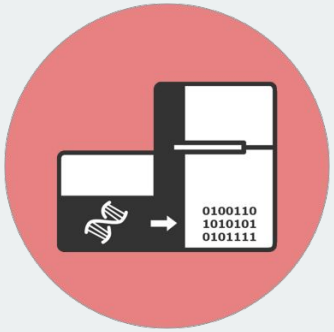
CRISPR HAS SEVERAL COMPUTATIONAL PROBLEMS

WHAT ARE WE ACTUALLY TRYING TO PREDICT ANYWAY?

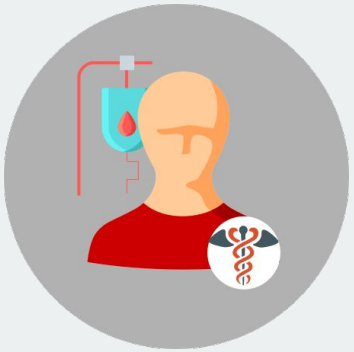




Biological Importance



Instrument Signal



Patient Outcome

RECURRING CRISPR PROBLEMS

USER ANALYTICS REVEALED COMMON PROBLEMS

	HUMAN	MACHINE
Guide selection	Get tired of choosing many guides for each gene	Considers all guides, choses consistently
Scoring function(s)	Undue weight given to some scoring functions	Weights of features carefully controlled
Genotype data	Considers only reference genome	Considers actual genome sequence
Overall objective	Few “winning” guides	Balanced, orthogonal training set

SELECTION OF BIOCHEMISTRY BASED FEATURES

SEVERAL MACRO & CONTEXTUAL FEATURES IDENTIFIED FROM BIOCHEMISTRY LITERATURE

DESIGN RULE	TYPE	RANGE	CONSIDERS	RESULT
NAG PAM (Control)	Negative	{0,1}	(PAM) Sequence	✔
GC%	Negative	[0,1]	Sequence	✔
Homopolymer (N4)	Negative	{0,1}	Sequence	✔
SNP Collision	Negative	{0,1}	Location	✔
UUU Triplet	Negative	{0,1}	Sequence	✔
Non-constitutive Transcript	Negative	{0,1}	Location	✔
1st third CDS	Positive	{0,1}	Location	✖
Functional domain	Positive	{0,1}	Location	✔
Truncated guide	Positive	{0,1}	Sequence	✖
Microhomology	Positive	[0,1]	Sequence	✖
Specificity (Hsu, 2013)	Negative	[0,1]	Sequence	?

GUIDE RNA SEQUENCE “WORDS”

SEQUENCES EMBEDDED INTO VECTOR SPACE USING ONE-HOT ENCODING OF K-MER@POSITION

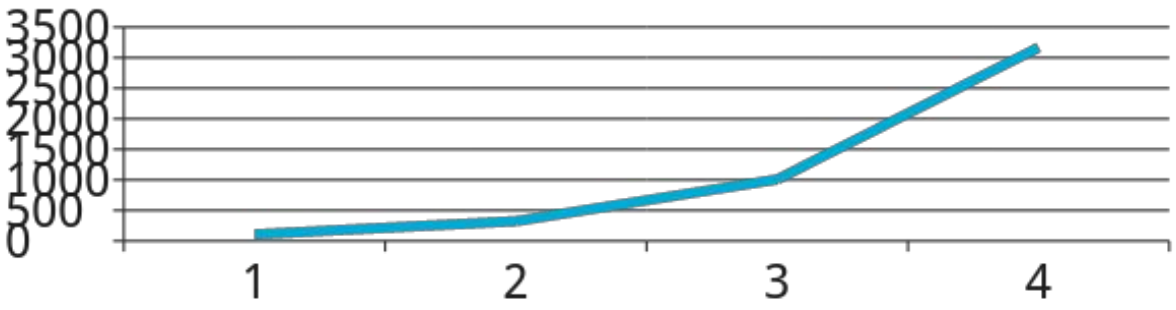
Number of **non-overlapping, position-dependent** sequence features is:

where *k* = feature size (nt) and *n* is length of sequence

$$\sum_{k=1}^{n} 4^k \left(\frac{n}{k} \right)$$

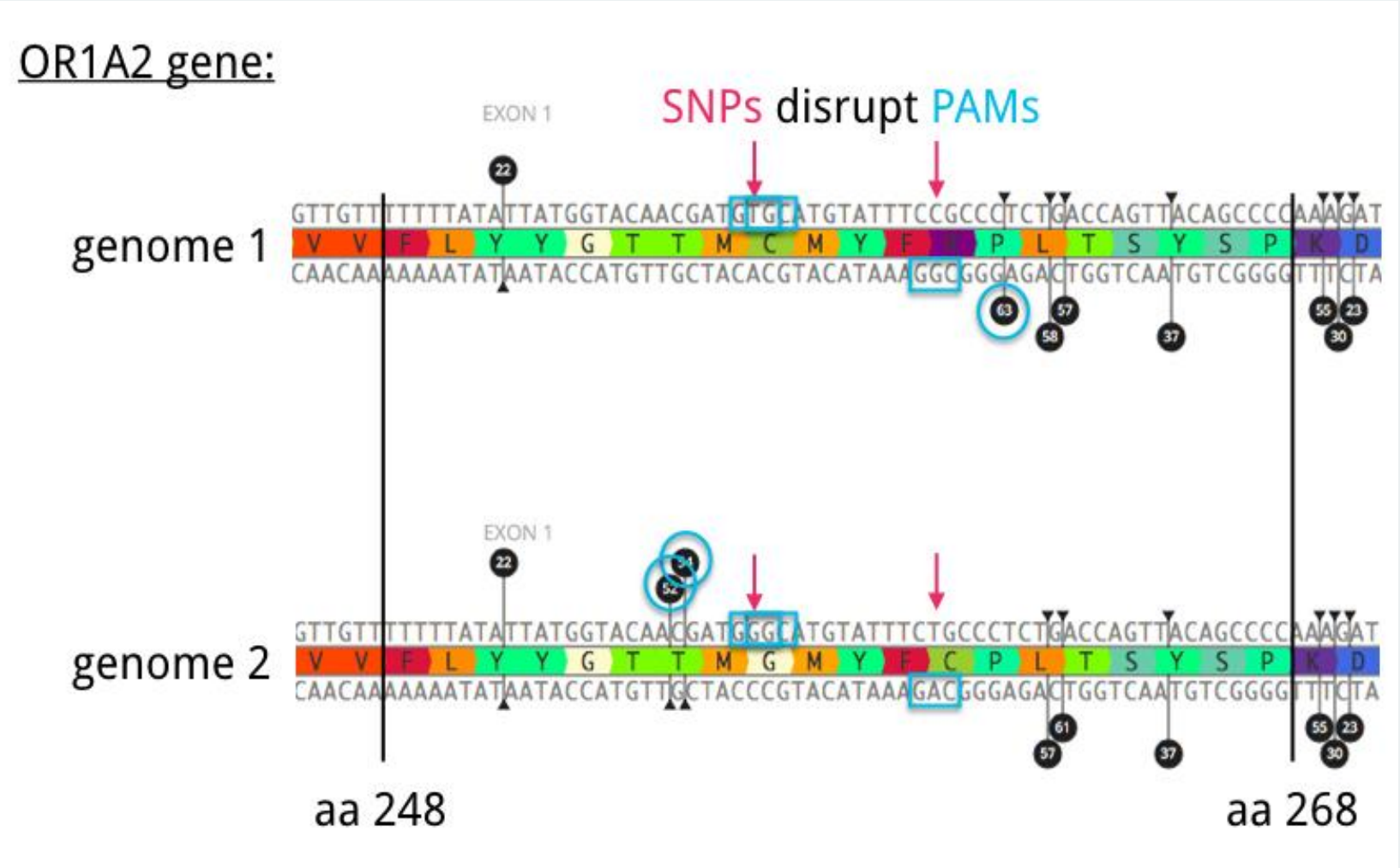
4 States: A → [1000], C → [0100], G → [0010], T → [0001]
at each position in n; repeat for all k-mers.

- We used k [1,3] for ~4700 features total
- Resulting embedding is very sparse.
- Too many dimensions + insufficient data = **over fitting**



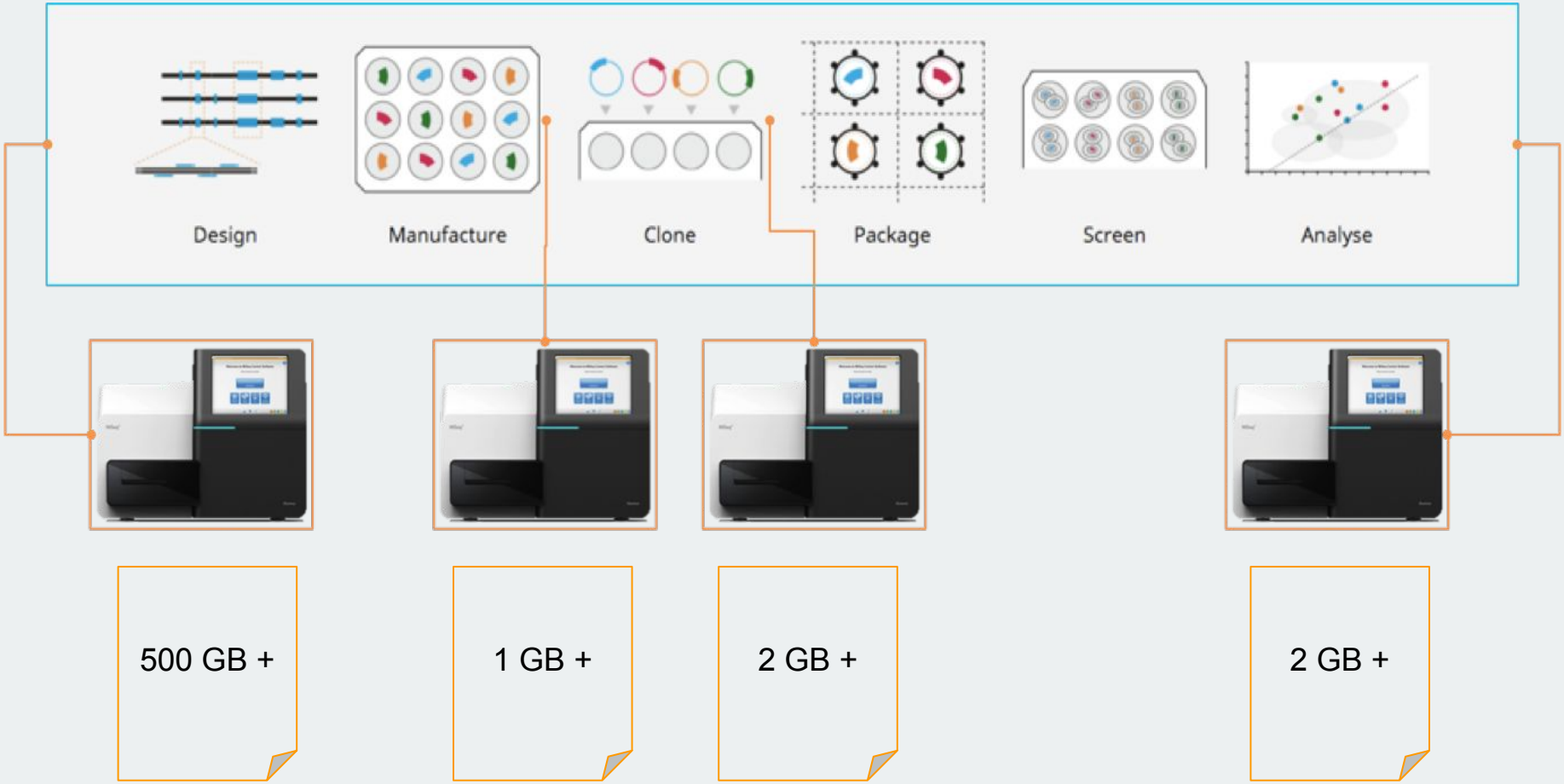
REAL GENOMES HAVE MUTATIONS

INDIVIDUAL GENOME VARIANTS CAN GENERATE NOISE



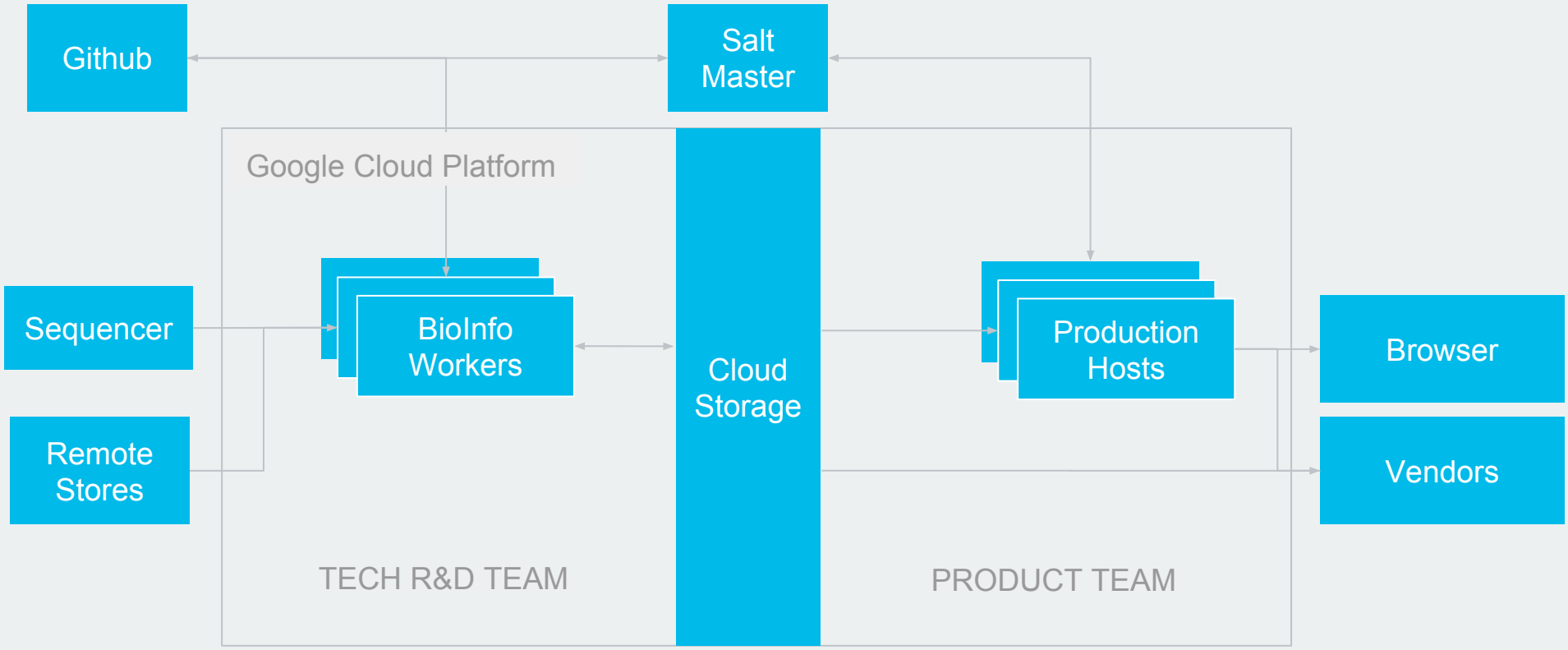
GENOME SEQUENCING IS DATA INTENSIVE

OUR SYSTEM NEEDS TO HANDLE LARGE VOLUMES OF DATA



DESKGEN INFRASTRUCTURE

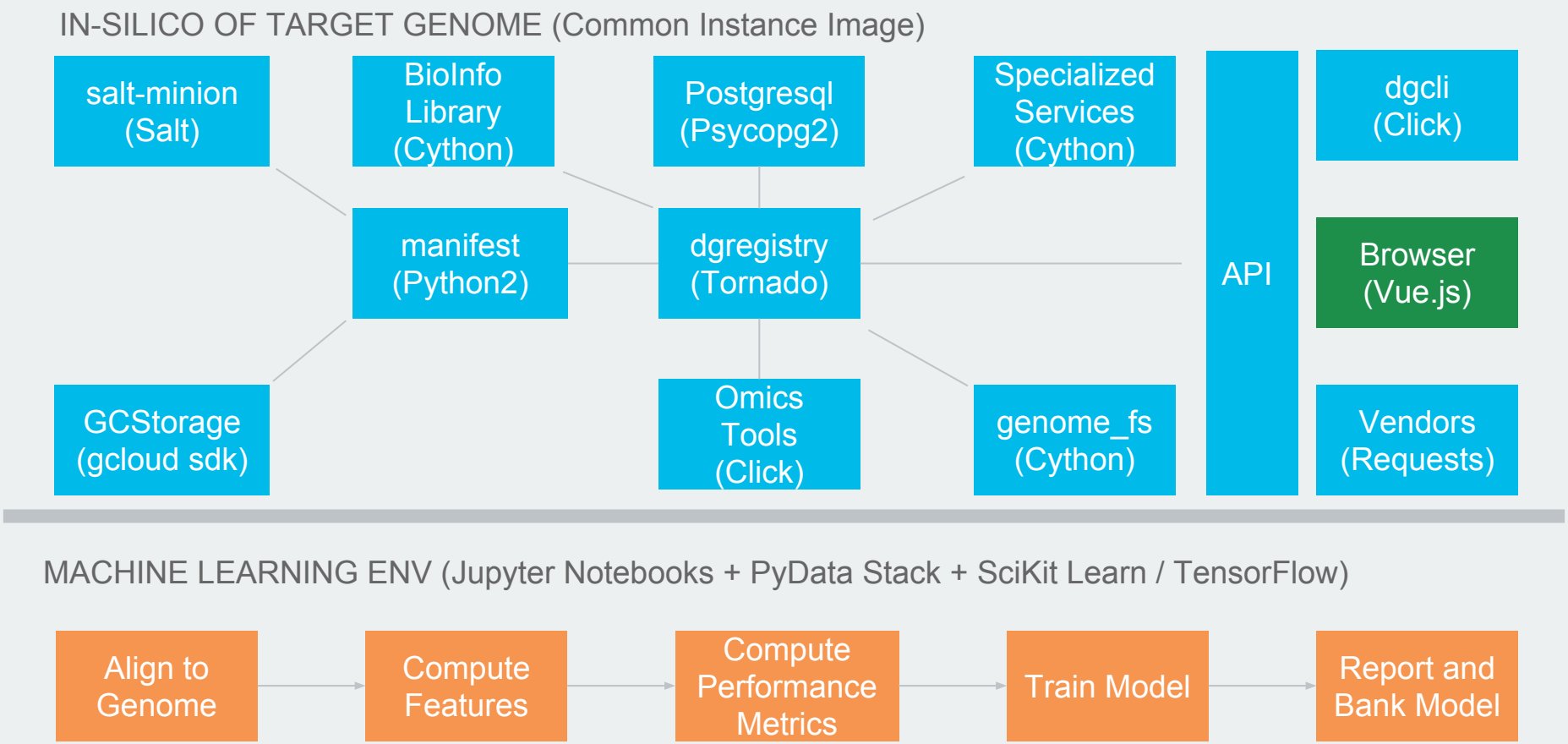
HANDLING GENOME DATA AT SCALE



SaltStack Control Layer orchestrates instance groups in both development and production environments.

DESKGEN HOST LEVEL ARCHITECTURE

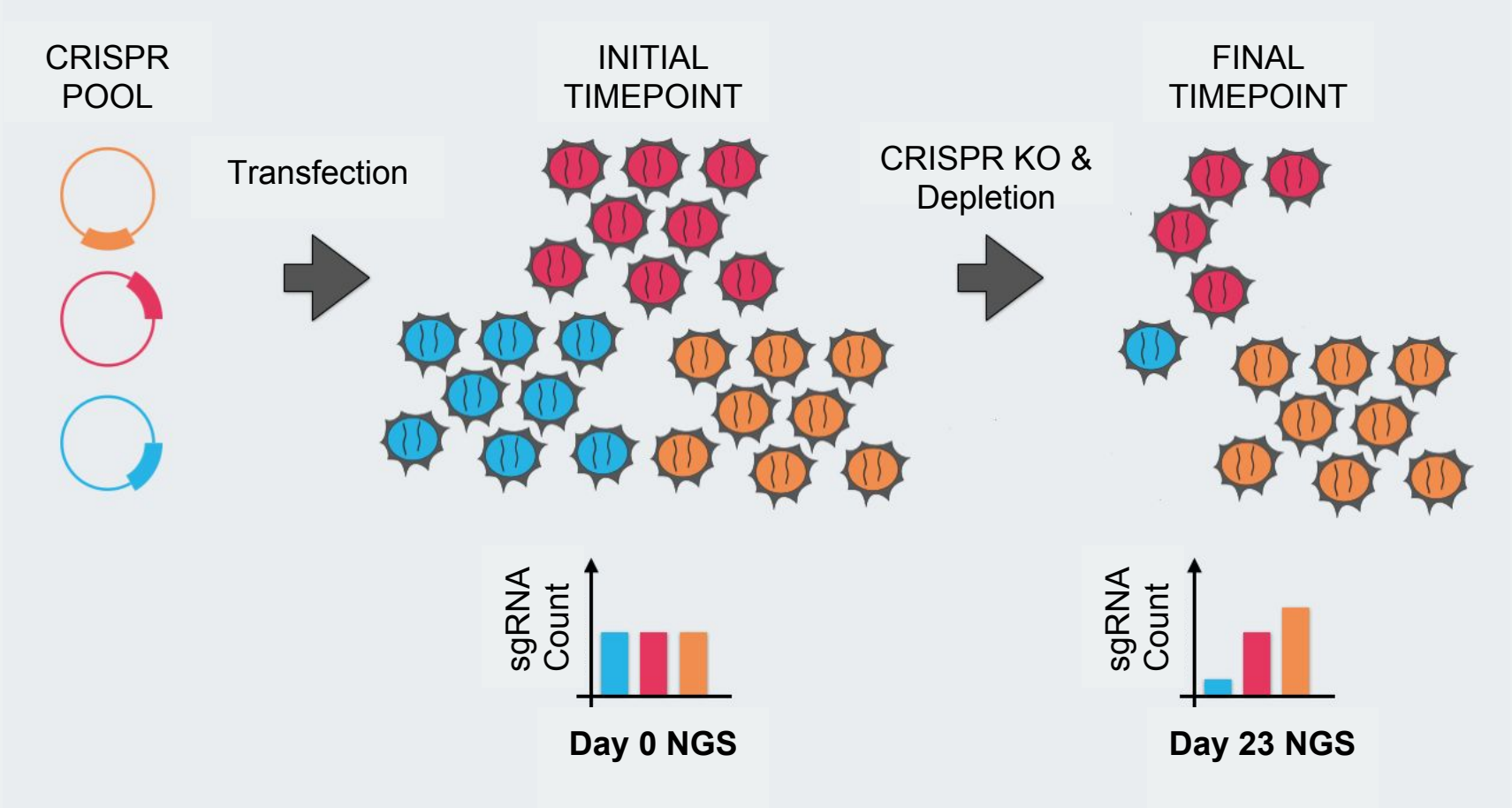
GENOME CONTEXT MADE AVAILABLE ACROSS STEPS OF ML PIPELINE



ML PIPELINE either imports Python code directly or uses CLI commands.

MEASURING GUIDE PERFORMANCE

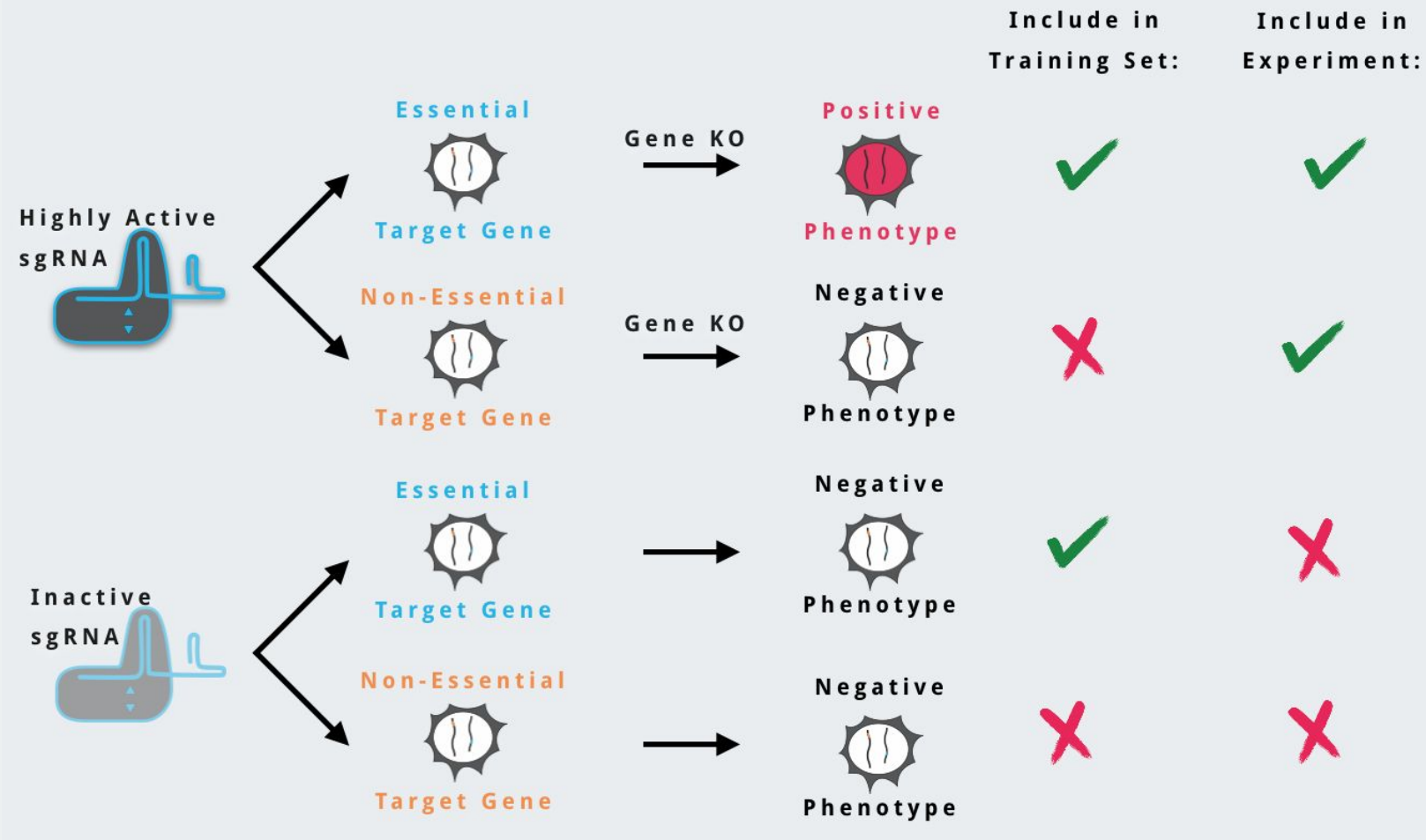
EVOLUTION SAYS GUIDES ACTIVE AGAINST ESSENTIAL GENES SHOULD KILL CELLS



GUIDE SCORING

NON-ESSENTIAL GENE TARGETS RESULT IN UNDETECTABLE GUIDES

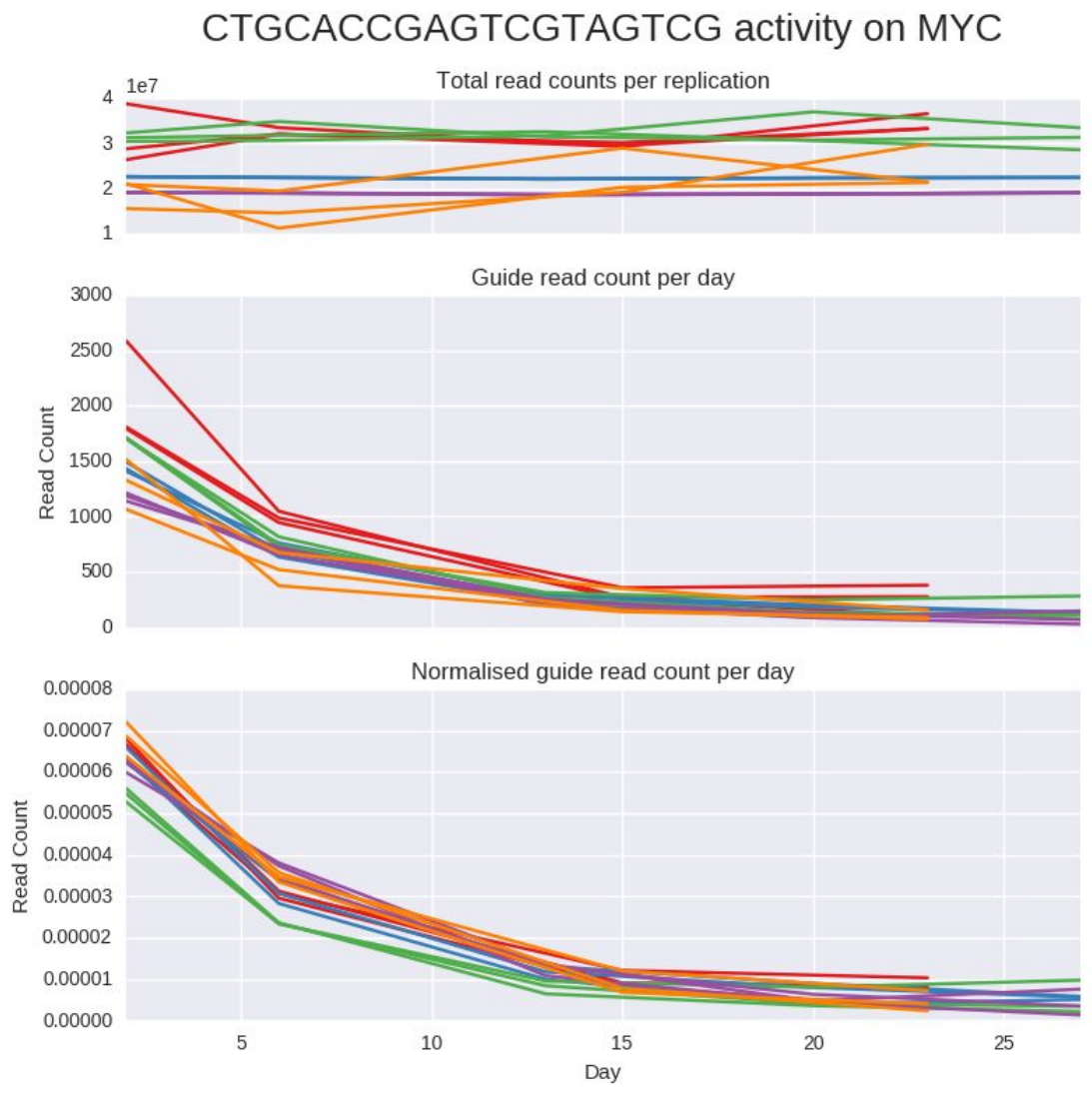
- Remove non-essential genes from analysis as sgRNA activity cannot be detected.



VARIANCE OF THE SAME GUIDE

AN ACTIVE GUIDE

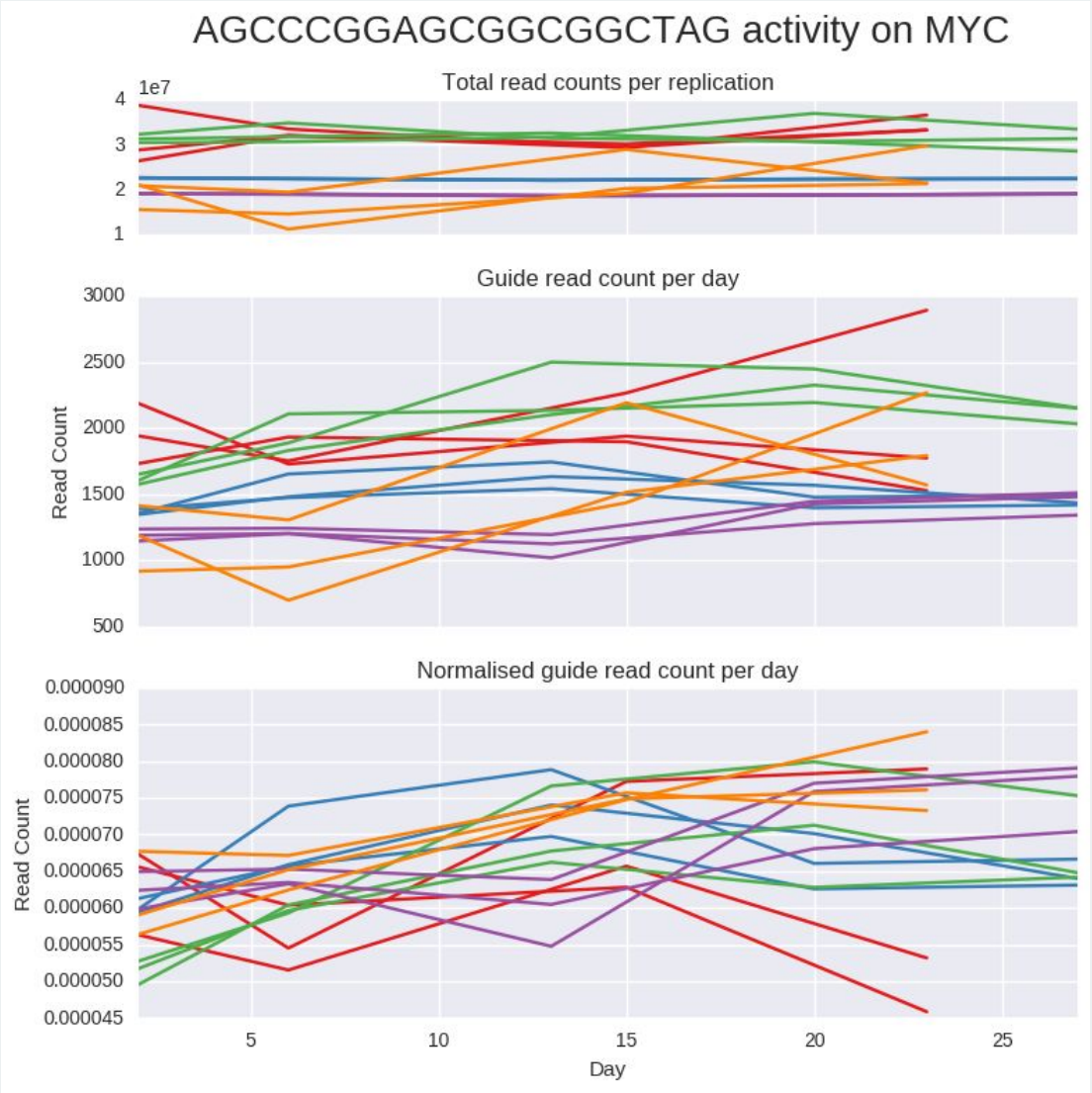
In active guides, there is little variance between biological replicates, and different experiments.



VARIANCE OF THE SAME GUIDE

AN INACTIVE GUIDE

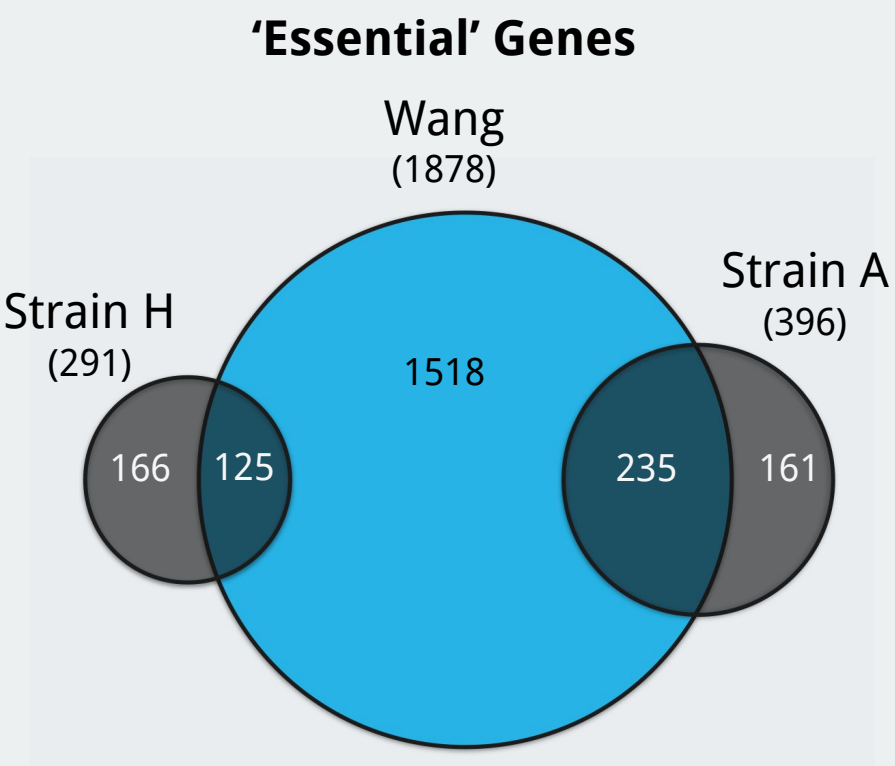
In inactive guides - there is large variance between biological replicates, and different experiments

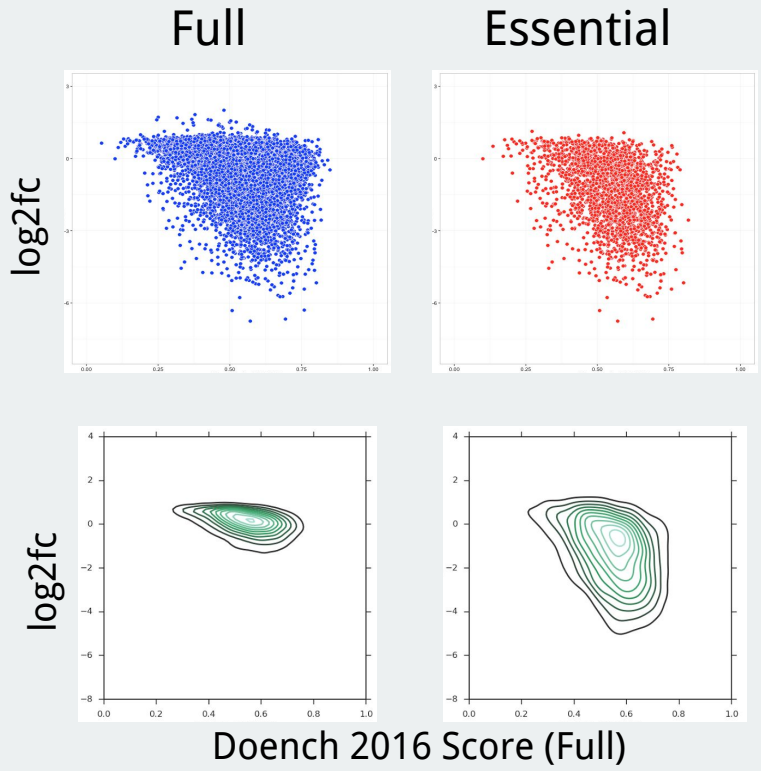


GUIDE SCORING

REMOVING NON-ESSENTIAL GENES INCREASES ROBUSTNESS OF GUIDE ACTIVITY DETECTION

Wang et al. (2015): Conducted CRISPR screen in the near-haploid human KBM7 chronic myelogenous leukemia (CML) cell line and confirmed essentiality using gene-trap.





Sabatini data: Wang et al. Science. 2015 Nov 27;350(6264):1096-101

DATA ANALYSIS PIPELINE

POST-PROCESSING AND NORMALIZATION CRITICAL TO MODEL

1. Normalization
   1.1. Normalized so that read count across columns was consistent per experiment

2. Selection
   2.1. Removed rows where there was a read count < 30
   2.2. Removed rows where gene was 'NA' or null
   2.3. Removed guides targeting non-coding regions
   2.4. Selected guides targeting essential genes using MAGeCK
      2.4.1. Human: 6509 guides (5.61% of dataset)
      2.4.2. Mouse: 8006 guides (5.58% of dataset)

3. Scoring derived from first-order kinetic rate law

$$GuideActivity = \frac{-log_2 \frac{1}{n} \sum_{i=0}^{n} \frac{count_2}{count_1}}{\delta t}$$

n = number of replicants - 3 in this case
δt = experiment run time - as close to 20 as possible

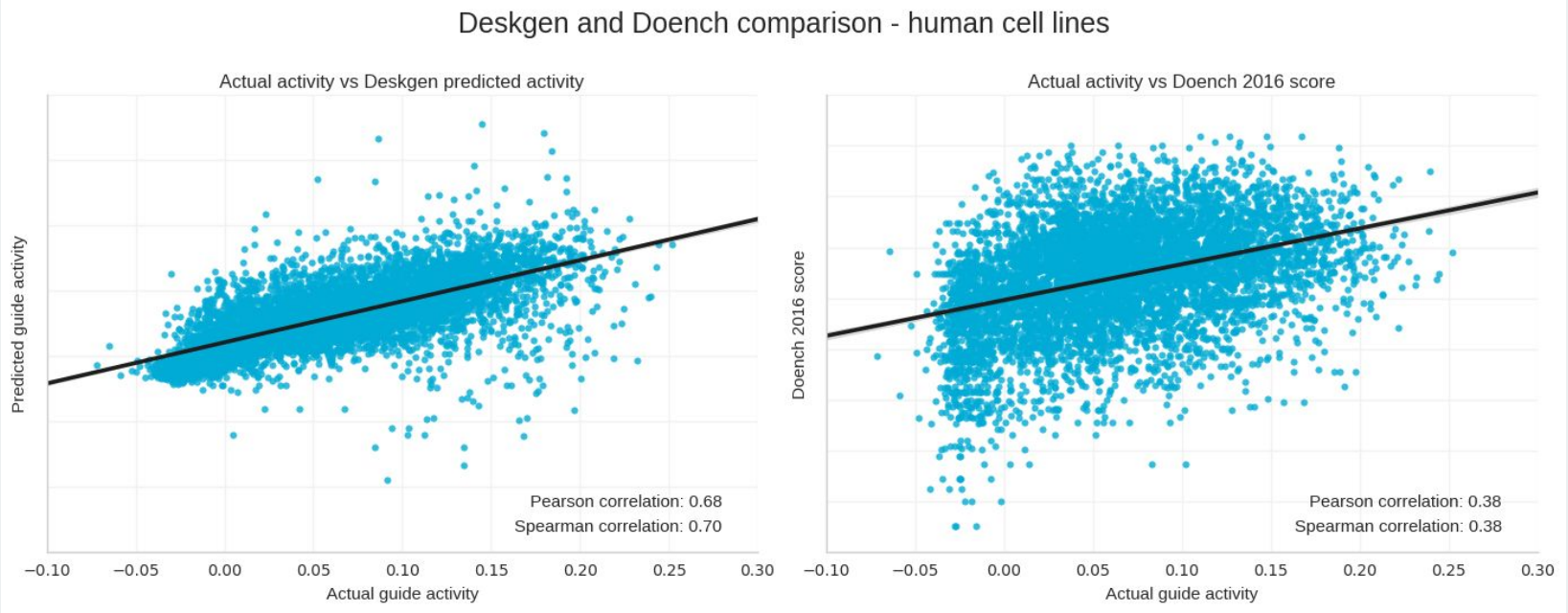
DESKTOP
GENETICS

3. OUR CRISPR DESIGN PROCESS

LINEAR MODEL PERFORMED SURPRISINGLY WELL

BOTH PEARSON AND SPEARMAN METRICS IMPROVED

Comparison of performance between DTG and Doench 2016 models

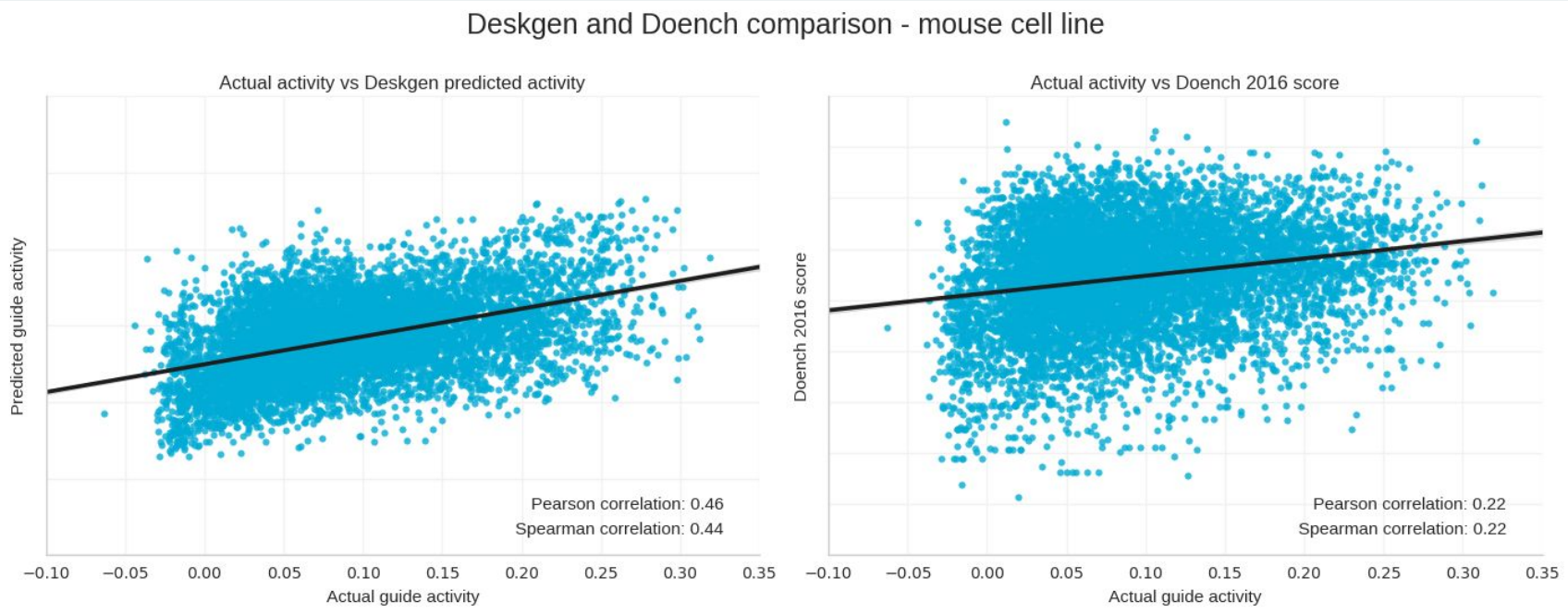


- Executing this algorithm found DTG's model is an 84% improvement over state of the art (Doench 2016)
- Generalized Linear Model performed as well as ConvNet and RandomForest

MODEL DOES NOT GENERALIZE ACROSS SPECIES

MOUSE PERFORMANCE ALSO IMPROVED BUT IS NOT AS GOOD AS HUMAN MODEL

Comparison of performance between DTG and Doench models



- Executing this algorithm found DTG's model is an 100% improvement over Doench 2016
- No literature list of essential genes available for Mouse
- Still unclear why performance is different

MODEL COEFFICIENTS CONFIRM POSITION-DEPENDENT SEQUENCE EFFECT

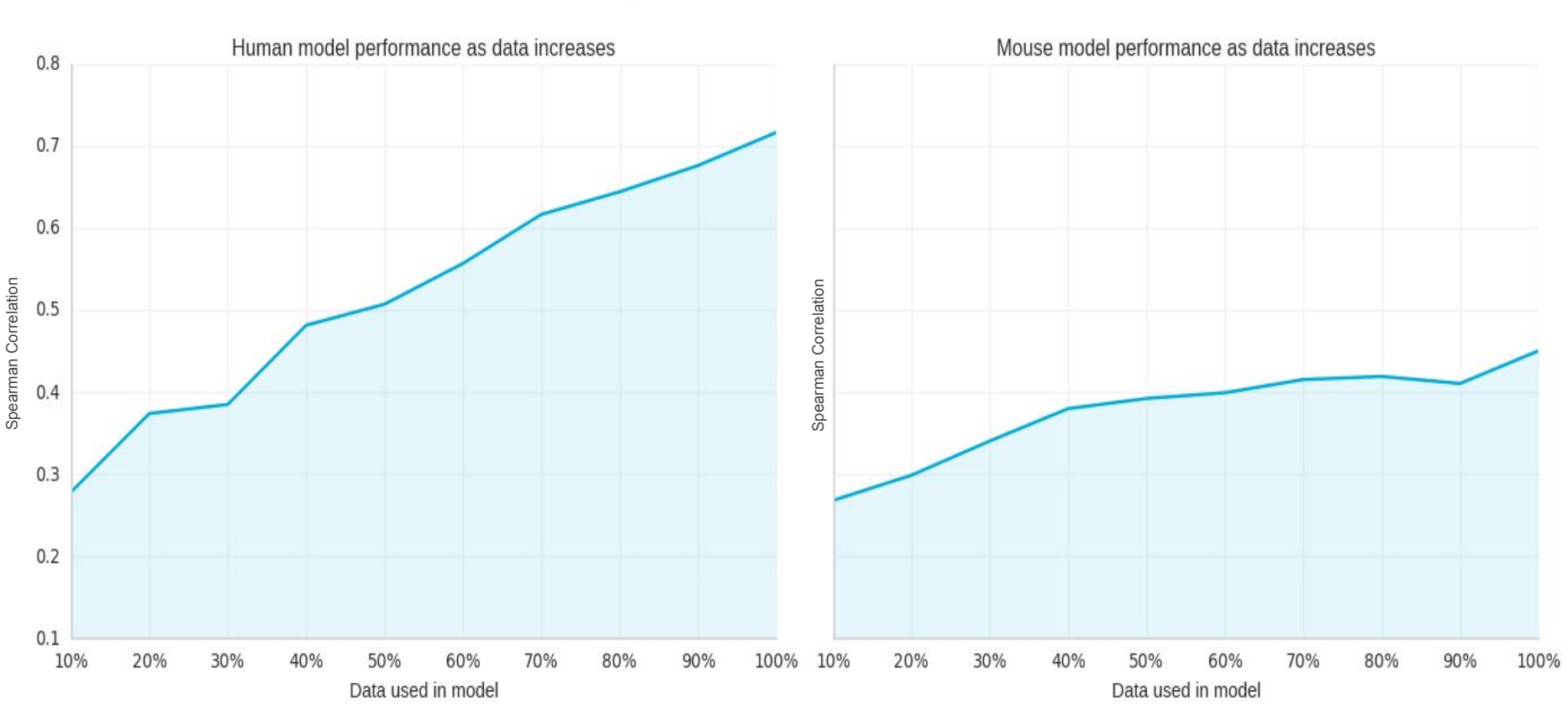
PRIOR WORK EXTENDED INTO NEW TRAINING DATA

- We examined the coefficients of the ridge regression model
- We determined the importance of single bases varies a lot of the range of the flank

MARGINAL BENEFIT OF ADDITIONAL DATA

HUMAN AND MOUSE MODELS BOTH IMPROVE AS FURTHER WET LAB DATA ADDED

- Relationship between model performance and data used = more data will help build a better model



DESKTOP
GENETICS

4. THE PATH FORWARD

CONCLUSIONS

SIGNIFICANTLY MORE ACCURATE GUIDE ACTIVITY PREDICTIONS WERE POSSIBLE

1. De-noising and normalization of the training data and feature engineering resulted in a linear model which outperformed more complex models.
2. Linear model currently predicts guide performance up to current noise level seen experimentally.
3. Model generalized across cell lines but not across species. We are currently unsure why.
4. Prior knowledge about essential genes and target genome significantly improved the model (ie. human genome better curated than mouse).
5. Model performance increased linearly with more training data, but less rapidly for mouse than human.

LESSONS LEARNED

ETL PIPELINE, FEATURES, AND DATA PROCESSING WERE CRITICAL TO SUCCESS

1. Task queues (Celery), microservices, containers (Docker, Kubernetes), and Postgresql **significantly increased dev-ops burden**, dependencies, code maintenance requirements, and learning curve **without increasing developer productivity**. **Pure python code nearly always ended up getting used more.**

2. Scikit Learn Model serialization (cPickle) is **not portable** as ABI breaks between minor and patch versions. Significant source of errors in production. **Acute need for better way to serialize more complex models.**

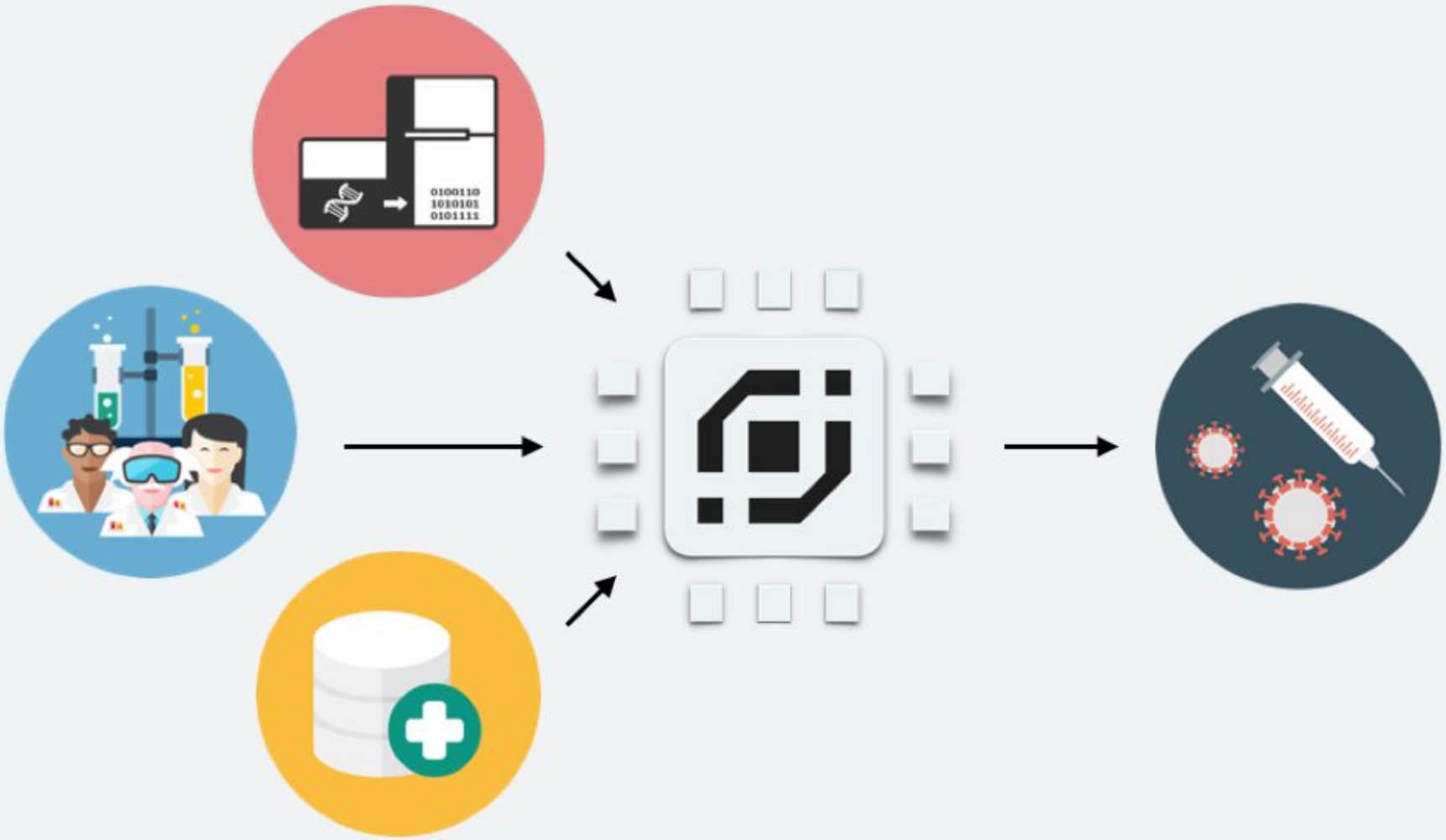
3. Docker Containers did not provide a “silver bullet” replacement for Python packaging, dependency management, or model portability. Instead they introduced significant learning curve as most bioinformatics tools expect direct access to a shared filesystem.

4. Data Science and BioInformatics team strongly preferred working with Conda environment vs. PyEnv + VirutualEnv.

5. Google Cloud Storage critical to working with large genomic data sets.

TAKING CRISPR AI TO THE CLINIC

EXTENDING APPROACH TO IMPROVE GENOME EDITING SAFETY AND EFFICACY



Further Resources

WHERE TO LEARN MORE

1. What can I edit? https://www.omim.org
2. “The” genomics library? https://github.com/samtools/htslib
3. Working with htslib in Python

   https://github.com/pysam-developers/pysam
4. Where to get genome data?
   a. Curated data: http://www.ensembl.org/
   b. Raw data: https://www.ncbi.nlm.nih.gov/sra
   c. Actual people’s genomes: http://personalgenomes.org
5. No lab, no problem!
   a. Transcriptic Client: https://github.com/transcriptic/transcriptic
   b. Antha: https://github.com/antha-lang/antha

GETTING INVOLVED WITH CRISPR

OPTIMISE AND IMPROVE

1. Dataset available on GitHub – try it yourself

   https://github.com/DeskGen/guide-cluster

2. Larger dataset with API coming 2017

   https://github.com/DeskGen/dgcli

3. Hiring full time at Desktop Genetics

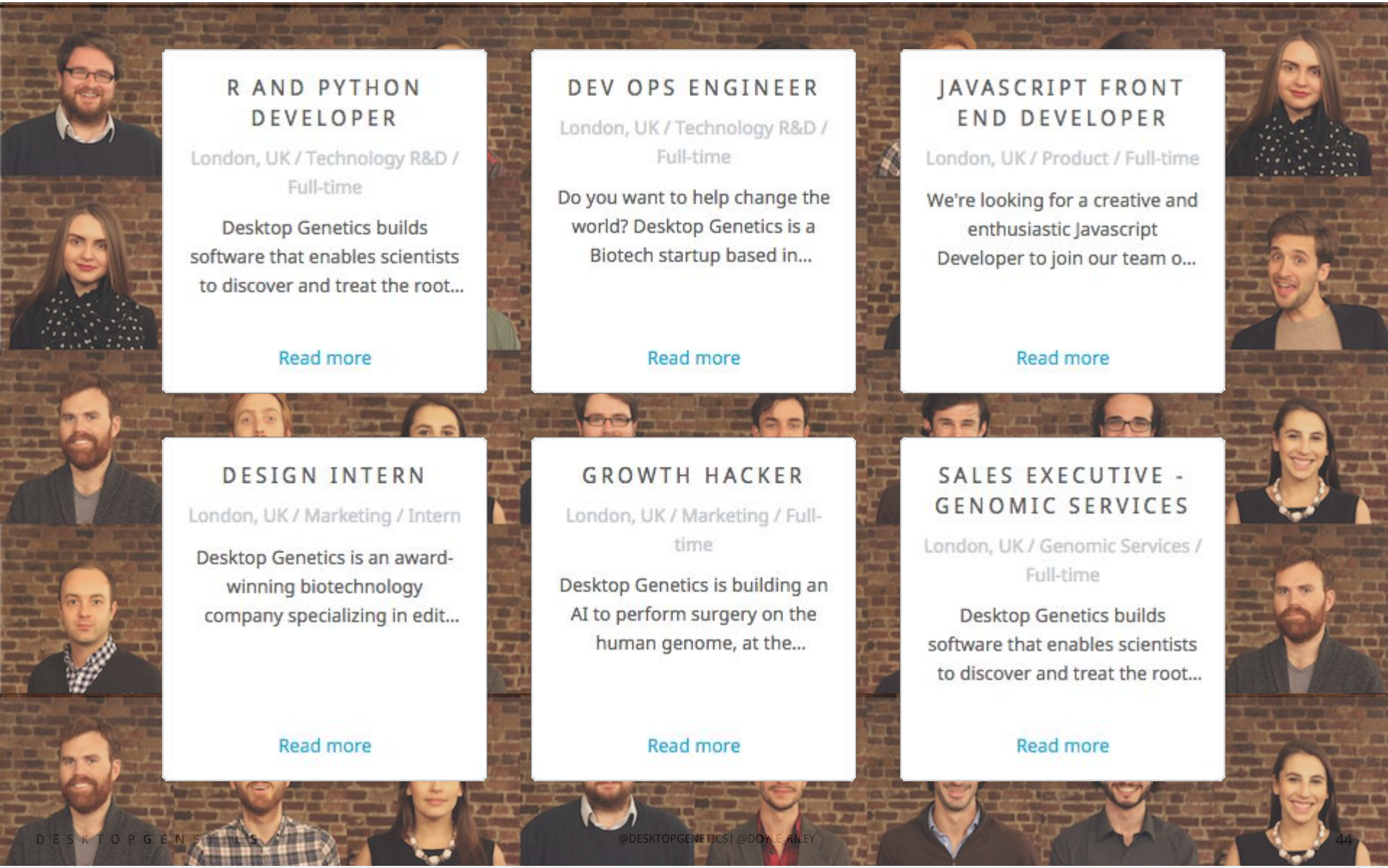
   https://www.deskgen.com/landing/company#about-careers

4. More detailed blog post

   https://www.deskgen.com/landing/blog/machine-learning-crispr-guide-design

JOBS AT DESKTOP GENETICS HQ

JOIN US IN LONDON - TELL YOUR FRIENDS!



RECOGNITION

TECH, BIOTECH AND EVERYTHING IN BETWEEN

genomeweb

Forbes

K'50

Co.DESIGN

BBC

LABIOTECH.eu
The European Biotech News Website

TEDx

GET EVERYTHING YOU JUST HEARD AND MORE

SLIDES, FUTURE MEETUPS, CRISPR RESOURCES, JOB OPPORTUNITIES

Send an empty email to

PyCon@deskgen.com

DESKTOP

GENETICS

pycon@deskgen.com