Paper TT07

Deep Sequencing (DS)

Koen Ven, Janssen R&D, Beerse, Belgium
Chantal Kersten, Janssen R&D, Beerse, Belgium
Ilse Augustyns, Janssen R&D, Beerse, Belgium

ABSTRACT

Deep sequencing is a collective name for methodologies to sequence whole genomes by trying to get multiple different reads for each base in the sequence. This improves reliability and allows to identify single nucleotide polymorphisms. At Johnson&Johnson (J&J) - Janssen R&D we applied these methods in infectious disease trials to better understand a disease and how new treatments may interact with a disease agent on both and proteomic level (including possible drug resistance).
Here we will show how deep sequencing helped us in answering multiple research questions but also how we solved some of the technical challenges we faced to provide these answers. Examples are the large size of the datasets and fact that the sequence was done against multiple reference sequences, namely a universal and a genotype specific reference.
The amount of outputs and the complexity of the topic made a close collaboration between statistical programmers, statisticians and scientists a must.

INTRODUCTION

At J&J Infectious Diseases we collect data in our clinical trials from techniques such as Deep Sequencing to find answers to research questions, like presence of baseline genetic variations which may result in reduced on-treatment response, and/or the emergence of (new) genetic variations in subjects with treatment failure. By doing so, we aim to get a better understanding of the viruses we hope to eliminate.

Deep sequencing is a technique that is used to represent a large number of genomic or proteomic sequences. It is often used in cancer research, but also in virus research it can be used to map genetic variations (substitutions, deletions, insertions, gaps) of the viral genome and proteome. This technique will read each region multiple times, sometimes thousands of times. This allows for elimination of transcript errors and the possibility to read Single Nucleotide Polymorphisms (SNPs), which are often difficult to distinguish from transcription errors. These sequenced genomes (or RNA) are then compared with a reference sequence.

Two techniques are used, Sanger Sequencing and Next Generation Sequencing (NGS). The Sanger technique, more widely known through the Human Genome Project, is an older technique, but still used widely. It has a technical cut-off of 15% to detect variations and will only show absence or presence of a variation. When for a given position the observed variation is not the same as the reference variation, this observed variation will be reported.
The NGS technique can detect genetic variations that are present in as low as 1% of the reads and will show all variations including the wild-type. This technique also gives the frequency of the variation and is better suited to detect SNPs. A full comparison and a more detailed explanation of how the sequencing works is outside the scope of this paper.

Each of these variations will be given a position number. For the genome this is a continuous number for the entire genome and in case of a circular genome, starting from a reference start point. Protein sequences will start at 1 for each different protein, which is often called a region or domain. Subregions are also possible. See figure 1 for an overview of the genome sequence, the derived protein sequence and how the comparison with a reference sequence is performed.

Per region:
example output

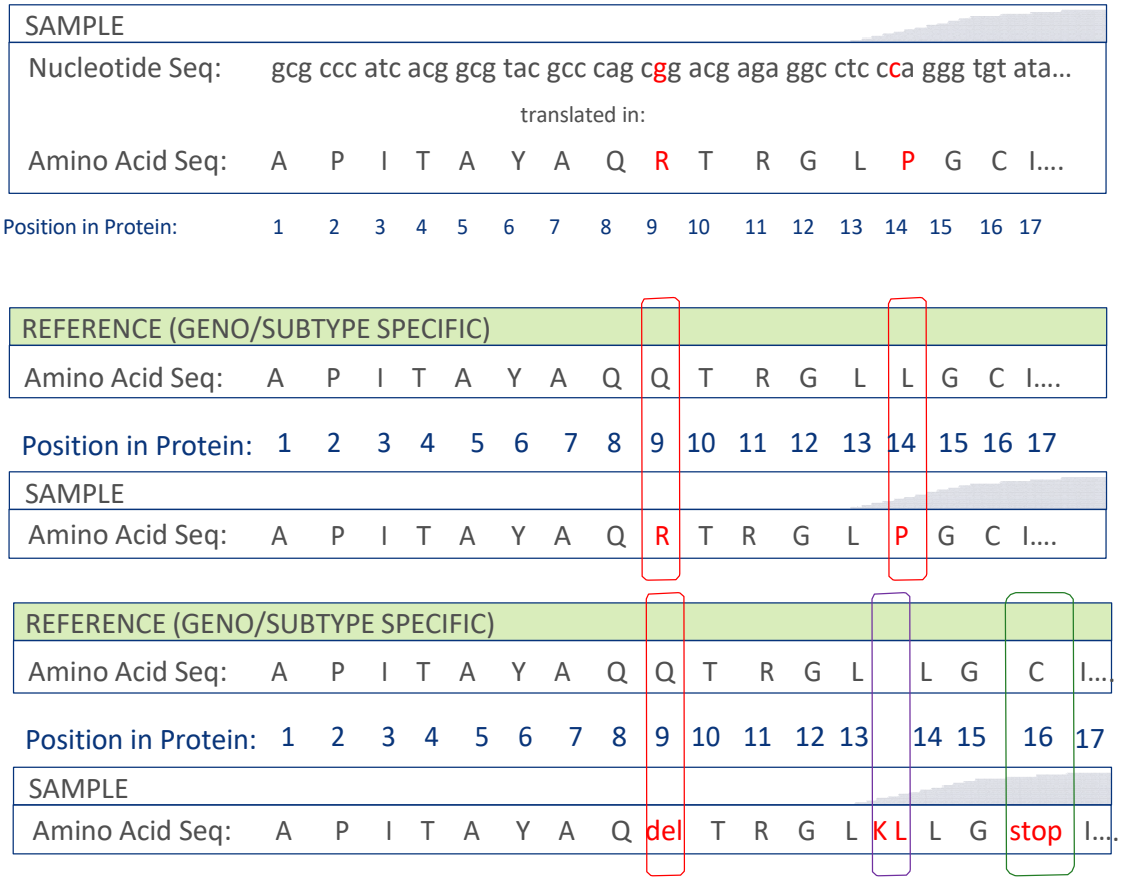


Figure 1: sample of sequencing and comparison with reference

In the dataset you will see the reference variant, the position and the observed variant. Using the example above, you can get Q9R, L14P, Q9del, L13insKL and C16stop. For NGS sequencing, also the frequency with which these variations are present will also be available in the dataset. An example of how the dataset will look like can be seen below. For NGS sequencing all the variations at one position (including wild type) will normally add up to 100%, although there might be variations below the threshold of 1% and that can make that the total is slightly less

SubjID	Visit	Gene	Variant	Read frequency
XXXX	Baseline	NS3	Q80K	0.80
XXXX	Baseline	NS3	Q80Q	0.15
XXXX	Baseline	NS3	R155K	0.70
XXXX	Baseline	NS3	R155R	0.25

Table 1: example of a dataset

In case a virus has multiple genotypes, reference sequences are often available for all different genotypes. Where this is the case, the comparison is often done against the different reference sequences and against a common reference sequence. Therefore, it is possible that a variation against one reference sequence isn't one against another sequence. These reference sequences are taken from genome sequence banks.

Data is collected from different visits. First, a baseline sample is taken, to see which variations are present at the start of the study. Comparing the data at baseline with those after treatment started, we could encounter 4 scenarios.

Variation Type	Definition
Same as baseline	Nothing happens, the variation that was present still is present with a similar frequency (above or below a threshold)
Disappearing variation	A variation disappears or falls below the threshold after baseline
Emerging variation	A variation that was not present appears
Enriched variation	A variation that was present at very low levels (below threshold) now is present at much higher levels

Table 2: possible options for variations after baseline

IMPLEMENTATION

GENERAL INTRODUCTION

After the sequencing is done, we will receive the data in the standard SDTM (Study Data Tabulation Model) format, in the PF (Pharmacogenomics Findings) and SUPPPF domains. For more information about the SDTM datasets for genetic findings, please see the CDISC PGx model (https://www.cdisc.org/standards/foundational/pgx).

First of all, we transformed this SDTM dataset into an ADaM-like (Analysis Data Model) dataset, adding all the population flags and creating extra variables we're going to need for creating the outputs. For this transformation and programming of the outputs we used the SAS programming language.

	DOMAIN	PFSEQ	PFGRPID	PFLNKID	PFTESTCD	PFTEST	PFCAT	PFSCAT	PFORRES	PFORRESU	PFSTRESC	PFSTRESN	PFSTRESU	PFRESCAT	PFSTAT	PFREASND	PFXFN	PFSPEC	PFMETHOD	PFBLFL	PFLLOQ	VISITNUM	VISIT
121	PF	34		620310746 1-28	GENETVAR	Genetic Variation	AMINO ACID		M		M190M			REFERENCE					NEXT GENERATION SEQUENCING			20010.00	TREA
122	PF	274		620310746 1-28	GENETVAR	Genetic Variation	AMINO ACID		S/R		S210S/R			SUBSTITUTION					SANGER SEQUENCING			20010.00	TREA
123	PF	138		620310746 1-28	GENETVAR	Genetic Variation	AMINO ACID		L		S193L			SUBSTITUTION					NEXT GENERATION SEQUENCING			20010.00	TREA
124	PF	278		620310746 1-28	GENETVAR	Genetic Variation	AMINO ACID		M		K130M			SUBSTITUTION					SANGER SEQUENCING			20010.00	TREA
125	PF	70		620310746 1-28	GENETVAR	Genetic Variation	AMINO ACID		I		I712I			REFERENCE					NEXT GENERATION SEQUENCING			20010.00	TREA
126	PF	282		620310746 1-28	GENETVAR	Genetic Variation	AMINO ACID		GAP		306_308GAP			GAP					SANGER SEQUENCING			20010.00	TREA

	RDOMAIN	IDVAR	IDVARVAL	QNAM	QLABEL	QVAL	QORIG	QEVAL
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	PFGENTYP	Type of Genetic ...	GENE	EDT	
6	PF	PFSEQ	1	PFMTHDDS	Method Description	ILLUMINA MISEQ	EDT	
7	PF	PFSEQ	1	PFNSPCES	Non-Host Species	HEPATITIS B VI...	EDT	
8	PF	PFSEQ	1	PFORREF	Genetic Variation...	K	EDT	
9	PF	PFSEQ	1	PFOSPEC	Original Specime...	PLASMA	EDT	
10	PF	PFSEQ	1	PFREFSEQ	Reference Sequ...	adw2	EDT	
11	PF	PFSEQ	1	PFRESDTC	Date/Time of Re...	2018-03-26	EDT	
12	PF	PFSEQ	1	PFRUNID	Run ID	UDS-SEQ00439	EDT	
13	PF	PFSEQ	1	PFTCOV	Total Coverage	23618	EDT	
14	PF	PFSEQ	1	PFVCOV	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	

Figure 2: example of a PF and SUPPPF SDTM datasets

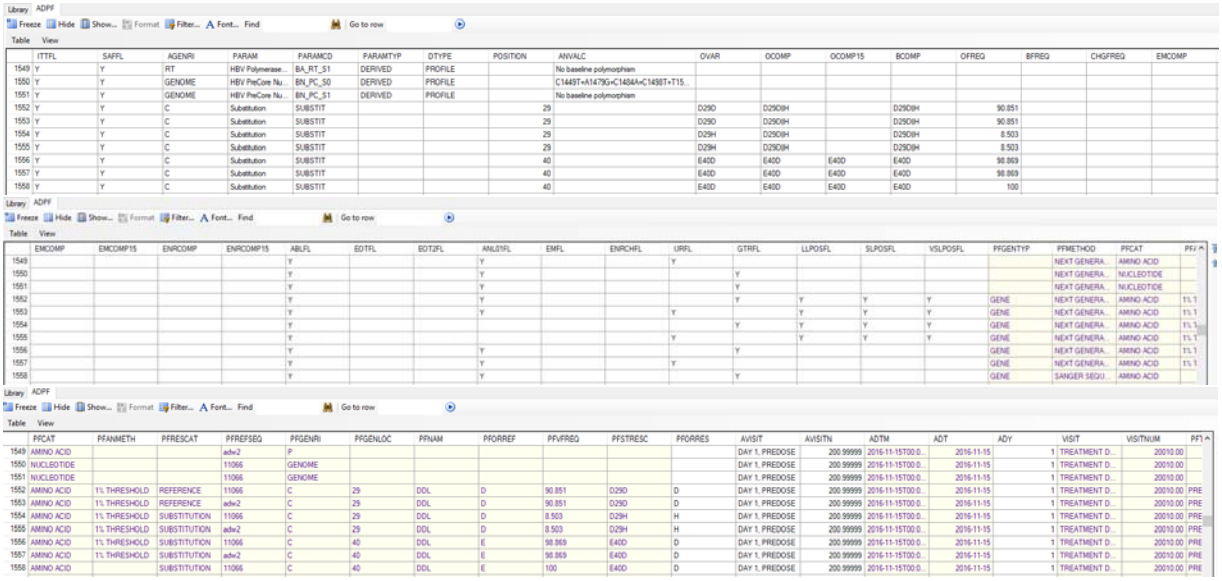


Figure 3: example of an ADPF ADaM dataset

CHALLENGES AND SOLUTIONS

During our work with Deep Sequencing data, we encountered some challenges. Here we will go into more detail on some of them as well as on some solutions we came up with.

DATASET SIZE

Due to the large number of variables and the often large amount of genetic variations encountered, even in small viral genomes, the datasets can be very large. On top of that if you have two sequencing techniques and two reference sequences (one general and one genotype specific), you get four blocks of records for each sample. In one of our smaller phase 1 studies, the full SDTM dataset was around 116.000 records. Fortunately, in the more recent studies we limit to Next Generation Sequencing, but even if the size of the datasets is going down, they are still some of the largest SDTM datasets in the trial.

PGx is a fairly new topic to programmers, our main challenge was a mix of learning the basics of the collected data and analysis requirements, and once the data is to our hands, tracing to the information of interest. During programming, the main challenge was finding things. If during programming and/or QC one had an error in one or a few records, finding them was not always easy, making debugging the program a tedious effort. Another challenge with the large datasets (and the large intermediate datasets during any run of programs) was the often long runtimes.

REPORTING LEVELS

Our customers, clinical virologists, are interested in data summaries at different levels of the genome. Eg, there are outputs dealing with 1/ a single variation at a single position (example the presence of G1896A pre-core variation in HBV), 2/ with all variations at a single position (example all variations at position 1896 pre-core variation in HBV), 3/ with variations at multiple positions in a region (example the positions of interest, see further below) or even 4/ with entire regions or subregions. There were also outputs comparing variations to the general reference sequence and the genotype specific sequences.

TVIBL001A1: Number (%) of Subjects with Substitutions on Amino Acid Level Versus the Universal HBV Reference Sequence at Baseline by Position Considering the HBV Core Protein - Next Generation Sequencing; Intent-to-treat Population

TVIBL001A2: Number (%) of Subjects with Substitutions on Amino Acid Level Versus the Universal HBV Reference Sequence at Baseline by Position Considering the HBV Core Protein - Sanger Sequencing; Intent-to-treat Population

TVIBL001B2: Number (%) of Subjects with Substitutions on Amino Acid Level Versus the HBV-genotype Specific Reference Sequence at Baseline by Position Considering the HBV Core Protein - Sanger Sequencing; Intent-to-treat Population

TVIBL001B1: Number (%) of Subjects with Substitutions on Amino Acid Level Versus the HBV-genotype Specific Reference Sequence at Baseline by Position Considering the HBV Core Protein - Next Generation Sequencing; Intent-to-treat Population

table 3: example of output titles showing the different levels of output for the same table

TVIBL001A1: Number (%) of Subjects with Substitutions on Amino Acid Level Versus the Universal HBV Reference Sequence at Baseline by Position Considering the HBV Core Protein - Next Generation Sequencing; Intent-to-treat Population

	Session 8	Session 9	Session 10	Session A	Session 11	All	Placebo	Total
Analysis set: Intent-To-Treat	8	8	9	9	7	41	16	57
All HBV genotypes	8	8	9	8	7	40	16	56
Subjects with Sequencing Data (Core)	8	7	8	6	7	36	16	52
No Baseline Polymorphism	0	0	0	0	0	0	0	0
1 or More Baseline Polymorphism	8 (100.0%)	7 (100.0%)	8 (100.0%)	6 (100.0%)	7 (100.0%)	36 (100.0%)	16 (100.0%)	52 (100.0%)
0003	0	0	0	0	0	0	4 (25.0%)	4 (7.7%)
0005	1 (12.5%)	1 (14.3%)	2 (25.0%)	0	1 (14.3%)	5 (13.9%)	0	5 (9.6%)
0009	0	1 (14.3%)	0	0	0	1 (2.8%)	0	1 (1.9%)
0012	5 (62.5%)	2 (28.6%)	6 (75.0%)	3 (50.0%)	7 (100.0%)	23 (63.9%)	11 (68.8%)	34 (65.4%)

TVIBL010A1: Number (%) of Subjects with Substitutions at Baseline on Amino Acid Level Versus the Universal HBV Reference Sequence by Position Considering 28 HBV Core Protein Positions of Interest - Next Generation Sequencing; Intent-to-treat Population

	Session 8	Session 9	Session 10	Session A	Session 11	All	Placebo	Total
Analysis set: Intent-To-Treat	8	8	9	9	7	41	16	57
All HBV genotypes	8	8	9	8	7	40	16	56
Subjects with Sequencing Data (Core)	8	7	8	6	7	36	16	52
No Baseline Polymorphism	3 (37.5%)	3 (42.9%)	6 (75.0%)	3 (50.0%)	5 (71.4%)	20 (55.6%)	5 (31.3%)	25 (48.1%)
1 or More Baseline Polymorphism	5 (62.5%)	4 (57.1%)	2 (25.0%)	3 (50.0%)	2 (28.6%)	16 (44.4%)	11 (68.8%)	27 (51.9%)
0023	0	0	0	0	0	0	1 (6.3%)	1 (1.9%)
0024	0	1 (14.3%)	0	0	0	1 (2.8%)	2 (12.5%)	3 (5.8%)
0029	0	0	0	0	0	0	1 (6.3%)	1 (1.9%)
0038	3 (37.5%)	2 (28.6%)	2 (25.0%)	3 (50.0%)	0	10 (27.8%)	6 (37.5%)	16 (30.8%)
0102	0	0	0	0	1 (14.3%)	1 (2.8%)	0	1 (1.9%)
0105	2 (25.0%)	0	0	1 (16.7%)	1 (14.3%)	4 (11.1%)	7 (43.8%)	11 (21.2%)
0109	0	0	1 (12.5%)	1 (16.7%)	2 (28.6%)	4 (11.1%)	4 (25.0%)	8 (15.4%)
0118	0	1 (14.3%)	0	0	0	1 (2.8%)	0	1 (1.9%)

TVIBL013B1: Number (%) of Subjects with Substitutions on Amino Acid Level Versus the Universal HBV Reference Sequence at Baseline by HBV Genotype and by Position and by Substitution Considering 28 HBV Core Protein Position - Next Generation Sequencing; Intent-to-treat Population

	Session 8	Session 9	Session 10	Session A	Session 11	All	Placebo	Total
Analysis set: Intent-To-Treat	8	8	9	9	7	41	16	57
All HBV genotypes	8	8	9	8	7	40	16	56
Subjects with Sequencing Data (Core)	8	7	8	6	7	36	16	52
No Baseline Polymorphism on position 38	5 (62.5%)	5 (71.4%)	6 (75.0%)	3 (50.0%)	7 (100.0%)	26 (72.2%)	10 (62.5%)	36 (69.2%)
1 or More Baseline Polymorphism on position 38	3 (37.5%)	2 (28.6%)	2 (25.0%)	3 (50.0%)	0	10 (27.8%)	6 (37.5%)	16 (30.8%)
Y38F	2 (25.0%)	2 (28.6%)	1 (12.5%)	3 (50.0%)	0	8 (22.2%)	5 (31.3%)	13 (25.0%)
Y38Y/F	0	0	1 (12.5%)	0	0	1 (2.8%)	1 (6.3%)	2 (3.8%)
Y38Y/H	1 (12.5%)	0	0	0	0	1 (2.8%)	0	1 (1.9%)

TVIBL020A1: Number (%) of Subjects with Substitutions on Amino Acid Level versus the Universal HBV Reference Sequence at Baseline by Substitution Profile Considering 28 HBV Core Protein Positions of Interest - Next Generation Sequencing; Intent-to-treat Population

	Session 8	Session 9	Session 10	Session A	Session 11	All	Placebo	Total
Analysis set: Intent-To-Treat	8	8	9	9	7	41	16	57
All HBV genotypes	8	8	9	8	7	40	16	56
Subjects with Sequencing Data (Core)	8	7	8	6	7	36	16	52
No Baseline Polymorphism	3 (37.5%)	3 (42.9%)	6 (75.0%)	3 (50.0%)	5 (71.4%)	20 (55.6%)	5 (31.3%)	25 (48.1%)
1 or More Baseline Polymorphism	5 (62.5%)	4 (57.1%)	2 (25.0%)	3 (50.0%)	2 (28.6%)	16 (44.4%)	11 (68.8%)	27 (51.9%)
Y38F	2 (25.0%)	2 (28.6%)	1 (12.5%)	2 (33.3%)	0	7 (19.4%)	3 (18.8%)	10 (19.2%)
I105V	2 (25.0%)	0	0	0	0	2 (5.6%)	0	2 (3.8%)
D29H	0	0	0	0	0	0	1 (6.3%)	1 (1.9%)
F23Y+Y38F+I105T+T109M	0	0	0	0	0	0	1 (6.3%)	1 (1.9%)
F24Y	0	1 (14.3%)	0	0	0	1 (2.8%)	0	1 (1.9%)
F24Y+I105T	0	0	0	0	0	0	1 (6.3%)	1 (1.9%)
F24Y+I105V	0	0	0	0	0	0	1 (6.3%)	1 (1.9%)
I105I/L+T109T/M	0	0	0	0	0	0	1 (6.3%)	1 (1.9%)
I105I/V+T109M	0	0	0	0	0	0	1 (6.3%)	1 (1.9%)
T109C	0	0	0	0	1 (14.3%)	1 (2.8%)	0	1 (1.9%)
W102W/STOP+I105L+T109M	0	0	0	0	1 (14.3%)	1 (2.8%)	0	1 (1.9%)

Figure 4: example of different outputs (from top to bottom: all positions, positions of interest, per position the exact variations, using profiles) (note: these are parts of the tables, not the full tables)

This was mostly solved by using macros in our output programs which we could run several times with different parameters. With two methods and different genotypes, we used nested macros quite often in the creation of the outputs, as you can see in the example below.

```
%macro _for1 ( method= );

…

%_for ( expr = %str(partn eq 2), pop =ittfl , geno = Genotype A );
%_for ( expr = %str(partn eq 2), pop =ittfl , geno = Genotype B );
%_for ( expr = %str(partn eq 2), pop =ittfl , geno = Genotype C );
%_for ( expr = %str(partn eq 2), pop =ittfl , geno = Genotype D );
%_for ( expr = %str(partn eq 2), pop =ittfl , geno = Genotype E );
%_for ( expr = %str(partn eq 2), pop =ittfl , geno = Genotype F );
%_for ( expr = %str(partn eq 2), pop =ittfl , geno = Genotype H );

%mend _for1;


/*****************************************************************************
 [04] Generate the final output
*****************************************************************************/

%_for1 ( method= NEXT GENERATION SEQUENCING);

%rmtgentbl          ( fileid   = TVIBL001A1
                    , orient   = landscape
                    , rtxtwdth = 2
                    );

%_for1 ( method= SANGER SEQUENCING);
```

```
%rmtgentbl          ( fileid   = TVIBL001A2
                    , orient   = landscape
                    , rtxtwdth = 2
                   );
```

Figure 5: example of macro call code.

POSITIONS OF INTEREST

As a result of previous research and available literature, there are positions/regions from the genome in which virologists are more focused for the virus at study. These result into another layer of analysis outputs, to focus towards these regions/positions, either one position at a time or looking at them together. For certain regions there were multiple lists (long list, short list, very short list,…).

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

Table 4: example of lists of positions of interest for Hepatitis B Virus proteins (note: Core protein positions for genotype G are 12 higher due to a 12 amino acid insertion in this genotype)

To keep track if a position was a position of interest, several flags were created in the ADaM dataset, mostly for the ease of selecting them later for outputs and further manipulations (like the profiles outlined below).

COMBINING RECORDS

The input dataset has one record for each variation, but for the outputs we needed one record for each position, even if there were multiple variations at the same position. Therefore, we had to combine information from multiple records in a position into one record/variable per position (and per subject/visit). Also, the NGS method has the capability to report up to 1%, for analysis, a threshold is introduced to only report on variations above this threshold. Typically this was chosen as 15% per analogue of the Sanger threshold. In order to do this, we created multiple variables, one with all the variations and one with only those variations above 15%. The same had to be done for emerging and enriched variations (see scenarios 3 and 4 in Table 2). The result is an overview of all the variations that look as follows: N158D/L||N/S 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 /. A similar overview was made for the deletions (N185DEL) and the insertions (N185.1D, N185.2T)

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

Figure 6: example of the combined variations in a dataset

An example is given in the above figure. For position 184 there are two variations and the wild type. As the Q to K variation is present in 77% of the reads, it is put before the 15% divide line. The wild type and the STOP variation are below 15%, so they are put on after the line, resulting in Q184K||Q/STOP.

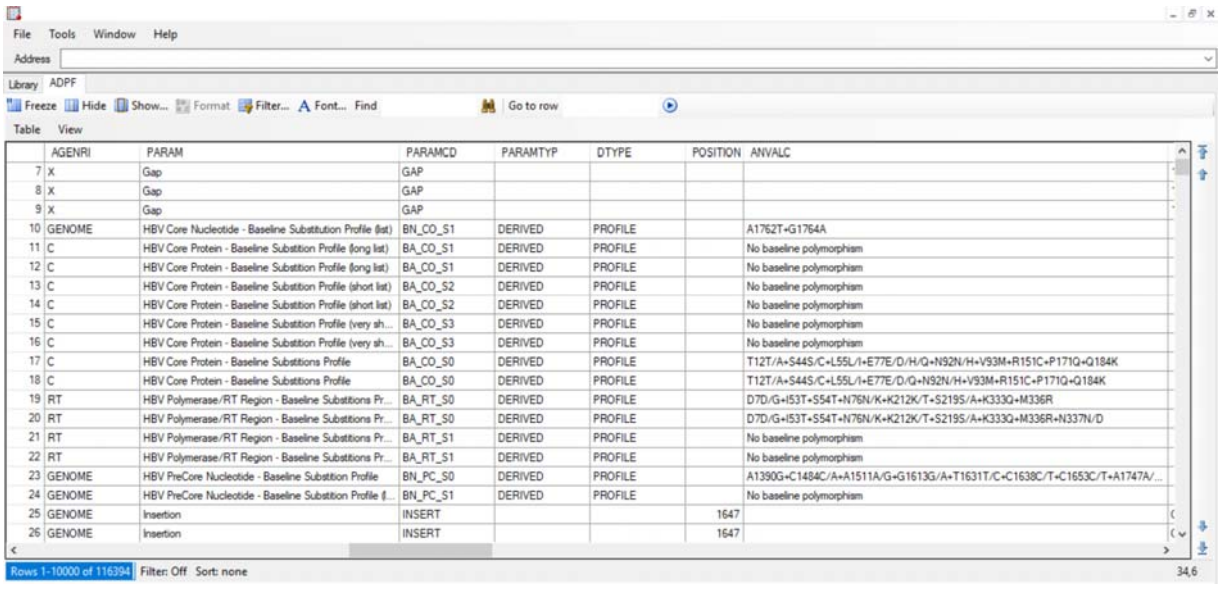
We created multiple of these variables. The first was called OCOMP and was simply the observed variations (Observed COMPonent) at a certain position at a timepoint. We also created a variable called EMCOMP where we

only take the variations which are emerging (scenario 3 from table 2). These were identified earlier in the process with a flag. Another variable called ENRCOMP did the same with the enriched variations (scenario 4 from table 2), also previously identified with a flag. For each of these variables we also had a variant that only listed those above the analysis threshold of 15% (called OCOMP15, EMCOMP15 and ENRCOMP15 respectively). Example: the OCOMP15 for the example in figure 2 would be Q184K.

As our data contained both the Sanger sequencing and NGS methodology, we tried to handle it through a single algorithm. To accomplish this, we tweaked the Sanger data in some way, by acting as if the observed frequency of a reported Sanger variation was 100%. This ”fake” Sanger frequency was only used during the data processing part, it was never used on outputs.

PROFILES

Till now, we only looked within a single position. As a next step in the analysis, combining information across different positions is of a research interest. Here we concatenated, within a subject and sample, all the overviews per position as given in the previous point into one long string (referred to as profiles). In these profiles we included not only the variations, but also the other genetic variations (deletions, insertions). In case no genetic variations exist at a visit, the variable reflects this, eg content <No variation present>. Profiles were created for each region/subregion of interest as well as for the different lists of positions of interest Also, different versions of profiles where created for all observed variations as well for restricting to emerging or enriched variations. This results in a variety of profile parameters (PARAM/PARAMCD) in our ADAM dataset (see Table 4). For these profiles we limited to those genetic variations above the threshold of 15%. For listings, however, also variations below the 15% threshold were to be reported, so yet other versions need to be considered (not discussed further here).


File Tools Window Help
Address
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 Substition Profile (long list)	BA_CO_S1	DERIVED	PROFILE		No baseline polymorphism
12	C	HBV Core Protein - Baseline Substition Profile (long list)	BA_CO_S1	DERIVED	PROFILE		No baseline polymorphism
13	C	HBV Core Protein - Baseline Substition Profile (short list)	BA_CO_S2	DERIVED	PROFILE		No baseline polymorphism
14	C	HBV Core Protein - Baseline Substition Profile (short list)	BA_CO_S2	DERIVED	PROFILE		No baseline polymorphism
15	C	HBV Core Protein - Baseline Substition Profile (very sh...	BA_CO_S3	DERIVED	PROFILE		No baseline polymorphism
16	C	HBV Core Protein - Baseline Substition Profile (very sh...	BA_CO_S3	DERIVED	PROFILE		No baseline polymorphism
17	C	HBV Core Protein - Baseline Substitions 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 Substitions Profile	BA_CO_S0	DERIVED	PROFILE		T12T/A+S44S/C+L55L/I+E77E/D/Q+N92N/H+V93M+R151C+P171Q+Q184K
19	RT	HBV Polymerase/RT Region - Baseline Substitions 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 Substitions 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 Substitions Pr...	BA_RT_S1	DERIVED	PROFILE		No baseline polymorphism
22	RT	HBV Polymerase/RT Region - Baseline Substitions Pr...	BA_RT_S1	DERIVED	PROFILE		No baseline polymorphism
23	GENOME	HBV PreCore Nucleotide - Baseline Substition Profile	BN_PC_S0	DERIVED	PROFILE		A1390G+C1484C/A+A1511A/G+G1613G/A+T1631T/C+C1638C/T+C1653C/T+A1747A/...
24	GENOME	HBV PreCore Nucleotide - Baseline Substition Profile (l...	BN_PC_S1	DERIVED	PROFILE		No baseline polymorphism
25	GENOME	Insertion	INSERT			1647	
26	GENOME	Insertion	INSERT			1647	

Rows 1-10000 of 116394 Filter: Off Sort: none 34,6

Figure 7: example of the profiles

Profiles were created for baseline records, but also for emerging and enriched records, using the xxCOMP15 variables created earlier.

The PARAMCD variable was build using a simple logic, to be as self-explanatory as possible. An XX-YY-ZZ structure is used, with each position having a meaning. An overview of some of the ways it was build can be seen in the below table.

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

Table 5: overview of PARAMCD structure

SUBREGIONS

Recalculate positions for subregions. For instance, in the Hepatitis B virus (HBV), the scientists were interested in the Reverse Transcriptase (RT) subregion of the Polymerase (PO) region. While the calculation itself is straightforward, care must be given that the RT subregion doesn't start at the same position for every genotype specific sequence.

HBV genotype	RT-domain
A	349-692
B, C, F, H	347-690
D	336-679
E, G	345-689

Table 6: start and end amino acid location of RT subregion in the Pol region in Hepatitis B.

DATA DELAYS

As sequencing takes quite a lot of time and samples are often shipped in bulk near the end of the trial, the data can arrive quite close to Database Lock, or even after. Especially in short, early development trials this is the case. Testing of the algorithms mostly takes place on partial data (often only screening, so without the ability to test emerging and enriched variations) or even on SDTM-like excel files. This means that a lot of work is needed after the final data arrives, making it difficult to put the data into the Clinical Study Report (CSR). Mostly we provide the results in a separate virology report.

CONCLUSION

Deep sequencing is a very powerful technique to see how variations emerge, disappear or get enriched during a trial. While some of them will be random variations, there will be variations due to the treatment pressure. Emerging variations can point to treatment resistance, while disappearing variations mean those are more vulnerable to the treatment.

It is however a technique that has its challenges for which solutions must be found, requiring a lot of resources to analyse the results and good cooperation between all parties involved. For all challenges there are multiple solutions, as we outlined some examples above.

REFERENCES

PharmaSUG 2017 - Paper DS05: Implementation of STDM Pharmacogenomics/Genetics Domains on Genetic Variation Data, Linghui Zhang, Merck & Co., Inc., Upper Gwynedd, PA USA
CDISC PGx model: https://www.cdisc.org/standards/foundational/pgx

ACKNOWLEDGMENTS

I would like to thank the J&J Id&V virology department and specially Leen Vijgen and Thierry Verbinnen for helping to understand the scientific background of Deep Sequencing and for always taking time to answer our questions.

CONTACT INFORMATION

(In case a reader wants to get in touch with you, please put your contact information at the end of the paper.)
Your comments and questions are valued and encouraged. Contact the author at:

Koen Ven
Janssen Research & Development, A Division of Janssen Pharmaceutica NV
Turnhoutseweg 30
2340 Beerse
Belgium
Tel. +32 14 607034
E-mail: kven1@its.jnj.com