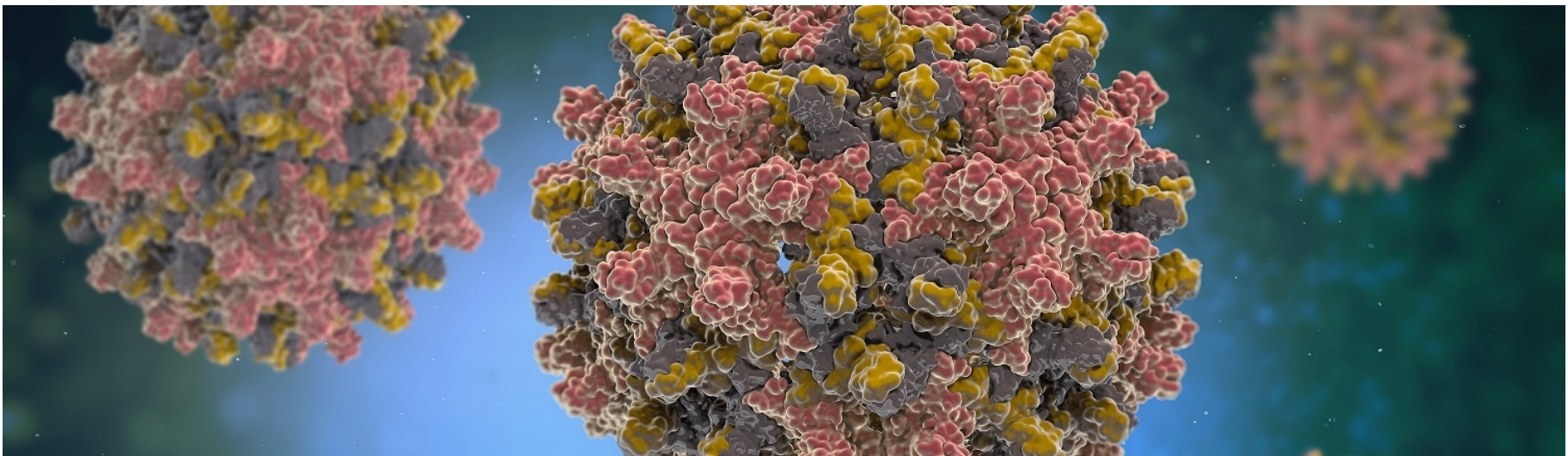




# Deep sequencing

Phuse EU Connect 2019

Hepatitis B virus

Koen Ven and Chantal Kersten, IDAR  
Janssen Research and Development  
Monday November 11th

# Agenda

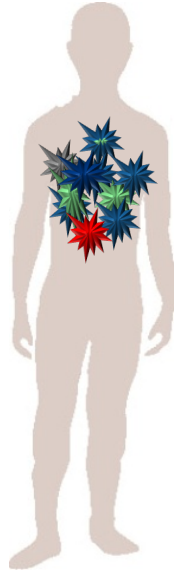
- Concepts
  - What is Deep Sequencing
  - Massively parallel sequencing
  - Research Questions/applications
  - References (universal versus genotype specific)
  - Emerging and enriched variations
- Challenges and solutions
  - Structure of the SDTMs
  - general
  - Outputs
  - Programming
- conclusion

# 1. What is Deep Sequencing?



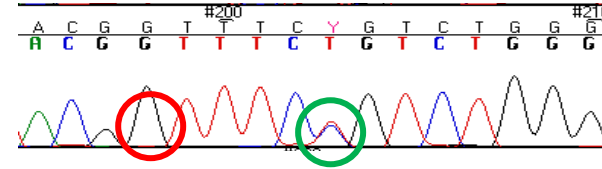
- Deep sequencing is a **technique** that is used **to represent** a large number of genomic (DNA or RNA) or proteomic **sequences**
- Goal is to **map genetic variations** in a host (substitutions, deletions, insertions, gaps) to reference genome and proteome
- This technique will **read** each region multiple times, sometimes thousands of times
- Next generation sequencing (**NGS**) is a **more sensitive** technique than Sanger sequencing

# Massively parallel sequencing



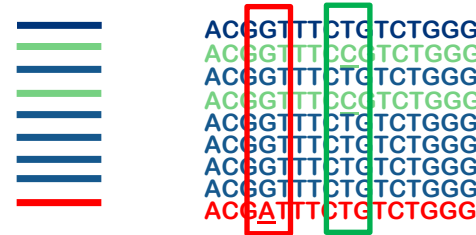
Viral population

Sanger sequencing



- detection limit: 20-30%

Deep sequencing



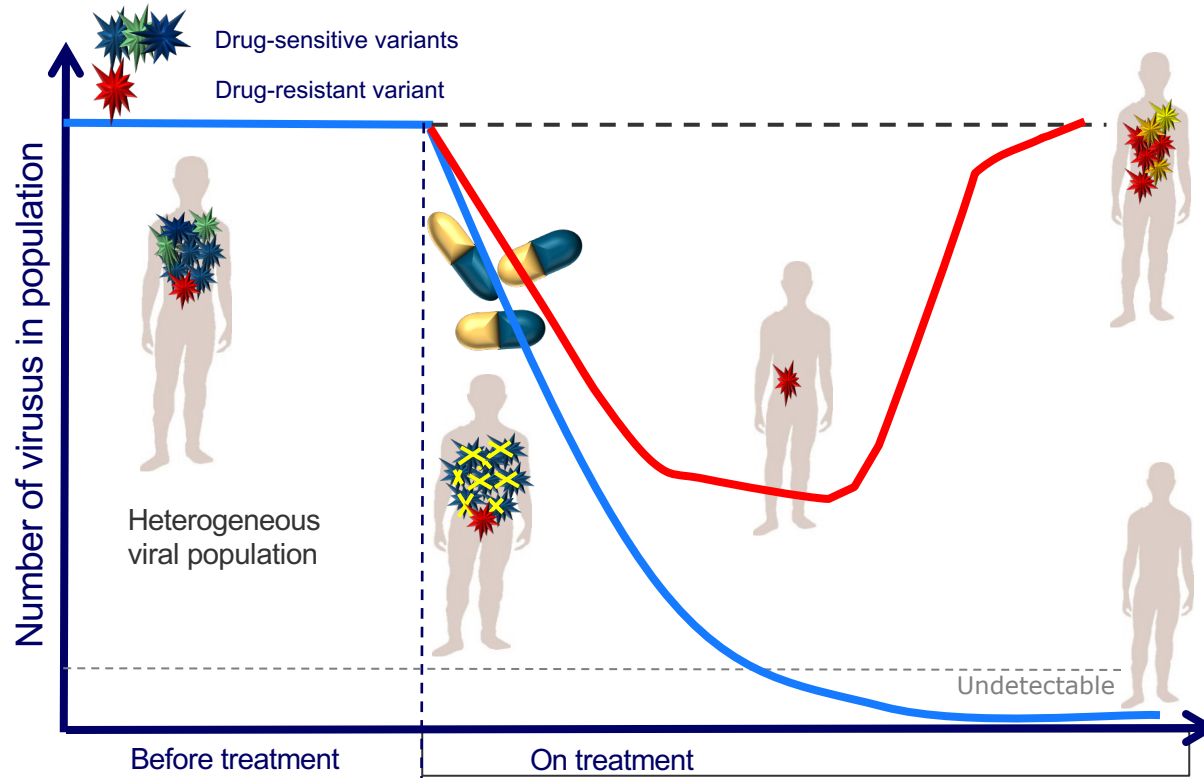
- detection limit << 20%

## 2. Research questions/ applications



- Deep sequencing is
  - a very powerful technique to see & show variations.
  - often used in cancer research to know in advance if a subject is more prone to survive....
- BUT in infection diseases we are more interest to trace emerging variations;  
**some** of them will be random variations,  
**others** will be variations due to the treatment pressure.  
**Emerging variations** can point to treatment resistance,  
while **disappearing variations** indicate those more vulnerable to the treatment.

# Emerging resistance to antiviral agents



# Alignment

## Comparison with a reference sequence

Per region:  
example output

| SAMPLE               |                                                                        |   |   |   |   |   |   |   |   |    |    |    |    |    |    |        |
|----------------------|------------------------------------------------------------------------|---|---|---|---|---|---|---|---|----|----|----|----|----|----|--------|
| Nucleotide Seq:      | gcg ccc atc acg gcg tac gcc cag cgg acg aga ggc ctc cca ggg tgt ata... |   |   |   |   |   |   |   |   |    |    |    |    |    |    |        |
| translated in:       |                                                                        |   |   |   |   |   |   |   |   |    |    |    |    |    |    |        |
| Amino Acid Seq:      | A                                                                      | P | I | T | A | Y | A | Q | R | T  | R  | G  | L  | P  | G  | C I... |
| Position in Protein: | 1                                                                      | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 17  |

| REFERENCE (GENO/SUBTYPE SPECIFIC) |   |   |   |   |   |   |   |   |   |    |    |    |    |    |    |    |       |
|-----------------------------------|---|---|---|---|---|---|---|---|---|----|----|----|----|----|----|----|-------|
| Amino Acid Seq:                   | A | P | I | T | A | Y | A | Q | Q | T  | R  | G  | L  | L  | G  | C  | I.... |
| Position in Protein:              | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17    |

| SAMPLE          |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |       |
|-----------------|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|-------|
| Amino Acid Seq: | A | P | I | T | A | Y | A | Q | R | T | R | G | L | P | G | C | I.... |

| REFERENCE (GENO/SUBTYPE SPECIFIC) |   |   |   |   |   |   |   |   |     |    |    |    |    |    |    |    |      |      |
|-----------------------------------|---|---|---|---|---|---|---|---|-----|----|----|----|----|----|----|----|------|------|
| Amino Acid Seq:                   | A | P | I | T | A | Y | A | Q | Q   | T  | R  | G  | L  | L  | G  | C  | I... |      |
| Position in Protein:              | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9   | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17   |      |
| SAMPLE                            |   |   |   |   |   |   |   |   |     |    |    |    |    |    |    |    |      |      |
| Amino Acid Seq:                   | A | P | I | T | A | Y | A | Q | del | T  | R  | G  | L  | K  | L  | G  | stop | I... |

- Wild type (WT)  
refers to phenotype as occurs in nature
- Types of genetic variance (mutations)
  - Substitutions
  - Deletions
  - Insertions
  - Gaps
- Comparison against multiple reference sequences
  - versus universal reference
  - versus genotype specific reference

# Emerging and enriched variations

Comparing mutations during or after treatment with baseline

Similar frequencies

Emergence of  
mutations

Disappearance of  
mutations

Enrichment of  
mutations

# Challenges and Solutions

# Data structure

## SDTM domains

- Sequencing method
- Test category
- Mutation type
- Genetic region
- Reference sequence
- Sequence result
- Variant frequency

## SUPPPF

|    | RDOMAIN | IDVAR | IDVARVAL | QNAM      | QLABEL               | QVAL              | QORIG | GEVAL |
|----|---------|-------|----------|-----------|----------------------|-------------------|-------|-------|
| 1  | PF      | PFSEQ | 1        | PFANMETH  | Analysis Method      | 1% THRESHOLD      | EDT   |       |
| 2  | PF      | PFSEQ | 1        | PFFRSC    | Forward/Reverse...   | 1.0295            | EDT   |       |
| 3  | PF      | PFSEQ | 1        | PFGENLOC  | Genetic Location     | 7                 | EDT   |       |
| 4  | PF      | PFSEQ | 1        | PFGENRI   | Genetic Region o...  | PreS1             | EDT   |       |
| 5  | PF      | PFSEQ | 1        | PFGENVT   | Type of Genetic ...  | GENE              | EDT   |       |
| 6  | PF      | PFSEQ | 1        | PFMTHDDS  | Method Description   | ILLUMINA MISEQ    | EDT   |       |
| 7  | PF      | PFSEQ | 1        | PFNSPACES | Non-Host Species     | HEPATITIS B VI... | EDT   |       |
| 8  | PF      | PFSEQ | 1        | PFORREF   | Genetic Variation... | K                 | EDT   |       |
| 9  | PF      | PFSEQ | 1        | PFOSPEC   | Original Specimen    | BI ASMA           | EDT   |       |
| 10 | PF      | PFSEQ | 1        | PFREFSEQ  | Reference Sequ...    | adw2              | EDT   |       |
| 11 | PF      | PFSEQ | 1        | PFRESBTC  | Date/Time of Re...   | 2018-03-28        | EDT   |       |
| 12 | PF      | PFSEQ | 1        | PFRUNID   | Run ID               | UDS-SEQ00439      | EDT   |       |
| 13 | PF      | PFSEQ | 1        | PFTCOV    | Total Coverage       | 23618             | EDT   |       |
| 14 | PF      | PFSEQ | 1        | PFCOV     | Variant Coverage     | 19125             | EDT   |       |
| 15 | PF      | PFSEQ | 1        | PFVFREQ   | Variant Frequency    | 80.976            | EDT   |       |
| 16 | PF      | PFSEQ | 1        | PFVFREQU  | Variant Frequenc...  | %                 | EDT   |       |
| 17 | PF      | PFSEQ | 1        | PFVREFID  | Vendor Internal R... | P393M0092         | EDT   |       |

## PF

|     | DOMAIN | PFSEQ | PFGRPID | PFLNKID       | PFTSTCD  | PFTST             | PFCAT      | PSCAT | PFORRES | PFORRESU | PFSTRESC   | PFSTRESN | PFSTRES | PFRESCAT     | PFSTAT | PFREASND | PFXFN | PFSPEC | PFMETHOD                   | PFLFL | PFLLOQ | VISITNUM | VISIT |
|-----|--------|-------|---------|---------------|----------|-------------------|------------|-------|---------|----------|------------|----------|---------|--------------|--------|----------|-------|--------|----------------------------|-------|--------|----------|-------|
| 121 | PF     | 34    |         | 6203107461-28 | GENETVAR | Genetic Variation | AMINO ACID |       | M       |          | M198M      |          |         | REFERENCE    |        |          |       |        | NEXT GENERATION SEQUENCING |       |        | 20010.00 | TREA  |
| 122 | PF     | 274   |         | 6203107461-28 | GENETVAR | Genetic Variation | AMINO ACID |       | S/R     |          | S210S/R    |          |         | SUBSTITUTION |        |          |       |        | SANGER SEQUENCING          |       |        | 20010.00 | TREA  |
| 123 | PF     | 138   |         | 6203107461-28 | GENETVAR | Genetic Variation | AMINO ACID |       | L       |          | S193L      |          |         | SUBSTITUTION |        |          |       |        | NEXT GENERATION SEQUENCING |       |        | 20010.00 | TREA  |
| 124 | PF     | 278   |         | 6203107461-28 | GENETVAR | Genetic Variation | AMINO ACID |       | M       |          | K130M      |          |         | SUBSTITUTION |        |          |       |        | SANGER SEQUENCING          |       |        | 20010.00 | TREA  |
| 125 | PF     | 70    |         | 6203107461-28 | GENETVAR | Genetic Variation | AMINO ACID |       | I       |          | I712I      |          |         | REFERENCE    |        |          |       |        | NEXT GENERATION SEQUENCING |       |        | 20010.00 | TREA  |
| 126 | PF     | 282   |         | 6203107461-28 | GENETVAR | Genetic Variation | AMINO ACID |       | GAP     |          | 306_308GAP |          |         | GAP          |        |          |       |        | SANGER SEQUENCING          |       |        | 20010.00 | TREA  |

# ADaM structure

## ADPF domain

### Steps in ADPF development:

- Add variables (merge with covariates, population flags and treatment groups)
  - Add timing variables (ady, avisit(n) and period)
  - Add genotype (can be derived from reference used)
- } By subject
- Create one record per subject, visit, per position\*
    - By genetic region
    - By reference
    - By amino acid/ gene
- Add ovar (observed variant)
  - Add ocomp/emcomp/enrcomp and ocomp15/emcomp15/enrcomp15 \*
  - Create variation profiles per genetic region per visit
- } By subject, visit and position  
} By subject and visit (all positions)

\* different approach for NGS/ Sanger

## Challenges and solutions

- Large dataset **size** leading to
  - Difficulties finding specific code in many loops during (self)QC
  - long run times
- Integration of **post baseline sequencing** samples into SDTMs just before lock (sequencing data often need to be blinded until end of trial)
  - Start as soon as baseline samples are in the trial
  - Prepare derivations for emerging/enriched variations on raw data[excel format] or dummy data
- Subregions to calculate (example: Reverse Transcriptase (RT) subregion of the Polymerase region in Hepatitis B)
  - Start at different positions for some genotypes
- Positions of Interest

|                                      |                                                                                                                                                                                                                            |
|--------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Longlist Core Protein                | 18, 19, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 37, 38, 102, 103, 105, 106, 107, 109, 110, 111, 114, 115, 116, 117, 118, 119, 122, 123, 124, 125, 126, 127, 128, 129, 131, 132, 133, 134, 136, 137, 138, 139, 140, 141 |
| List Reverse Transcriptase subregion | 169, 173, 180, 181, 184, 194, 202, 204, 236, 250                                                                                                                                                                           |

## Challenges and solutions: outputs

- Multiple reporting levels (eg. Position, position of interest, region, profiles, reference sequence)
  - Multiple outputs
- For new viruses / new reference sequences the team is still **learning** and scope not clear; amount of permutations will multiply the amount of outputs
  - Close collaboration made it possible to prepare SAP/DPS and select KEY outputs for CSR (other outputs in virology report)
- Two sequencing methods (Sanger and NGS)
  - Duplicating of outputs
  - Sanger is used less in future

# Challenges and Solutions: Programming derivation of observed- variable and component

- OVAR consists of reference positions OBSERVED (Q184K)
- OCOMP will contain information about all observed frequencies
  - OCOMP format: Q184K||Q/STOP with the first letter being the amino acid of the reference sequence, then the position number in the protein, then all variations above 15% separated by / and after the || all variations lower than 15% separated by /.
- OCOMP15 will contain information about frequencies above 15%\*

|    | POSITION | ANVALC | OVAR        | OCOMP         | OCOMP15 | BCOMP         | OFREQ  |
|----|----------|--------|-------------|---------------|---------|---------------|--------|
| 89 | 184      |        | Q184K       | Q184K         | Q184K   | Q184K         | 100    |
| 90 | 184      |        | Q184K       | Q184KIQ/STOP  | Q184K   | Q184KIQ/STOP  | 77.373 |
| 91 | 184      |        | Q184Q       | Q184KIQ/STOP  | Q184K   | Q184KIQ/STOP  | 11.636 |
| 92 | 184      |        | Q184STOP    | Q184KIQ/STOP  | Q184K   | Q184KIQ/STOP  | 10.841 |
| 93 | 186      |        | STOP186Q    | STOP186STOPIQ |         | STOP186STOPIQ | 1.76   |
| 94 | 186      |        | STOP186STOP | STOP186STOPIQ |         | STOP186STOPIQ | 97.848 |

- For all post-baseline visits: derivate EMCOMP(15) and ENRCOMP(15)

\* Sanger technique has a 15% analysis threshold so we kept this as threshold for both techniques

# Challenges and Solutions: Programming derivation of variation profiles

| position X1         | position X2    | positions YY                         | positions ZZ         |
|---------------------|----------------|--------------------------------------|----------------------|
| <b>B (Baseline)</b> | A (Amino acid) | CO (Core region)                     | S0 (all variations)  |
| <b>M (eMerging)</b> | N (Nucleotide) | RT (Reverse Transcriptase subregion) | S1 (long list)       |
| <b>N (eNriched)</b> |                | PC (Pre-Core region)                 | S2 (short list)      |
|                     |                | PO (Polymerase region)               | S3 (very short list) |
|                     |                | S1 (Pre-S1 region)                   |                      |
|                     |                | S2 (Pre-S2 region)                   |                      |

| Library ADPF                                                                                                              |        |                                                               |          |          |         |          |                                                                |
|---------------------------------------------------------------------------------------------------------------------------|--------|---------------------------------------------------------------|----------|----------|---------|----------|----------------------------------------------------------------|
| Freeze          Hide          Show...          Format          Filter...          Font...         Find          Go to row |        |                                                               |          |          |         |          |                                                                |
| Table                                                                                                                     | View   |                                                               |          |          |         |          |                                                                |
|                                                                                                                           | AGENRI | PARAM                                                         | PARAMCD  | PARAMTYP | DTYPE   | POSITION | ANVALC                                                         |
| 7                                                                                                                         | X      | Gap                                                           | GAP      |          |         |          |                                                                |
| 8                                                                                                                         | X      | Gap                                                           | GAP      |          |         |          |                                                                |
| 9                                                                                                                         | X      | Gap                                                           | GAP      |          |         |          |                                                                |
| 10                                                                                                                        | GENOME | HBV Core Nucleotide - Baseline Substitution Profile (list)    | BN_CO_S1 | DERIVED  | PROFILE |          | A1762T+G1764A                                                  |
| 11                                                                                                                        | C      | HBV Core Protein - Baseline Substitution Profile (long list)  | BA_CO_S1 | DERIVED  | PROFILE |          | No baseline polymorphism                                       |
| 12                                                                                                                        | C      | HBV Core Protein - Baseline Substitution Profile (long list)  | BA_CO_S1 | DERIVED  | PROFILE |          | No baseline polymorphism                                       |
| 13                                                                                                                        | C      | HBV Core Protein - Baseline Substitution Profile (short list) | BA_CO_S2 | DERIVED  | PROFILE |          | No baseline polymorphism                                       |
| 14                                                                                                                        | C      | HBV Core Protein - Baseline Substitution Profile (short list) | BA_CO_S2 | DERIVED  | PROFILE |          | No baseline polymorphism                                       |
| 15                                                                                                                        | C      | HBV Core Protein - Baseline Substitution Profile (very sh...  | BA_CO_S3 | DERIVED  | PROFILE |          | No baseline polymorphism                                       |
| 16                                                                                                                        | C      | HBV Core Protein - Baseline Substitution Profile (very sh...  | BA_CO_S3 | DERIVED  | PROFILE |          | No baseline polymorphism                                       |
| 17                                                                                                                        | C      | HBV Core Protein - Baseline Substitutions Profile             | BA_CO_S0 | DERIVED  | PROFILE |          | T12T/A+S44S/C+L55L/I+E77E/D/H/Q+N92N/H+V93M+R151C+P171Q+Q184K  |
| 18                                                                                                                        | C      | HBV Core Protein - Baseline Substitutions Profile             | BA_CO_S0 | DERIVED  | PROFILE |          | T12T/A+S44S/C+L55L/I+E77E/D/H/Q+N92N/H+V93M+R151C+P171Q+Q184K  |
| 19                                                                                                                        | RT     | HBV Polymerase/RT Region - Baseline Substitutions Pr...       | BA_RT_S0 | DERIVED  | PROFILE |          | D7D/G+I53T+S54T+N76N/K+K212K/T+S219S/A+K333Q+M336R             |
| 20                                                                                                                        | RT     | HBV Polymerase/RT Region - Baseline Substitutions Pr...       | BA_RT_S0 | DERIVED  | PROFILE |          | D7D/G+I53T+S54T+N76N/K+K212K/T+S219S/A+K333Q+M336R+N337N/D     |
| 21                                                                                                                        | RT     | HBV Polymerase/RT Region - Baseline Substitutions Pr...       | BA_RT_S1 | DERIVED  | PROFILE |          | No baseline polymorphism                                       |
| 22                                                                                                                        | RT     | HBV Polymerase/RT Region - Baseline Substitutions Pr...       | BA_RT_S1 | DERIVED  | PROFILE |          | No baseline polymorphism                                       |
| 23                                                                                                                        | GENOME | HBV PreCore Nucleotide - Baseline Substitution Profile        | BN_PC_S0 | DERIVED  | PROFILE |          | A1390G+C1484C/A+A1511A/G+G1613G/A+T1631T/C+C1653C/T+A1747A/... |
| 24                                                                                                                        | GENOME | HBV PreCore Nucleotide - Baseline Substitution Profile (l...  | BN_PC_S1 | DERIVED  | PROFILE |          | No baseline polymorphism                                       |
| 25                                                                                                                        | GENOME | Insertion                                                     | INSERT   |          |         | 1647     |                                                                |
| 26                                                                                                                        | GENOME | Insertion                                                     | INSERT   |          |         | 1647     |                                                                |

# conclusion

- Powerful technique
  - Identify treatment related mutations (emerging, disappearing)
  - treatment pressure on viral population
- Challenges
  - Operational challenges
  - Need team cooperation (programmers/scientists/data management/statistician/...)

