LEIDEN UNIVERSITY MEDICAL CENTER

Research selection

Jeroen F.J. Laros
Leiden Genome Technology Center
Department of Human Genetics
Center for Human and Clinical Genetics

Introduction

About me

The last decade:

- Master Computer Science.
- Ph.D. Mathematics and Natural Sciences.
- Post-doctoral researcher LUMC.
- Senior researcher.
    - Coördinator Bioinformatics LGTC.

Introduction

About me

The last decade:

- Master Computer Science.
- Ph.D. Mathematics and Natural Sciences.
- Post-doctoral researcher LUMC.
- Senior researcher.
    - Coördinator Bioinformatics LGTC.

Currently active in:

- Variant databases / formal descriptions.
- Next Generation Sequencing.
- Metagenomics.
- Forensics.

Introduction

Activities

Collaborations:

- Within the LUMC: Human Genetics, Clinical Genetics, Forensic lab, Medical Microbiology, Dermatology, Hematology, ICT.
- Leiden University: Leiden Institute of Advanced Computer Science, Faculty of Social Sciences.
- Hogeschool Leiden.
- SURFSara.
- Commercial: GenomeScan, BaseClear, PhenoSystems.

Introduction

Activities

Other activities:

- Communication with external partners.
- Innovation / research and development.
- Implementation in diagnostics.
- Design and policy HPC infrastructure.
- Design and policy good research practice.
- Algorithm design.
- Audits.
- Education.

Next generation sequencing

Sequencers

Figure 1: HiSeq 2500.

Figure 2: Ion proton.

Next generation sequencing

Next generation sequencing data

```
1   @SGGPP:4:101
2   TTCGGGGGCTGGCAAATCCACTTCCGTGACACGCTACCATTCGCTGGTGGT
3   +
4   -'+4589,53330-0&07+03:54/2362-+.488587>@/25440++0(+
5   @SGGPP:4:102
6   CGGTAAACCACCCTGCTGACGGAACCCTAATGCGCCTGAAAGACAGCGTTC
7   +
8   34/--0'+.000(.55:;:99(0(+2(22(0316;185;;0;:<<>=AA59
9   @SGGPP:4:106
10  TCGTTAACGACTTTGTTCGCCACCGCAACCGCCTGTTTCGGGTCACAGGCA
11  +
12  09875;5?<;?@A4?B:BBB<AA>CCC>C>BB0.->=0488+3444:@5@<
13  @SGGPP:4:112
14  TTGATGAATATATTATTTCAGGGAATAATTATGACACCTTTAGAACGCATT
15  +
16  70<<@::5:<;==7;>>/79<:.:494.8(,,8:753/5@5??C>B???B7
```

Listing 1: A FastQ file.

Next generation sequencing

Data analysis

Very diverse.

Align to a reference genome (resequencing):

- Variant detection.
- Phylogenetic reconstruction.

Next generation sequencing

Data analysis

Very diverse.

Align to a reference genome (resequencing):

- Variant detection.
- Phylogenetic reconstruction.

Or to multiple references:

- Antibiotic resistance testing.

Next generation sequencing

Data analysis

Very diverse.

Align to a reference genome (resequencing):

- Variant detection.
- Phylogenetic reconstruction.

Or to multiple references:

- Antibiotic resistance testing.

De novo assembly:

- First step to make a reference genome.
- Finding large rearrangements.

Next generation sequencing

Metrics for NGS data files

Comparing k-mer profiles.

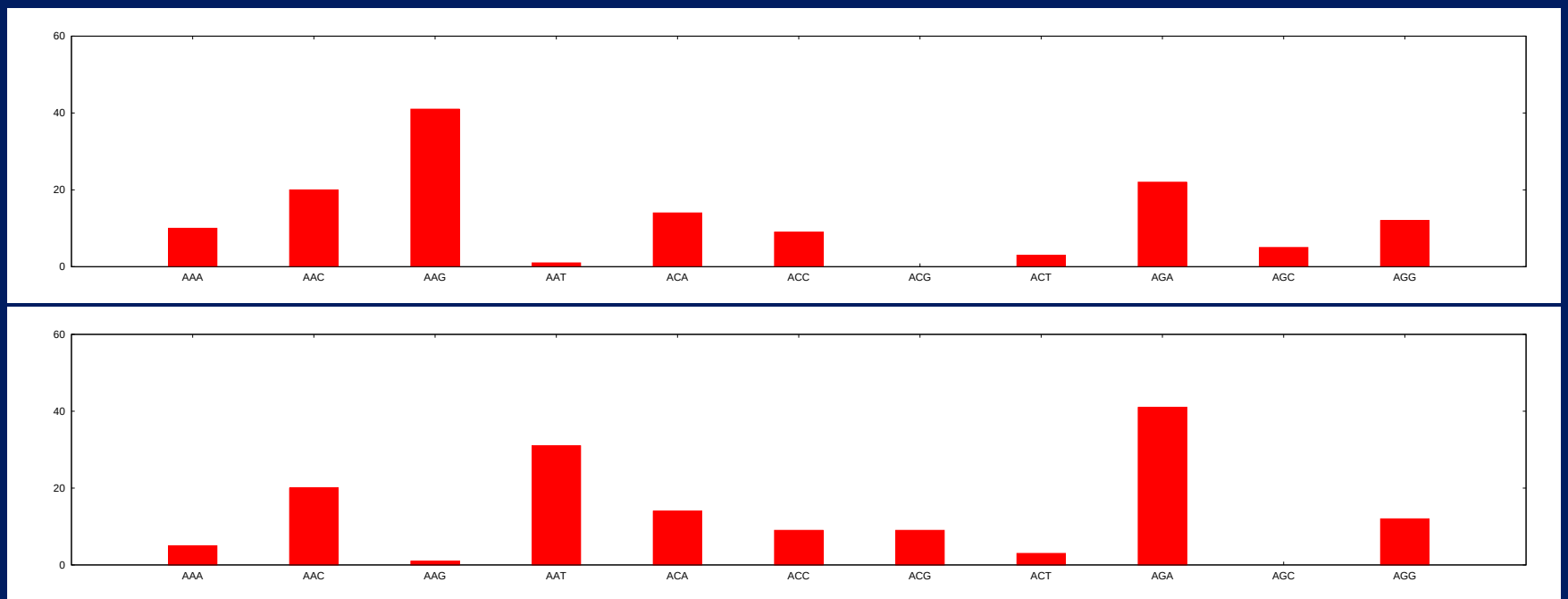

Figure 3: Two k-mer profiles.

Next generation sequencing

Forensics

STR profiling:

- Look deeper into STRs by using sequencing.
- Semi-global alignment of flanking sequences.
- Regular expressions for known alleles.
- Classification of new alleles.

Next generation sequencing

Forensics

STR profiling:

- Look deeper into STRs by using sequencing.
- Semi-global alignment of flanking sequences.
- Regular expressions for known alleles.
- Classification of new alleles.

SNP profiling:

- Highly variable regions in a certain population.
- Easier to work with than with STRs.

Variant databases and formal descriptions

Effect prediction

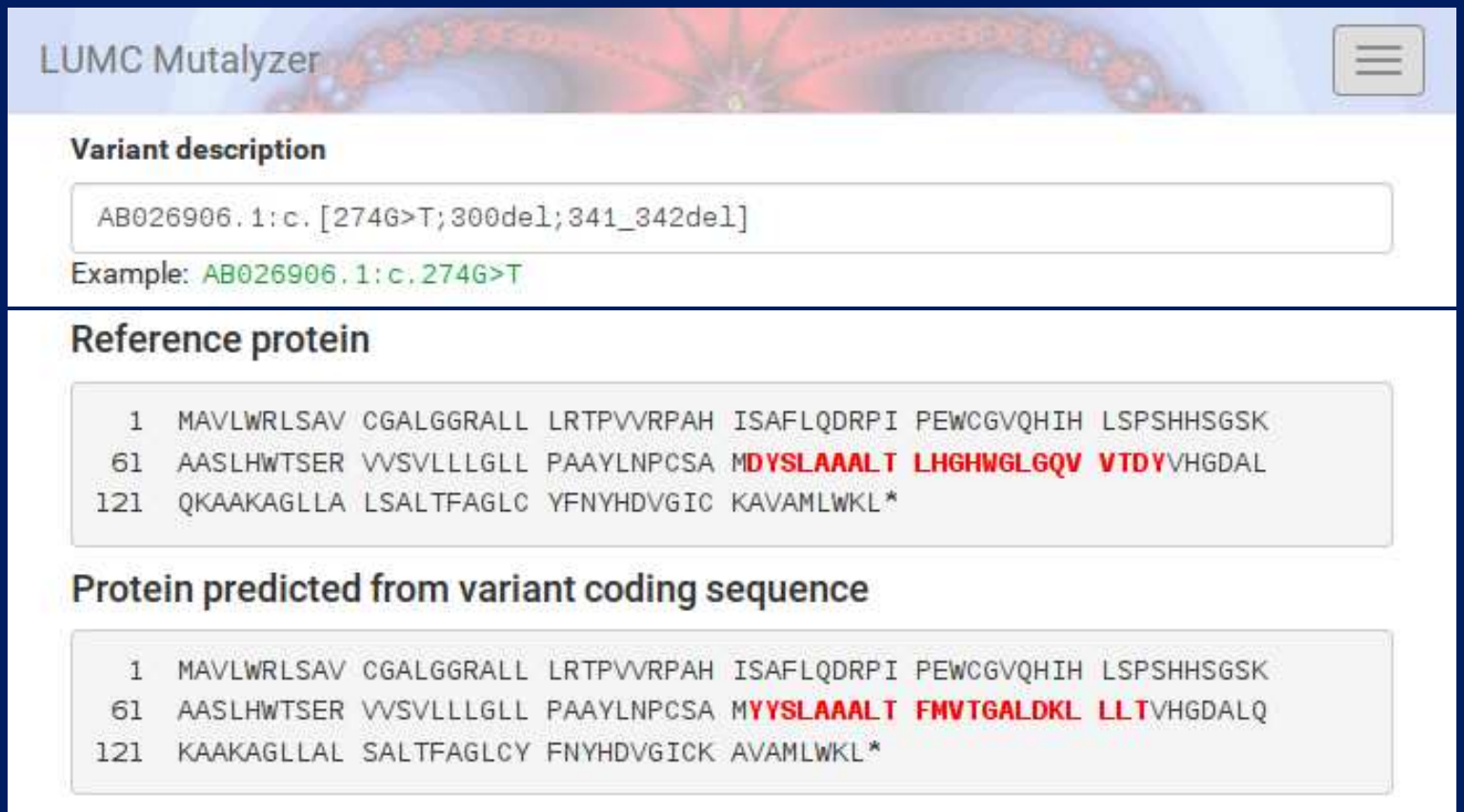

Figure 4: Mutalyzer.

Variant databases and formal descriptions

A "human" way of finding a description

Observation:

- There is always a default way of describing a variant (`delins`).
- A `delins` may be split in smaller parts.

Variant databases and formal descriptions

A "human" way of finding a description

Observation:

- There is always a default way of describing a variant (`delins`).
- A `delins` may be split in smaller parts.

Outline:

- Find the *area of change*.
- Describe this as a `delins`.
- Find the largest overlap in this area of change, splitting the area in two.
- Describe the two sub areas, and see whether this description is smaller than the one we have.

Variant databases and formal descriptions

Outline of the algorithm

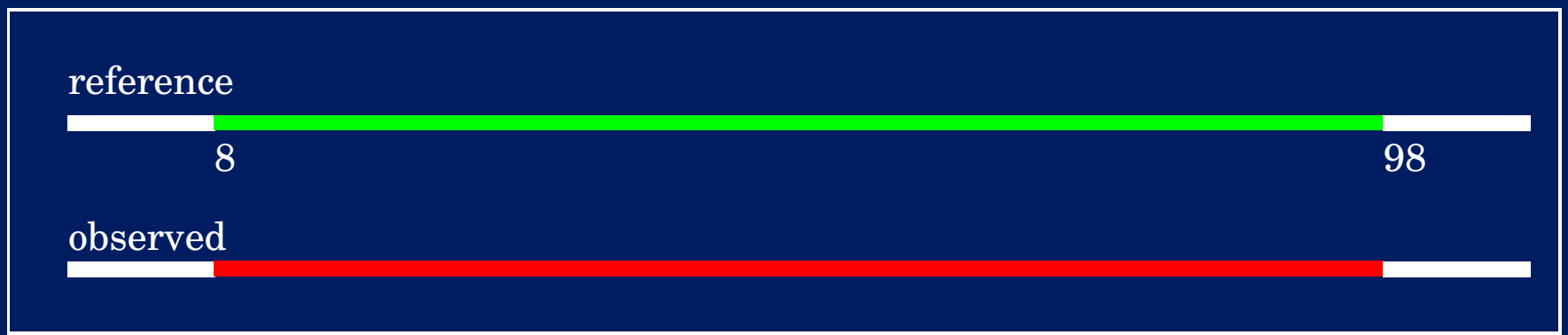

Figure 5: How would a human do it?

```
8_98delinsAGATGCGATAGATTAGCTATATAGGATCG...
```

Variant databases and formal descriptions

Outline of the algorithm

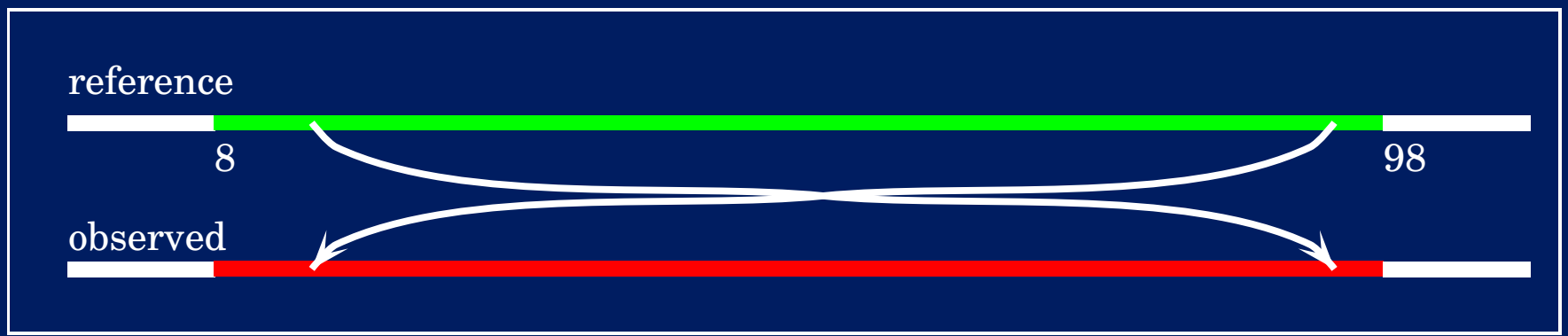

Figure 5: How would a human do it?

`8_98delinsAGATGCGATAGATTAGCTATATAGGATCG...`

Variant databases and formal descriptions

Outline of the algorithm

Figure 5: How would a human do it?

```
8_98delinsAGATGCGATAGATTAGCTATATAGGATCG...
[8_12delinsAGATG;13_96inv;97_98delinsTG]
```

Variant databases and formal descriptions

Outline of the algorithm

Figure 5: How would a human do it?

```
8_98delinsAGATGCGATAGATTAGCTATATAGGATCG...
[8_12delinsAGATG;13_96inv;97_98delinsTG]
[8G>A;12C>G;13_96inv;97_98delinsTG]
```

Variant databases and formal descriptions

Finding common sub strings

How would a computer do it?

	A	T	G	A	G	C	G
A	1	0	0	1	0	0	0
T	0	2	0	0	0	0	0
C	0	0	0	0	0	1	0
A	1	0	0	1	0	0	0
G	0	0	1	0	2	0	1
C	0	0	0	0	0	3	0
A	1	0	0	1	0	0	0

Table 1: LCS dynamic programming.

Variant databases and formal descriptions

Accuracy vs. speed

AGAGGACG $\Rightarrow$ **AG AG GA CG**

GAGGACA $\Rightarrow$ **GA AG GG GA AC CA**

Variant databases and formal descriptions

Accuracy vs. speed

AGAGGACG $\Rightarrow$ **AG AG GA CG**
GAGGACA $\Rightarrow$ **GA AG GG GA AC CA**

	A **G**	**A** **G**	**G** **A**	**C** **G**
GA	0	0	1	0
AG	1	1	0	0
GG	0	0	0	0
GA	0	0	2	0
AC	0	0	0	0
CA	0	0	0	0

Table 2: Rough method to find large strings.

Variant databases and formal descriptions

Accuracy vs. speed

AGAGGACG $\Rightarrow$ **AG AG GA CG**
GAGGACA $\Rightarrow$ **GA AG GG GA AC CA**

	A **G**	**A** **G**	**G** **A**	**C** **G**
GA	0	0	1	0
AG	1	1	0	0
GG	0	0	0	0
GA	0	0	2	0
AC	0	0	0	0
CA	0	0	0	0

Table 2: Rough method to find large strings.

Variant databases and formal descriptions

Accuracy vs. speed

	A A G C	
	G G A G	
GA	0 0 1 0	
AG	1 1 0 0	
GG	0 0 0 0	
GA	0 0 2 0	
AC	0 0 0 0	
CA	0 0 0 0	

Table 3: "Zoom out" $k = 2$.

	A G	
	G G	
	A A	
GAG	0 0	
AGG	0 0	
GGA	0 1	
GAC	0 0	
ACA	0 0	

Table 4: "Zoom out" $k = 3$.

Variant databases and formal descriptions

Accuracy vs. speed

	A A G C
	G G A G
GA	0 0 1 0
AG	1 1 0 0
GG	0 0 0 0
GA	0 0 2 0
AC	0 0 0 0
CA	0 0 0 0

Table 3: "Zoom out" $k = 2$.

	A G
	G G
	A A
GAG	0 0
AGG	0 0
GGA	0 1
GAC	0 0
ACA	0 0

Table 4: "Zoom out" $k = 3$.

We find all common sub strings larger than k.

Variant databases and formal descriptions

Accuracy vs. speed

	A A G C
	G G A G
GA	0 0 1 0
AG	1 1 0 0
GG	0 0 0 0
GA	0 0 2 0
AC	0 0 0 0
CA	0 0 0 0

Table 3: "Zoom out" $k = 2$.

	A G
	G G
	A A
GAG	0 0
AGG	0 0
GGA	0 1
GAC	0 0
ACA	0 0

Table 4: "Zoom out" $k = 3$.

We find all common sub strings larger than k.
The length of these strings are at least ℓk and at most $\ell k + (k - 1)$ long.

Questions?

Acknowledgements:

Martijn Vermaat
Jonathan Vis
Lusine Khachatryan
Yahya Anvar
Kristiaan van der Gaag
Peter de Knijff
Johan den Dunnen