



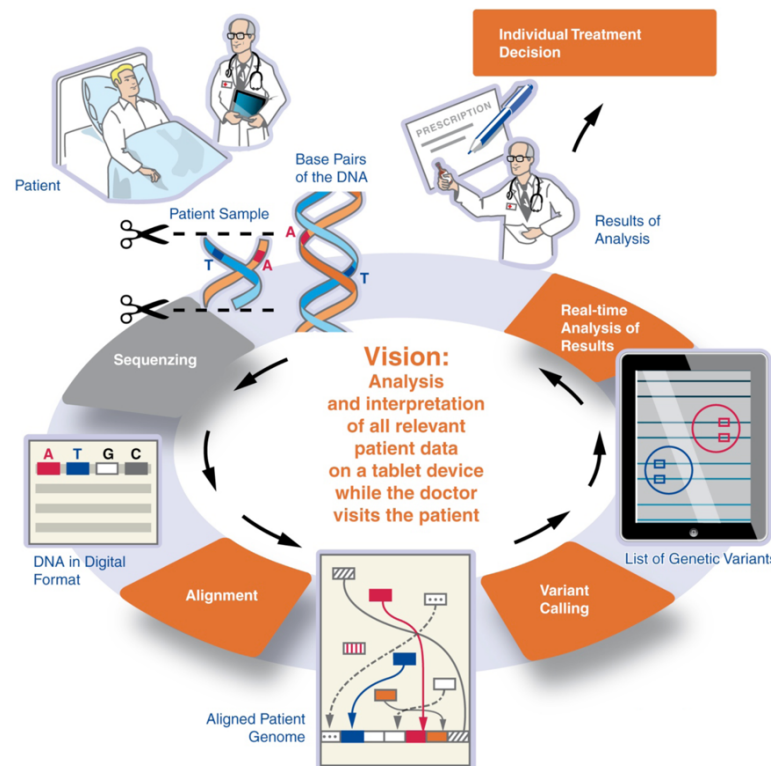
Identifying Changes: Variant Calling

Dr. Matthieu-P. Schapranow
Code of Life

November 2016

From Raw Genome Data to Analysis

- **DNA Sequencing:** Transformation of analogues DNA into digital format
- **Alignment:** Reconstruction of complete genome with snippets
- **Variant Calling:** Identification of genetic variants
- **Data Annotation:** Linking genetic variants with research findings

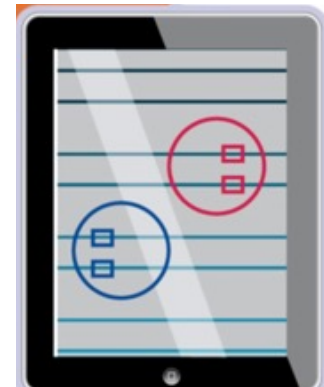


**Identifying Changes:
Variant Calling**

Schapranow, Nov 2016

Variant Calling Overview

- Purpose: Variant Calling := Detect variations within a genome
- Input:
 - Mapped DNA reads, i.e. output of alignment process
 - Reference genome
- Output: List of variants

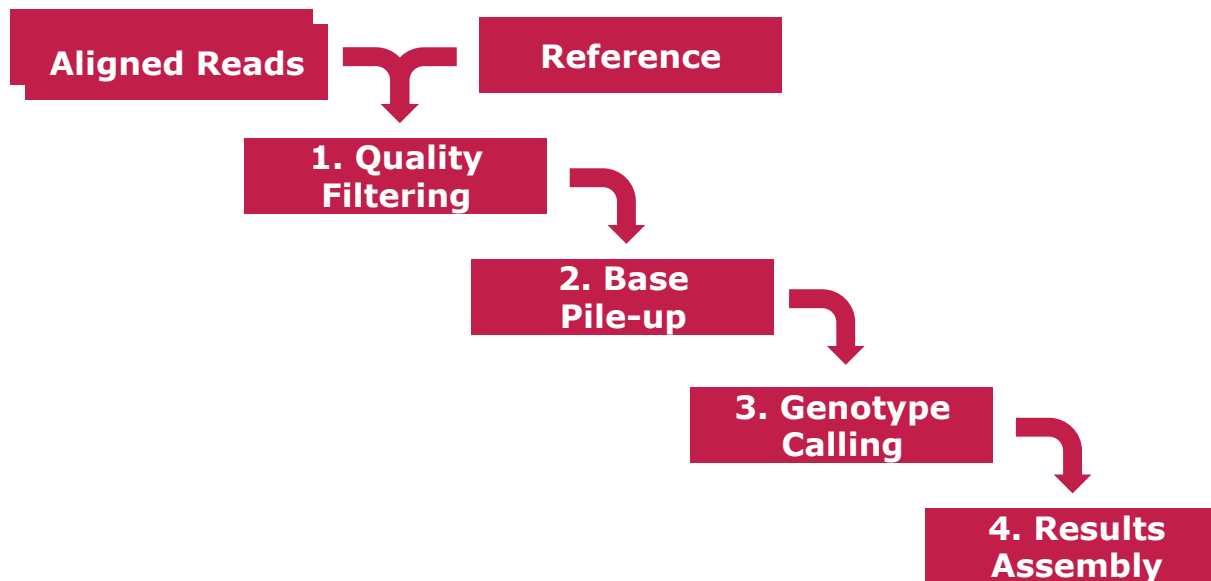


- Bear in mind:
 - Read depth at pos_i := Number of nucleotides storing information about pos_i

**Identifying Changes:
Variant Calling**

Schapranow, Nov 2016

Variant Calling Process



**Identifying Changes:
Variant Calling**

Schapranow, Nov 2016

1. Quality Filtering

- Extract locations from mapped reads where mapping issues were detected

- Reference

A	C	G	C	R	A	G	A	T	A
---	---	---	---	---	---	---	---	---	---

- Read

-	-	G	C	A	T	G	A	T	A
---	---	---	---	---	---	---	---	---	---

- CIGAR

2D	8M
----	----

Op	BAM	Description
M	0	alignment match (can be a sequence match or mismatch)
I	1	insertion to the reference
D	2	deletion from the reference
N	3	skipped region from the reference
S	4	soft clipping (clipped sequences present in SEQ)
H	5	hard clipping (clipped sequences NOT present in SEQ)
P	6	padding (silent deletion from padded reference)
=	7	sequence match
X	8	sequence mismatch

**Identifying Changes:
Variant Calling**

Schapranow, Nov 2016

2. Base Pile-up

- Reference (FASTA)
- Aligned read 1

A	C	G	C	R	A	G	A	T	A
		G	C	A	T	G	A	T	A

**Identifying Changes:
Variant Calling**

Schapranow, Nov 2016

2. Base Pile-up

■ Reference (FASTA)

■ Aligned read 1

...

■ Aligned read 8

■ Alleles

A	C	G	C	R	A	G	A	T	A
		G	C	A	T	G	A	T	A
A	C	G	C	G	T	G	A	T	A
A	T	G	C	G	T	G	A		
A	C	G	C	G	A	G			
	C	G	C	A	T	G	A	T	A
A	C	G	C	G	T				
			C	A	T	G	A	T	A
	C	G	C	A	T	G	A	T	A

SNP or Read Error

a	4	5	7	8	4	1	7	6	5	5
b	0	1	0	0	4	7	0	0	0	0



aa



ab



bb

**Identifying Changes:
Variant Calling**

Schapranow, Nov 2016

3. Genotype Calling

- Compute genotype probability to eliminate impact of noise and poor reading quality

- Bayes' Theorem:

- D: All observation about current position $\{D_1, \dots, D_n\}$
- G: Genotype probability to calculate $\{AA, AC, AG, AT, CC, CG, CT, GG, GT, TT\}$
- $P(G_i)$: Prior probability
- $P(D|G_i)$: genotype likelihood

$$\begin{aligned} P(G|D) &= \frac{P(D|G)P(G)}{P(D)} \\ &= \frac{P(D|G) P(G)}{\sum_{i=1}^n P(D|G_i) P(G_i)} \end{aligned}$$

- Heng Li (2011): "A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from seq. data"

**Identifying Changes:
Variant Calling**

Schapranow, Nov 2016

4. Results Assembly

- Results are stored in Variant Calling Format (VCF)
- VCF is extensible, i.e. can store an arbitrary number of attribute/value pairs
- Result consists of:

- Header defining attributes

##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">

- One entry per variant (fixed number of attributes)

CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	SAMPLE
chr7	140753336	rs113488022	T	A	61	PASS	NS=1	GT	0/1

**Identifying Changes:
Variant Calling**

Schapranow, Nov 2016

What is Next?



- How to assess variants and their impact?
- How to identify variants relevant for treatment?



**Identifying Changes:
Variant Calling**

Schapranow, Nov 2016

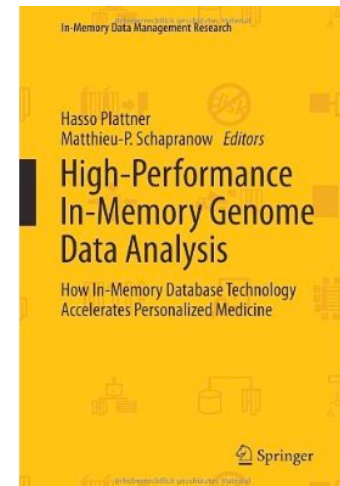
Keep in contact with us!



Dr. Matthieu-P. Schapranow
Program Manager E-Health & Life Sciences

Hasso Plattner Institute
August-Bebel-Str. 88
14482 Potsdam, Germany

schapranow@hpi.de
<http://we.analyzegenomes.com/>



**Identifying Changes:
Variant Calling**

Schapranow, Nov 2016