Biological Applications of Deep Learning
Lecture 8

Alexander Schönhuth

Bielefeld University
November 30, 2022

Contents Today

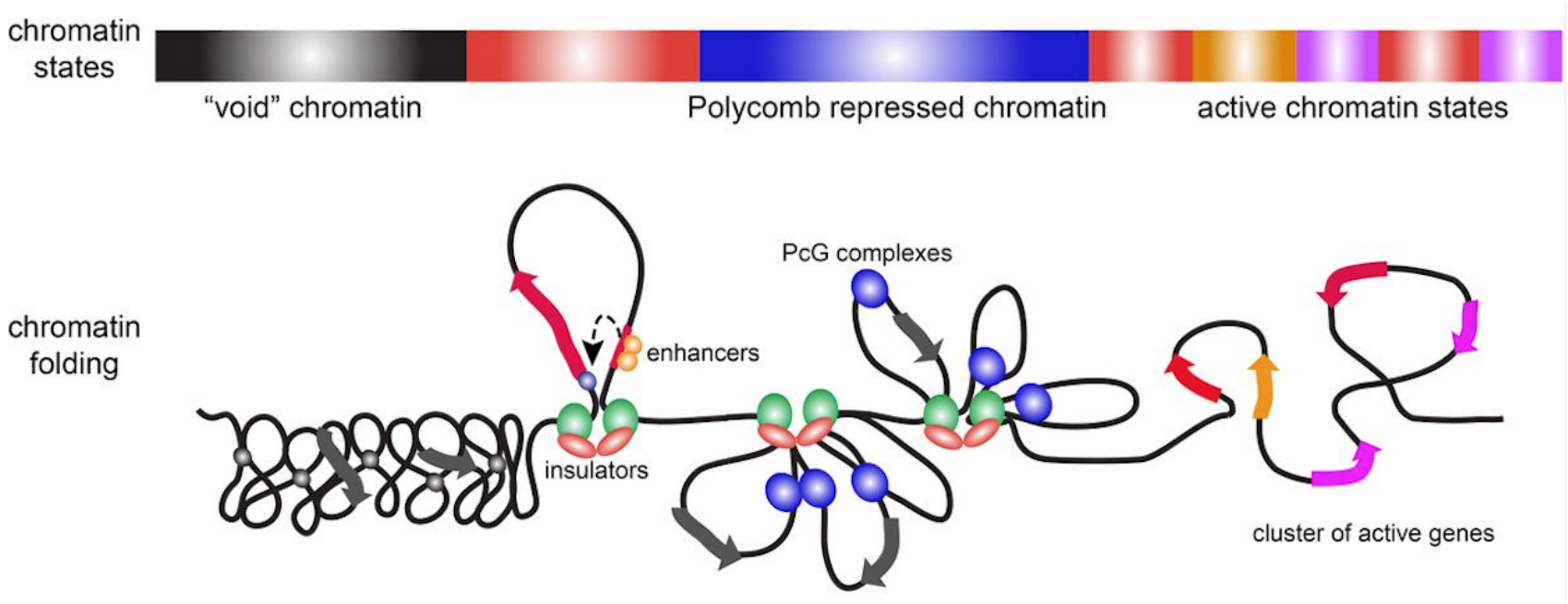
- Hilbert CNN
  - Predict epigenetic state of sequences
  - Turn sequences into rectangles using space filling curves
- Genetic Variant Primer
  - Single Nucleotide Polymorphisms (SNPs)
  - Structural Variants
  - Zygosity
- DeepVariant
  - Calling simple genetic variants in genomes
  - Turning alignment pile-ups into RGB image

Hilbert CNN

Reference

- B. Yin, M. Balvert, D. Zambrano, A. Schönhuth*, S. Bohte*
  An image representation based convolutional network for DNA classification
  International Conference for Learning Representations (ICLR) 2018
  *Joint last authorship

Genome Shape and Functional States



[Schwarz & Cavalli, Genetics, 2017]

Datasets

ID	♯ Samples	Description
Epi-1	14965	H3 occupancy
Epi-2	14601	H4 occupancy
Epi-3	27782	H3K9 acetylation
Epi-4	33048	H3K14 acetylation
Epi-5	34095	H4 acetylation
Epi-6	31677	H3K4 monomethylation
Epi-7	30683	H3K4 dimethylation
Epi-8	36799	H3K4 trimethylation
Epi-9	34880	H3K36 trimethylation
Epi-10	28837	H3K79 trimethylation
Splice	3190	Splice-junction gene sequences

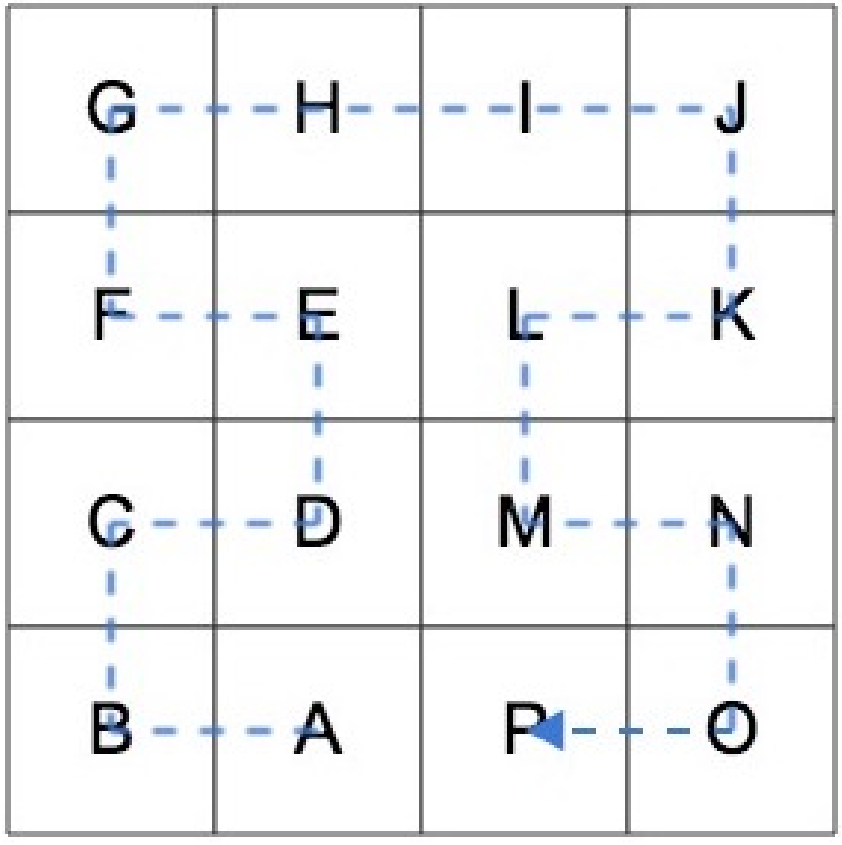
Prediction Task I

Predicting Shape Determinants of DNA

- Additional molecules (e.g. histones) and their chemical modifications determine the local shape of DNA
- There are 10 different determinants of shape, henceforth referred to as *Epi-1* to *Epi-10*, one would like to predict
- **Prediction Task I**:
  - *Input*: pieces of DNA of length approx. 500 letters
  - *Output*: for each shape determinant, 1 if it applies, 0 if not
- These are 10 different binary-valued predictions

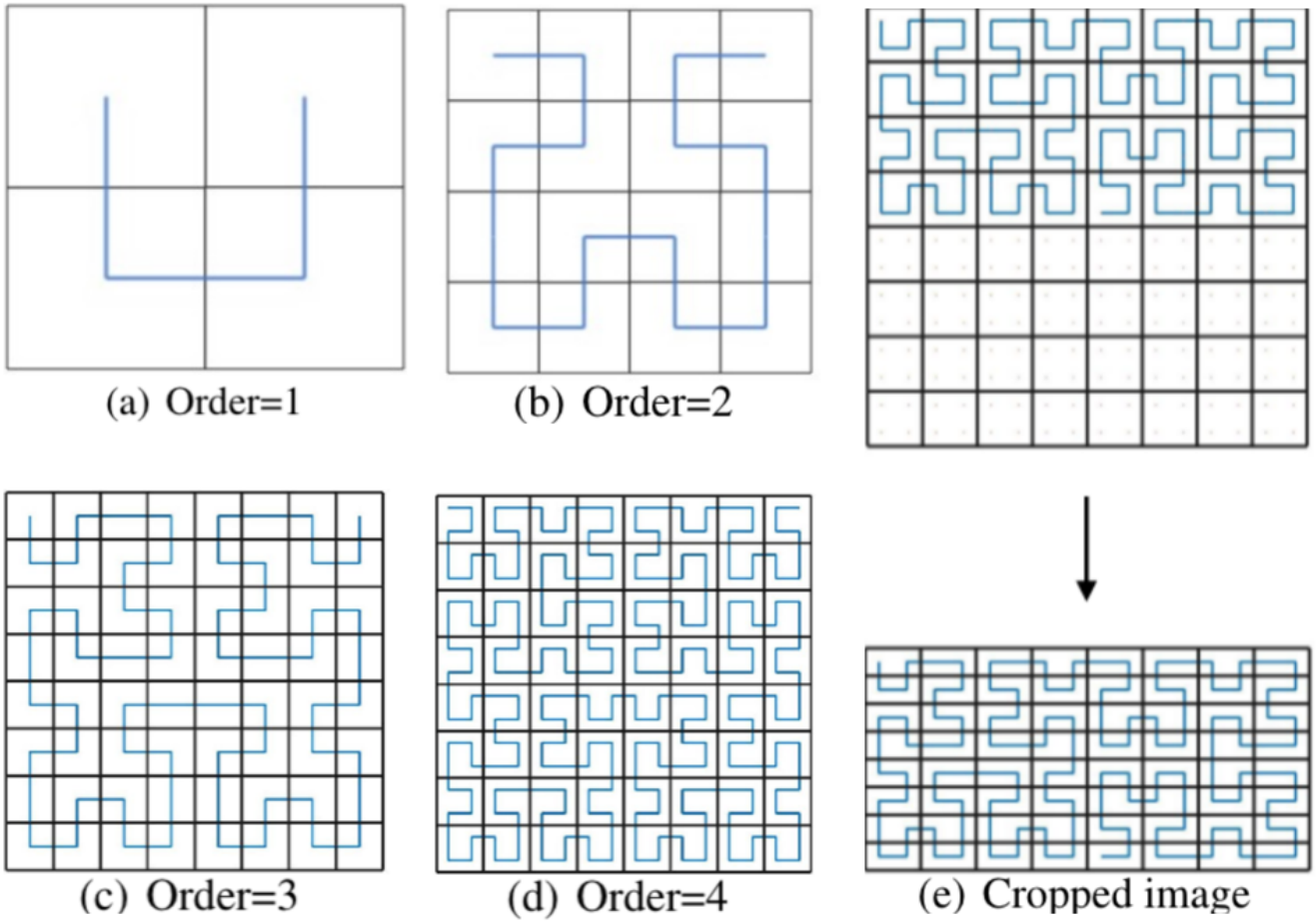
Hilbert Curves

Transform Sequences into Images



Hilbert Curves

Transform Sequences into Images



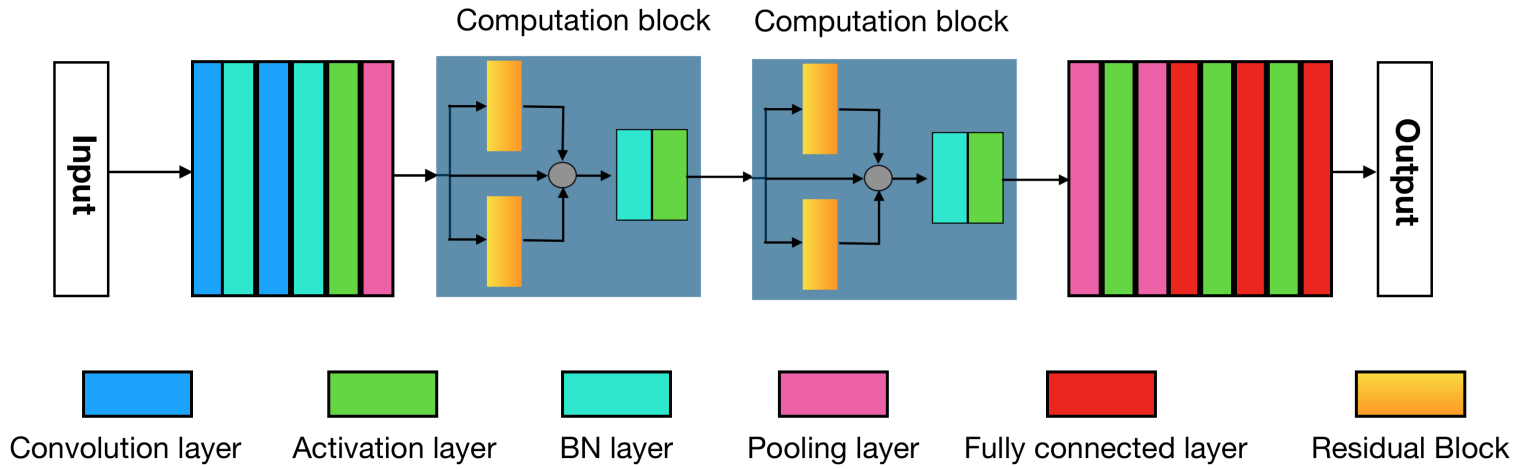
Hilbert Curves

Motivation

- *"Continuity Property"*: Continuity of distance in image w.r.t. distance in sequence
- *"Clustering Property"*: Of all space filling curve techniques, minimal number of subsequences per rectangle(!)
- Can optimally model distant relationships
- **Idea**:
  1. Transform DNA sequence of length approx. 500 into images using Hilbert curves
  2. Classify the resulting images using CNN's

Hilbert CNN

Architecture



- *BN* = Batch Normalization [Ioffe & Szegedy, 2015]
- *Residual* = ResNet residual block [He et al., 2015]
- *BN* and *Residual* prevent the gradient to vanish
- *Pooling* makes image smaller

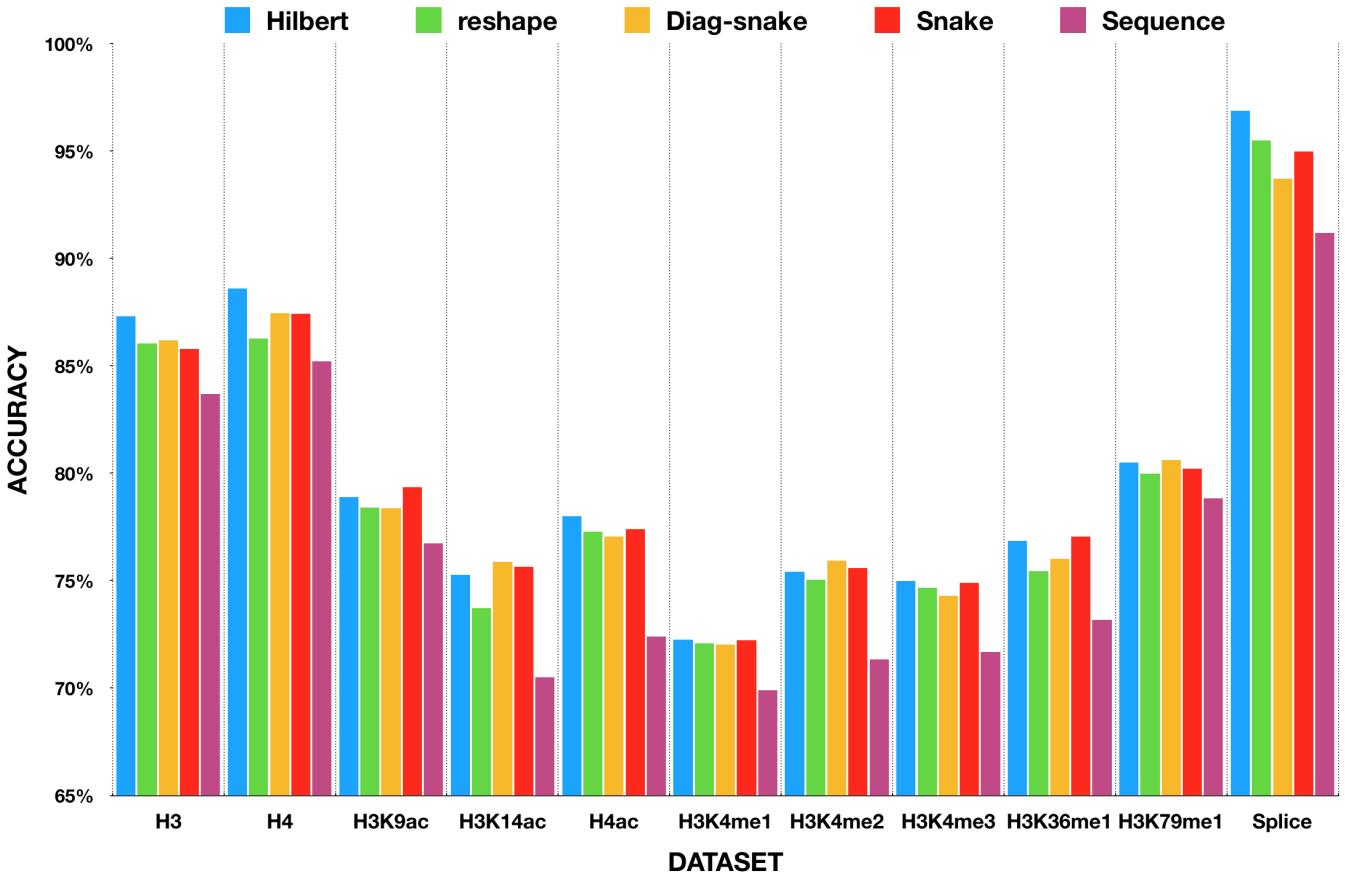
RESULTS

Data ID	SVM	Seq-CNN		Seq-HCNN		LSTM		HCNN	
	acc	acc	time	acc	time	acc	time	acc	time
Epi-1	86.5	79.3	95:23	81.0	3:45	64.1	35:43	**88.0**	3:40
Epi-2	87.8	81.9	95:53	87.1	5:32	63.8	45:32	**89.0**	4:02
Epi-3	75.1	68.8	173:18	75.8	6:12	63.1	76:09	**79.1**	8:40
Epi-4	73.3	68.3	180:56	72.5	6:09	59.3	81:21	**76.3**	10:01
Epi-5	72.1	64.8	181:33	73.8	6:05	60.6	93:32	**78.7**	10:32
Epi-6	69.7	62.6	192:20	67.5	7:12	60.4	93:44	**72.6**	11:12
Epi-7	69.0	62.4	188:13	72.4	7:04	61.5	94:22	**74.2**	9:03
Epi-8	68.6	62.3	162:32	70.7	6:54	58.0	96:03	**74.2**	9:54
Epi-9	75.2	72.2	161:12	73.2	6:34	60.8	93:48	**77.7**	9:45
Epi-10	80.6	75.1	158:34	78.1	5:43	63.8	64:28	**81.2**	9:13
Splice	94.7	91.8	35:12	91.2	2:32	96.2	6:42	**96.9**	1:30

SVM = Support Vector Machines [Higashihara et al., 2008]
Seq-CNN = CNN's on Sequence [Nguyen et al., 2016]
Seq-HCNN = CNN's on flattened Hilbert Curve
LSTM = Long-Short Term Memory NN
HCNN = CNN's on Hilbert image

Results II

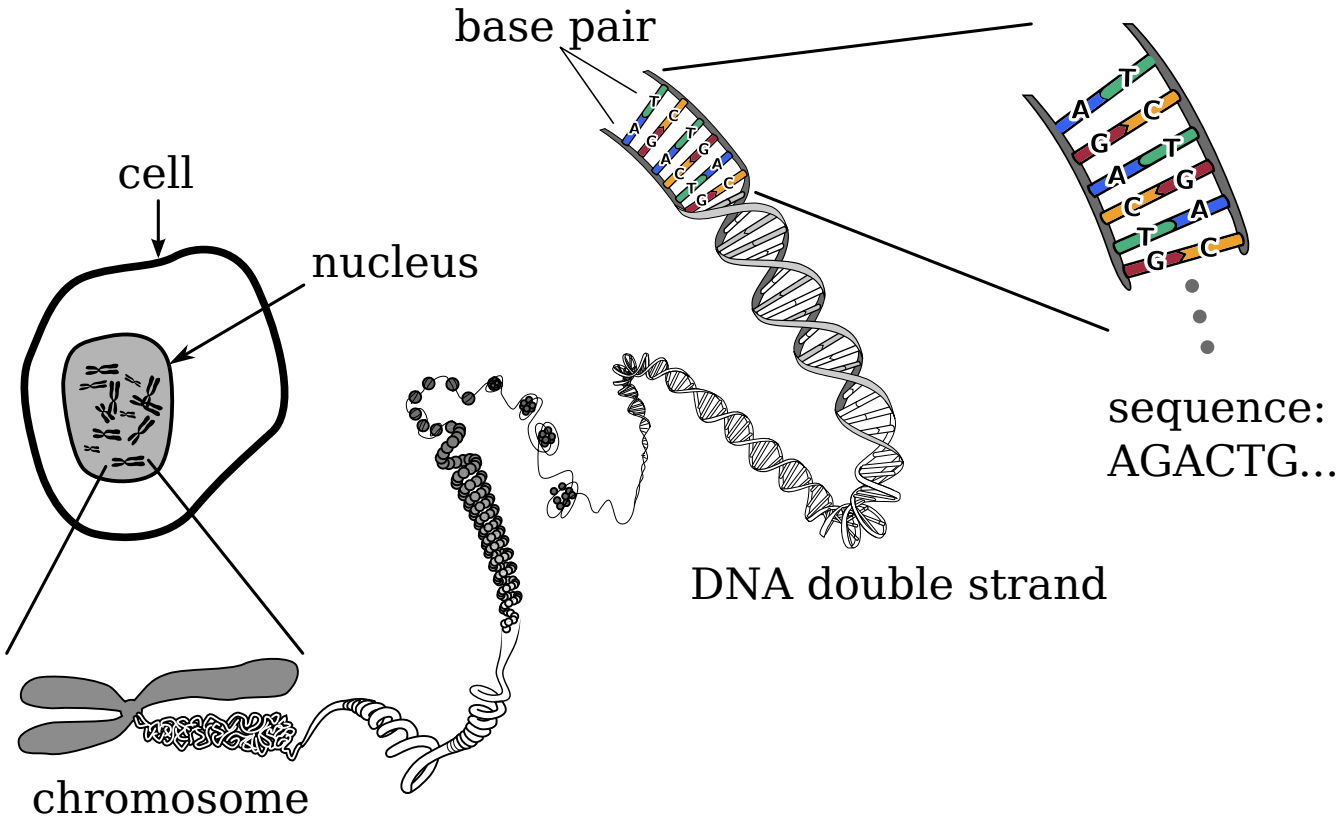
Space-Filling Curves



Hilbert-CNN architecture using different space filling curves

Genetic Variant Discovery

Cells and DNA



Human genome: string of length 3×10^9, letters: A,C,G und T.

Genetic Variants

Until 2006

Single nucleotide polymorphisms (SNPs)

Genetic Variants

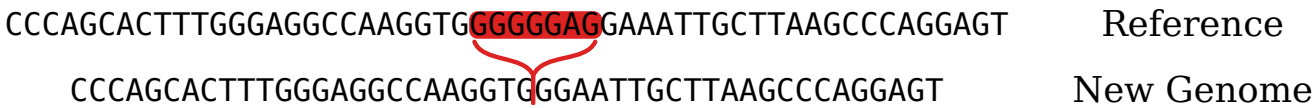
Until 2006

Single nucleotide polymorphisms (SNPs)

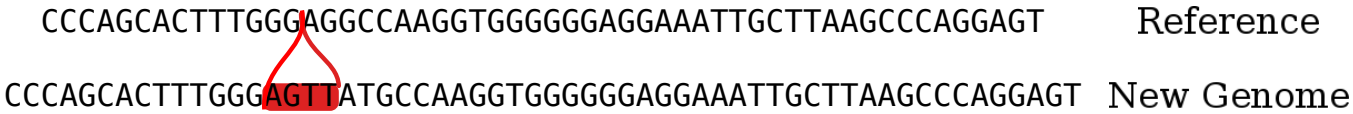
From 2006

Structural Variants

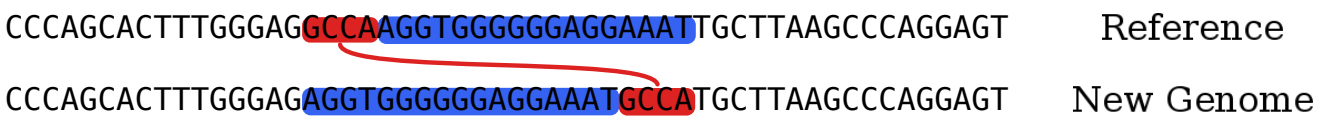
Deletion



Insertion



Translocation

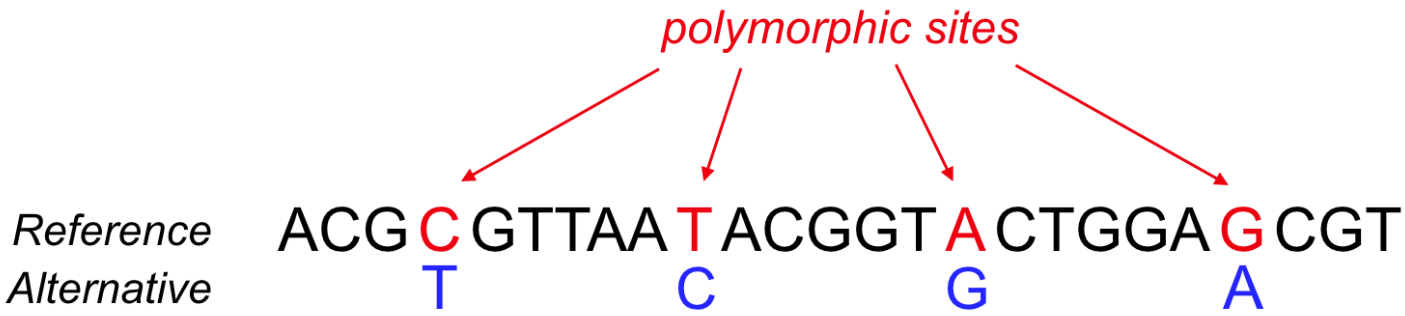


Further variations: inversions, duplications, ...

Genetics Primer

- The vast majority of genetic variants show in about a few million well-known positions, so called *polymorphic sites*
- For again the vast majority of them, there are two options
  - For a SNP for example an A or a G
- By convention, one refers to one of the options (usually the more predominant one) as *reference*, and the other one as *alternative*

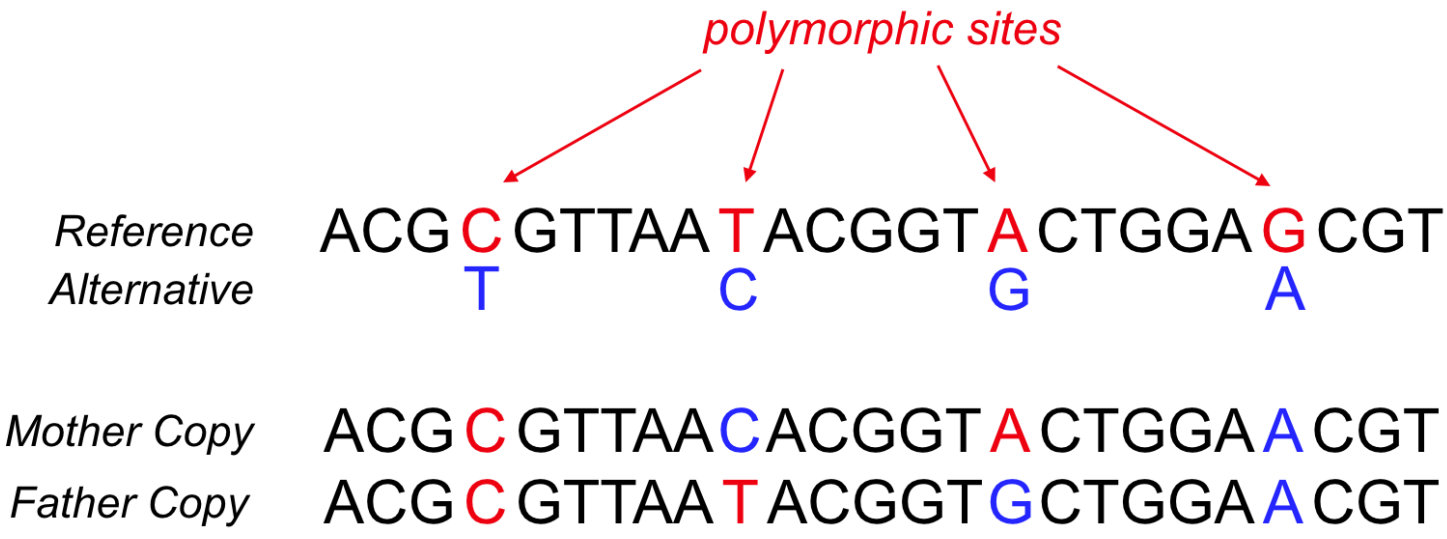
Individual Genotypes



Individual Genotypes

- Every individual human genome comes in two copies:
  - One inherited from the mother
  - One inherited from the father
- These copies can differ at polymorphic sites
  - Homozygous for reference *("Hom-Ref")*: both copies carry reference allele
  - Heterozygous *("Het")*: copies differ
  - Homzygous for alternative *("Hom-Alt")*: both copies carry non-reference allele

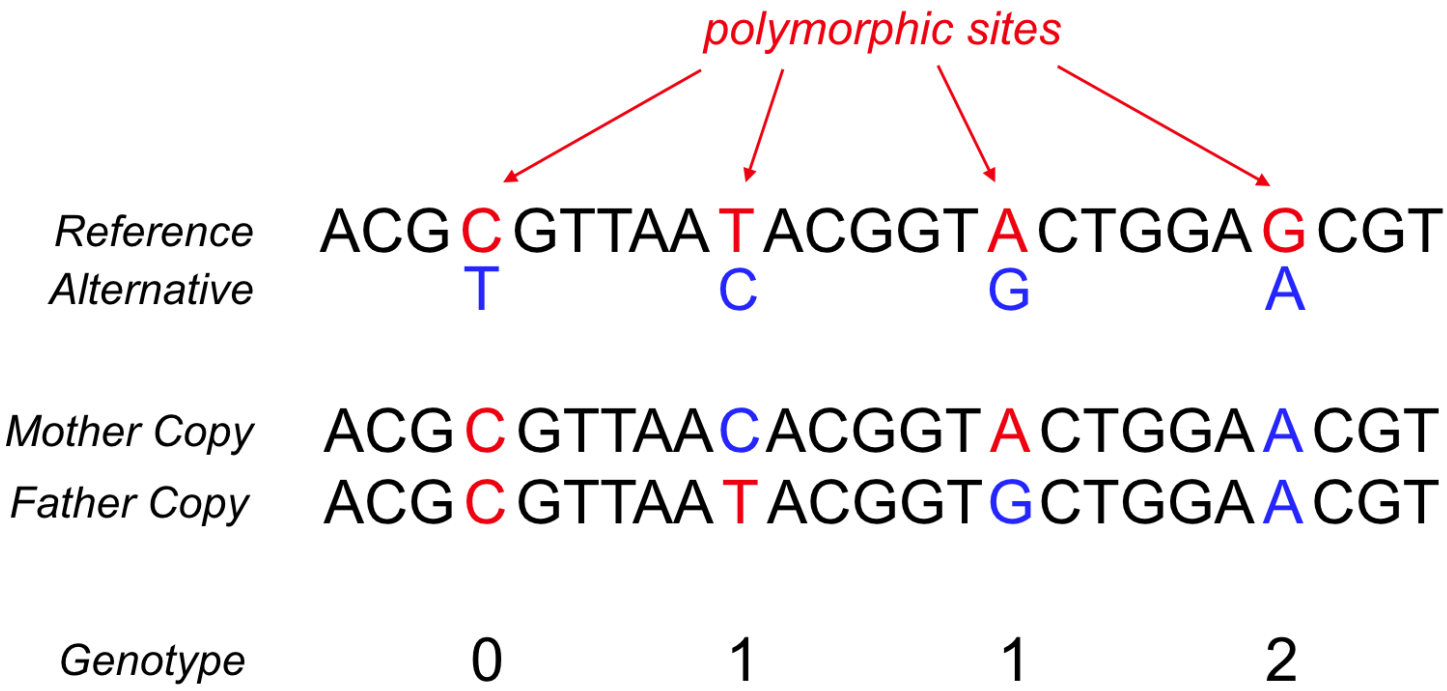
Individual Genotypes



Individual Genotypes

- The *genotype* of an individual is a vector whose length is the number of polymorphic positions
- The entries of such a vector are
  - 0 = *"Hom-Ref"*
  - 1 = *"Het"*
  - 2 = *"Hom-Alt"*

Individual Genotypes



How to Discover Variants?

Genetic Variants: Discovery Modes

Re-Sequencing

- Sequence DNA of genome of interest
- Align resulting reads against reference genome
- Note down all differences

Genetic Variants: Discovery Modes

Re-Sequencing

- Sequence DNA of genome of interest
- Align resulting reads against reference genome
- Note down all differences

De Novo Assembly

- Sequence DNA of genome of interest
- Connect resulting reads to form full-length genome
- Note down differences as per full-length comparison with reference genome

Genetic Variants: Discovery Modes

Re-Sequencing

- Sequence DNA of genome of interest
- Align resulting reads against reference genome
- Note down all differences

De Novo Assembly

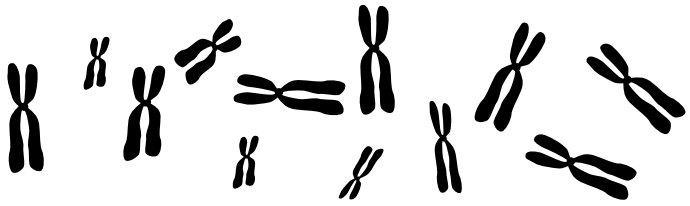
- Sequence DNA of genome of interest
- Connect resulting reads to form full-length genome
- Note down differences as per full-length comparison with reference genome

Somatic Variants

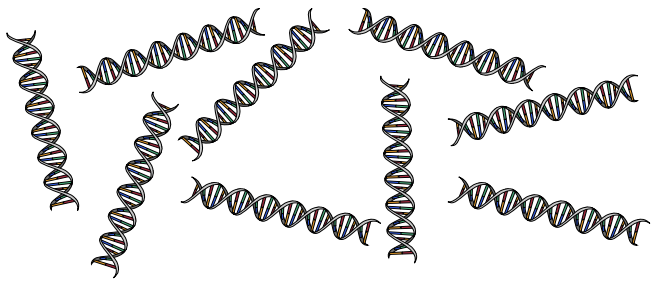
- Note down differences between cancer and control as well

Next Generation Sequencing

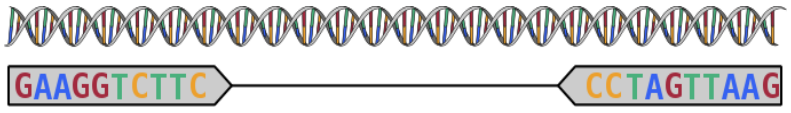
1. Extract Donor Genome DNA



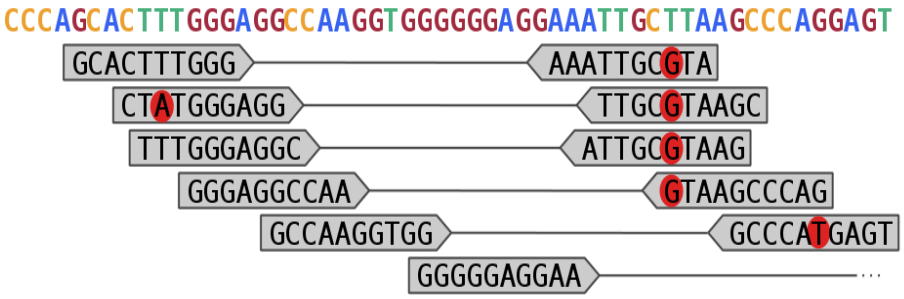
2. Break into fragments



3. Sequence fragments



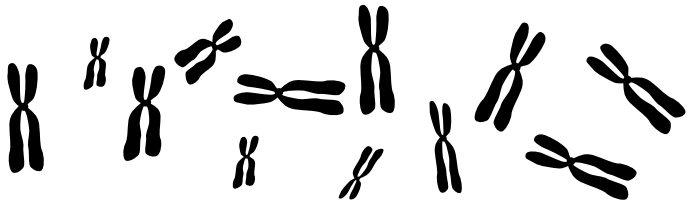
4. Map against reference genome



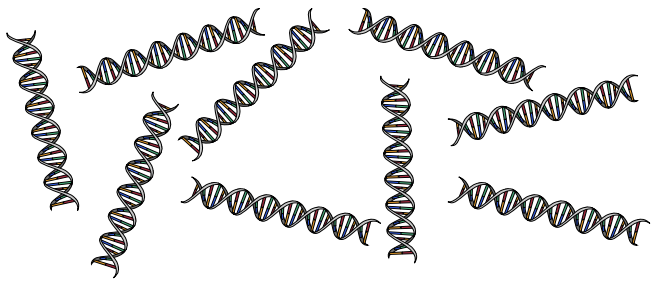
- For reference guided variant discovery, start from 4.
- For de novo assembly, start from 3.

Next Generation Sequencing

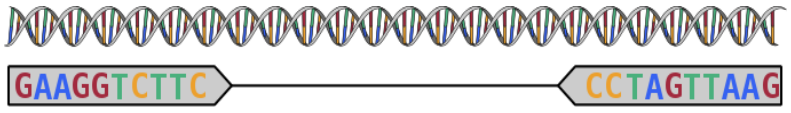
1. Extract Donor Genome DNA



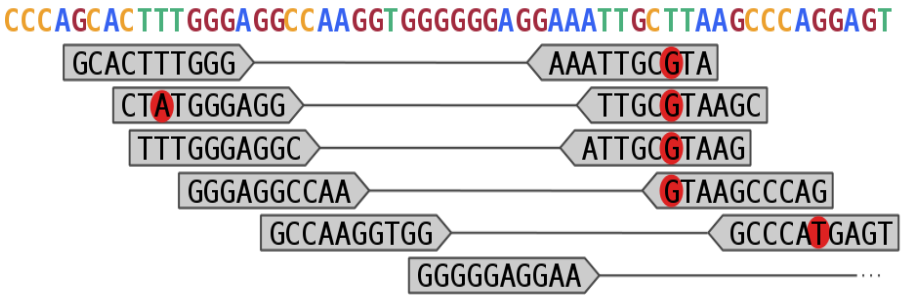
2. Break into fragments



3. Sequence fragments



4. Map against reference genome



- For reference guided variant discovery, start from 4.
- For de novo assembly, start from 3.

Re-Sequencing

Re-Sequencing: Variant Discovery

Evaluate signals emerging from aligned reads

SNP's and small insertions and deletions ("Indels")

- Look at alignments of reads with reference genome

Re-Sequencing: Variant Discovery

Evaluate signals emerging from aligned reads

SNP's and small insertions and deletions ("Indels")

- Look at alignments of reads with reference genome

Structural variants

- Variants may still yield signals in alignments directly
- Variants give rise to signals in paired-end alignments

Re-Sequencing: Variant Discovery

Evaluate signals emerging from aligned reads

SNP's and small insertions and deletions ("Indels")

- Look at alignments of reads with reference genome

Structural variants

- Variants may still yield signals in alignments directly
- Variants give rise to signals in paired-end alignments

Indels: Read Pair and Alignment Signals

Discovering Indels

Reference genome

CCCAGCACTTTGGGAGGCCAAGGTGGGGGGAGGAAATTGCTTAAGCCCAGGAGT

Discovering Indels

Reference genome

CCCAGCACTTTGGGAGGCCAAGGTGGGGGGAGGAAATTGCTTAAGCCCAGGAGT

Sequenced genome

CCCAGCACTTTGGGAGGCCAAAAATTGCTTAAGCCCAGGAGT

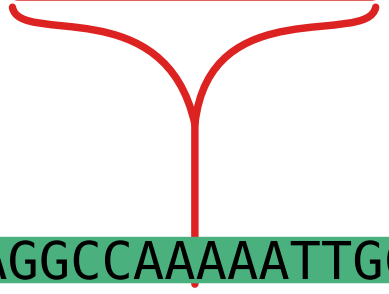
Discovering Indels

Reference genome

CCCAGCACTTTGGGAGGCCAAGGTGGGGGGAGGAAATTGCTTAAGCCCAGGAGT

Sequenced genome

CCCAGCACTTTGGGAGGCCAAAAATTGCTTAAGCCCAGGAGT



Fragment

Discovering Indels

Reference genome

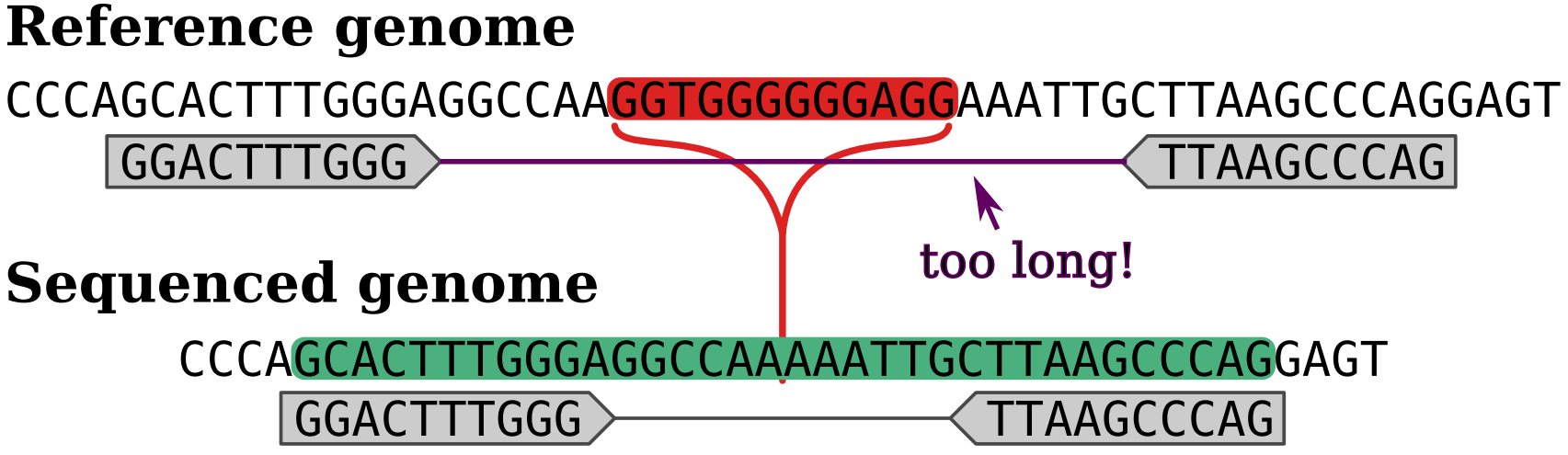
CCCAGCACTTTGGGAGGCCAAGGTGGGGGGAGGAAATTGCTTAAGCCCAGGAGT

Sequenced genome

CCCAGCACTTTGGGAGGCCAAAAATTGCTTAAGCCCAGGAGT

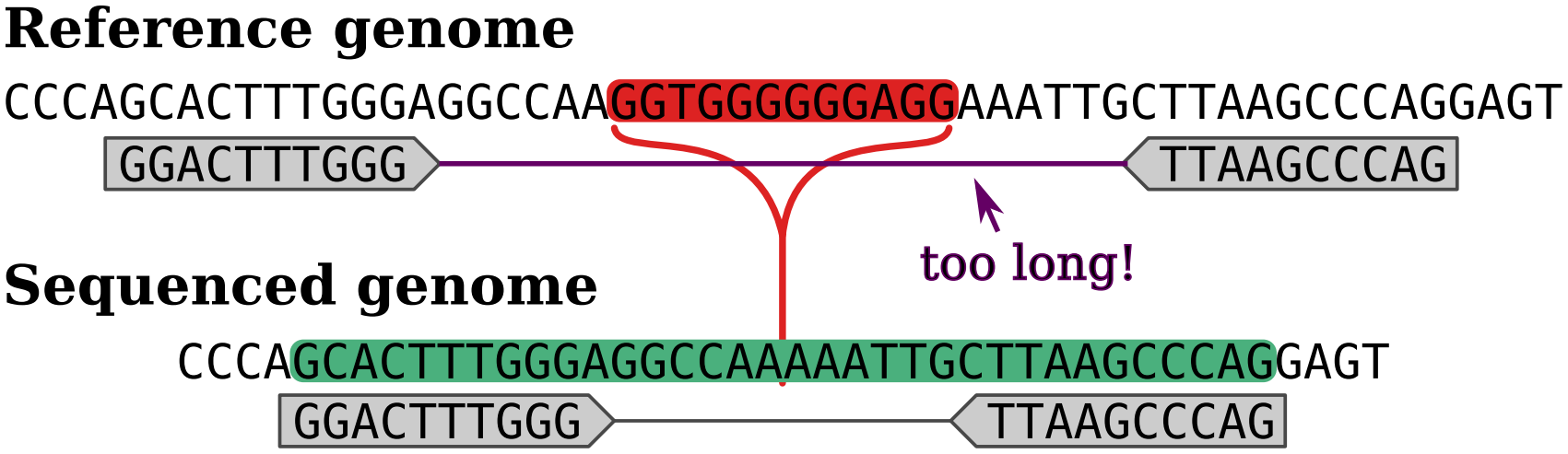
GGACTTTGGG — TTAAGCCCAG

Discovering Indels



- Insertions: alignment length too short

Discovering Indels: Signals



Read Pair Signal: Deviating Alignment Length

Alignment Signal: Gap

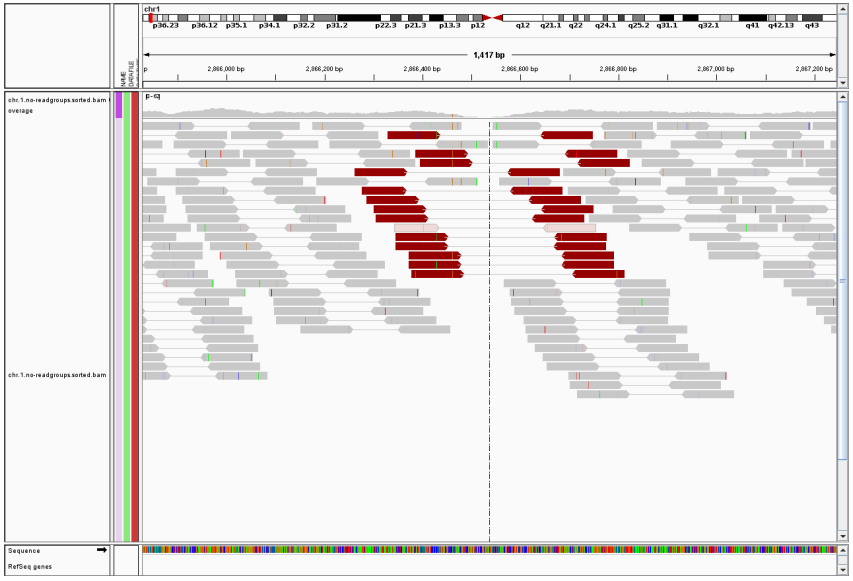
Deep Variant

Reference

- R. Poplin, [et al.], M.A. DePristo
  A universal SNP and small-indel variant caller using deep neural networks
  Nature Biotechnology, 2018
  *Joint last authorship

Motivation: Discovery of SNPs and Indels

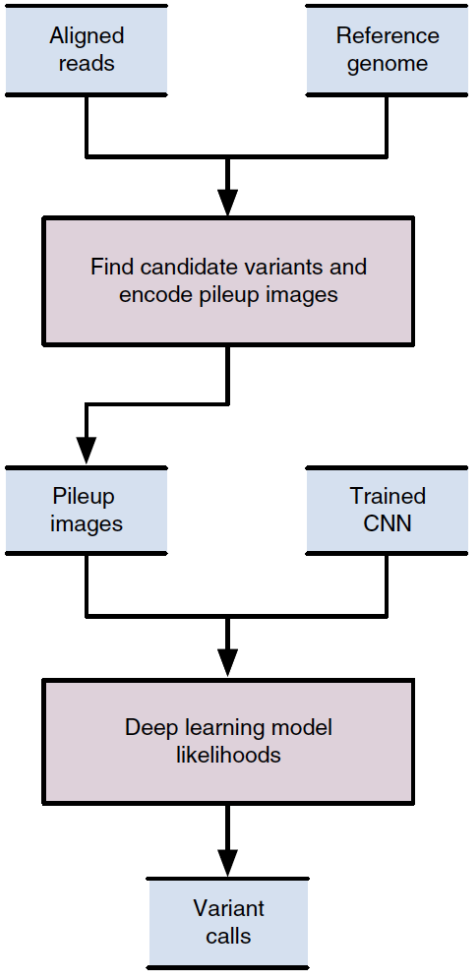
Motivation



Integrative Genome Viewer: normal alignments and repeat

- *Goal:* Predict SNPs and indels (insertions and deletions)
- *Idea:*
  - Turn alignment scenarios into RGB type images
  - Train deep NN with labeled images
  - Deep NN predicts occurrence of SNPs and indels
- *Challenge:* Encode alignment data into tri-channel information

Deep Variant: Workflow

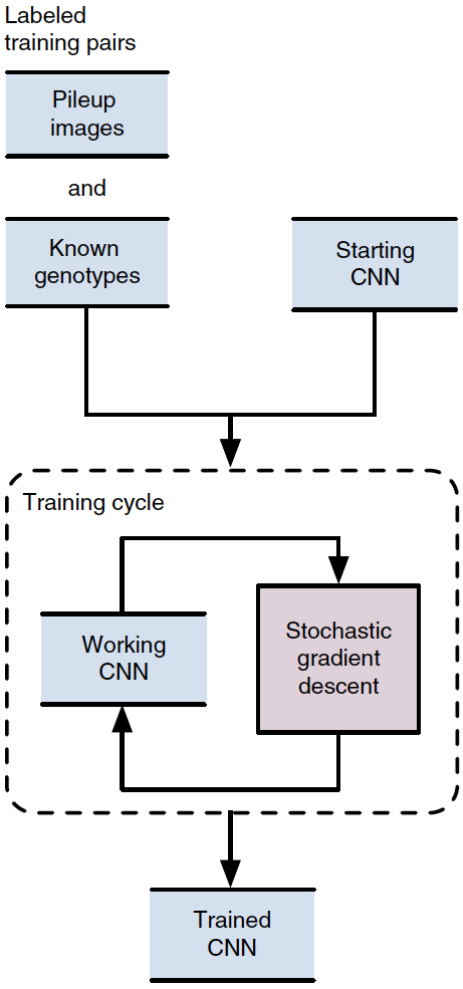


From [Poplin et al., 2018]

- *Input Data:*
  - Find candidate variant positions
    ☞ Screen every position in reference
  - Recruit corresponding alignments
  - Turn alignments + reference genome into RGB image
- *ML Model:* Standard CNN accepting RGB images
- *Predictions:* Variant zygosity
  - *Hom-Ref:* No variant
  - *Het:* Heterozygous variant
  - *Hom-Alt:* Homozygous variant

Deep Variant: Training

Deep Variant: Training

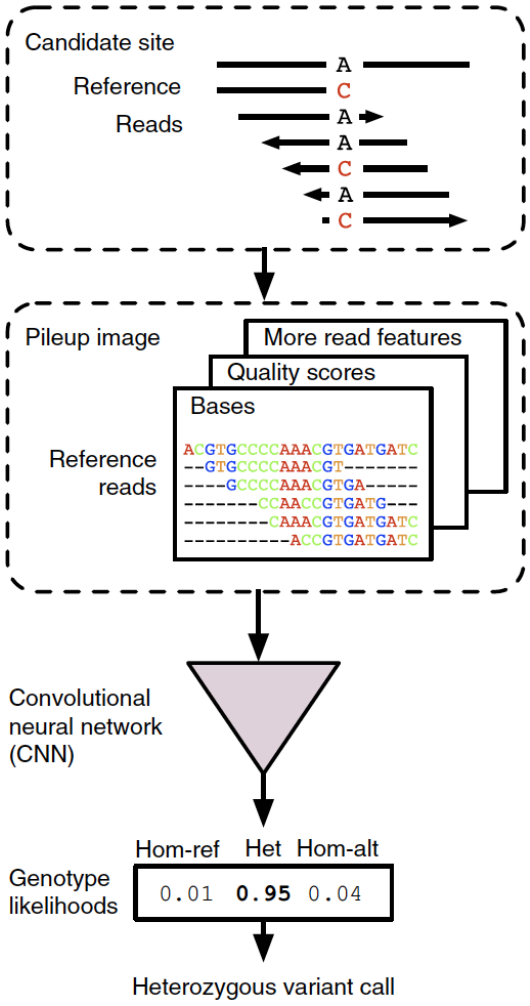


From [Poplin et al., 2018]

- *Training Data:*
  - RGB images from known (non-)variant positions
  - *Labels:* Applicable zygosity status
  - Image size: 299×299 pixels
- *Model:* Inception v2, from 2015
  - Off-the-shelf, no adaptations
  - Ensemble of 4 inception networks
  - `https://arxiv.org/abs/1512.00567`
- *Training:* Stochastic gradient descent
  - CNN pre-initialized
  - Batches of 32 images
  - Optimization: momentum based

Deep Variant: Creating Images

Deep Variant: Image Creation



From [Poplin et al., 2018]

- *Original Data:*
  - Reference genome: From .fasta file
  - Alignments: From .sam/.bam file
- *RGB Image Dimension:*
  - *Rows:* One per alignment
  - *Columns:* One per position
- *RGB Image Channels:*
  - *Channel 1:* Bases in reads
  - *Channel 2:* Per base quality score
  - *Channel 3:* Strand read stems from
    ☞ Positive (5′-3′) or negative (3′-5′)

Sequence Alignment/Map (SAM) Files

```
SRR081708.237649    163 1   10003   6   1S67M   =   10041   105 GACCCTGACCCTAACCCTGACCCTGACCCTAACCCTGACCCTGACCCTA
ACCCTGACCCTAACCCTAA    S=<====<<>=><?=?=?>==@??;?>@@@=??@@????@??@?>?@@<@>@'@=?=??=<=>?>?=Q    ZA:Z:<&;0;0;;308;6
8M;68><@;0;0;;27;;>MD:Z:5A11A5A11A5A11A13 RG:Z:SRR081708  NM:i:6  OQ:Z:GEGFFFEGGGDGDGGGDGA?DCDD:GGGDGDCFGFDDFFFCC
CBEBFDABDD-D:EEEE=D=DDDDC:
```

From `https://bioinformatics-core-shared-training.github.io/cruk-summer-school-2017/Day1/Session5-alignedReads.html`

Sequence Alignment/Map (SAM) Files

- Read identifier (1)
- Chromosome (3) and start position (6) within chromosome
- Mapping quality (5)
- CIGAR (Compact Idiosyncratic Gapped Alignment Report) string (6)

Sequence Alignment/Map (SAM) Files

```
SRR081708.237649    163 1   10003   6   1S67M   =   10041   105 GACCCTGACCCTAACCCTGACCCTGACCCTAACCCTGACCCTGACCCTA
ACCCTGACCCTAACCCTAA    S=<====<<>=><?=?=?>==@??;?>@@@=??@@????@??@?>?@@<@>@'@=?=??=<=>?>?=Q    ZA:Z:<&;0;0;;308;6
8M;68><@;0;0;;27;;>MD:Z:5A11A5A11A5A11A13 RG:Z:SRR081708  NM:i:6  OQ:Z:GEGFFFEGGGDGDGGGDGA?DCDD:GGGDGDCFGFDDFFFCC
CBEBFDABDD-D:EEEE=D=DDDDC:
```

From `https://bioinformatics-core-shared-training.github.io/cruk-summer-school-2017/Day1/Session5-alignedReads.html`

Sequence Alignment/Map (SAM) Files

- Position of mate in paired-end read (8)
- Sequence itself (10)
- Base qualities = Phred string (11)

Sequence Alignment/Map (SAM) Files

```
SRR081708.237649    163 1    10003    6    1S67M    =    10041    105 GACCCTGACCCTAACCCTGACCCTGACCCTAACCCTGACCCTGACCCTA
ACCCTGACCCTAACCCTAA     S=<====<<>=><?=?=?>==@??;?>@@@=??@@????@??@?>?@@<@>@'@=?=??=<=>?>?=Q     ZA:Z:<&;0;0;;308;6
8M;68><@;0;0;;27;;>MD:Z:5A11A5A11A5A11A13 RG:Z:SRR081708  NM:i:6  OQ:Z:GEGFFFEGGGDGDGGGDGA?DCDD:GGGDGDCFGFDDFFFCC
CBEBFDABDD-D:EEEE=D=DDDDC:
```

From `https://bioinformatics-core-shared-training.github.io/cruk-summer-school-2017/Day1/Session5-alignedReads.html`

Sequence Alignment/Map (SAM) Files

- Position of mate in paired-end read (8)
- Sequence itself (10)
- Base qualities = Phred string (11)

Deep Variant: Image Channels

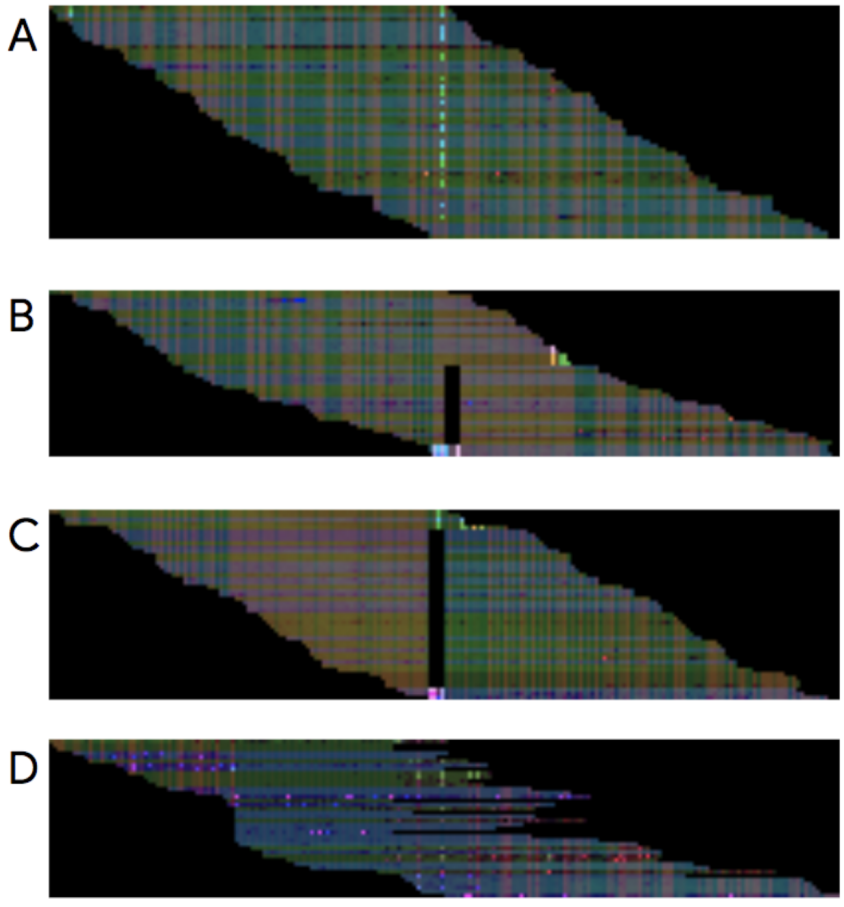
```
def get_base_color(base):
  base_to_color = {'A': 250, 'G': 180, 'T': 100, 'C': 30}
  return base_to_color.get(base, 0)


def get_quality_color(quality):
  return int(254.0 * (min(40, quality) / 40.0))


def get_strand_color(on_positive_strand):
  return 70 if on_positive_strand else 240
```

From [Poplin et al., 2018]

Deep Variant: Image Channels



Four different variant scnearios coded as DeepVariant images

From [Poplin et al., 2018]

Deep Variant: Results

Deep Variant: Results

Method	Type	F1	Recall	Precision
DeepVariant	Indel	0.95806	0.92868	0.98936
Strelka	Indel	0.95074	0.91623	0.98796
16GT	Indel	0.94010	0.90803	0.97452
GATK (raw)	Indel	0.93268	0.89504	0.97363
GATK (VQSR)	Indel	0.91212	0.84497	0.99087
FreeBayes	Indel	0.90438	0.83025	0.99305
SAMtools	Indel	0.86976	0.79089	0.96611
DeepVariant	SNP	0.99103	0.98888	0.99319
Strelka	SNP	0.98865	0.98107	0.99636
16GT	SNP	0.97862	0.98966	0.96782
FreeBayes	SNP	0.96910	0.94837	0.99075
GATK (VQSR)	SNP	0.96895	0.94542	0.99368
SAMtools	SNP	0.96818	0.94386	0.99378
GATK (raw)	SNP	0.96646	0.95685	0.97627

From [Poplin et al., 2018]

- *Recall:* Discovered true variants over true variants
- *Precision:* Disovered true variants over discovered variants

Outlook

- SVision
  - Calling complex genetic variants in genomes
  - Turn alignment patterns of long reads into RGB images
- ALSNet
  - Predicting ALS disease status from genotype profiles
  - Employ convolution on sequences directly
- Capsule Networks
  - Motivation
  - Tutorial
- Disease Capsule
  - Predicting ALS disease status using capsule networks
  - Biological Interpretation

Thanks for your attention