CeGaT

Research &
Pharma Solutions

On the Quest for a More Precise Exome

Tech Note

On the Quest for a More Precise Exome

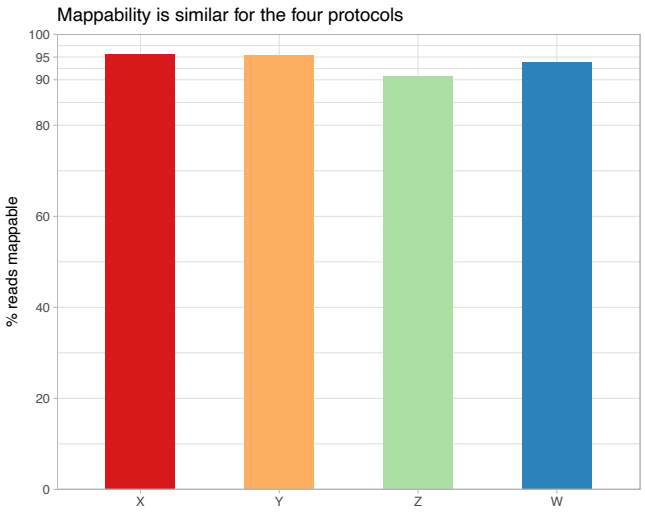
Whole exome sequencing (WES) has been extensively used for variant calling of protein-coding regions. Besides medical diagnostics for inherited diseases, WES can help address a broad spectrum of applications, ranging from patient's HLA genotype to treatment-related biomarkers like tumor mutational burden (TMB) and microsatellite instability (MSI).

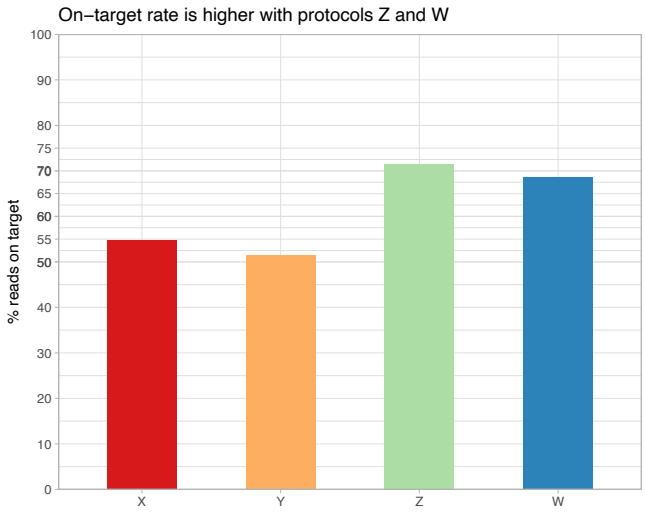
Technically, WES aims at a complete enrichment of all known coding regions and generation of highly uniform sequencing coverage to ensure optimal data quality for all applications. WES protocols are continuously being improved and the quest for a more precise exome that enables more confident variant calling is ongoing. Consequently, this TechNote is intended to address key quality metrics such as probe design efficiency, on-target rate, sequencing coverage, and uniformity of coverage of four WES protocols. We tested three newly released exome protocols on the market: *Agilent SureSelect Human All Exon V7* (as **Protocol Y** herein), *Illumina-IDT xGen Exome Research Panel v1.0* (as **Protocol Z**), and *Twist Human Core Exome* (as **Protocol W**). *Agilent SureSelect Human All Exon V6* (as **Protocol X**) was used to benchmark their performance. The same DNA sources were used for this test, all procedures were carried out according to vendor's original protocol, and reads were downsampled to mimic equivalent sequencing.

Based on the hg19 build, protocol **X**, **Y**, **Z**, or **W** is designed to target 38, 36, 39, or 33 Mb of highly conserved protein-coding regions, respectively. To enrich target regions, vendors design capture baits/probes either directly on or close to the desired regions. One parameter for evaluating vendors' designs is the efficiency of targeted versus baited/probed regions, i.e. how many Mb of genomic regions need to be baited/probed in order to enrich a desired target set. The more non-targeted regions are baited/probed, the more off-target reads will be produced, thus limiting the sequencing efficiency. This parameter varies greatly between protocols with **W** being the most efficient at 90%, which translates into minimal sequencing of regions outside of exons.

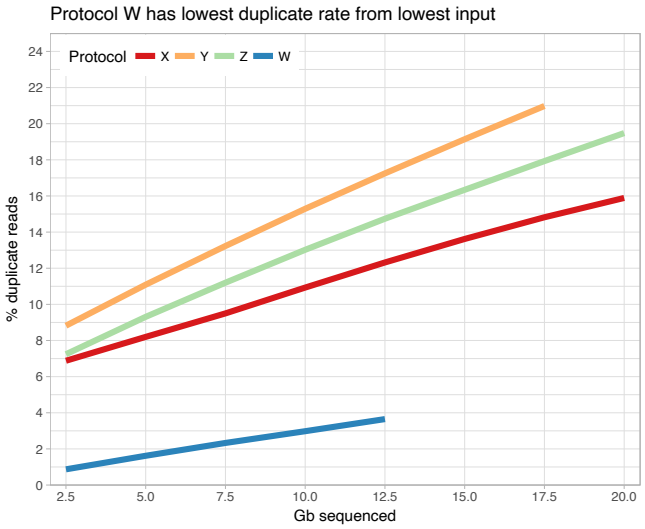
Protocol	X	Y	Z	W
Targeted Region [Mb]	38	36	39	33
Design efficiency [%]	64	72	76	90
Bait/Probe type	RNA	RNA	ssDNA	dsDNA
Input DNA (ng)	1000	1000	100	50
Pre-hybridized library (ng)	750	750	500	187.5

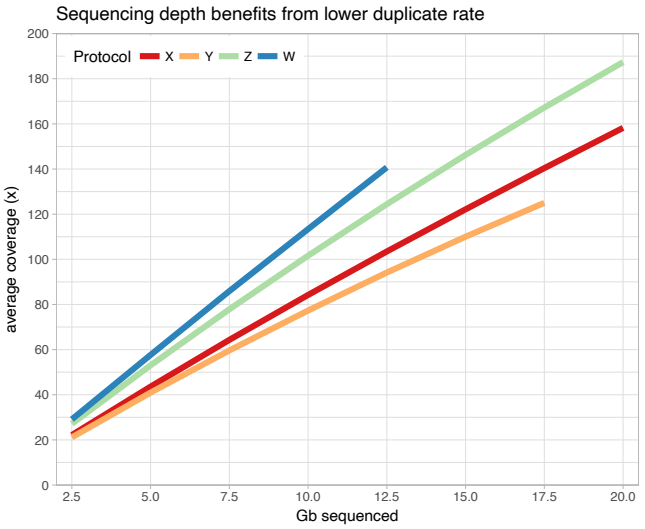
As for DNA input requirement, protocol **W** starts lowest at 50 ng compared to 1000 ng required by **X** and **Y**, and 100 ng by **Z**. Moreover, protocol **W** requires much less pre-hybridized library going into capture at 187.5 ng, instead of 500 or 750 ng required by its counterparts. Both lower than standard requirements are indicative of a higher library complexity that could be preserved by protocol **W**.



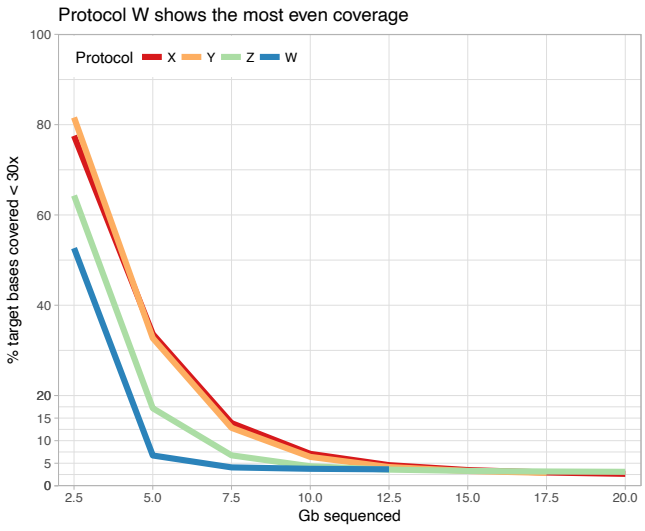


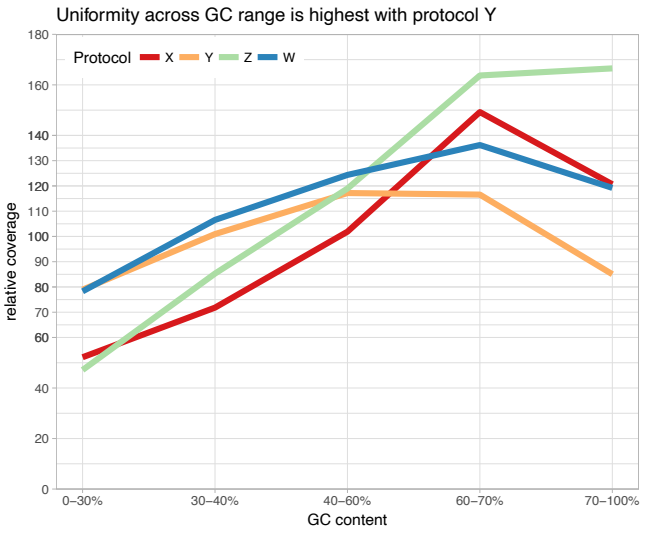
While the mappability is similar among these 4 protocols, the on-target rate is highest with protocol **Z**, followed by **W**, **X**, and **Y**. Both protocols **W** and **Z** have about 20% higher number of targets covered with 30 or more reads as compared to protocols **X** and **Y**.





More efficient read usage and higher library complexity of protocol **W** are reflected in the extremely low duplicate rate of <5% at 12 Gb of sequencing output. In contrast, protocols **X**, **Y**, and **Z** resulted in more than three-fold that duplicate rate at 12 Gb. As a result, protocol **W** could achieve higher coverage than other protocols at the same sequencing output, e.g. 7 or 12 Gb. As the slope of protocol **W** is steeper than the others, higher coverage could still be achieved beyond 12 Gb while the other protocols will reach saturation sooner.





The number of regions with coverage less than 30x is lowest with protocol **W**, followed by protocols **Z**, **X**, and **Y**. However, at 12 Gb of sequencing output this difference becomes minimal, meaning protocols **X**, **Y**, and **Z** would require much more sequencing in order to reach 30x of coverage on most targets. Protocol **Y** shows a higher uniformity across a GC range, closely followed by protocol **W**. Both protocols **X** and **Z** exhibited a strong GC bias whereas AT-rich regions are poorly covered.

In summary, protocol **W** has the lowest input requirement, high on-target rate, lowest duplicate rate, highest sequencing coverage, lowest underrepresented regions, and high uniformity of coverage. Compared to protocols **X**, **Y**, and **Z**, protocol **W** has superior conversion efficiency by producing highly complex library from 50 ng of starting material while evenly enriching regions independent of their GC content. Combining the aforementioned pros of protocol **W** with a very efficient probe design results in a high on-target rate that provides unprecedented coverage at minimal sequencing. Not only is this higher coverage providing more confidence in variant calling, but also making TMB estimation and MSI status more precise.

About Us

CeGaT was founded 2009 in Tübingen, Germany. Since then, we have been specialized in next-generation sequencing (NGS). Our scientists do not only provide NGS for genetic testing, but also a variety of sequencing services for research purposes.

Our dedicated Customer Solutions team of scientists and bioinformaticians works closely with you, to develop the best strategy for the realization of your project. Depending on its scope we select the most suitable library preparation and sequencer from our Illumina platforms.

We would be pleased to provide you with our service. Get in touch with us to start planning your project today.

CeGaT GmbH
Research & Pharma Solutions
Paul-Ehrlich-Str. 23
72076 Tübingen
Germany

Phone: +49 7071 56544-333
Fax: +49 7071 56544-56
Email: rps@cegat.com
Web: www.cegat.com

DAkkS Deutsche Akkreditierungsstelle D-PL-13206-01-00

Accredited by DAkkS according to DIN EN ISO/IEC 17025:2018

CAP ACCREDITED COLLEGE of AMERICAN PATHOLOGISTS

CLIA CERTIFIED ID: 99D2130225

2023.6.1