Universität Zürich[UZH]

hCNV Pipeline for Data Normalization in Oncogenomics

Hangjia Zhao
2023.11.30

elixir

pgxRpi

an interface API for analyzing Progenetix CNV data in R using the Beacon+ API

	All users	R users
Interface	 	 
Variant query	https://progenetix.org/beacon/biosamples/pgxbs-kftvh94d/g_variants	variants <- pgxLoader(type="variant",biosample_id="pgxbs-kftvh94d")
Output		


```
"results": [
    {
        "caseLevelData": [
            {
                "analysisId": "pgxcs-kftvu6cg",
                "biosampleId": "pgxbs-kftvh94d",
                "id": "pgxvar-5bab5837727983b2e0121e97"
            }
        ],
        "variantInternalId": "11:0-134452384:DEL",
        "variation": {
            "copyChange": "efo:0030067",
            "identifiers": {},
            "subject": {
                "interval": {
                    "end": {
                        "type": "Number",
                        "value": 134452384
                    },
                    "start": {
                        "type": "Number",
                        "value": 0
                    },
                    "type": "SequenceInterval"
                },
                "sequence_id": "refseq:NC_000011.10",
                "type": "SequenceLocation"
            },
            "variantAlternativeIds": []
        }
    },
    {
        "caseLevelData": [
            {
                "analysisId": "pgxcs-kftvu6cg",
                "biosampleId": "pgxbs-kftvh94d",
                "id": "pgxvar-5bab5837727983b2e0121e99"
            }
        ],
        "variantInternalId": "1:0-84699999:DEL",
        "variation": {
            "copyChange": "efo:0030067",
            "identifiers": {},
            "subject": {
                "interval": {
```

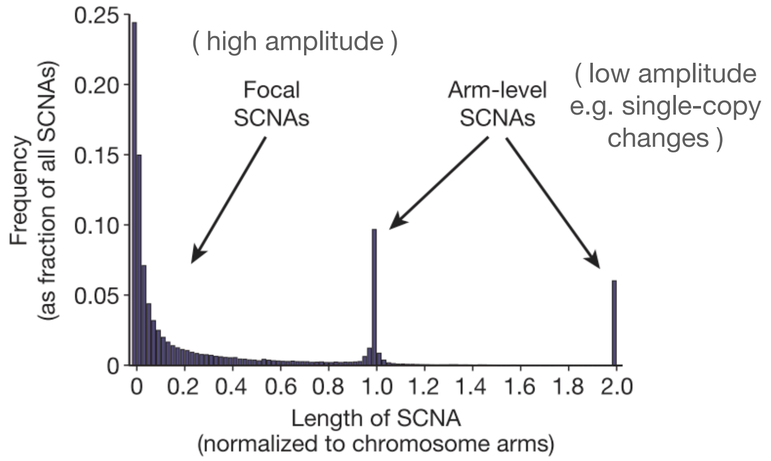

	variant_id	biosample_id	analysis_id	reference_genome	variant
1	pgxvar-5bab5837727983b2e0121e99	pgxbs-kftvh94d	pgxcs-kftvu6cg	refseq:NC_000001.11	1:0-84699999:DEL
2	pgxvar-5bab5837727983b2e0121e9a	pgxbs-kftvh94d	pgxcs-kftvu6cg	refseq:NC_000001.11	1:124300000-247249719:DEL
3	pgxvar-5bab5837727983b2e0121e9c	pgxbs-kftvh94d	pgxcs-kftvu6cg	refseq:NC_000002.12	2:12800000-61099999:DEL
4	pgxvar-5bab5837727983b2e0121e9d	pgxbs-kftvh94d	pgxcs-kftvu6cg	refseq:NC_000002.12	2:197100000-242951149:DEL
5	pgxvar-5bab5837727983b2e0121e94	pgxbs-kftvh94d	pgxcs-kftvu6cg	refseq:NC_000003.12	3:14700000-71799999:DEL
6	pgxvar-5bab5837727983b2e0121e8d	pgxbs-kftvh94d	pgxcs-kftvu6cg	refseq:NC_000004.12	4:35500000-191273063:DUP
7	pgxvar-5bab5837727983b2e0121e8e	pgxbs-kftvh94d	pgxcs-kftvu6cg	refseq:NC_000005.10	5:18500000-143099999:DUP
8	pgxvar-5bab5837727983b2e0121e91	pgxbs-kftvh94d	pgxcs-kftvu6cg	refseq:NC_000006.12	6:0-60499999:DEL
9	pgxvar-5bab5837727983b2e0121e92	pgxbs-kftvh94d	pgxcs-kftvu6cg	refseq:NC_000006.12	6:130400000-170899992:DEL

Github: https://github.com/progenetix/pgxRpi

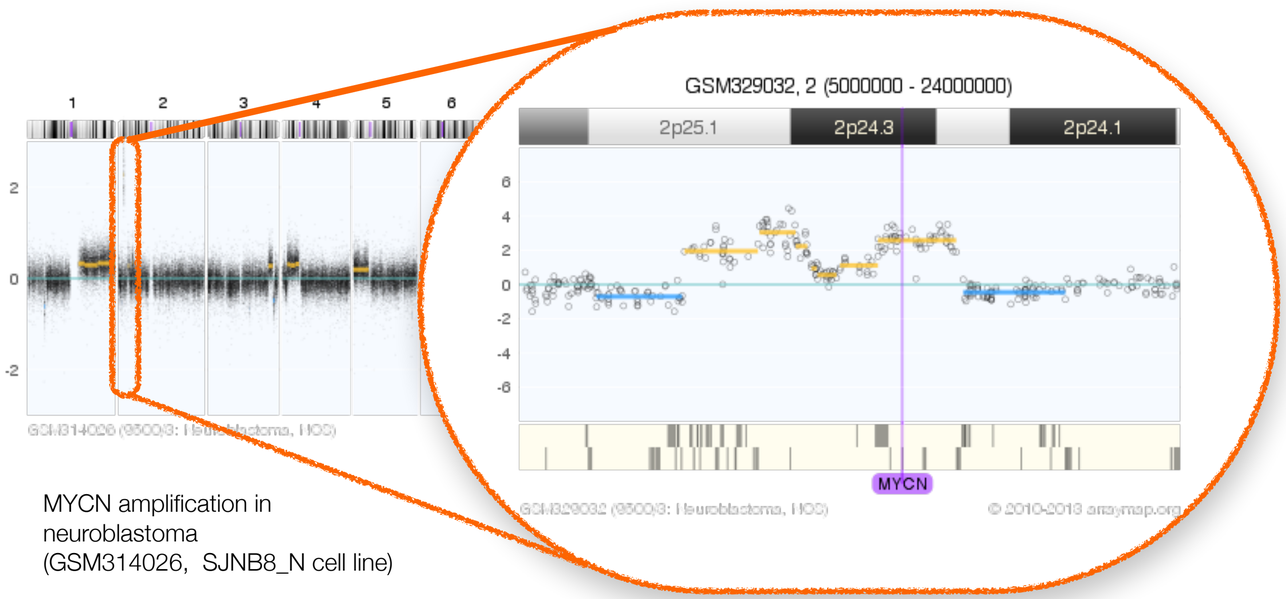
Bioconductor: https://bioconductor.org/packages/pgxRpi

CNV Categorization

different levels of CNV



Rameen et al 2010 Nature



MYCN amplification in neuroblastoma (GSM314026, SJNB8_N cell line)

GA4GH Variation Representation Specification

Global Alliance for Genomics & Health

Collaborate. Innovate. Accelerate.

CopyNumberChange

Copy Number Change captures a categorization of copies of a molecule within a system, relative to a baseline. These types of Variation are common outputs from CNV callers, particularly in the somatic domain where integral CopyNumberCount are difficult to estimate and less useful in practice than relative statements. Somatic CNV callers typically express changes as relative statements, and many HGVS expressions submitted to express copy number variation are interpreted to be relative copy changes.

Computational Definition

An assessment of the copy number of a Location or a Feature within a system (e.g. genome, cell, etc.) relative to a baseline ploidy.

Information Model

Some CopyNumberChange attributes are inherited from Variation.

Field	Type	Limits	Description
_id	CURIE	0..1	Variation Id. MUST be unique within document.
type	string	1..1	MUST be “CopyNumberChange”
subject	Location | CURIE | Feature	1..1	A location for which the number of systemic copies is described.
copy_change	string	1..1	MUST be one of “efo:0030069” (complete genomic loss), “efo:0020073” (high-level loss), “efo:0030068” (low-level loss), “efo:0030067” (loss), “efo:0030064” (regional base ploidy), “efo:0030070” (gain), “efo:0030071” (low-level gain), “efo:0030072” (high-level gain).

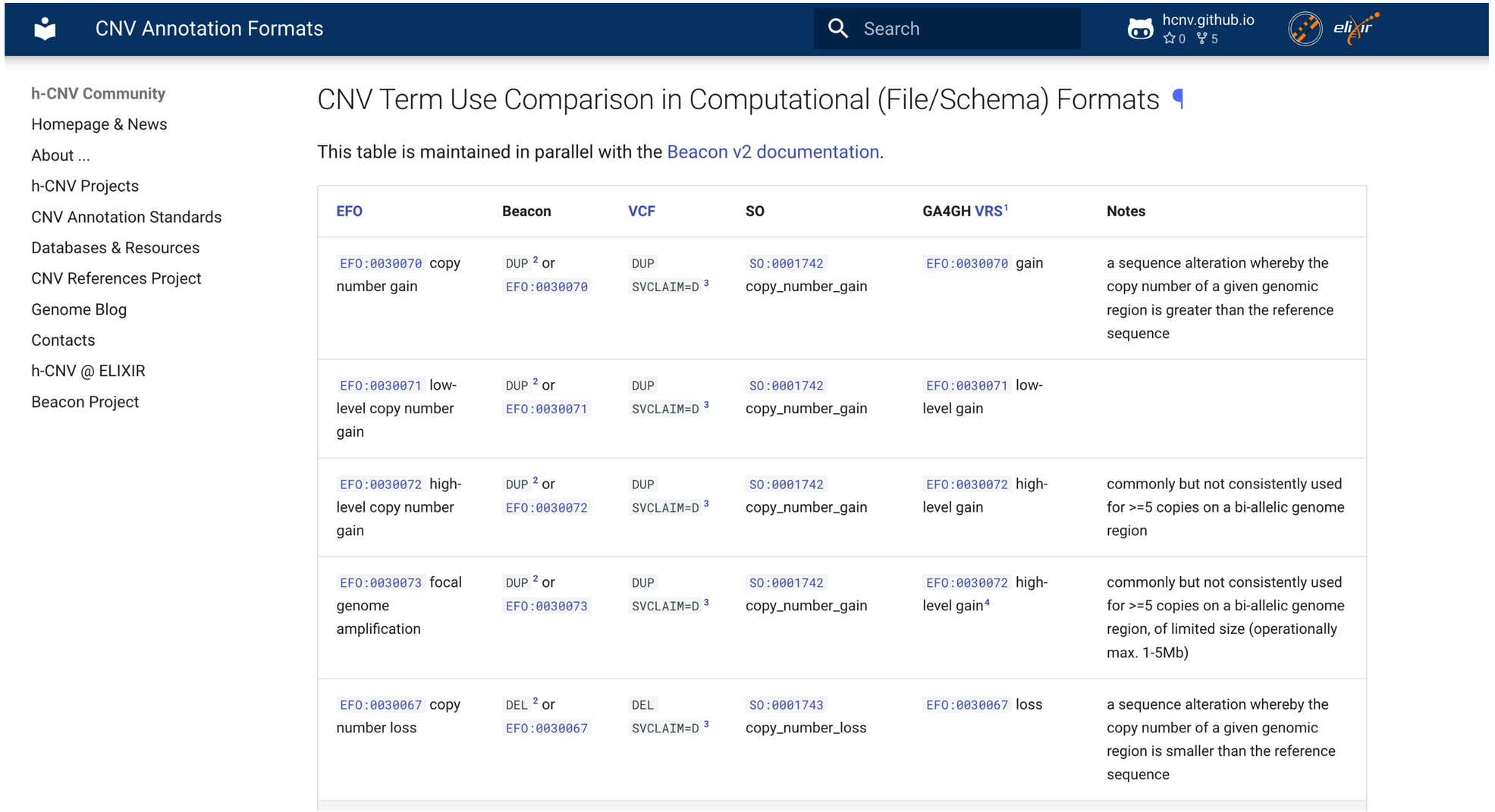
CNV Term Use Comparison

in computational (file/schema) formats

EFO	Beacon	VCF	SO	GA4GH VRS1.3
EFO:0030070 copy number gain	DUP or **EFO:0030070**	DUP SVCLAIM=D	SO:0001742 copy_number_gain	**EFO:0030070** gain
EFO:0030071 low-level copy number gain	DUP or **EFO:0030071**	DUP SVCLAIM=D	SO:0001742 copy_number_gain	**EFO:0030071** low-level gain
EFO:0030072 high-level copy number gain	DUP or **EFO:0030072**	DUP SVCLAIM=D	SO:0001742 copy_number_gain	**EFO:0030072** high-level gain
EFO:0030073 focal genome amplification	DUP or EFO:0030073	DUP SVCLAIM=D	SO:0001742 copy_number_gain	**EFO:0030072** high-level gain
EFO:0030067 copy number loss	DEL or **EFO:0030067**	DEL SVCLAIM=D	SO:0001743 copy_number_loss	**EFO:0030067** loss
EFO:0030068 low-level copy number loss	DEL or **EFO:0030068**	DEL SVCLAIM=D	SO:0001743 copy_number_loss	**EFO:0030068** low-level loss
EFO:0020073 high-level copy number loss	DEL or **EFO:0020073**	DEL SVCLAIM=D	SO:0001743 copy_number_loss	**EFO:0020073** high-level loss
EFO:0030069 complete genomic deletion	DEL or **EFO:0030069**	DEL SVCLAIM=D	SO:0001743 copy_number_loss	**EFO:0030069** complete genomic loss

CNV Term Use Comparison

in computational (file/schema) formats

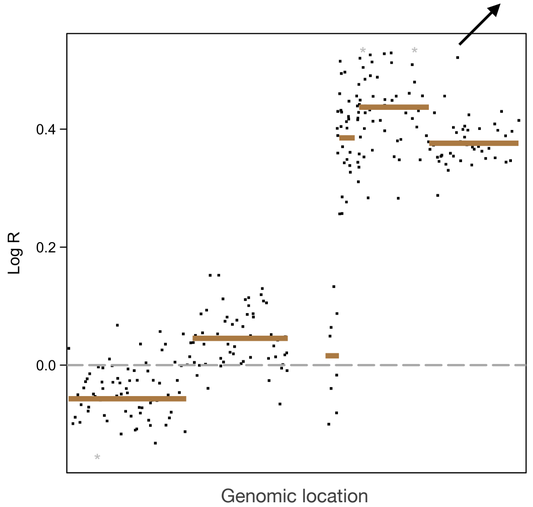


https://cnvar.org/resources/CNV-annotation-standards/#cnv-term-use-comparison-in-computational-fileschema-formats

labelSeg

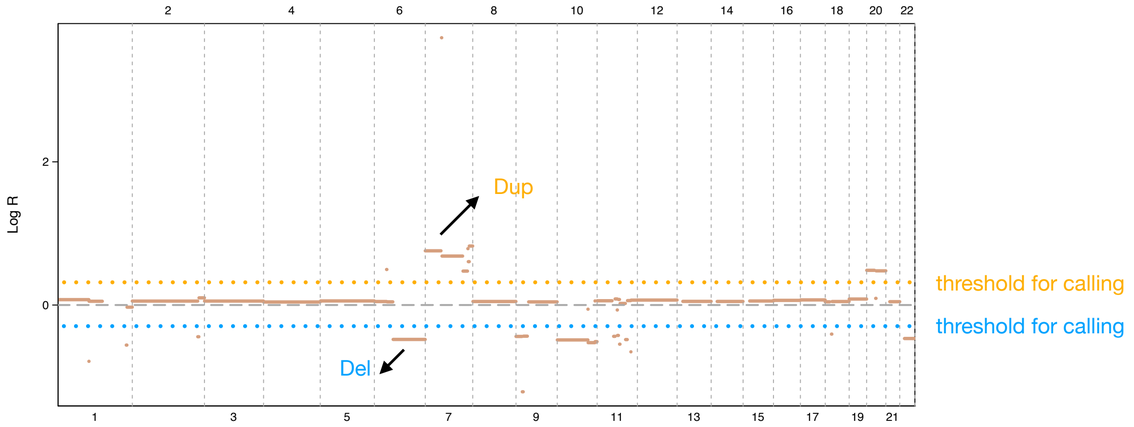
segment annotation for tumor copy number variation profiles

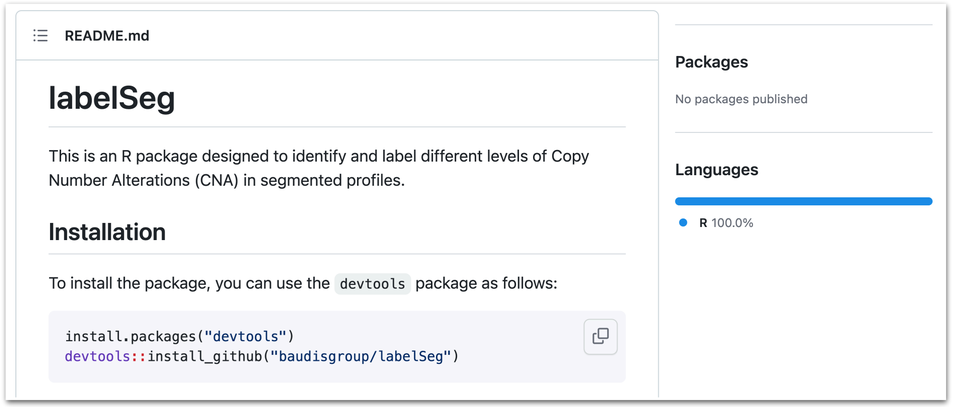
Signal from probes in microarray or from reads in NGS

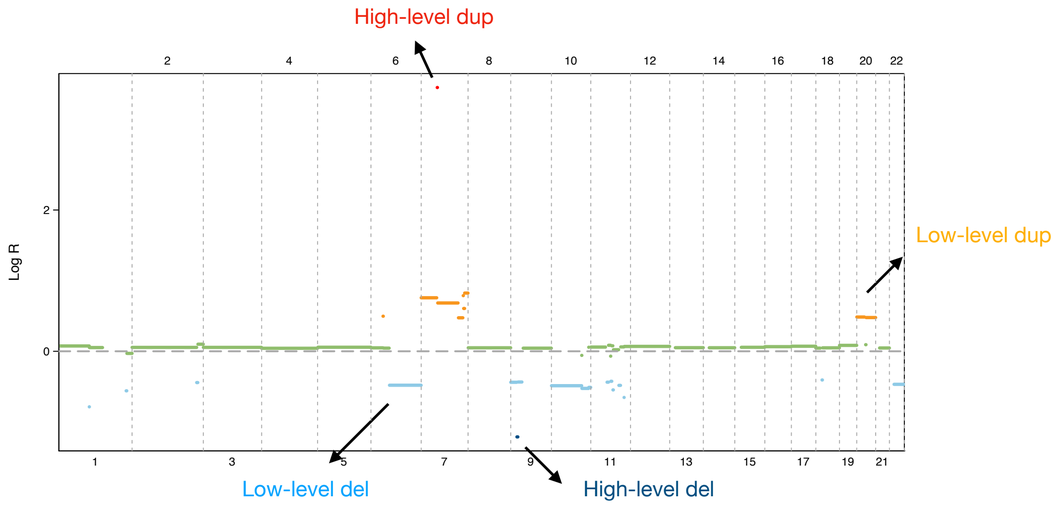


Segmentation

a step to split the chromosomes into regions of equal copy number that accounts for the noise in the data.







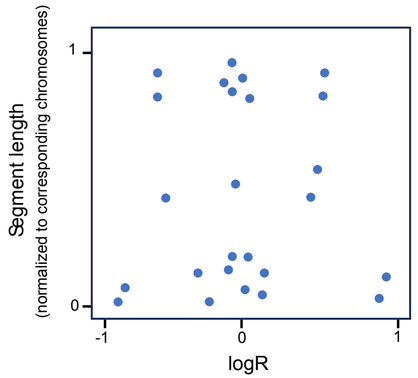
labelSeg

segment annotation for tumor copy number variation profiles

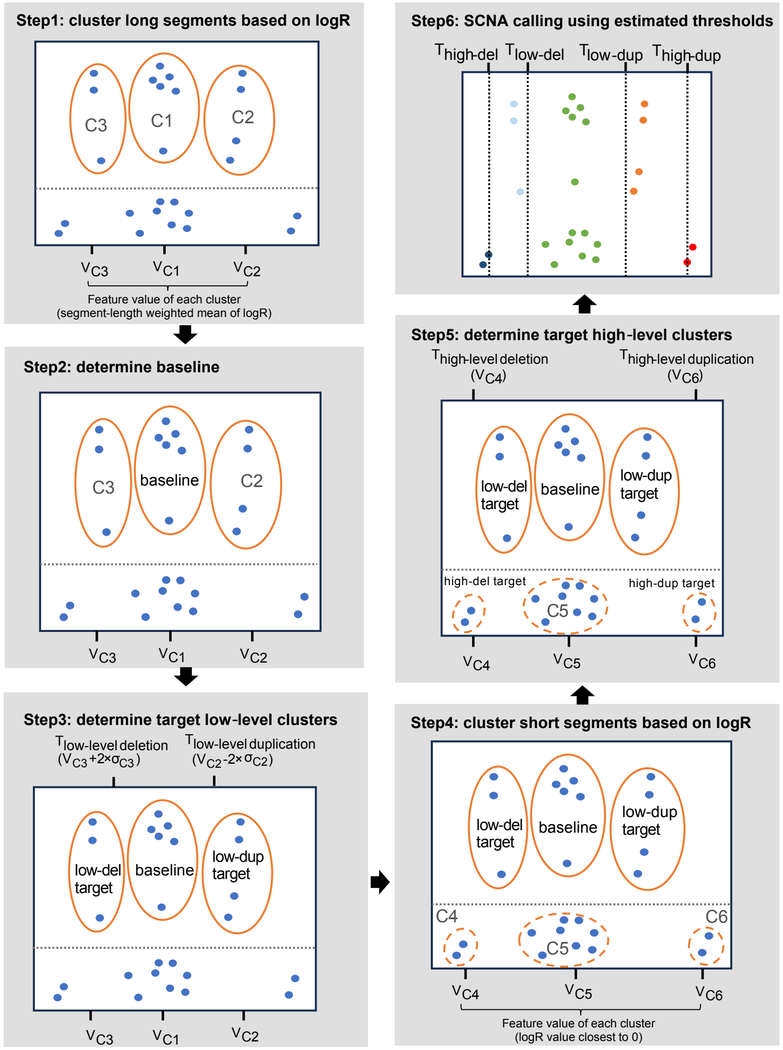
A

INPUT: individual .seg files

Smaple ID	Chr	Start	End	#Marker	logR
S1	1	9965785	10911995	450	0.306
…	…	…	…	…	…
S1	X	3236359	155677414	63741	0.005



B



C

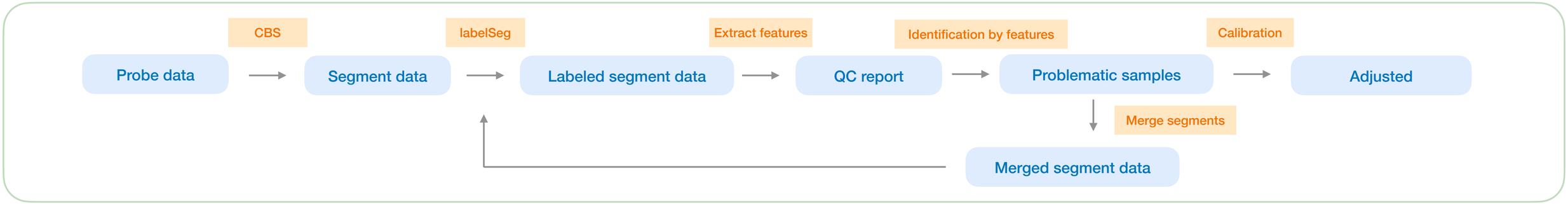
OUTPUT: individual .seg files with SCNA labels

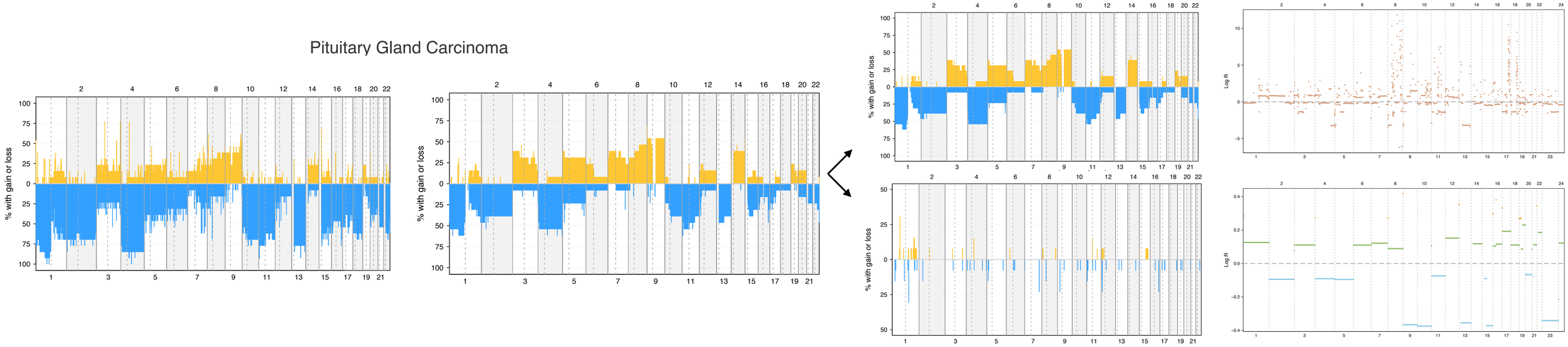
Smaple ID	Chr	Start	End	#Marker	logR	Label
S1	1	9965785	10911995	450	0.306	+1
…	…	…	…	…	…	
S1	X	3236359	155677414	63741	0.005	0

Pipeline Development

improve CNV calling in large numbers of heterogeneous cancer samples

nextflow





Pipeline Development

improve CNV calling in large numbers of heterogeneous cancer samples

Performance

- exclude false positive calls
- integrate/replace methods

Availability

- expansion
- workflow sharing