Utrecht University

Faculty of Science
Department of Information and Computing Science
Master of Business Informatics

Seminar Medical Informatics

Bioinformatics and Precision Medicine

Verónica Burriel Coll

v.burriel@uu.nl

19 March 2019

Agenda for today

Bioinformatics and Precision Medicine

- What is Bioinformatics? What is Precision Medicine?
- Genomic basis
- Bioinformatics pipeline
- Variant Calling Format (VCF) file
- Importance of bioinformatics
- Precision Medicine
- Personal genetics testing

Assignment for next March 25th: Bioinformatics Workshop

Rescheduling Data Science in Healthcare Workshop

Food for further thoughts

What is Bioinformatics?

"It is the emerging field that deals with the application of computers to the collection, organization, analysis, manipulation, presentation, and sharing of biologic data" (NCBI)

What is Precision (or personalized) Medicine?

"An emerging approach for disease treatment and prevention that takes into account individual variability in genes, environment, and lifestyle for each person" (NHI)

How these terms interrelate?

Bioinformatics

- **Biology**, **computer science** and **information technology** merge to form a single discipline

- Makes use of computer science to develop algorithms for facilitating the development and testing of biological hypothesis (e.g.):

  - Finding the genes responsible for certain disease

  - Predicting the structure or function of proteins and other biological molecules in specific pathways

  - Defining the function of genes

  - Examining evolutionary relationships

Bioinformatics

- Areas of biology where bioinformatics is being applied:

GENOMICS PROTEOMICS TRANSCRIPTOMICS METABOLOMICS

PHARMACOGENOMICS METAGENOMICS

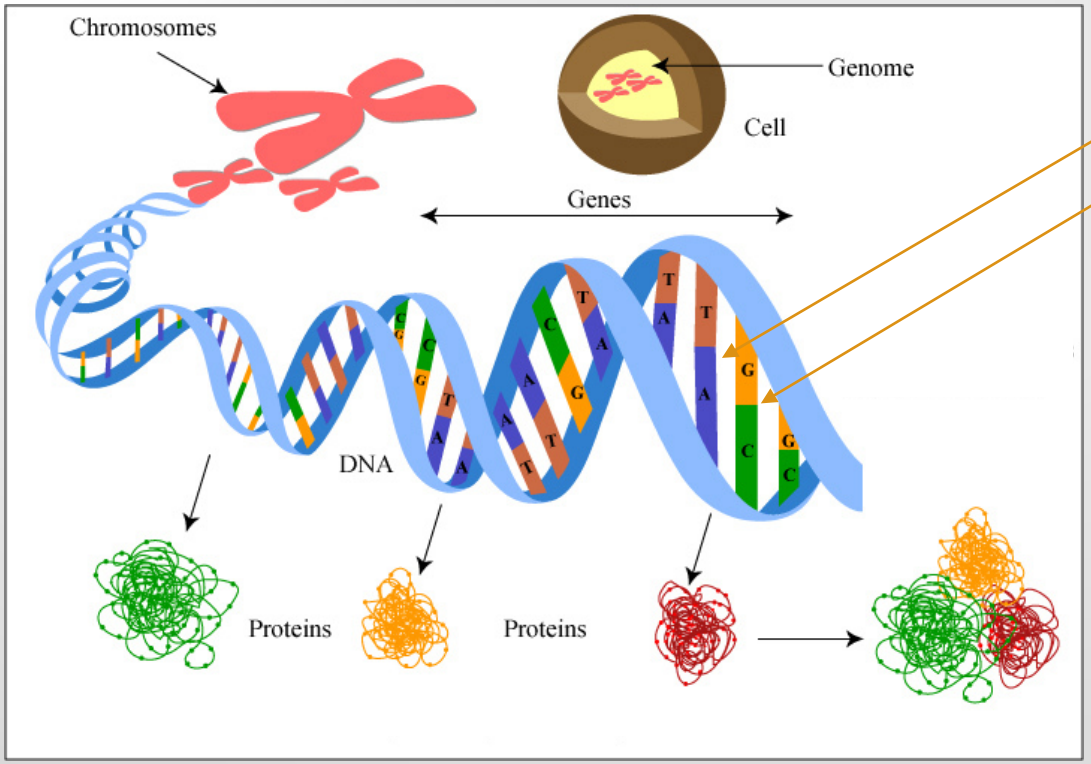
- **Translational bioinformatics:** Bioinformatics for human health

Genomic basis

Human body has 100 trillion cells

Each cell has 23 pairs of chromosomes

Double twisted helices of deoxyribonucleic acid (DNA)



Nucleotides
4 sugar-based blocks:
[A] Adenine
[T] Thymine
[C] Cytosine
[G] Guanine

Genes contain instructions for making proteins

Proteins act alone or in complexes to perform many cellular functions

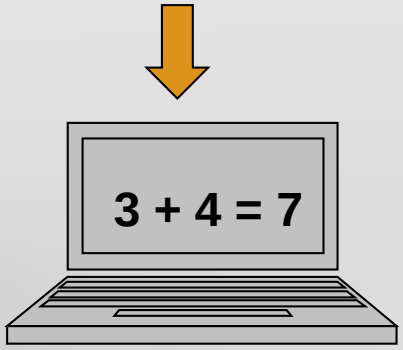
Genomic basis: how can we understand genome code?

00010011	00000111	00000011	00001000

Physical Level

ADD $7 $3 $8 ***Instruction Level***

Semantics: Add the values from the processor registers '3' and '7' and store the result in the register '8'



Representation Level

Genomic basis: how can we understand genome code?

AUG	GAA	CAC	GAC	GAG	UAA

Physical Level

START Glu His Asp Glu STOP

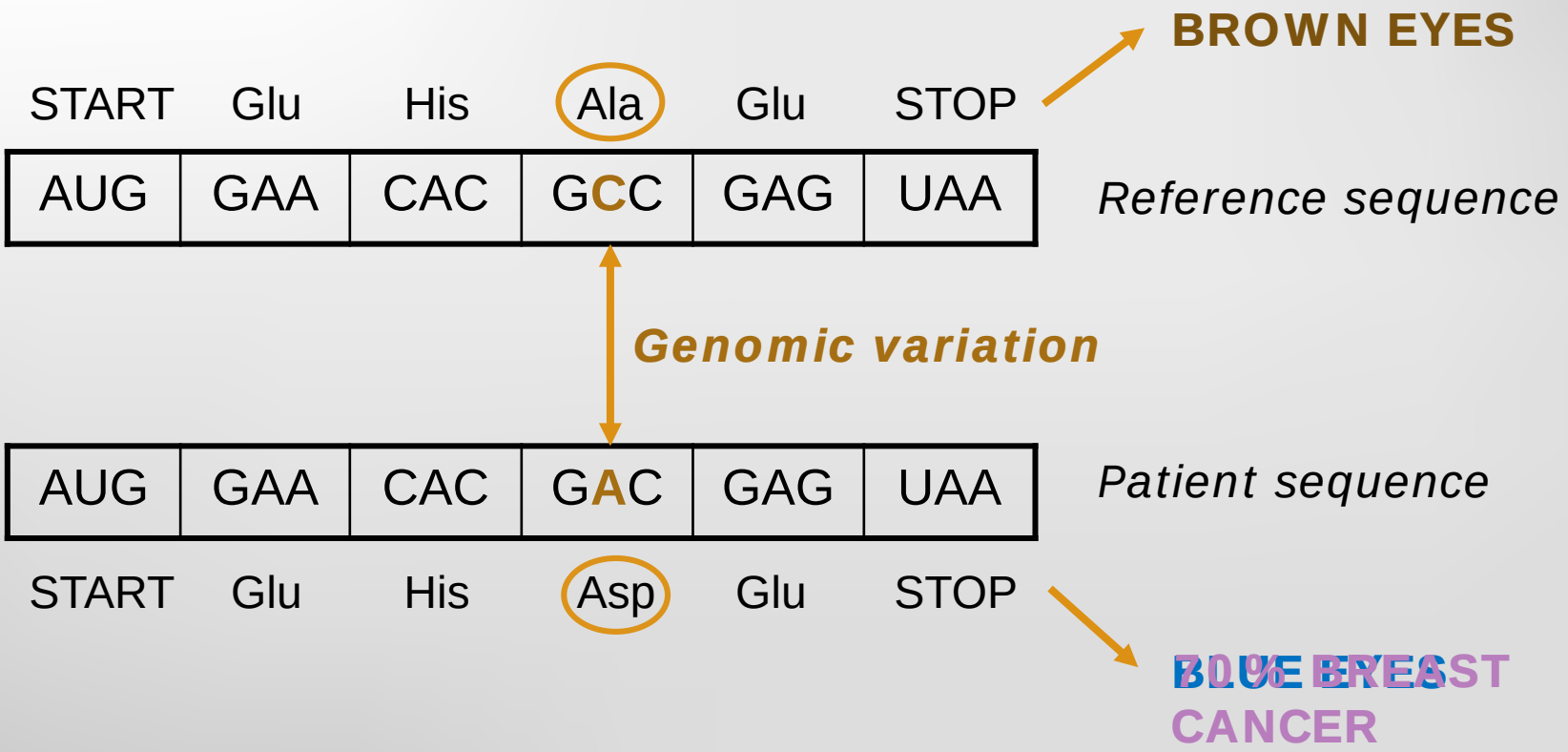
Instruction Level

Semantics: Process a protein with the four selected aminoacids

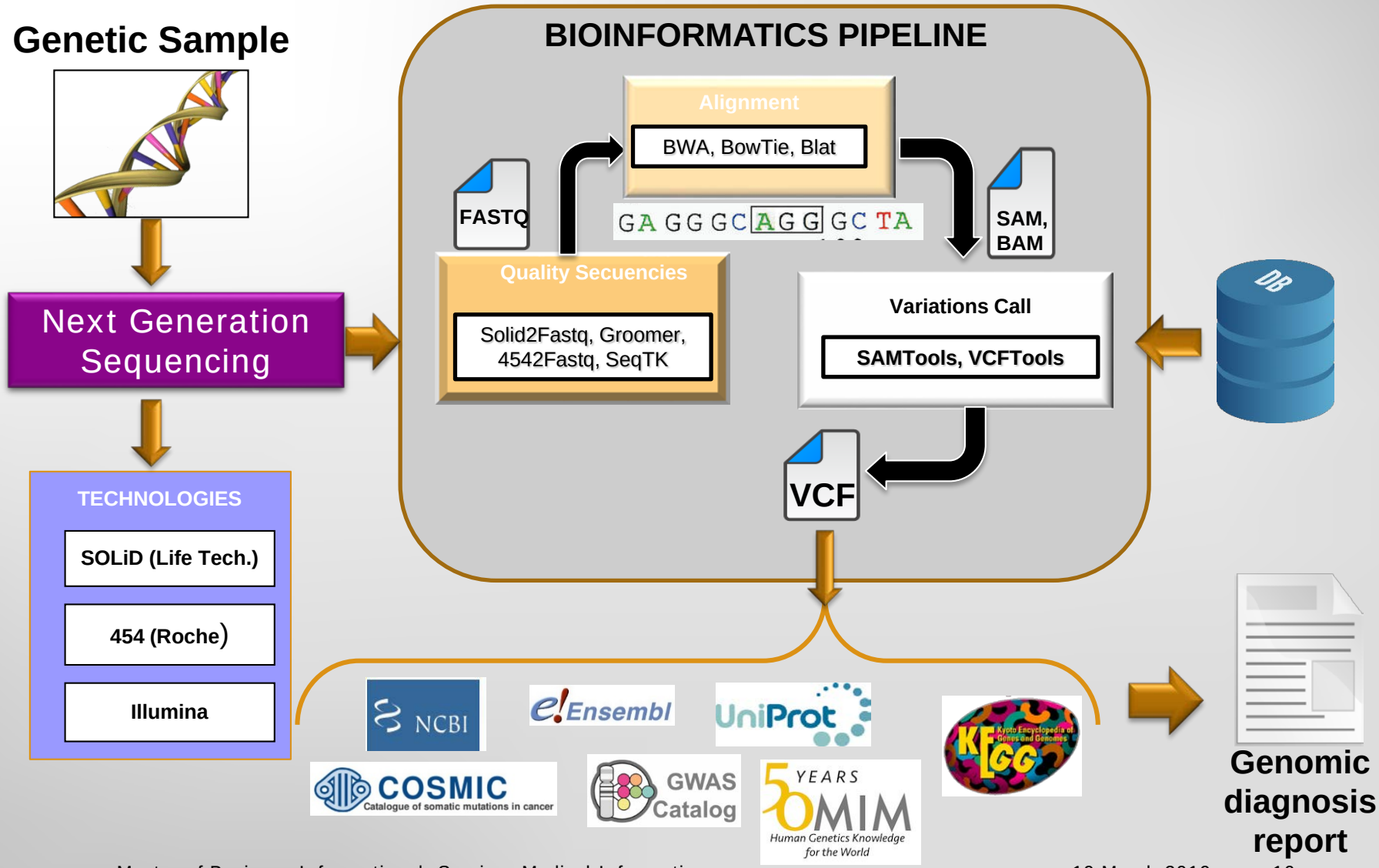
But…
Why?

Representation Level

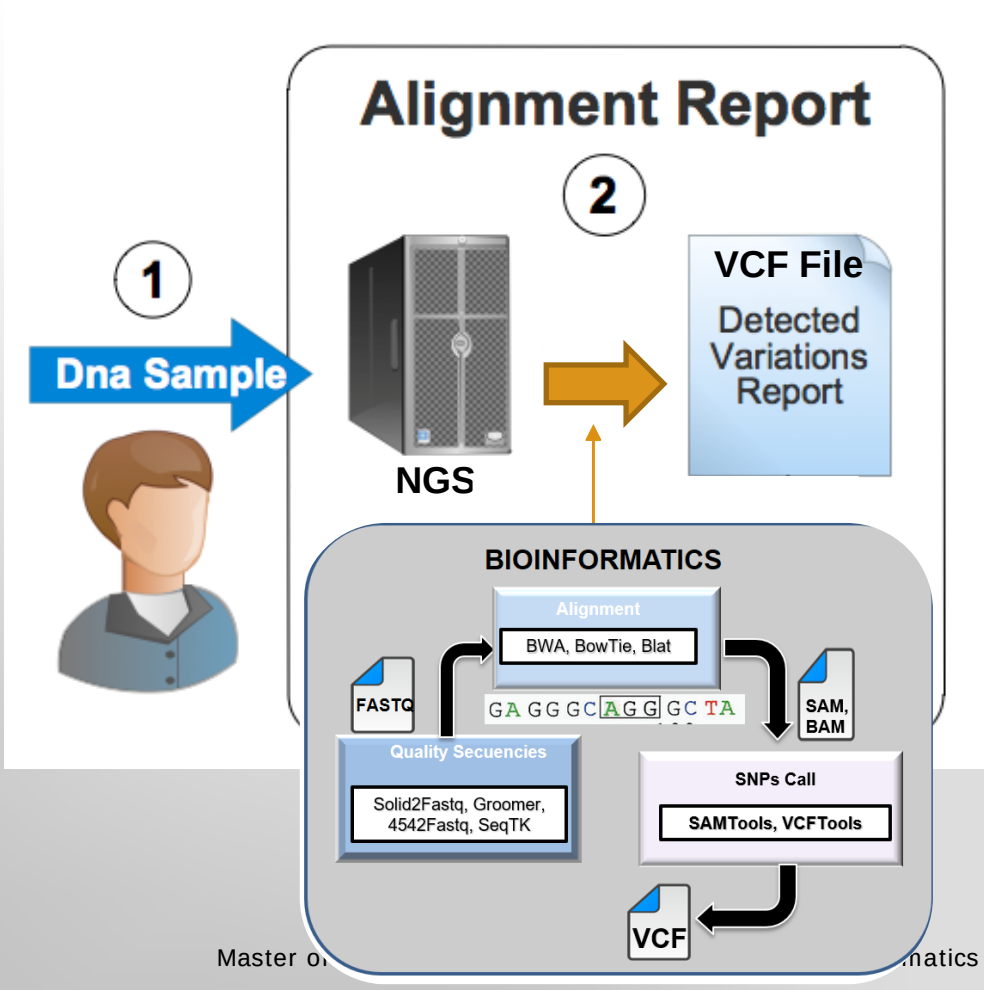
Genomic basis: how can we understand genome code?



Bioinformatics pipeline



Variation analysis process

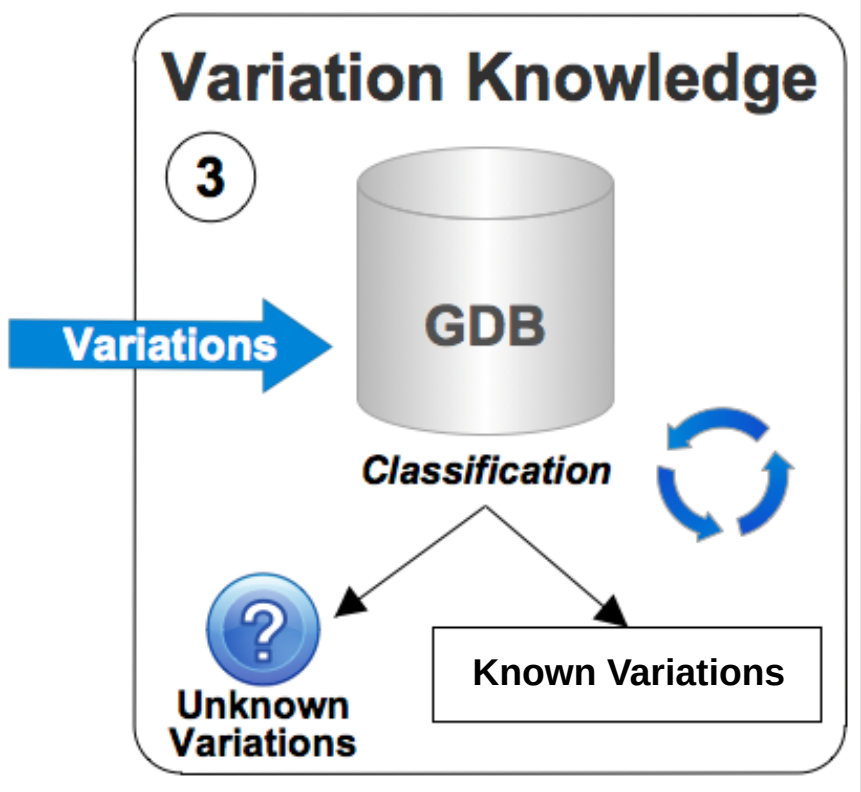


The Input of the process is introducing a DNA sample into a sequencing machine

The sequence obtained is aligned to a consensus reference sequence.

Each discovered difference is formalized as a record into VCF file

Variation analysis process

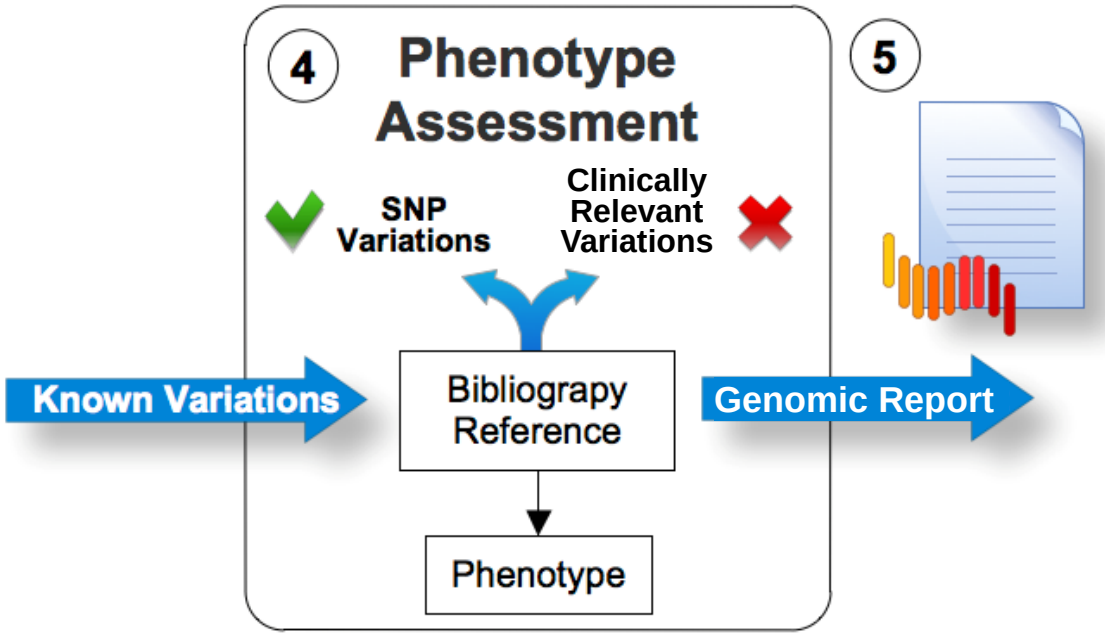


Founded variations are searched in several public genomic databases in order to find an associated phenotype

Known variations are classified into an specific type of sequence change: Insertion, Deletion, Indel, CNV, Inversion

Unknown variations are classified as “Silent” or “Non-Silent” using a prediction software

Variation analysis process



In order to assess the phenotype of an specific variation, a research publication is required

Variations with a pathogenic phenotype are classified as "Clinically relevant", jointly with "Non-silent" unknown variations

Finally, the information is gathered in a report to support the clinical diagnosis

Variant Calling Format (VCF) file

- Created by **1000 Genomes** and now updated and maintained by Global Alliance for Genomics and Health
- Need of store **information about variations** included in SAM/BAM files
- Allows to store information about **precise variations** such as SNPs and indels and also **imprecise variations** found in sequenced samples.
- Big **flexibility**, allowing to extend the file using metadata

Variant Calling Format (VCF) file

FILE FORMAT

RECOMMENDED DATA

REFERENCE SEQUENCE

INFO

FILTER

FORMAT

HEADER

VARIATIONS DATA

```
##fileformat=VCFv4.3
##fileDate=20170805
##source=myImputationProgramV3.1
##reference=file:///seq/references/1000GenomesPilot-NCBI36.fasta
##contig=<ID=20,length=62435964,assembly=B36,md5=f126cdf8a6e0c7f379d618ff66beb2da,species
="Homo sapiens",taxonomy=x>
##phasing=partial
##INFO=<ID=NS,Number=1,Type=Integer,Description="Number of Samples With Data">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total Depth">
##INFO=<ID=AF,Number=A,Type=Float,Description="Allele Frequency">
##INFO=<ID=DB,Number=0,Type=Flag,Description="dbSNP membership, build 129">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
##FILTER=<ID=q10,Description="Quality below 10">
##FILTER=<ID=s50,Description="Less than 50% of samples have data">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
#CHROM  POS     ID        REF ALT    QUAL FILTER INFO                              FORMAT      NA00001
20 14370   rs6054257 G A 29 PASS NS=3;DP=14;AF=0.5;DB;H2 GT:GQ:DP 0|1:48:1
20 17330   . T A 3 q10 NS=3;DP=11;AF=0.017 GT:GQ:DP 0|1:49:3
20 1110696 rs6040355 A G,T 67 PASS NS=2;DP=10;AF=0.333,0.667;DB GT:GQ:DP 1|2:21:6
```

Variant Calling Format (VCF) file

Header:

Define the content of each column of data

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	NA00001

- **CHROM:** Chromosome
- **POS:** Chromosomic position where the variation has been found
- **ID:** Variation identifier in a genomic database (e.g. dbSNP)
- **REF:** Reference base(s) in genomic reference sequence
- **ALT:** Alternative base(s) found in the sample
- **QUAL:** Phred-scaled quality score for the variation
- **FILTER:** Filter status. PASS if this variation has passed all filters
- **INFO:** (optional) Additional information about the variation
- **FORMAT:** (optional) Additional information about genotype data of a sample
- **SampleID:** Sample identifier assigned by the laboratory

Variant Calling Format (VCF) file

Variations data

Example	Sequence	Alteration
Ref	a t C g a	C is the reference base
1	a t G g a	C base is a G in some individuals
2	a t - g a	C base is deleted w.r.t. the reference sequence
3	a t CAg a	A base is inserted w.r.t. the reference sequence

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO
20	3	.	C	G	.	PASS	DP=100
20	2	.	TC	T	.	PASS	DP=100
20	3	.	C	CA	.	PASS	DP=100

Example	Sequence	Alteration
Ref	a t C g a	C is the reference base
1	a t G g a	C base is a G in some individuals
2	a t - g a	C base is deleted w.r.t. the reference sequence

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO
20	2	.	TC	TG,T	.	PASS	DP=100

Variant Calling Format (VCF) file

Variations data

Example	Sequence	Alteration
Ref	a t C g a	C is the reference base
1	a t - g a	C base is is deleted w.r.t. the reference sequence
2	a t -- a	C and G bases are deleted w.r.t. the reference sequence
3	a t CAg a	A base is inserted w.r.t. the reference sequence

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO
20	2	.	TCG	TG,T,TCAG	.	PASS	DP=100

Variant Calling Format (VCF) file

Variations data

How these variations would be represented in one VCF record?

Example	Sequence	Alteration
Ref	a t C g a	C is the reference base
1	a t T g a	C base is a T in some individuals

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO
20	3	.	C	T	.	PASS	DP=100

Example	Sequence	Alteration
Ref	a t C - - - g a	C is the reference base
1	a t C T A G g a	following the C base is an insertion of 3 bases

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO
20	3	.	C	CTAG	.	PASS	DP=100

Example	Sequence	Alteration
Ref	a T C G a	T is the (first) reference base
1	a T - - a	following the T base is a deletion of 2 bases

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO
20	2	.	TCG	T	.	PASS	DP=100

Example	Sequence	Alteration
Ref	a t c G C G - - a	G is the (first) reference base
1	a t c G - - - - a	following the G base is a deletion of 2 bases
2	a t c G C G C G a	following the G base is a insertion of 2 bases

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO
20	4	.	GCG	G,GCGCG	.	PASS	DP=100

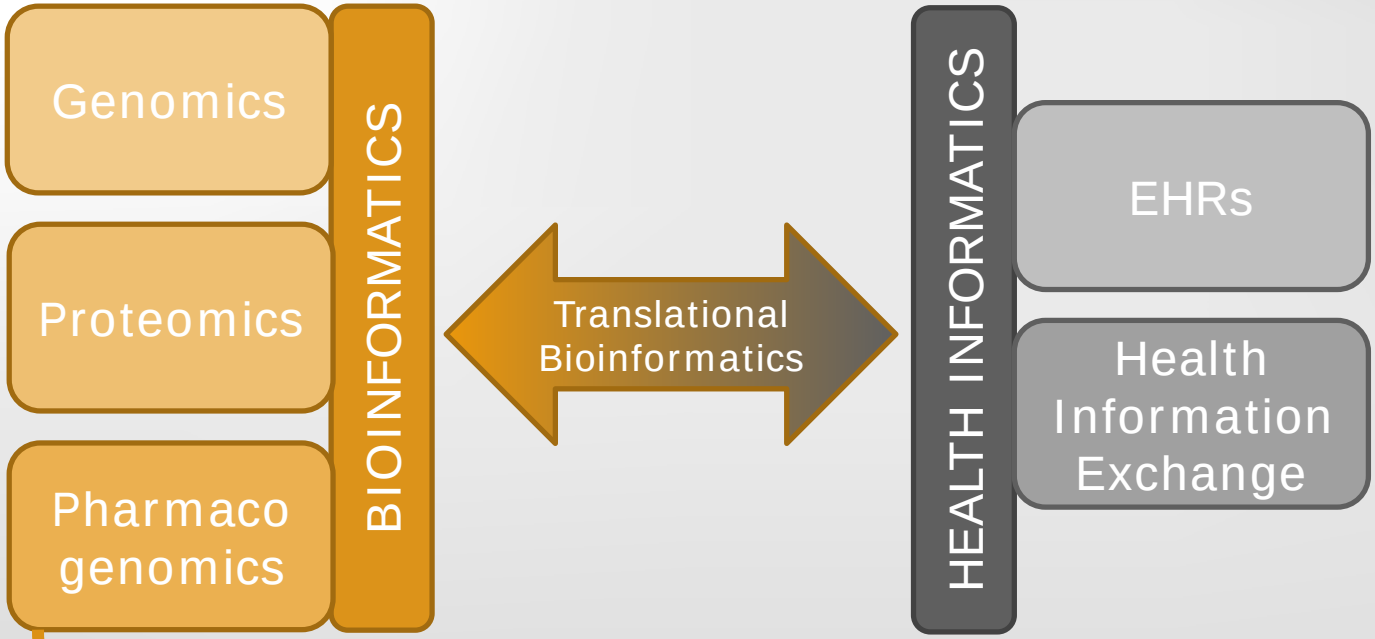
Variant Calling Format (VCF) file

INFO	**NS**, Number of samples	Number of samples with data
	DP, Read depth	Combined depth across samples
	AF, Allele Frequency	Allele frequency for each ALT allele in the same order as listed
FORMAT	**GQ**, Genotype quality	Conditional genotype quality, encoded as a phred-scaled quality $-10\log_{10} p$
	GT, Genotype	Genotype, encoded as allele values separated by either of / or |. The allele values are 0 for the reference allele, 1 for the first allele in ALT, 2 for the second and so on.
	DP, Read depth	Read depth at this position for this sample.
FILTER	##FILTER=<ID=s50,Description="Less than 50% of samples have data">	
	##FILTER=<ID=q10,Description="Quality below 10">	
	##FILTER=<ID=LowVariantFreq,Description="Low variant frequency < 0.20">	

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO
20	14370	rs6054257	G	A	29	PASS	NS=1;DP=14;AF=0.5;DB;H2

FORMAT	NA00001
GT:GQ:DP	0|1:29:14

Importance of bioinformatics



- New indications for an old drug
- New target for existing drugs
- Drugs to work better in certain patient groups
- Side effects of certain drugs in specific group of patients
- Decision support based on pharmacogenomics

Precision Medicine

Precision Medicine

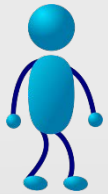
“An emerging approach for disease treatment and prevention that takes into account individual variability in genes, environment, and lifestyle for each person.“

- Enables physicians to tailor medical treatment for each patient
- Supports the development of molecularly targeted drugs based on biologic pathways
- Identifies at-risk populations for targeted prevention prior to disease onset

As health care becomes more expensive, there is greater interest in understanding which treatments work for which patients in which settings

Precision Medicine

The top ten highest-grossing drugs in the United States help between 1 in 25 and 1 in 4 of the people who take them



Abilify (Schizophrenia)	X 4
Nexium (Heartburn)	X 24
Humira (Arthritis)	X 3
Crestor (High Cholesterol)	X 19
Cymbalta (Depression)	X 7
Advair Diskus (Asthma)	X 19
Enbrel (Psoriasis)	X 3
Remicade (Crohn's disease)	X 3
Copaxone (Multiple Sclerosis)	X 15
Neulasta (Neutropenia)	X 12

Schork, Nature 2015, 520 (7549)

Precision Medicine: Key drivers

Genomic Sequencing Technologies

- Rapid drop in sequencing costs
- First human genome cost $2.7 billion
- Currently ~$1000/genome with the promise of the $100 genome in near future
- It takes ~3 days with HiSeq XTen or even 26h (still in research)

Genomic Data and Analytic Capabilities

- Creation of large genomic datasets
- Advanced analytics to identify novel disease associations and treatment strategies

Precision Medicine: Applications of genomics

Patients

- Identification of disease risk / susceptibility to support preventive medical care
- Targeted prescribing to increase adherence, improve drug response and reduce adverse events

Providers

- Data driven clinical decision support tools based on individual patient profiles
- Pharmacogenomic-informed prescribing using genetic profiles and companion diagnostics

Payers

- Effective preventive medical care to address disease risks before onset of chronic disease
- Targeted and effective treatment plans to improve patient care while reducing costs

Life Sciences

- Discovery of novel drug targets
- Improved clinical trial recruitment / execution
- Drug repurposing / repositioning
- Genomic diagnosis development

Personal Genetics Testing

ancestry™ — Determine ethnicity estimates and identify remote cousins

23andMe — Direct to consumer online genetic testing. Provides insights on your genetic health risks, carrier status, traits, wellness and ancestry

myriad® WHEN DECISIONS MATTER — Genetic testing for cancers with a hereditary component. Only for professional use

Assignment for next March 25th: Bioinformatics Workshop

Assigned students:

- Amber Brauer
- Hielke Koopstra
- Antoine Lyonnet
- Charles Vernerey
- Alquin Nooteboom (Telemedicine)

Each assigned student:

1. **Select a paper** of aprox. 8 pages about a Bioinformatics solution and send it to v.burriel@uu.nl **before Wednesday at 13.00.** During the afternoon all selected papers will be published on course's website.

2. **Prepare a presentation** of **7/8 minutes** about the paper and include some questions (at least 2) at the end of the presentation to challenge the audience and activate the discussion.

3. Join with the other assigned students and **prepare 1 or 2 group activities** to make during the last 30 minutes of the session. These activities should be related to the solutions presented.

Assignment for next March 25th: Bioinformatics Workshop

Each no-assigned student:

1. **Read all the selected papers** and **prepare some questions or comments** (at least 2) per paper to discuss them after the presentation. Try to be critical and/or creative.

2. Send the questions/comments using this form **before Monday**
https://goo.gl/forms/K69vlahNqxzFe1ZG3

Rescheduling Data Science in Healthcare Workshop

Options:

A. Tue. 26/03/2019 at 13:15 (before the Quiz)

B. Wed. 20/03/2019 at 9:00

C. Wed. 20/03/2019 at 11:00

D. Wed. 27/03/2019 at 9:00

E. Wed. 27/03/2019 at 11:00

Food for further thought: Bioinformatics and Precision Medicine

There is nothing more personal than your genome

Dawn Barry | TEDxSanDiego

https://youtu.be/M3SLHhWYxiY

1. Would you like to have your genome analyzed? Why or why not?

2. What would you do if your genome says that you have a high probability of suffering a chronic disease, such as Parkinson, Alzheimer or Cancer?

3. What do you learn from her argumentation? Do you think that genomic analysis is a revolution into medicine and health care? Why?

See you next week!