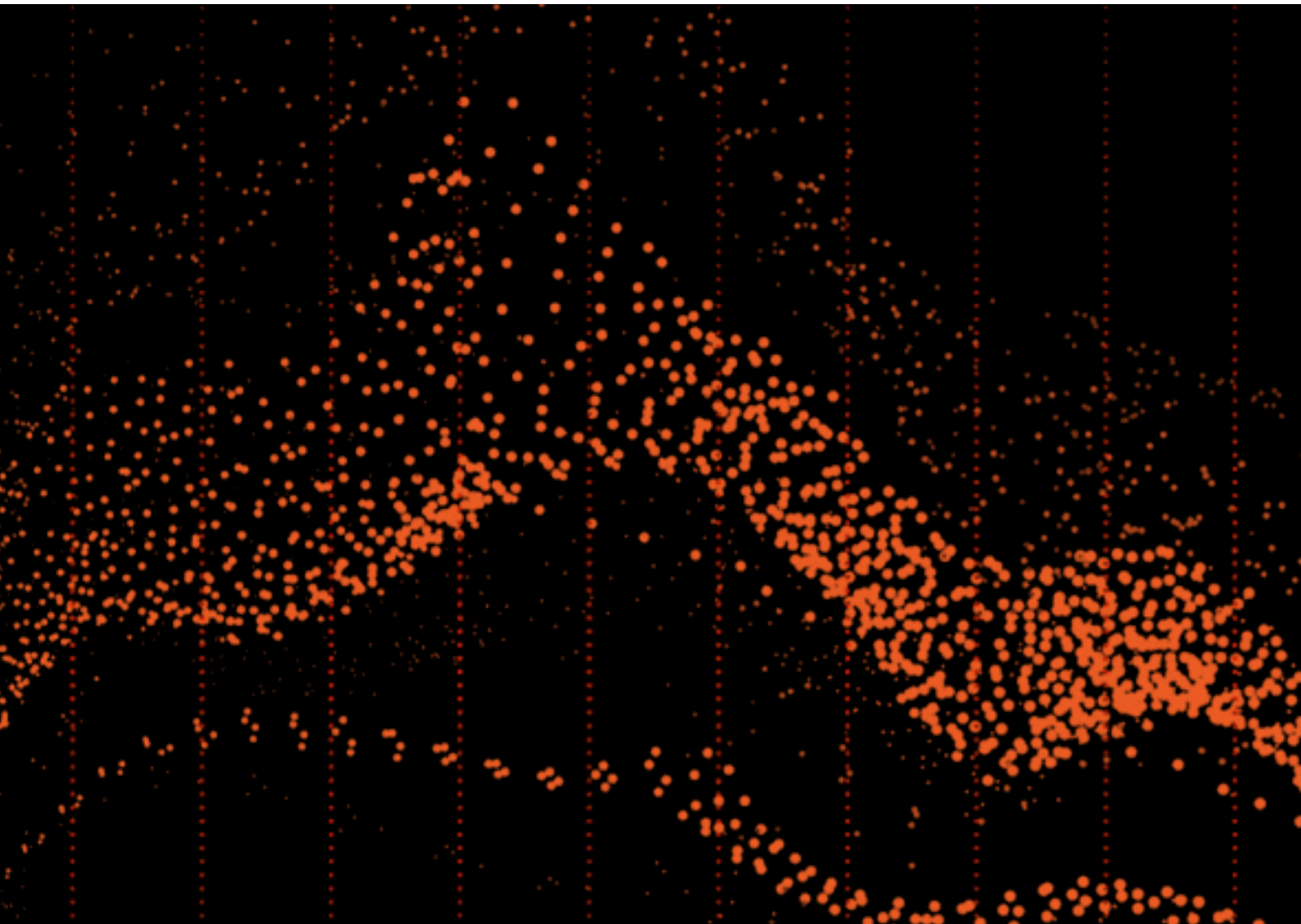
EuroClonality-NDC Assay

Analysis Guide

Release: September 2021

For **Research** Use Only. Not for use in diagnostic procedures.



EuroClonality-NDC Assay

This document is intended to guide the users through the analysis of the results generated by the EuroClonality-NDC assay and associated bioinformatics tools provided as part of ARResT/Interrogate[1,2]. Although every effort has been made to keep the information correct and up to date, we cannot guarantee that there are no unwilling errors or omissions. We encourage our customers to provide feedback which may lead to modifications and improvements in future versions of this document.

Table of Contents

Introduction

Intended Use

The EuroClonality-NDC is a capture-based targeted next generation sequencing (NGS) assay with customised bioinformatics pipeline for research use only. It has been designed and analytically validated for the integrated and qualitative detection of clonal immunoglobulin (IG) and T cell receptor (TR) rearrangements, translocations, copy number variation (CNV) and somatic mutations from suspected cases of lymphoproliferative disorders or lymphoid populations. The EuroClonality-NDC pipeline has been validated for use with genomic DNA (gDNA) isolated from fresh/frozen and formalin fixed paraffin embedded (FFPE) tissue. However, the CNV tool has been validated only for high molecular-weight DNA from fresh specimens such as peripheral blood, bone marrow or separated cells.

Accessing the EuroClonality-NDC bioinformatics pipeline

The EuroClonality-NDC bioinformatics pipeline is available from the ARResT/Interrogate platform and requires an account which is provided to users of the EuroClonality-NDC assay. The transfer of files is by FTPS (File Transfer Protocol Secure) and therefore you will need an FTP client, e.g. FileZilla, Cyberduck. You may need permissions and specific instructions from your institution, make sure you have these in place before proceeding.

IMPORTANT: to reduce security risks, do not share your user-name and password with anyone and always delete data/results after you have processed/downloaded them.

Please contact the team at contact@arrest.tools if you face any issues with these steps:

1. Go to the latest ARResT/Interrogate platform at http://arrest.tools/interrogate-latest/ preferably using Chrome.
2. Enter your username and password and click the 'log in' icon.
3. Once logged in, make sure the interface is displaying the Interrogate.ECNDC user mode as shown here:

4. To upload samples to analyse:
    a. Open your FTP client application.
    b. Establish a new connection (give it an informative name) and specify the encryption requirement to explicit authentication over TLS, e.g. if using FileZilla, select 'Require explicit FTP over TLS' under Encryption; if using Cyberduck, select 'FTP-SSL (Explicit AUTH TLS)'.
    c. Enter 'file.arrest.tools' under 'Host' or 'Server'.
    d. Enter your username and password (either in the client or when requested). You will need to click 'OK' to accept the certificate when prompted to continue. Make sure you have assessed the risks and you have obtained permission from your institution to do this.
    e. Upload the FASTQ files (in compressed .gz format) that you want to analyse in a single run to the 'data' folder. Restrict sample name characters to letters and numbers and use underscores/dashes instead of spaces. Note: if possible, do not analyse more than 24 samples at a time.
5. Switch to the 'processing' tab of ARResT/Interrogate - you should see your files under the 'samples' tab, if not, click on the 'refresh samples' button. If this still does not work, you may have to log out and log in again.
6. Name the run with letters and numbers and use underscores/dashes instead of spaces. If you want to overwrite a previous run with the same name (e.g. if there have been errors) tick the corresponding box.
7. You can choose one or more (comma separated) email addresses to receive notifications when the run is complete or if any errors occur.
8. When everything is ready click on the blue "test it" button and wait until all the samples are tested - make sure that you see the samples you intended to analyse in the test log (in the 'process output' tab). TIP: you can click on specific samples to run in the 'samples' tab. If everything is OK, you should be able to click on the green "process" button to initialise the pipeline.
9. Progress of the bioinformatics pipeline can be followed in real-time in the 'process output' tab or you can wait for an email if you have provided at least one correct email address – you may also log off / close the browser, the pipeline will keep running.
10. When the run is complete, you can download the output files from the 'results' folder on the FTPS server.

File Descriptions

EuroClonality-NDC output files

A series of run-specific and sample-specific outputs are available to download (file name/extensions in bold):

1. Run-specific outputs:
    a. **log**: summary of the run and analytical output per sample.
    b. **gathered.metrics.tsv**: summary metrics per sample (e.g., percentage of duplicate reads or percentage of bases on/near target) as well as some ARResT/Interrogate-specific metrics.
    c. **gathered.tsv & gathered.xlsx**: gathered results for IG/TR rearrangements (R), structural variants (SV) and sequence variants (VAR) for all samples in the run. The .xlsx file contains individual tabs in Excel for each of the samples included in the run, the .tsv file contains all the gathered data for all the samples.
2. Sample-specific outputs:
    a. **bam & bai**: alignment files and their indices required to visualise the data on any genome viewer such as IGV.
    b. **complete.vcf**: VCF file (v4.3) containing the complete list of sequence variants and corresponding annotation. All variants are included here (after removing known artefacts), irrespective of location, functionality, or population frequency.
    c. **curated.vcf**: VCF file curated to remove artefacts, most non-functional variants as well as most of the common polymorphic variants and synonymous variants.
    d. **CNV.*.html**: CNV plots for IG, TR and cancer related genes (onco.html) for each sample. For each CNV plot there is a choice of FFPE or HMW, and users are encouraged to use the one corresponding to the type of sample analysed. Only HMW has been analytically validated for CNV analysis.

Overview of the EuroClonality-NDC 'gathered' files

These files are the main output of the EuroClonality-NDC assay and contain information on all events apart from CNV. The following data items are included in the 'gathered' files:

name	R example	SV example	VAR example	description
sample	sample1	sample2	sample3	sample name as per FASTQ file
in FASTQs	6986785	12169894	3768296	total reads in sample FASTQs
duplicated	30.97%	34.38%	34.06%	percentage of reads duplicated
mapped	98.91%	98.88%	98.72%	percentage of reads mapped
on/near target	78.20%	76.21%	79.65%	percentage of bases on or near target
usable[a]	1599	2671	514	sum of fragments with rearrangement events
class1	R	SV	VAR	event class, R = rearrangement, SV = structural variant, VAR = sequence variant
class2	VJ:Vh-(Dh)-Jh	translocation	SNV	junction class (for R) or variant type (for SV/VAR)
locus	IGH			IG or TR locus (only for R)
gene	IGHV1-45	MYC	PIK3CA	gene involved in the event
gene partner	IGHJ6	IGHswc		partner gene involved in R/SV event
coord chr	14	8	3	chromosome containing event
coord pos	106506996	127736912	179234297	chromosomal coordinate
coord chr partner	14	14		partner chromosome
coord pos partner	105863195	105710469		partner chromosomal coordinate
% or ratio	19.0909	33.7657	28.9044	relative abundance, in fragments, in locus (for R/SV) or variant allele frequency (for VAR)
numerator	21	755	124	number of fragments containing the event
denominator	110	2236	429	sum of fragments in locus (for R) or total depth (for SV/VAR)
in class	19.2661			relative abundance, in fragments, in class2 R (junction class)
of genes	95.4545			relative abundance, in fragments, in locus, of the clonotype genes involved in the R (event1)
event1	V1-45 J6	MYC / IGHswc	PIK3CA	clonotype genes (for R) or genes involved in the SV (for SV) or gene containing the sequence variant (for VAR)
event2	-0/7/-4		c.3140A>G p.His1047Arg	clonotype segmentation (for R) or HGVS (for VAR); empty for SV
event3	CAR*GE#TS-DYGMDVW	t(8;14)	PASS;ECNDC.v2.PON=curated	clonotype amino acid junction sequence (for R) or translocation summary (for SV) or variant annotation category[b] (for VAR)
event comments	[...][c]		[...][c]	functionality call and safety tags (for R) or variant annotation (for VAR); empty for SV

name	R example	SV example	VAR example	description
junction	CAR*GE#TS-DYGMDVW	8:127,736,912 / 14:105,710,469	H/R	clonotype amino acid junction sequence (for R) or genomic breakpoint coordinates (for SV) or amino acid change (for VAR)
junction nt seq	tgtgcaagata GGGAGAG ....#ctacttctg actacggtatgg acgtctgg	GCCTTATGAA TATATTCACGC TGACTCCCG GC GCTCAGG TCAGCCCAGT CCAGCTCAAT CCA	cAt/cGt	clonotype nucleotide junction sequence (for R) or clipped (30nt before/after) genomic breakpoint sequence (for SV) or nucleotide triplet change (for VAR)
sequence context	[...][c]	[...][c]	A>G	clonotype nucleotide sequence (for R) or genomic breakpoint sequence (for SV) or Ref/Alt sequence change (for VAR)[d]

Shaded rows indicate the value may have different meaning, depending on the type of variant. [a] 'usable' is currently based on fragments with rearrangements, and is an indicator for quality and information content of the sample (see 'ANALYSIS: IG and TR rearrangement analysis' section below). [b] Variant annotation category contains the values available for 'event3' when analysing sequence variants (VAR): they provide a combination of overall quality (e.g. PASS, f0.02, etc) and whether the variant has been classified as 'curated' or 'complete' (see Analysis section). [c] These tend to be long strings of text, for examples please refer to the downloadable files at www.univ8genomics.com. [d] Note these sequences may not be accurate as they are based on consensus contigs from the aligned reads.

Analysis

IG and TR rearrangement (R) analysis

Each rearrangement (class1 = 'R') row in the 'gathered' files represents a specific clonotype for either a V-D-J, V-J, D-J, Kde, D-D or V-D event. The analysis provides all types of rearrangements, including complete, incomplete and even those that occur between different loci (i.e. TRA-TRD, which can also appear as SV).

As opposed to PCR-based NGS methods, where the V-D-J rearrangement sequences have the same start-end positions (i.e. primer locations), the EuroClonality-NDC data is generated upon hybridisation of a shotgun library preparation and, therefore, clonal fragments from different cells will have different start and end positions, making it more complex to consistently assign reads to clonotypes. This results in multiple clonotypes sometimes being assigned for an actual single clonal event. For example, a sample with a clonal TRGV4-TRGJ2 rearrangement, could have a number of entries for such rearrangement with different clonotypes (including 'X' if the junction is not fully sequenced), each of them with varying number of reads (i.e. 'numerator' in the 'gathered' files). Users should check the 'of genes' and 'event1' columns, which the report uses to provide information on combined frequencies, by assuming that all clonotypes with the same genes (e.g. TRGV4-TRGJ2) come from the same event, irrespective of the sequence. Although this may result in potentially missing a second rearrangement with identical genes and different CDR3, these are rare events and given the additional information for all other R, SV and VAR events included in the report, should not lead to confusion in most cases.

Based on the EuroClonality-NDC validation performed by the EuroClonality-NGS working group, a minimum of 6 reads/fragments are required to assign clonality to a given rearrangement. This means that users can focus on analysing only events with 'numerator' ≥ 6 for the purposes of clonality. This value assumes similar unique depth as achieved during the validation work, which generally corresponds to 'usable' reads values of between 1,000 and 3,000.

Thanks to the unique design and integrated analysis of the EuroClonality-NDC, we can estimate relative abundance of clonal rearrangements within each locus (e.g. IGK) and within each class (e.g. either Vk-Jk, Vk-Kde or intron-Kde). These values (i.e. '% or ratio' and 'of genes' for locus and 'in class' for class) provide valuable evidence on the likelihood that the events are clonal, and should be used alongside the total number of reads and clonotype checks.

For analysis of IG/TR rearrangement repertoires, where each rearrangement is expected to occur in only 1 or 2 reads, no filtering should be performed, other than perhaps by 'class2' to omit non-functional/incomplete rearrangements. In cases where low-level clonal populations are expected (e.g. immuno-oncology studies or TIL analyses in solid tumours), particularly when analysing lymph nodes or any other lymphocyte-rich specimen, some rearrangements may be present in relatively high numbers, although as opposed to clonal lymphoproliferations, their relative abundance may be low.

Structural variant (SV) analysis

Analysis of SV is relatively easy as the number of events per sample with ≥ 6 fragments is generally very low and most of which are 'double calls' of the same event. This is generally due to a slightly different breakpoint coordinate or annotation being identified because of alternative mapping.

SV can be confirmed visually using IGV or any other genome visualisation tool and we recommend performing this step in all instances. It can also help ascertain the exact makeup of the breakpoint and whether it is a balanced translocation or not.

The main limitation of SV analysis is the capture of at least one of the regions involved in the translocation. In particular, the EuroClonality-NDC can only detect SV in the regions described in the Quick Reference Guide. For example, MYC translocations can only be detected if the partner is included in the panel (e.g. IGH, IGK, IGL) or if the breakpoint in MYC falls within the region covered. Additionally, there are areas that are not fully covered or where the breakpoints may fall outside of the regions covered, in particular IGH switch regions, where up to 10-15% of breakpoints could be potentially missed.

Sequence variant (VAR) analysis

There are two main ways of analysing sequence variants with the EuroClonality-NDC assay, one is using the VCF files produced and the other one is using the 'gathered' files. There are two VCF v4.3 format files, one labelled 'complete' containing all variants (after removal of potential artefacts), and the other labelled 'curated' containing the more functionally relevant variants (see above under 'output files' section).

Both are annotated using VEP on hg38 (ensembl.org) and including the MANE SELECT transcript (or MANE CLINICAL) when available and appropriate (note that for BRAF we do not use MANE SELECT due to the disparity with the historical classifications and clinical and scientific literature).

Analysis of the sequence variants can be carried out by focusing on the 'VAR' class1 in the 'gathered' files. All variants included in the 'complete' VCF file are included in the 'gathered' files, although there is a flag under 'event3' that allows users to filter by only 'curated' variants. The variants with the 'curated' flag represent the most likely relevant variants for most analyses, including those that are functional, non-synonymous and present at low allele-frequency in the global population. Conversely, the variants with the 'complete' flag represent non-functional (or at least not clearly functional, such as intronic, regulatory, etc.), and/or synonymous and/or highly polymorphic variants. While the 'curated' flag is recommended for most applications the 'complete' flag can be used for specific applications or troubleshooting requiring 'identification' such as investigating contaminations, chimerism, or MRD and cfDNA analysis. The 'curated' variants include splice donor/acceptor sites within 2bp of the exon. If you need to analyse wider splice regions then use the 'complete' list. Note that the 'complete' VCF files contains all variants, while the 'complete' flag in the 'gathered' files contain only the non-'curated' variants.

Variants are reported here down to > 1% variant allele frequency (VAF), although the EuroClonality-NDC assay has been validated to report variants at > 4% VAF. Variants present at 1-4% VAF are likely the result of background noise and should not be analysed routinely. Nonetheless, the file provides these low-level variants for users to consider when expecting low-level variants or when analysing follow-up samples or known hotspot variants.

Copy number variant (CNV) analysis

For each sample, there are six CNV plots in web-browser-loadable HTML format. These are designed for users to be able to perform CNV in high molecular weight (HMW) and formalin-fixed paraffin-embedded (FFPE) sample types. Performance metrics are only available for CNV in HMW samples (see below).

The plots highlight each location in different colours, according to the likelihood of CNV being present: green are regions likely to be lost (logR: 0.7-0.8), blue regions are highly likely deleted (logR < 0.7), orange represent potential gains (logR: 1.3-2) and red are areas likely to be amplified (logR >2). There are three CNV plots per sample type: IG, TR and Onco.

IG and TR: these plots display CNV in the IG and TR genes, respectively. Since the process of V-(D)-J recombination results in the deletion of the intervening genes, these CNV plots are useful to assess clonality and confirm the specific genes involved in the rearrangement (for examples of these plots visit www.univ8genomics.com).

Due to limitations of the capture-based NGS technology and the lack of some polymorphic IG/TR genes in the reference human genome, not all possible genes are displayed in these plots. Some have been taken out due to lower performance and some of these (but not all) may be associated with polymorphic variants. Some polymorphic variants, however, do remain in the CNV plots, which sometimes can cause confusion when visualising the plots - i.e. if the specimen contains the polymorphic loss. In those instances where the rearrangement involves genes not included in the display, the CNV plots should show the respective copy number loss of the genes up until the gene closest to the one reported in the 'gathered' files.

It is normal to see artefactual copy number gains in TR genes in cases with high number of IG rearrangements (either clonal or polyclonal) and vice versa, due to the way the baseline copy number values are generated for this analysis. It is of course possible that these gains or even amplifications can be seen in cases with aneuploidy in the respective chromosomes or chromosomal regions. Generally, however, gains of these regions are of no significant scientific or clinical relevance.

Onco: these plots display CNV in oncogenes and tumour suppressor genes, as well as some additional, clinically-relevant genomic locations included in the panel (e.g. 13qMDR, 11q, 17p, etc.).

A validation study carried out in HMW samples for four genomic regions (11q, 13q, chr12 and 17p) resulted in overall >95% sensitivity and specificity. This performance is limited for samples with CNV present in >20% of the DNA and therefore the recommended minimum tumour content is 40%; this can be reduced to 20% if the alterations are homozygous.

The main limitations of the CNV analysis for cancer-related genes are: i) they have only been validated in HMW-DNA; ii) they can be highly affected by DNA quality and purity (ddH_2O elution is highly recommended); iii) CNV analysis is significantly affected by hyperdiploidy and CNV analysis should not be performed in samples with >20 regions altered; iv) under- or over-clustering can have a negative effect in CNV analysis (e.g. if using significantly more or less samples than recommended).

Analysis tips

VCF compatibility: the VCF files contain only sequence variant information in v4.3 format.

Integrated analysis: the 'gathered' files contain all the results apart from CNV. Users can filter and sort by any given header, to simplify analysis. It can be helpful to see all the types of variants together, to have a holistic view of the entire sample. Due to the large number of individual events, for suspected clonal specimens, this can be best achieved as follows:

- Filter your results by 'numerator' ≥ 6 to show events supported by at least 6 fragments/reads (for higher coverage samples or if high clonality expected, this can be raised to >10 or even higher)
- For VAR, filter 'event3' to omit any variant that does not contain the keyword 'complete' (this will show only 'curated' VAR, without filtering out R/SV events)

These settings will generally provide a list of approximately 20-40 variants per sample, including R, SV and VAR. Further refinement can be achieved by setting more stringent filters (e.g. '% or ratio' > 4 will only show VAR with VAF>4% and R present in >4% in class, plus all SV) at the risk of missing low-level or sub-clonal events.

False positive clonal R: due to the way capture-based data is analysed and the limited diversity of some genes, there are two types of rearrangements that can lead to potential false positives being reported, these are IGHD7-27 and IGKde-IGKIntron rearrangements. The former is a result of the close proximity of IGHD7-27 to IGHJ1 and can result in sometimes reporting IGHD7-27 with IGHJ2 as a clonal rearrangement. In the case of Kde-Intron rearrangements, this is due to the limited diversity existing in this type of rearrangements. In both cases, the number of total reads and relative abundance data can be analysed to ensure they are high enough and in keeping with other rearrangements in the same sample, to ensure they are true positives (note that these two types of rearrangement are not generally subjected to somatic hypermutation). No clonality calls should be made on the basis of either of these rearrangements in isolation.

Example data files

Please visit the Resources section of www.univ8genomics.com to download example files from a number of cell lines analysed with the EuroClonality-NDC assay

Performance specifications

The validation and performance characterisation of the EuroClonality-NDC was carried out on 280 samples comprising a range of different B and T cell malignancies using a different version of library preparation and hybridisation methodologies. The EuroClonality-NDC assay and analysis workflows have been updated to improve performance and target coverage.

Overall performance of the EuroClonality-NDC assay was assessed by identifying underperforming regions using DNA extracted from a panel of 91 normal tissue samples. Regions were classed as underperforming if coverage was >2 standard deviations (SD) below the mean in ≥50% samples. Only 10 underperforming coding regions have been identified that could affect mutation detection in the genes and exons shown in the table below.

Gene Name	Chromosome	Exon(s)
ARID1A	chr1	6
CARD11	chr7	12
JAK2	chr9	15
TRAF2	chr9	1
PAX5	chr9	10
STAT5B	chr17	18
TCF3	chr19	7
BTK	chrX	18
PHF6	chrX	6
PHF6	chrX	3

Expected Metrics

The table below shows the approximate expected metrics of the EuroClonality-NDC assay based on 22 samples per hybridisation reaction (plus controls) and run on a NextSeq500/550 (Mid-Output 150 cycles cartridge) at 2 x 75bp. These can vary depending on DNA quality and quantity, technical factors in each laboratory, differences in equipment, etc.

Parameter	Expected HMW-DNA	Expected FFPE-DNA
Library yield (ng)	4,000 ng (range 1,000-8,000)	
Library average fragment size (bp)	330 bp (range: 275 – 450 bp)	290 bp (range: 245 – 330 bp)
Average on/near target bases (%)	80% (range: 70 - 85%)	
Duplicate reads (%)	40% (range: 30 – 50%)	
Mean unique target coverage (reads)	1,000x (500 – 1,500x)	

This improved version of the EuroClonality-NDC assay and workflow underwent a verification bridging study with gDNA extracted from 22 FFPE and 44 HMW samples which were included in the original validation study, and which comprised a broad range of B and T cell malignancies. The following table represents the observed analytical performance:

	Sensitivity	Specificity	Limit of Detection
IG/TR Rearrangements	>97%	>97%	5% VAF
Translocation	>95%	>99%	5% VAF
SNV	>99%	100%	4% VAF
CNV*	>95%	>97%	20%**

*SNV: single nucleotide variant; CNV: copy number variation. *CNV have been validated in HMW-DNA for trisomy 12 and deletions of 11q, 13q and 17p. No performance data available for FFPE or other genomic regions.**Deletions can be detected if present in ≥ 20% of the DNA. The threshold baseline fold-change for deletions detected by EuroClonality-NDC is 20% (i.e. 0.8) for losses and 30% (i.e. 1.3) for gains, e.g. trisomy.*

Abbreviations

CNV	Copy number variation
FFPE	Formalin fixed paraffin embedded
HMW	High-molecular weight
IG	Immunoglobulin
SNV	Single nucleotide variant
TR	T cell receptor
SV	Structural variant

References

1. Stewart JP, Gazdova J, Darzentas N, Wren D, Proszek P, Fazio G, Songia S, Alcoceba M, Sarasquete ME, Villarese P, van der Klift MY, Heezen KC, McCafferty N, Pal K, Catherwood M, Kim CS, Srivastava S, Kroeze LI, Hodges E, Stamatopoulos K, Klapper W, Genuardi E, Ferrero S, van den Brand M, Cazzaniga G, Davi F, Sutton LA, Garcia-Sanz R, Groenen PJTA, Macintyre EA, Brüggemann M, Pott C, Langerak AW, Gonzalez D; EuroClonality-NGS Working Group. Validation of the EuroClonality-NGS DNA capture panel as an integrated genomic tool for lymphoproliferative disorders. Blood Adv. 2021 Aug 24;5(16):3188-3198. doi: 10.1182/bloodadvances.2020004056. PMID: 34424321.

2. Bystry V, Reigl T, Krejci A, Demko M, Hanakova B, Grioni A, Knecht H, Schlitt M, Dreger P, Sellner L, Herrmann D, Pingeon M, Boudjoghra M, Rijntjes J, Pott C, Langerak AW, Groenen PJTA, Davi F, Brüggemann M, Darzentas N; EuroClonality-NGS. ARResT/Interrogate: an interactive immunoprofiler for IG/TR NGS data. Bioinformatics. 2017 Feb 1;33(3):435-437. doi: 10.1093/bioinformatics/btw634. PMID: 28172348.

Revision History

Release	Summary of change	Date
2100012 r01	First released	September 2021

Technical Assistance

Univ8 Genomics Ltd
63 University Road
Belfast
Northern Ireland

Email: contact@univ8genomics.com
Web: univ8genomics.com