There is more to the genome: increasing the range of actionable high-confidence neoantigen discovery by whole genome sequencing

Next-generation sequencing has become indispensable in current clinical diagnostics settings. In cancer settings in particular, sequencing enables the exhaustive molecular characterisation of tumours, paving the way for better subtype classification, tailored therapies based on the presence of select driver mutations, and even fully personalised therapies leveraging the molecular make-up of the tumour against itself.

Historically, sequencing-based diagnostics have relied on either whole-exome sequencing (WES) or targeted panels, interrogating part or all of currently annotated coding exons. Continuous improvements of these technologies have drastically reduced cost and downstream processing requirements over time, resulting in a broad adoption. Now, with further drops in sequencing costs, the question has arisen whether it is feasible and advantageous to transition from WES to whole genome sequencing (WGS). In the latter, the entire target genome is sequenced, instead of the limited exonic fraction (1-3 %) provided by WES. Although based on a shared set of technologies, both techniques differ in their requirements and in their ouput, and therefore several practical and conceptual considerations have to be taken into account when choosing the correct approach (Venter et al., 2001; Sakharkar et al. 2004).

The detailed outline below clearly confirms the benefits of exchanging WES by WGS in clinical personalised variant identification workflows. WGS is not only more sensitive and generates more reliable calls, but also allows the identification of more variant types in comparison with WES. Especially the increased detection rate of structural variants and fusion events in WGS is likely to have a significant effect on personalised cancer therapies, since peptides from such features have the potential of being more immunogenic than SNV/indel-generated peptides, due to their higher degree of dissimilarity from self. In addition, the ability of WGS to elucidate variants in non-coding regions of the genome, which have been shown to produce peptides, will further increase the pool of potential new therapeutic candidates.

Especially in the context of low tumour mutational burden (TMB) tumours these advantages have a significant impact since these tumour types are characterised by extremely low levels of SNVs and indels, making it difficult to identify suitable targets for personalised therapy. Hence, by harnessing the larger enrichment region associated with WGS, it is possible to maximise the chance of finding potent therapeutic candidates in these hard-to-treat cancers.

General differences between WES and WGS

WES and WGS practical differences boil down to cost, time, input quantity, and sample preparation method. While the price has dramatically reduced in recent years, **per-sample cost** of the WGS technique remains two to four times higher than for WES. **Input-wise**, there is not a lot of difference between WES and PCR-free WGS methods, both commonly require between 50ng and 1µg of input material. In contrast, input requirements of PCR-based WGS approaches are much lower and range from 1ng up to 500ng. In preparation of the WES sequencing itself, preparation kits are commonly used that enrich their targets using probes capturing all or a selection of coding sequences. An example of the latter is the Illumina Trusight One Sequencing panels which together target approximately 6700 disease-related genes. In contrast, **sample preparation** for WGS does not require target enrichment and is commonly performed by means of PCR-free approaches. Since WES kits focus on a limited region of the genome in comparison with WGS platforms, **downstream processing** of WES data is also more straightforward and comes with less IT requirements such as storage and computing power. In general, it is estimated that WGS sequencing data requires on average 5 to 20 times more storage capacity than WES, and that data processing takes

approximately 1 to 2 and 4 to 5 days for WES and WGS, respectively. An overview of the most relevant differences between WES and PCR-free WGS is summarised in **Table 1**.

Coverage differences between WES and WGS

Probe-based capture is known to underperform for high-GC sequences. As such, WES sequencing data is prone to **drops in coverage at GC rich targets** (Carss et al., 2017; Trudsø et al., 2020). The same holds true for PCR-based WGS approaches where the PCR amplification step tends to result in reduced coverage in regions having GC percentages below 20-30% or 60% and above (Björn N & Sahlén, 2018; Meienberg et al., 2016, Bailey et al., 2020). In contrast, PCR-free methods are able to completely cover all GC-rich exons (WES covers only 93.6% of such exons) and generate more uniform coverage, evenly distributed across the complete genome. This results in a higher detection rate of variants in GC-rich sequences, i.e. 18% for WGS vs 5% for WES (Trudsø et al., 2020). Another reason for suboptimal exon coverage in WES is that these platforms typically only target a subset of the known exons (e.g. no UTRs), commonly the ones associated with the most abundant transcripts. For example, the Agilent SureSelect v5 + UTR capturing kit only covers 98.25% of the exons linked to genes recommended by the American College of Mecical Genetics for mutation screening (Meienberg et al., 2016). Some of these exons might contain potentially deleterious variants that could be strong targets for cancer therapy, and as such would be missed by WES.

Table 1. Most important differences between WES and WGS PCR-free based library preparation workflows.

	WES	**WGS (PCR-free)**
Practical considerations		
Input requirements	50 – 1000 ng	50 – 1000 ng
Time (sample prep)	6.5 h	4 h
Time (IT processing on 60 cores)	1-2 days	4-5 days
Cost	400 EUR	1600 EUR *(4-6x higher than WES)*
Variant calling		
Rate of True Positive calls *(for WGS/WES unique calls)*	SNV : 10% Indel : 50-55% CNV : 20%	SNV : 75% Indel : 50-55% CNV : 90%
Sensitivity	SNV : > 95% in exome, 5% outside exome Indel : 72%	SNV : > 97% in exome, >95% outside exome Indel : 20%
Coverage (average)	120x *(usually ~2-3x WGS coverage)*	40x
Variant types		
CNV	Possible, but less reliable	Possible
Fusion genes	Possible, but limited to exonic breakpoints and less reliable	Possible
Structural variation	Possible, but limited to exonic breakpoints and less reliable	Possible

Since **coverage depth** has a profound impact on the accuracy and sensitivity of calling variants, it is one of the major aspects taken into account by studies comparing WGS and WES workflows from a diagnostic viewpoint. When focusing on variants in protein-coding regions, results have shown WES coverage data to be skewed towards lower coverage, while WGS commonly displays a normal-like coverage distribution. This means that in general, WES generates a higher percentage of low(er) covered variants in comparison

to WGS as described by Belkadi *et al.* (Belkadi et al., 2015). They observed that the fraction of low-covered SNV calls (DP < 8) was higher for WES (4.3%) in comparison to WGS (0.4%). However, when only looking at SNVs within the WES target regions, a recent study by Barbitoff *et al.* in Scientific Reports describes no differences in variant coverage between the two platforms and reports that WES can achieve similar coverage statistics as WGS when the average coverage of the former is 2 to 3 times higher (Barbitoff et al., 2020; Clark et al., 2011; Parla et al., 2011).

Variant differences between WGS and WES

Also regarding **variant quality**, conclusions are conflicting. The study by Barbitoff *et al.* claims that the current best performing WES platforms are indistinguishable from WGS when focusing on SNV calling quality (Barbitoff et al., 2020). Only for indels, slightly fewer low-quality calls were generated by WGS in comparison with WES platforms. In contrast, older studies show that WGS outperforms WES for variant detection, with WGS resulting in more calls with better quality when considering both variant coverage and genotype quality (Belkadi et al., 2015; Björn N & Sahlén, 2018). The study by Belkadi *et al.* showed that variants called by both types of methods overall have a very good genotype concordance (SNV : ~96%, deletions : ~74% and insertions : ~85%) (Belkadi et al., 2015) , which was confirmed in the study of Bailey *et al.* showing an overall concordance of 76.7%, with lower concordance for indels (57%) in comparison with SNVs (79%) (Bailey et al. 2020). The majority of the variants having a discordant genotype in WGS and WES were called as homogeneous by WES and heterogeneous by WGS. The latter might indicate some sort of allelic bias in WES platforms, where probes preferentially bind the wild-type sequence of an allele. However, this is contradictory to the observations of Barbitoff *et al.* who saw no differences in allelic bias between the WGS and WES platforms they compared (Barbitoff et al., 2020). When looking at the discordant calls (present in WGS but not WES, or vice versa), Belkadi *et al.* and Björn *et al.* observed that most SNVs uniquely called using WES are false positives (91%), while this was less the case for WGS-unique calls (25%) – as confirmed by Sanger sequencing for a subset of such variants (Belkadi et al., 2015; Björn N & Sahlén, 2018). Looking at the genomic context of these WES false positive hits, the majority – approximately 57% according to Belkadi *et al.* - are located in exon flanking regions that have been shown to generate lowered covered variants (Belkadi et al., 2015). Similarly, Bailey *et al.* indicated that – in addition to GC content bias – subclonal variants and variants with low allele frequencies account for the majority of the WGS- and WES-unique variants, and that false positive rates can greatly be reduced by combining multiple variant calling methods (Bailey et al., 2020). In such contexts, the more uniform coverage is expected to favor WGS methods. For indels, 52.2% and 55.2% of the calls unique to WGS and WES, respectively, could be validated by Sanger sequencing. This means, combined with the fact that on average 6-7 times more SNVs and 2-3 times more indels are uniquely called by WGS in comparison with WES, that WGS will typically generate a higher number of true-positive calls (Belkadi et al., 2015). In the context of personalised therapies, this implicates that WGS based pipelines will generate more and a broader scope of variants than WES, thereby significantly increasing the possibility of identifying potent candidates for therapy.

In addition to the quality of the variant calls, WGS and WES also differ in the **number of variant types** they can detect. In the context of copy number variants (CNV) and fusion gene detection, WGS has a clear advantage over WES (Carss et al., 2017). Breakpoints for such features can often be located in regions outside the ones enriched by WES platforms, as indicated by the findings of Belkadi *et al.* showing that 97.3% of the WGS CNV calls had breakpoints outside WES targets (Belkadi et al., 2015). This means that more fusion genes and CNVs are commonly detected using WGS in comparison with WES, with WES-based CNV analysis also being more complex and less accurate (Gilissen et al., 2014; Meienberg et al., 2015). Concordance between the two platform types is also quite low according to Belkadi *et al.*, with 93% and 14.7 – 22.7% of the called CNVs already known from literature for WGS and WES, respectively (Belkadi et al., 2015). This indicates that CNVs called by WGS are potentially more accurate.

WGS outperforms WES when combined with RNA sequencing

Because of the limited target size of WES, variation in non-coding regions -such as located in promotor-, silencer-, and enhancer regions as well as deep intronic mutations that could potentially impact splicing - will only be detected by WGS (Carss et al., 2017). This is also one of the reasons why WGS is preferred above WES in workflows combining multiple sequencing platforms. For example, a WGS identified deep intronic mutation could help substantiate an alternative splicing event elucidated by RNA sequencing (RNA-seq), which is not possible when using WES. The same holds true for fusion genes were the breakpoints fall outside the targets enriched by WES. According to Rusch *et al.*, the WGS/RNA-seq is overall also able to identify more variants in comparison with the WES/RNA-seq combo, especially in the case of CNVs (Rusch et al., 2018). Data from the study show that the highest impact can be observed when combining WES and RNA-seq, resulting in a 205% increase in indentifiable structural variants. However, the WGS/RNA-seq combination is still the most sensitive approach, able to elucidate 94.8% of the variants (91.2% for WES/RNA-seq), with the highest increase for indels (14.8%) in comparison to the WGS-only workflow. In addition to an increased sensitivity, extending WGS or WES-workflows with RNA-seq analyses further allows better identification of novel neoisoforms (thus validating identified splice variants), and enables filtering of variants located in expressed genomic regions,

WGS benefits in neoantigen discovery for cancer treatment

In the context of neoantigen elucidation for cancer treatment, the benefits of WGS over WES are multiple. Besides the better variant quality in targets covered by both WGS and WES, WGS results in more identified structural variants and fusion events, and enables the discovery of variations in yet uncharacterised coding regions or non-coding regions having protein-coding potential (**Figure 1**). Especially the latter is important as more and more evidence is emerging that pseudogenes, and long non-coding (lncRNA) as well as circular (circRNA) RNAs are also sources of (micro)peptides translated from short open-reading fragments (Ji et al., 2015; Xiang et al., 2021). Examples include the 46 amino acid peptide translated from the LINC00948 locus and proteins encoded by circRNAs circ-ZNF609 and circ-SHPRH (Anderson et al., 2015; Pamudurti et al., 2017; Zhang et al., 2018). Taking into account the higher degree of tissue-specificity of the non-coding host genes, such peptides might become a powerful new source of neoantigens for cancer treatments.

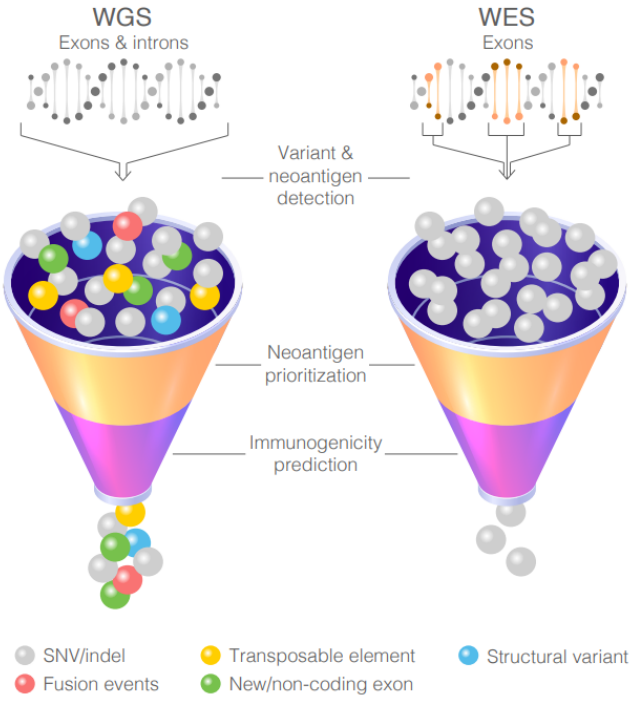

Figure 1. Comparison of the antigen search space covered by WGS versus WES.

About myNEO

myNEO (Ghent, Belgium) developed a platform enabling genomic-informed drug discovery in the key therapeutic areas of oncology and immunology. The data-driven ImmunoEngine identifies the most efficacious targets (epitopes) for each cancer patient, uniquely presented on the tumour cells and capable of redirecting a patient's immune system, leading to elimination of the cancer cells. The discovery platform enables targets to be identified even in hard-to-treat tumours with a cold/lowly mutated profile. Similarly, the company has applied its technology to identify immunogenic sequences in infectious diseases, capable of protecting populations with strong broad immune responses. myNEO is one of the companies that emerged from the Novalis biotech incubator fund at the end of 2018, founded by two leading entrepreneurs already known for several successes in the biotech industry: Wim Van Criekinge, professor of computational biology at Ghent University, and childhood friend Jan Van den Berghe.

Contact us

Interested in more information about myNEO? Contact us!

Lien Lybaert, PhD
Scientific Alliance Manager
lien.lybaert@myneotx.com

Anderson D *et al.*, A micropeptide encoded by a putative long noncoding RNA regulates muscle performance, Cell, 2015
Bailey M et al., Retrospective evaluation of whole exome and genome mutation calls in 746 cancer samples, Nature Communications, 2020
Barbitoff Y *et al.*, Systematic dissection of biases in whole-exome and whole-genome sequencing reveals major determinants of coding sequence coverage, Scientific Reports, 2020
Belkadi A *et al.*, Whole-genome sequencing is more powerful than whole-exome sequencing for detecting exome variants, Proceedings of the National Academy of Sciences of the United States of America, 2015
Björn N, *et al.*, Comparison of Variant Calls from Whole Genome and Whole Exome Sequencing Data Using Matched Samples, Next Generation Sequencing & Applications, 2018
Carss K *et al.*, Comprehensive Rare Variant Analysis via Whole-Genome Sequencing to Determine the Molecular Pathology of Inherited Retinal Disease, American Journal of Human Genetics, 2017
Clark M *et al.*, Performance comparison of exome DNA sequencing technologies, Nature Biotechnology, 2011
Gilissen C *et al.*, Genome sequencing identifies major causes of severe intellectual disability, Nature, 2014
Ji Z *et al.*, Many lncRNAs, 5' UTRs, and pseudogenes are translated and some are likely to express functional proteins, eLife, 2015
Meienberg J *et al.*, New insights into the performance of human whole-exome capture platforms, Nucleic Acids Research, 2015
Meienberg J *et al.*, Clinical sequencing : is WGS the better WES?, Human Genetics, 2016
Pamudurti N *et al.*, Translation of CircRNAs, Molecular Cell, 2017
Parla J *et al.*, A comparative analysis of exome capture, Genome Biology, 2011
Rusch M *et al.*, Clinical cancer genomic profiling by three-platform sequencing of whole genome, whole exome and transcriptome, Nature Communications, 2018
Sakharkar M *et al.*, Distributions of exons and introns in the human genome, In Silico Biology, 2004
Trudsø L *et al.*, A comparative study of single nucleotide variant detection performance using three massively parallel sequencing methods, PLoS ONE, 2020
Venter G *et al.*, The sequence of the human genome, Science, 2001
Xiang R *et al.*, Increase expression of peptides from non-coding genes in cancer proteomics datasets suggests potential tumor neoantigens, Communications Biology, 2021
Zhang M *et al.*, A novel protein encoded by the circular form of the SHPRH gene suppresses glioma tumorigenesis, Oncogene, 2018