Applications of deep learning in medical genetics and cancer diagnostics

Tomasz Stokowy

Department of Clinical Science
and
Computational Biology Unit
University of Bergen, Norway

Medical University of Gdańsk

Norwegian Consortium for Sequencing and Personalized Medicine (NorSeq)

Plan of the presentation

1. **Introduction: human DNA and the method to read it**

2. **Methods: machine learning methods outperform currently used algorithms**

3. **Examples of possible machine learning applications in medicine:**

   - **Rare genetic disorders**
   - **Hereditary cancer (ca 10% of cancer cases)**
   - **Sporadic cancer (ca 90% of cases)**

4. **Conclusions**

Motivation

Use human genome data to fight rare disorders and cancer

Human DNA and the method to read it

3×10^9 A C T G pairs = Big Data

Human DNA and the method to read it

3 x 10^9 A C T G pairs = Big Data

Illumina HiSeq 4000 – www.illumina.com

Human DNA and the method to read it

3×10^9 A C T G pairs = Big Data

We can accurately read ca. 150 bases,
so we sequence many such reads
Schneeberger, Nat Rev Genetics 2014

Human DNA and the method to read it

3×10^9 A C T G pairs = Big Data

We can accurately read ca. 150 bases, so we sequence many such reads
Schneeberger, Nat Rev Genetics 2014

1.2% difference in the sequence between human and chimp

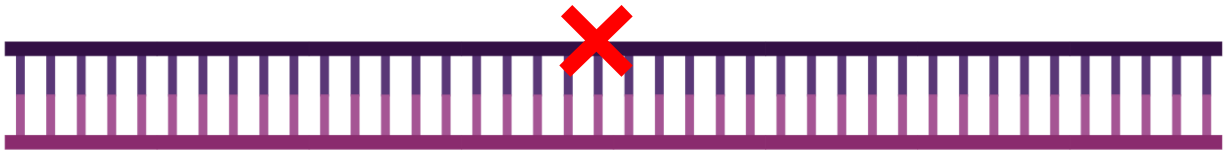
0.1% difference in the sequence between humans

Diseases of human genome

	Rare/Monogenic Disorder	Hereditary Cancer	Sporadic Cancer	Polygenic Disorder
Number of causative mutations	1 or a few	1 or a few	?	Many
Effect of the mutation	Strong, damaging whole organism	Medium, increased cancer prevalence during lifetime	DNA repair mechanism does not work in a group of cells, uncontrolled cell divisions, metastasis	Weak, but cumulative effect of many mutations can be strong
Examples of the disease	Cystic fibrosis, Fragile X syndrome, sickle cell disease	BRCA1/BRCA2 induced hereditary breast cancer, Lynch syndrome	90% of cancer cases	Diabetes (excluding MODY), psychosis

Diseases of human genome

	Rare/Monogenic Disorder	Hereditary Cancer	Sporadic Cancer	Polygenic Disorder
Number of causative mutations	1 or a few	1 or a few	?	Many
Effect of the mutation	Strong, damaging whole organism	Medium, increased cancer prevalence during lifetime	DNA repair mechanism does not work in a group of cells, uncontrolled cell divisions, metastasis	Weak, but cumulative effect of many mutations can be strong
Examples of the disease	Cystic fibrosis, Fragile X syndrome, sickle cell disease	BRCA1/BRCA2 induced hereditary breast cancer, Lynch syndrome	90% of cancer cases	Diabetes (excluding MODY), psychosis



Diseases of human genome

	Rare/Monogenic Disorder	Hereditary Cancer	Sporadic Cancer	Polygenic Disorder
Number of causative mutations	1 or a few	1 or a few	?	Many
Effect of the mutation	Strong, damaging whole organism	Medium, increased cancer prevalence during lifetime	DNA repair mechanism does not work in a group of cells, uncontrolled cell divisions, metastasis	Weak, but cumulative effect of many mutations can be strong
Examples of the disease	Cystic fibrosis, Fragile X syndrome, sickle cell disease	BRCA1/BRCA2 induced hereditary breast cancer, Lynch syndrome	90% of cancer cases	Diabetes (excluding MODY), psychosis

Diseases of human genome

	Rare/Monogenic Disorder	Hereditary Cancer	Sporadic Cancer	Polygenic Disorder
Number of causative mutations	1 or a few	1 or a few	?	Many
Effect of the mutation	Strong, damaging whole organism	Medium, increased cancer prevalence during lifetime	DNA repair mechanism does not work in a group of cells, uncontrolled cell divisions, metastasis	Weak, but cumulative effect of many mutations can be strong
Examples of the disease	Cystic fibrosis, Fragile X syndrome, sickle cell disease	BRCA1/BRCA2 induced hereditary breast cancer, Lynch syndrome	90% of cancer cases	Diabetes (excluding MODY), psychosis

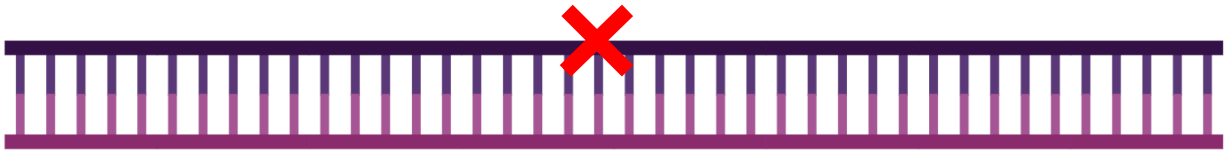
unique pattern in each tumor

Part 2

Methods

Problem statement

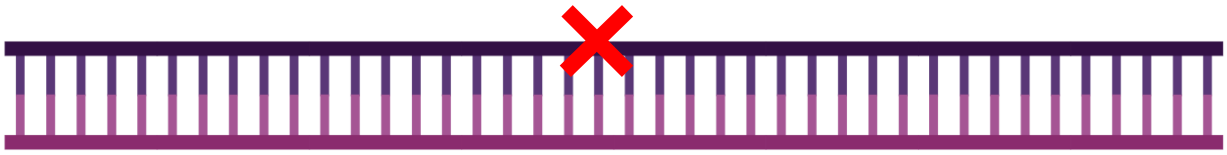
Precisely indicate variant positions,
to tell the patient that he is carrier of the disease
(diagnostics)



Recognize the mutation pattern to indicate the cause of the disease
and choose appropriate treatment
(personalized medicine)

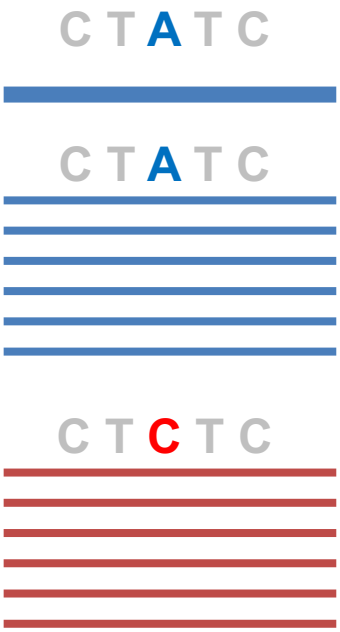
Problem 1

Precisely indicate variant positions,
to tell the patient that he is carrier of the disease
(diagnostics)



Variant calling (problem 1)

Classic approach: base counting



Reference (A)

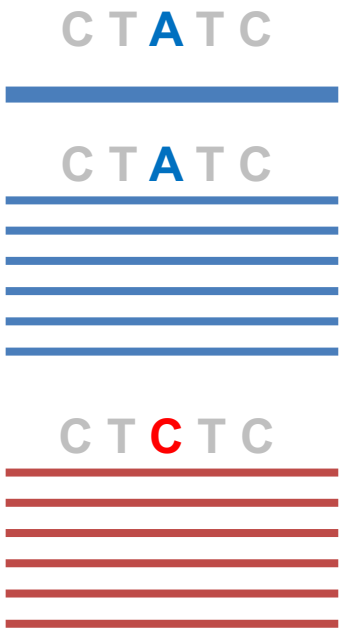
Reference reads (A)

Alternative reads (C)

Conclusion: A/C heterozygous

Variant calling (problem 1)

Classic approach: base counting



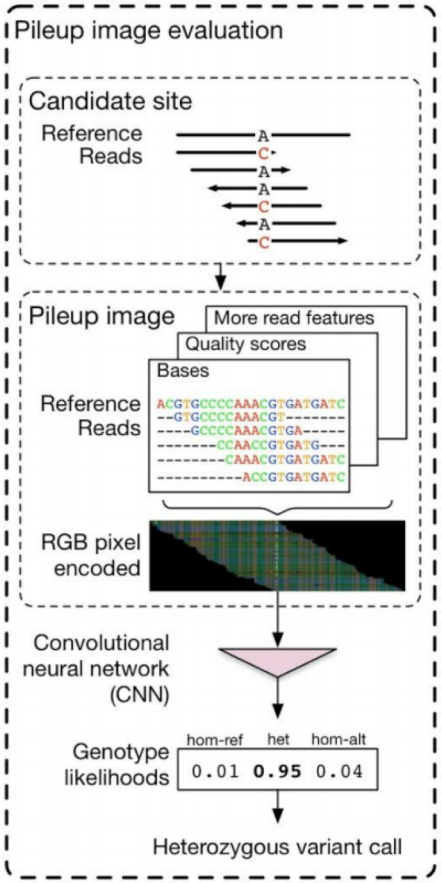
Reference (A)

Reference reads (A)

Alternative reads (C)

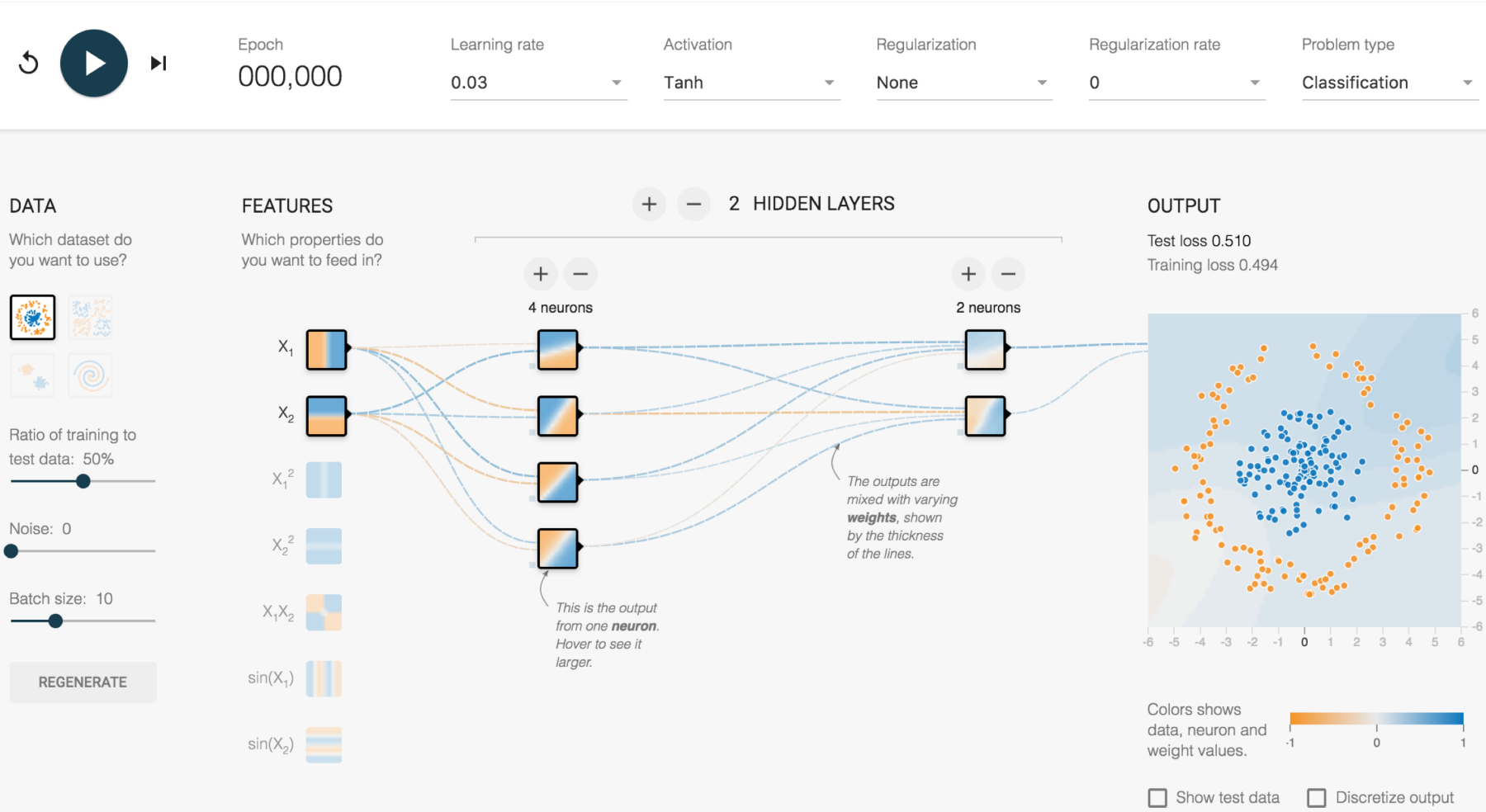
Conclusion: A/C heterozygous

Deep learning approach



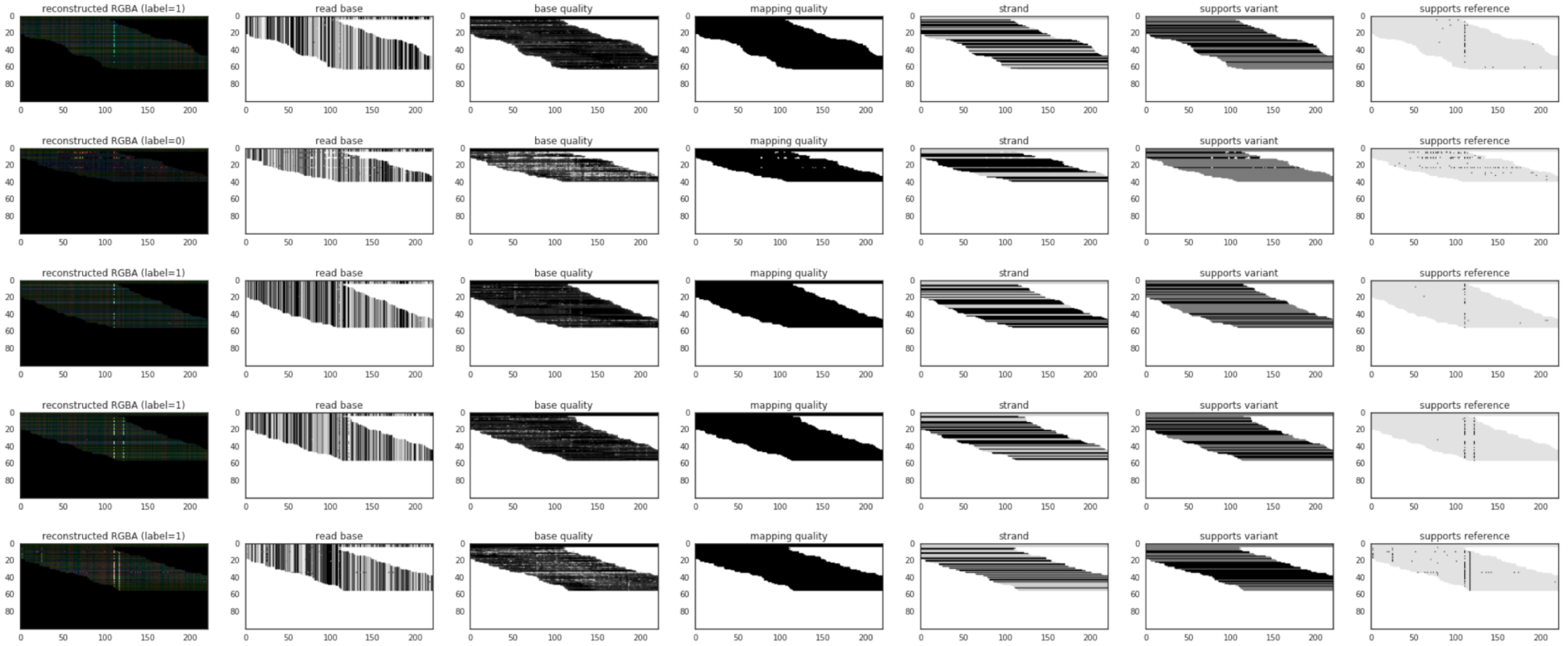
Poplin et al., Nat Biotechnol 2018

TensorFlow: does the variant really exist or it is a false positive/artefact?



http://playground.tensorflow.org

DeepVariant variant calling



Poplin et al., Nat Biotechnol 2018

Performance of variant calling methods: FDA Truth Challenge

Label	Submitter	Organization	SNP-Fscore	SNP-recall	SNP-precision	INDEL-Fscore	INDEL-recall	INDEL-precision
rpoplin-dv42	Ryan Poplin et al.	Verily Life Sciences	99.9587 %	99.9447 %	99.9728 %	98.9802 %	98.7882 %	99.1728 %

Variant calling based on machine learning is more accurate than other methods

SNV	True positive SNV calls	False negative SNV calls	False positive SNV calls	Genotype mismatch	Total number of SNV calls	SNV calling precision	SNV recall	F1 Score
SpeedSeq	2942217	100572	38107	11869	3802913	0.987223	0.966947	0.97698
DeepVariant 0.4.1	2948290	94499	22902	19595	3714945	0.992294	0.968943	0.98048
GATK 4.0 - WDL	2952605	90184	41684	12579	3814443	0.986082	0.970361	0.978159
INDEL	True positive INDEL calls	False negative INDEL calls	False positive INDEL calls	Genotype mismatch	Total number of INDEL calls	INDEL calling precision	INDEL recall	F1 Score
Speedseq	383930	115767	32263	13310	619159	0.923499	0.768326	0.838796
DeepVariant 0.4.1	460271	39426	16122	8147	816456	0.967406	0.9211	0.943685
GATK 4.0 - WDL	429859	69838	24191	9251	764422	0.948269	0.860239	0.902112

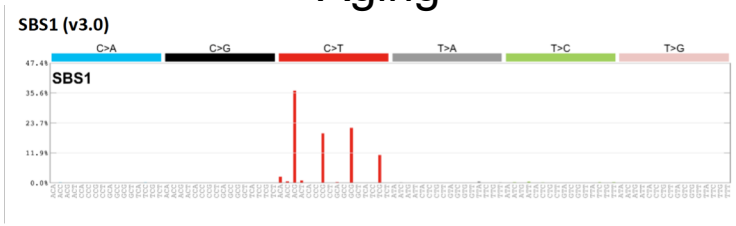
Supernat, Stokowy et al., Scientific Reports 2018

Problem 2

Recognize the mutation pattern to indicate the cause of the disease
and choose appropriate treatment
(personalized treatment)

Mutational signatures in cancer (problem 2)

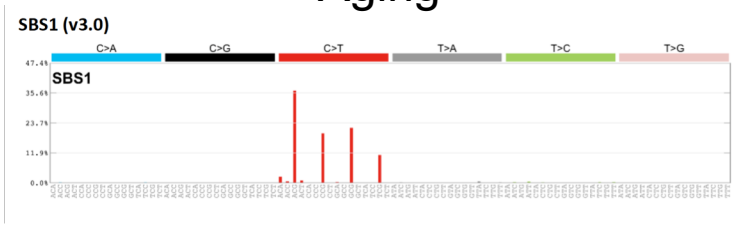
Aging



Alexandrov et al., Nature 2020

Mutational signatures in cancer (problem 2)

Aging

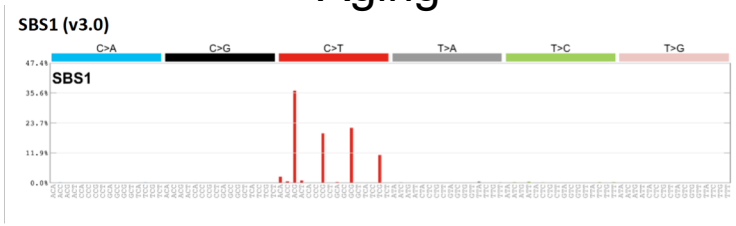


Smoking

Alexandrov et al., Nature 2020

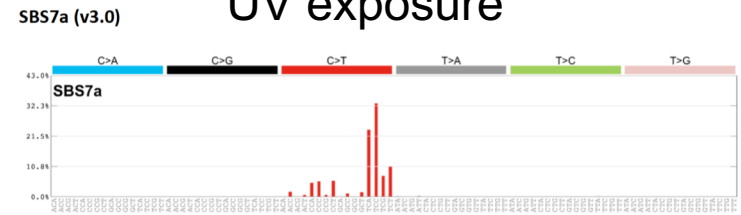
Mutational signatures in cancer (problem 2)

Aging



Smoking

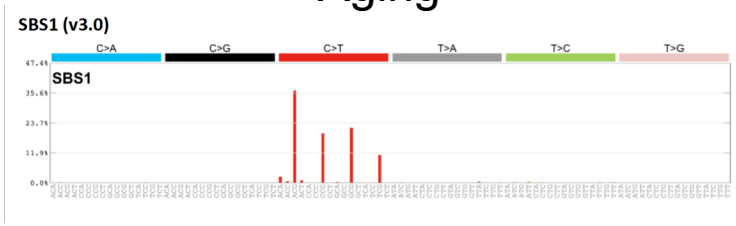
UV exposure



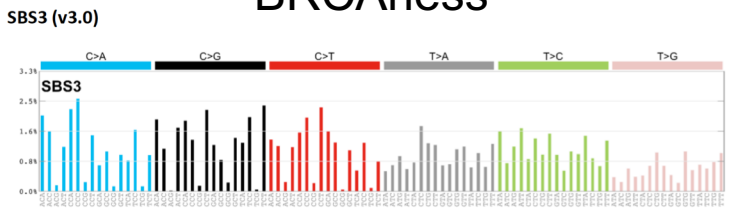
Alexandrov et al., Nature 2020

Mutational signatures in cancer (problem 2)

Aging

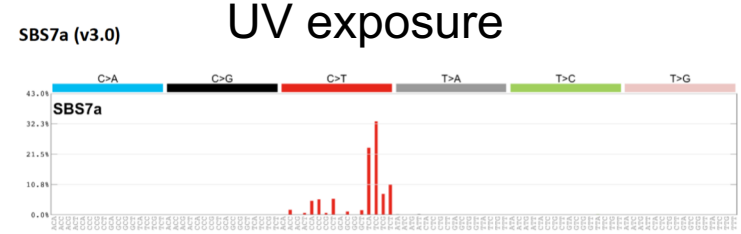


BRCAness



Smoking

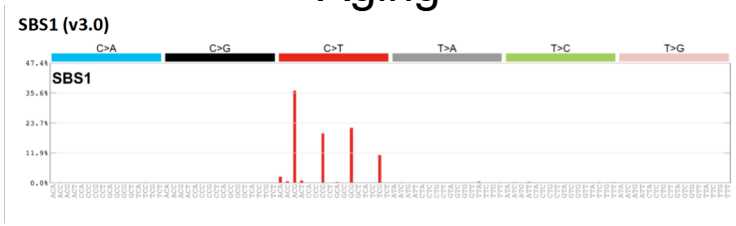
UV exposure



Alexandrov et al., Nature 2020

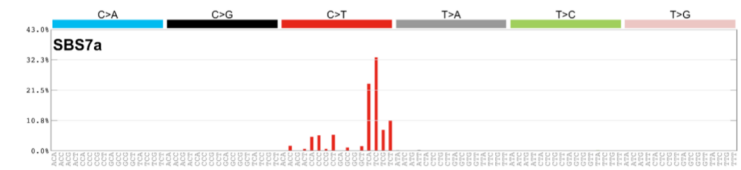
Mutational signatures in cancer (problem 2)

Aging

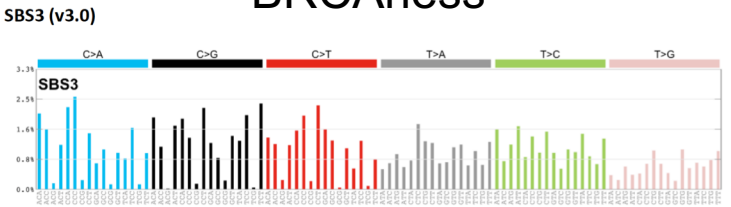


Smoking

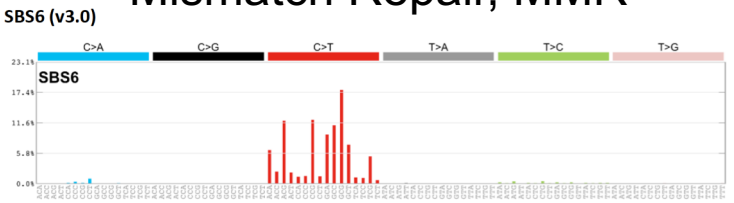
UV exposure



BRCAness



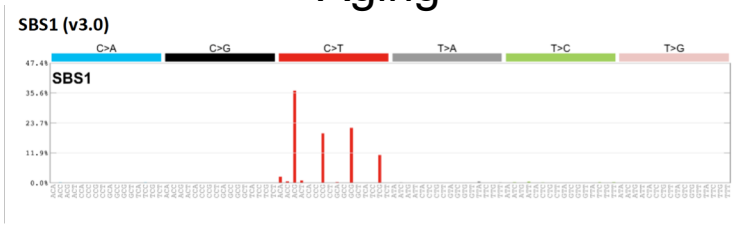
Mismatch Repair, MMR



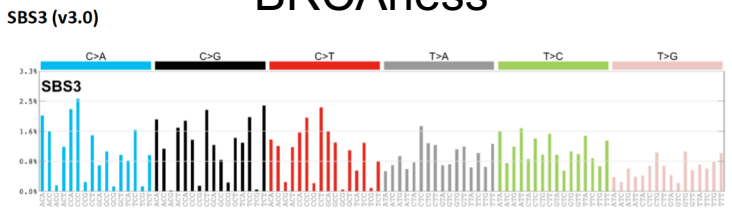
Alexandrov et al., Nature 2020

Mutational signatures in cancer (problem 2)

Aging

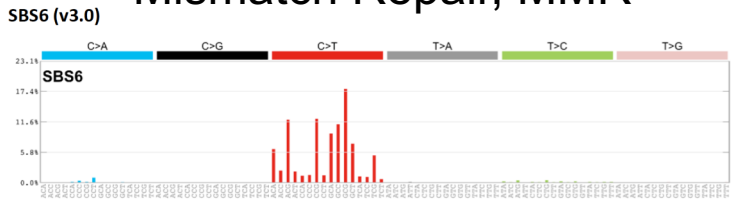


BRCAness

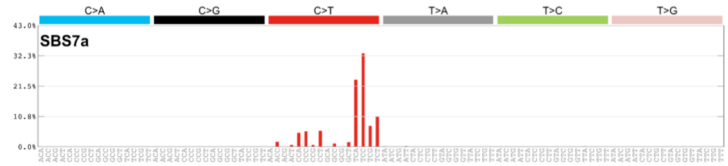


Smoking

Mismatch Repair, MMR



UV exposure

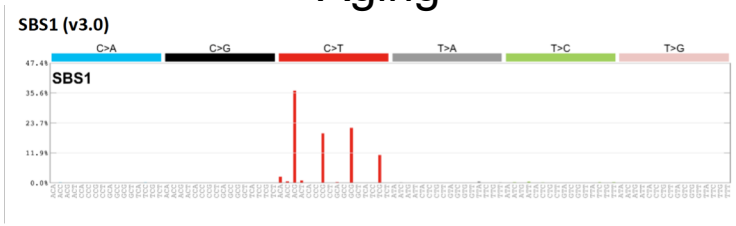


Total number of currently known signatures: 59
Many patients have mixed signatures

Alexandrov et al., Nature 2020

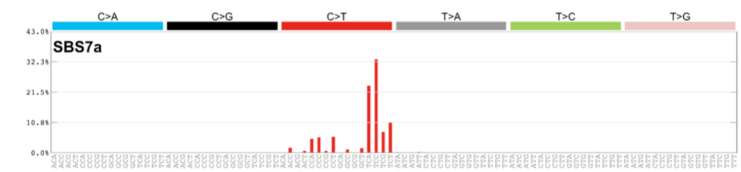
Mutational signatures in cancer (problem 2)

Aging

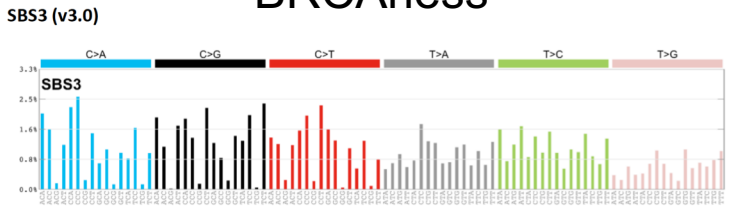


Smoking

UV exposure

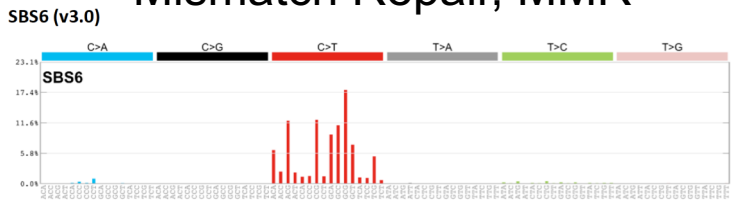


BRCAness



Cisplatin, Olaparib, PARPi inhibitors

Mismatch Repair, MMR



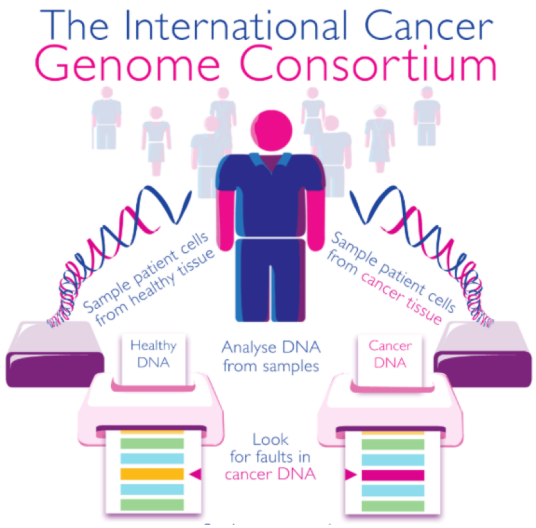
PD-1 blockade immunotherapy

Total number of currently known signatures: 59
Many patients have mixed signatures

Alexandrov et al., Nature 2020

Data

Training set:
25000 cancer genomes

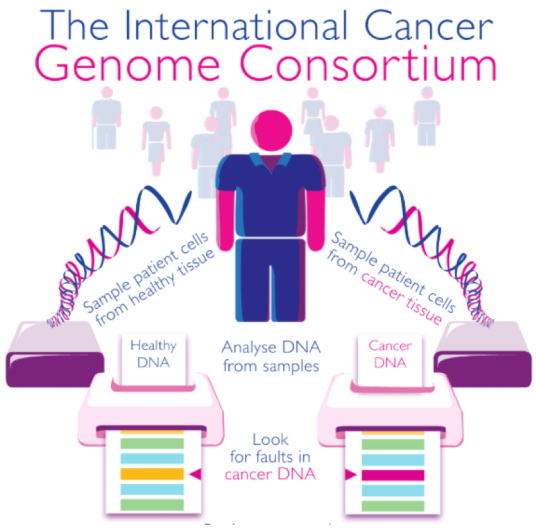


Alexandrov et al., Nature 2020

Testing set:
patient in the clinic

Cancer data

Training set:
25000 cancer genomes



Alexandrov et al., Nature 2020

Testing set:
patient in the clinic

Work in progress, MNM Diagnostics, Poznań

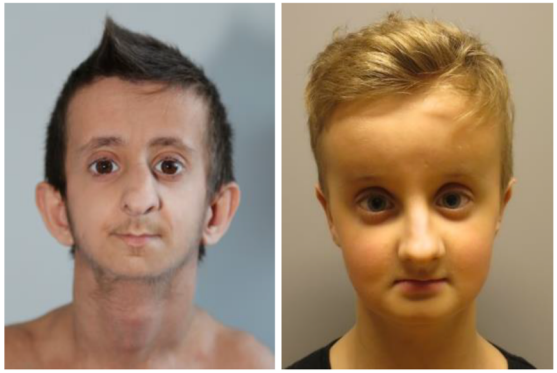
Part 3

Examples of successful application: genome sequencing, big data and IA

Rare genetic disorders

progeria – Penttinen syndrome,
Bredrup et al., EJHG, 2018

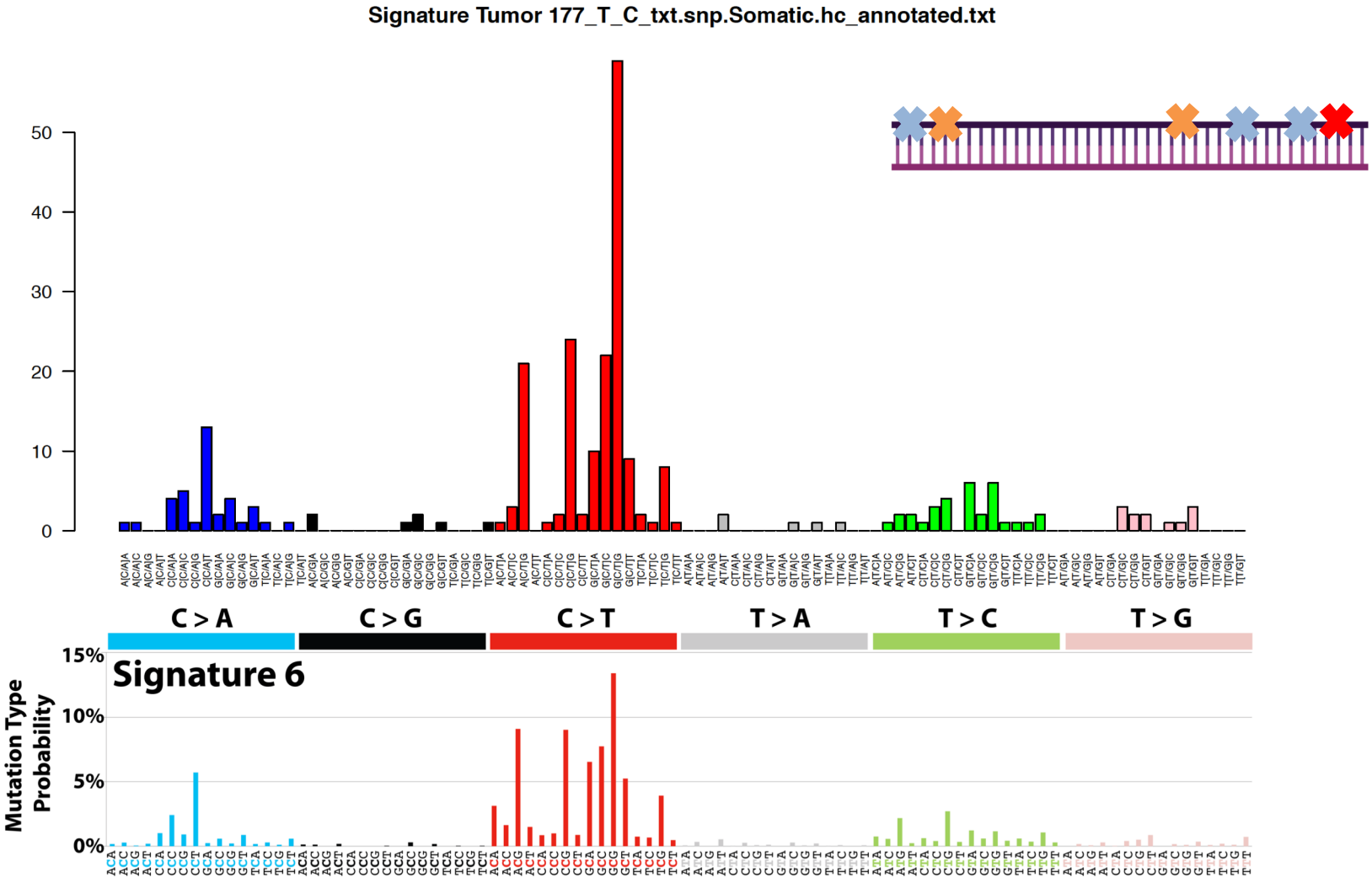
Keratolythic winter erythema,
Ngcungcu et al., AJHG, 2017



progeria-like envelopathy,
Marbach et al., AJHG, in press

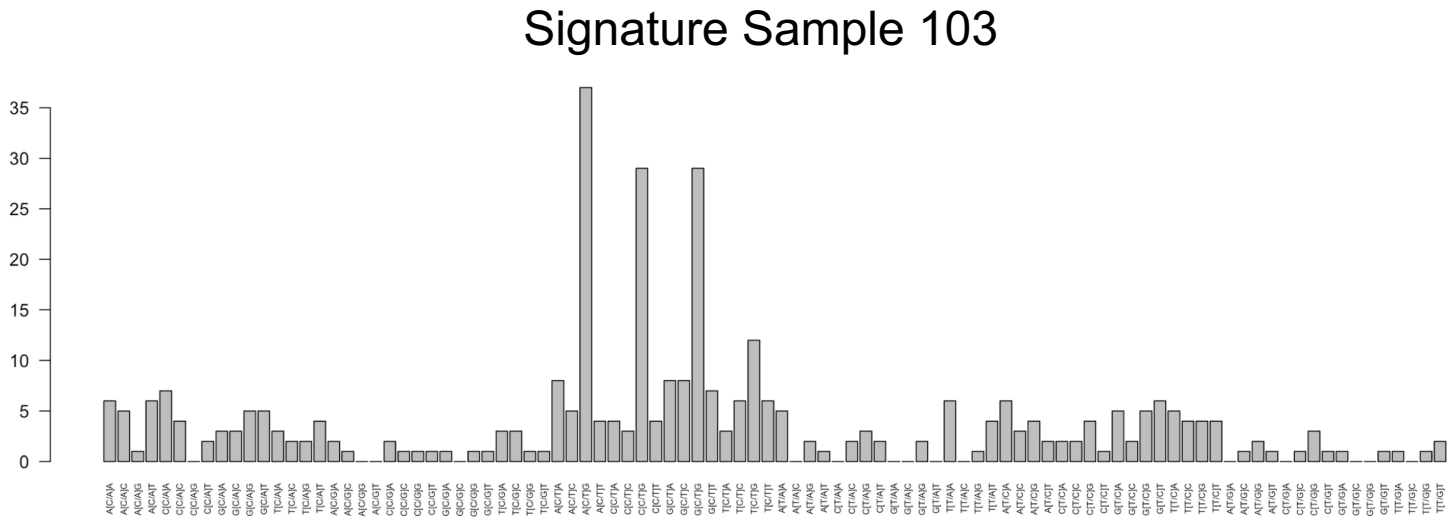
One variant, severe consequence

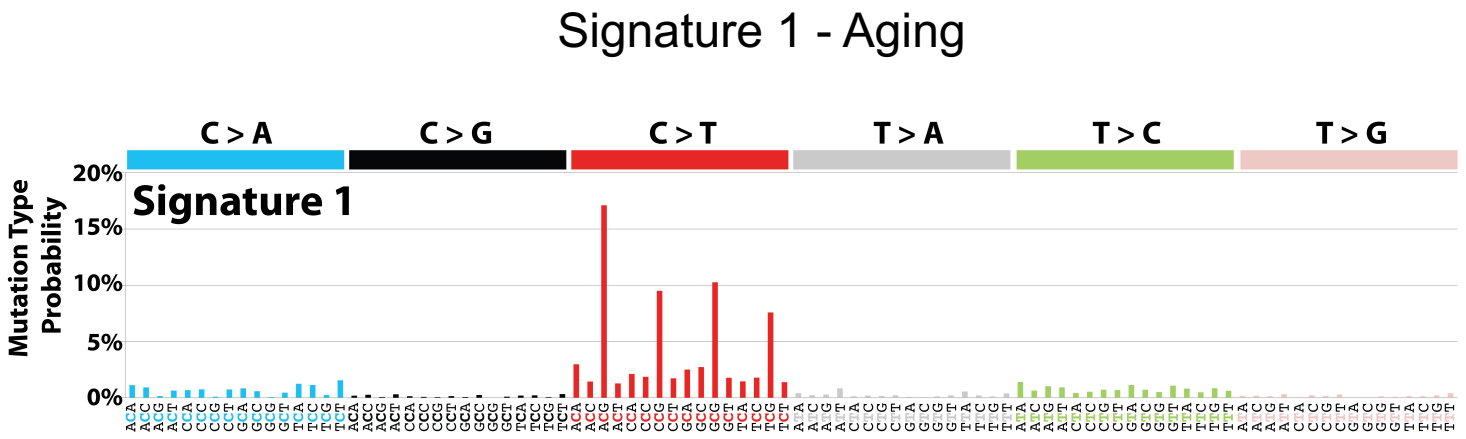
Mutational signatures and pattern classification in colon cancer: therapy – PD-1 blockade



Signature fit obtained using either Non-negative matrix factorization (NNMF) or correlation

Mutational signatures and pattern classification in colon cancer: quite likely no effective therapy





Conclusions and future work

1. The most accurate genetic variant calls can be obtained using TensorFlow based tool – DeepVariant.

2. We have discovered novel causes of several monogenic diseases (progeria, skin disorder and immunodeficiency syndrome) using genome sequencing data and big data analysis techniques. All results were verified using independent experimental methods.

3. Machine learning could help to choose the therapy (i.e. Olaparib and Niraparib for breast and ovarian cancer cases, PD-1 blockade immunotherapy in colon cancer).

Acknowledgements

Department of Clinical Science, University of Bergen, Norway:
Vidar M. Steen
Gunnar Houge
Rita Holdhus
Hans-Richard Brattbakk
Torunn Fiskerstrand

Computational Biology Unit, University of Bergen, Norway:
Inge Jonassen
Kjell Petersen
Oskar Vidarsson

Medical University of Gdańsk, Poland
Anna Supernat
Krzysztof Pastuszak

Memorial Sloan Kettering Cancer Center, New York, USA
Dominik Głodzik

MNM Diagnostics
Paweł Zawadzki
Paweł Sztromwasser

Many clinical collaborators

Thank you for your attention