CNV detection from targeted next-generation sequencing data: whole exome and gene panels

THE HUMAN VARIOME PROJECT
an NGO official partner of UNESCO
sharing data · reducing disease

МГНЦ
медико-генетический научный центр

Anna Benet-Pagès

Focus

- the key features for CNV calling tools using NGS data
- the key factors to consider before and after pipeline design
- examples combined-tool approach for accurate CNV calling in a routine diagnostics set up

Detection of structural variants and human disease

B/B$^+$

Calvin Bridges, 1936

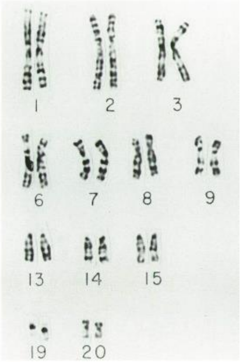
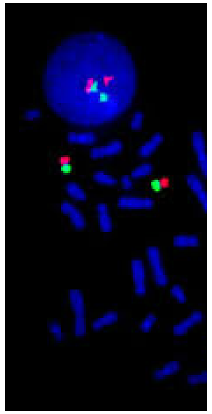
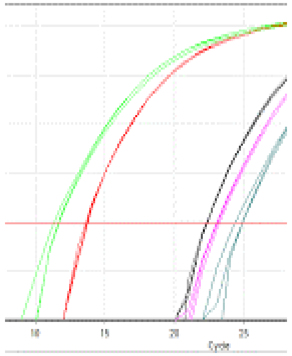


FIGURE 1B: G-banding karyotype of 46, XY individual.

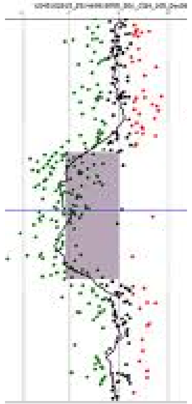
G-banding Rowley et al., 1973



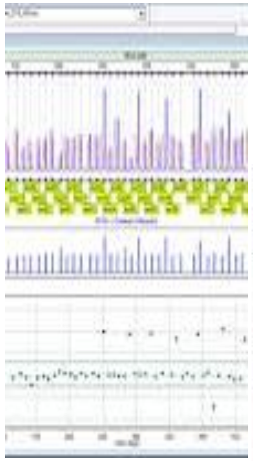
FISH Langer et al., 1982



Quantitative PCR Porcher et al., 1992

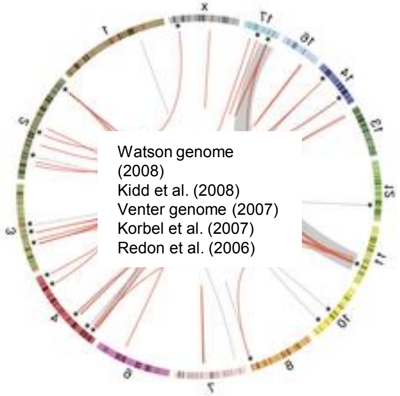


Array CGH Pinkel et al., 1998c



MLPA Schouten et al., 2002

SNP array Komura et al., 2006

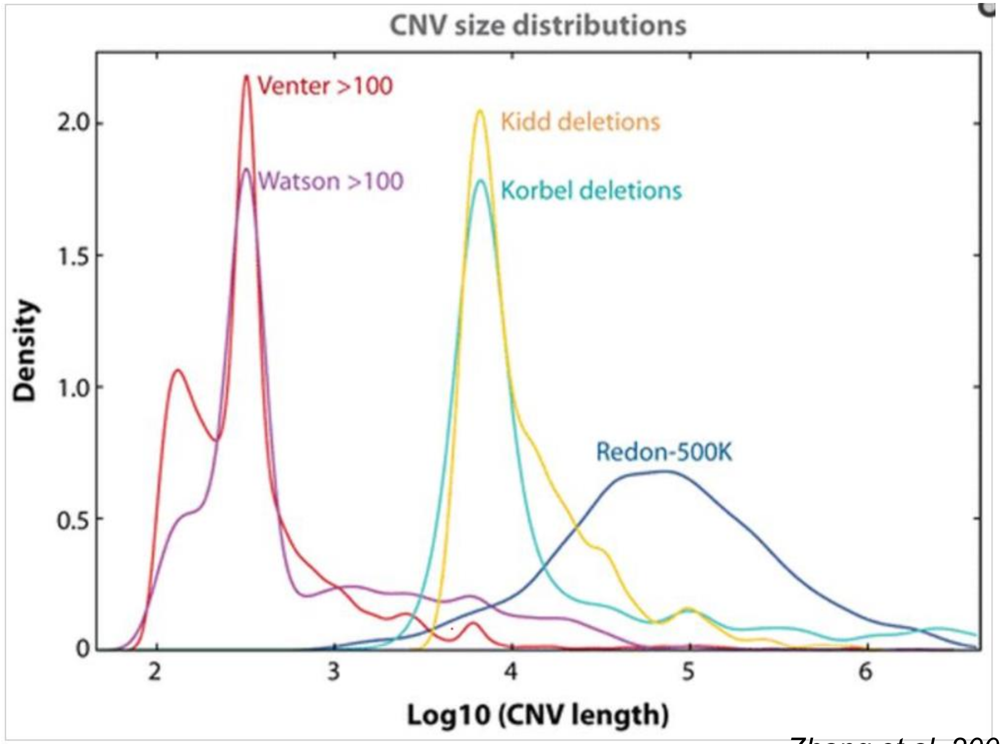


Watson genome (2008)
Kidd et al. (2008)
Venter genome (2007)
Korbel et al. (2007)
Redon et al. (2006)

NGS

NGS technologies reveal smaller-size CNVs

- ***Smaller structural variants are the most frequent***



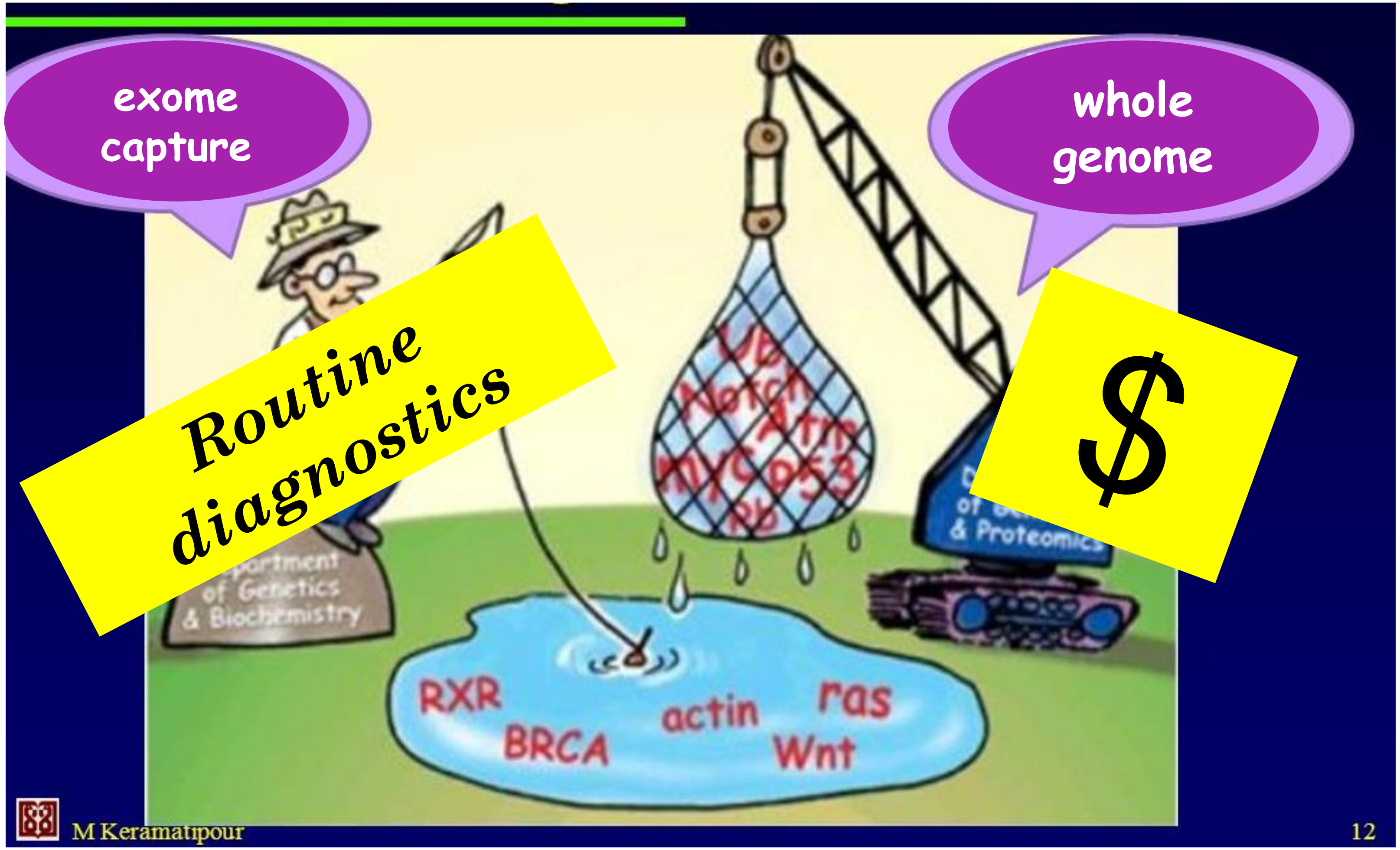
Zhang et al. 2009

Size distribution of copy number variations (CNVs) larger than 100 bp

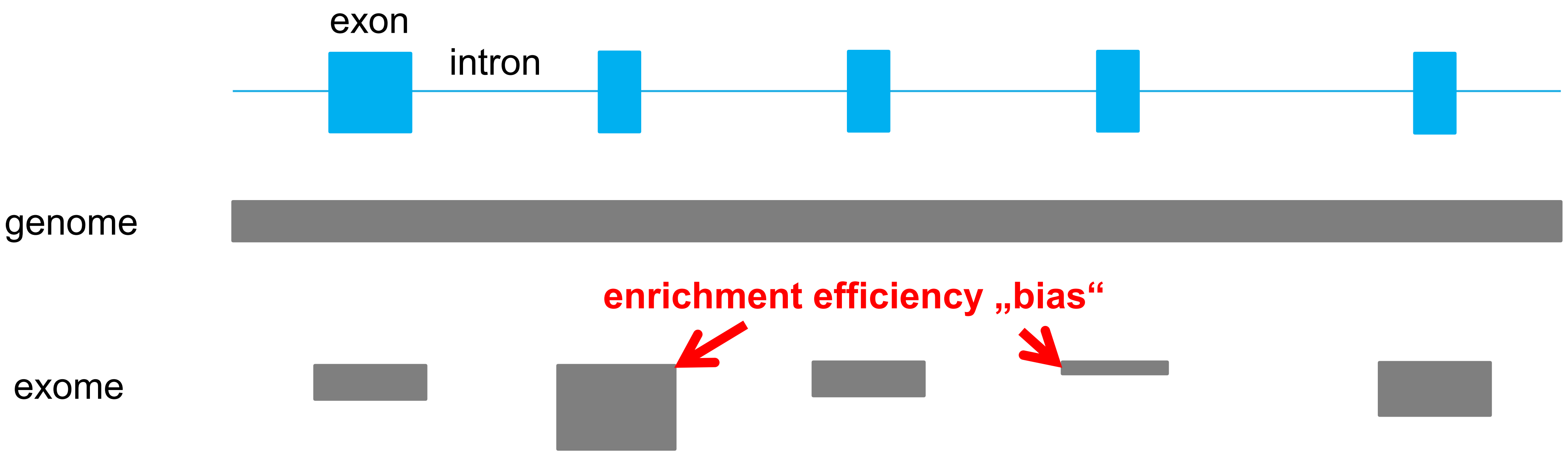
Advantages of the NGS approach:

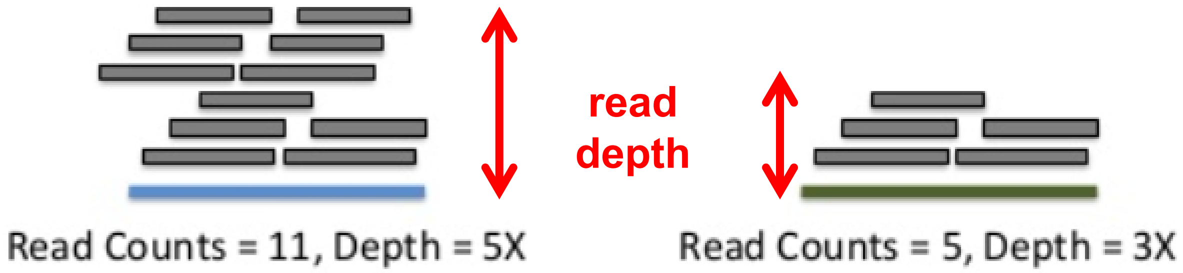
- *higher coverage and resolution*
- *more accurate estimation of copy numbers*
- *more precise detection of breakpoints*
- *higher capability to identify novel CNVs*

Genome vs. Exome



Approaches to detect CNVs from targeted-capture data





Read depth method:

“counts the number of reads mapped to each genomic region”

CNV detection from targeted-capture data

Challenges:

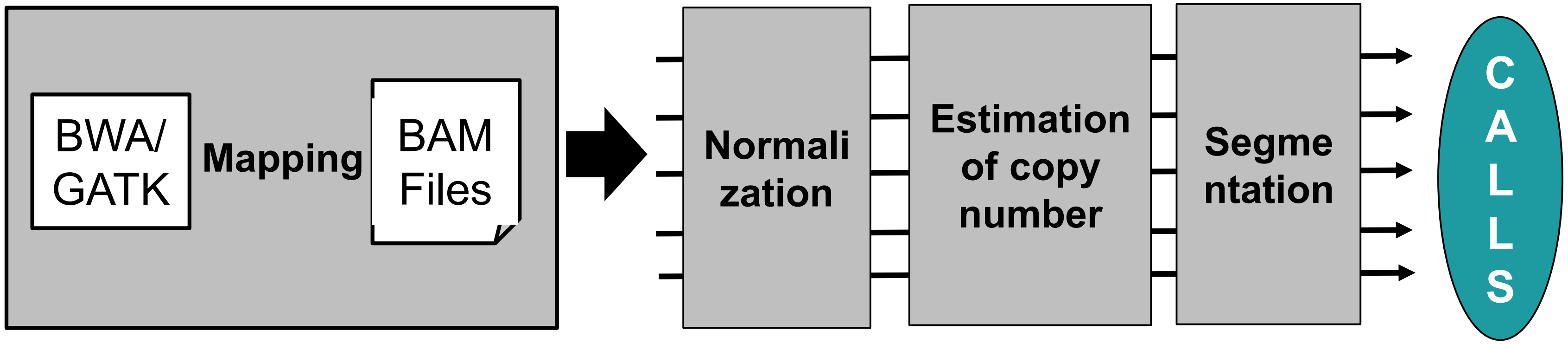
- Inconsistent capture efficiency, the depth from different genomic regions may vary substantially
- Coverage bias inter- and intra-sequencing runs
- Assumption of normal distribution of data may no longer be valid
- Control individuals are difficult to obtain (reference set/ validation)

CNV detection from targeted-capture data

Limitations:

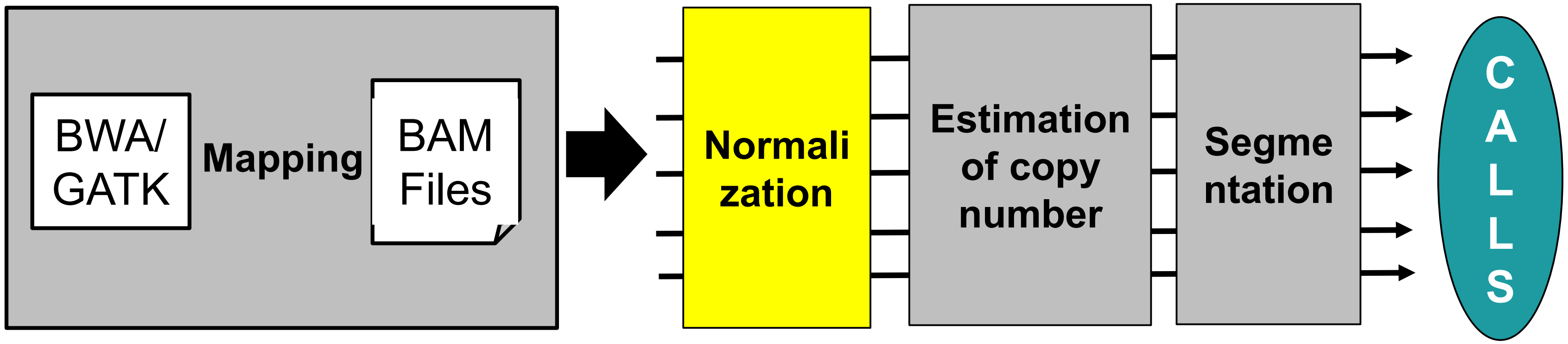
- The full spectrum of CNVs and breakpoints may not be completely characterized
- Large CNVs and cross-chromosome events may not be detected
- Single exon events are difficult to detect (false negatives)
- Duplications/Gains call ratio much higher than deletions (false positives)
- Validation is expensive (need several samples for a comprehensive CNV dataset)
- Longer analysis time (compared to SNV) - more IT infrastructure

CNV Pipeline Structure



- ➢ mapping of short reads to the reference genome

CNV Pipeline Structure



BWA/ GATK

Mapping

BAM Files

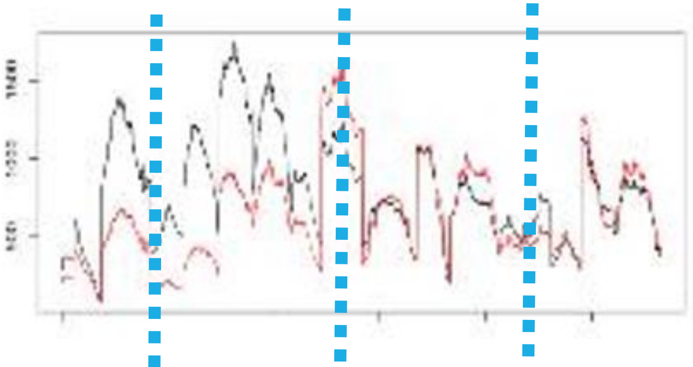
Normalization

Estimation of copy number

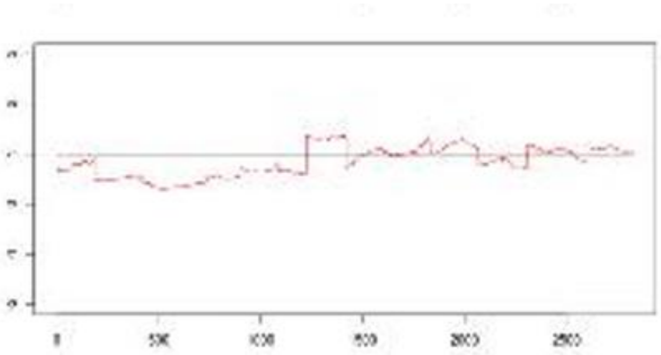
Segmentation

CALLS

- ➢ breakdown of the target region exons/ windows and read depth is calculated according to the number of mapped reads
- ➢ correction of potential biases in read depths mainly caused by GC contents, repeat genomic regions, and homologous regions



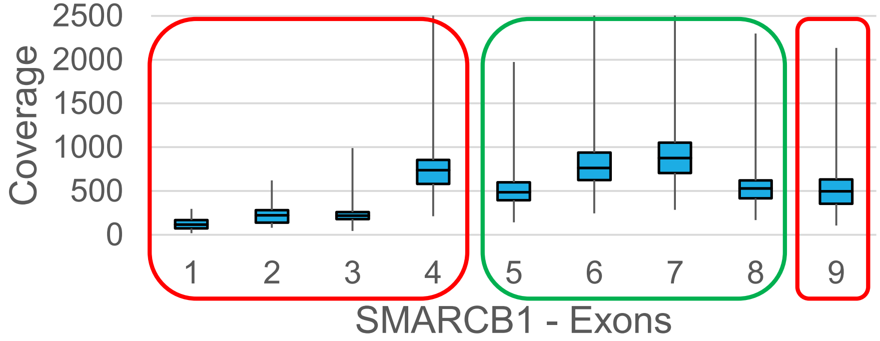
GC
Mapping qual
# reads
Ref. Set



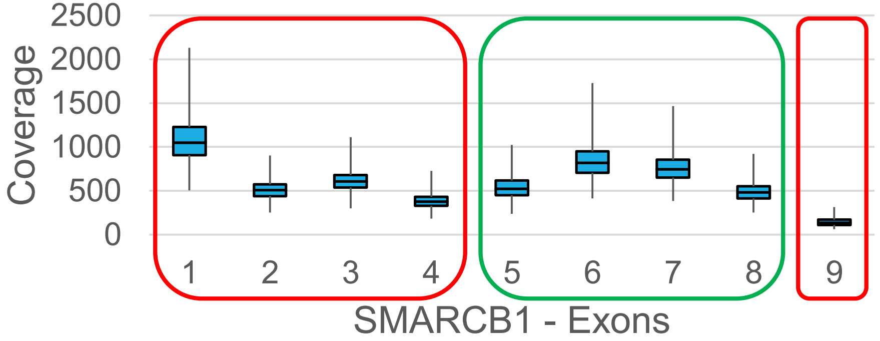
Reference Sets and data normalization

- different reference sets for different kits / enrichment methods
- normalization against samples from the same sequencing run to improve robustness against workflow bias

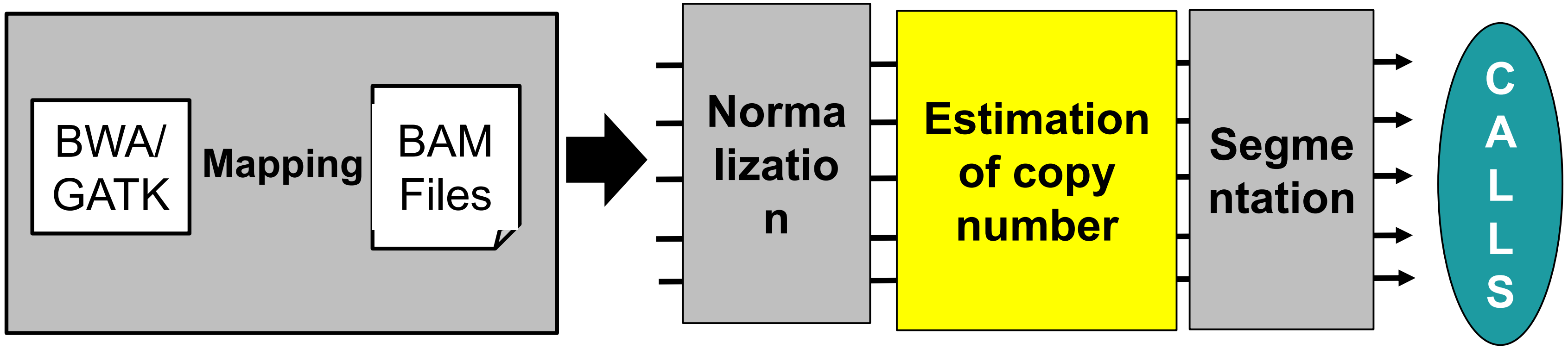
Illumina



Agilent

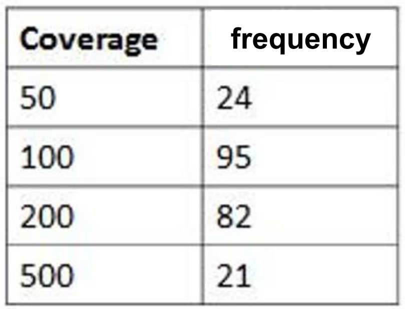


CNV Pipeline Structure

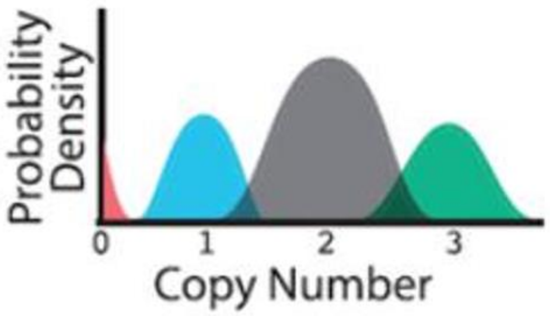


- ➢ estimate the accurate copy number along the chromosome to determine the gain or loss

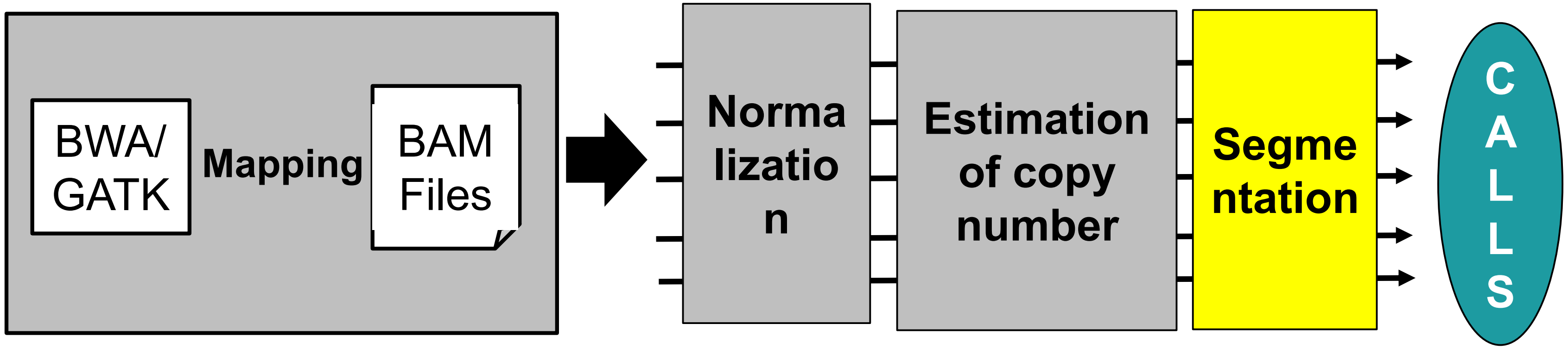
Poison distribution
Beta binomial
Negative binomial
Normal distribution



Coverage	frequency
50	24
100	95
200	82
500	21



CNV Pipeline Structure



BWA/ GATK

Mapping

BAM Files

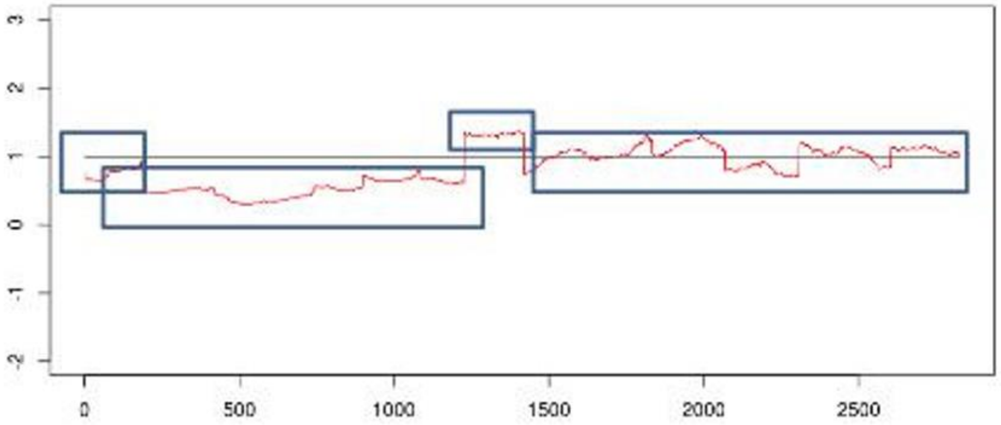
Normalization

Estimation of copy number

Segmentation

CALLS

- grouping areas (exon/window) with the same prediction (gain / loss / normal)



CNV detection methods general considerations

- **Which tool should I choose?**

AGE, BicSeq, BreakDancer, Breakpointer, Breakseq, Canoes, Clamms, Clever, ClipCrop,

Cn.MOPS, CNAnorm, CNAseg, CND, CNV_TV, Cnvator, CNVer, CNVer, HugeSEQ,

hydra, inGAP_sv, JointSLM, Matchclip, modil, mogul, mrcanavar, Patchwork, pemer

, ReadDepth, rSW_seq, segseq, seqcbs, CNVer, cnvHiTSeq, cnvrd, CNV-seq,

conserting, CONTROL_FREEC, cops, copySeq, crest, ERDS, codex

EWT_RDXplorer, GasvPRO, GENSENG, XHMM

CNV detection methods general considerations

- **Which tool should I choose?**

- applicable to capture data
- easy to integrate (take bam files as input)
- easy handling (installation / running time)
- multi-sample usage (possibility to normalize against reference set)
- Tools should use different statistic models

CNV detection methods

- ➢ Use a combination of several detection tools

AGE, BicSeq, BreakDancer, Breakpointer, Breakseq, Canoes, Clamms, Clever, ClipCrop,
Cn.MOPS, CNAnorm, CNAseg, CND, CNV_TV, Cnvator, CNVer, CNVer, HugeSEQ,
hydra, inGAP_sv, JointS Patchwork, pemer
, ReadDepth, rSW_seq, segseq, seqcbs, CNVer, cnvHiTSeq, cnvrd, CNV-seq,
conserting, CONTROL_FREEC, cops, copySeq, crest, ERDS, codex
EWT_RDXplorer, GasvPRO, GENSENG, XHMM

„combined-CNV-caller“

Noll et al., Npjgenmed 2016

Pipeline: Combined-CNV tools

- **ExomeDepth**

  extremely sensitive and robust against samples that do not correlate with the reference

- **Canoes**

  has a high sensitivity for small deletions, high performance in low coverage regions and with few reference samples

- **Clamms**

  corrects for GC content and mappability, divides large exons into smaller regions and calls also common CNVs

- **Codex**

  corrects for GC content and mappability, calls also common CNVs, uses no HMM for segmentation (all other tools use HMMs)

- **In-house method**

  is well adapted on in-house data, screens for heterozygosity, corrects for GC content, exon score depends on previous analyses

Performance of single tools

- Training set: true set of 146 CNV calls detected via MLPA
- Sensitivity and precision not sufficient for routine diagnostics

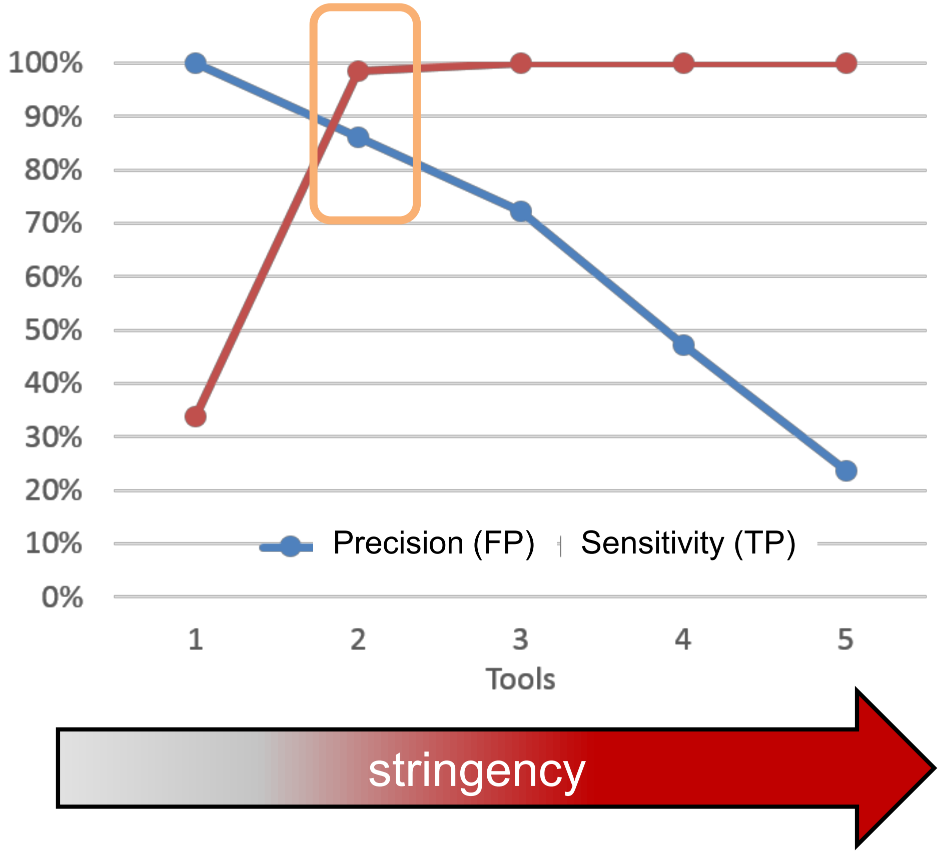
	Exome Depth	Clamms	Canoes	Codex	In-house
Precision	45.63%	68.57%	96.77%	64.75%	40.82%
Sensitivity	90.38%	46.15%	57.69%	63.46%	76.92%

How could accuracy be improved?

create a combined pipeline using all 5 tools

Performance of tool combinations

- What is the minimum number of concordance predictions required to consider a CNV a reliable call? (minimum number of tools that call the same variant)

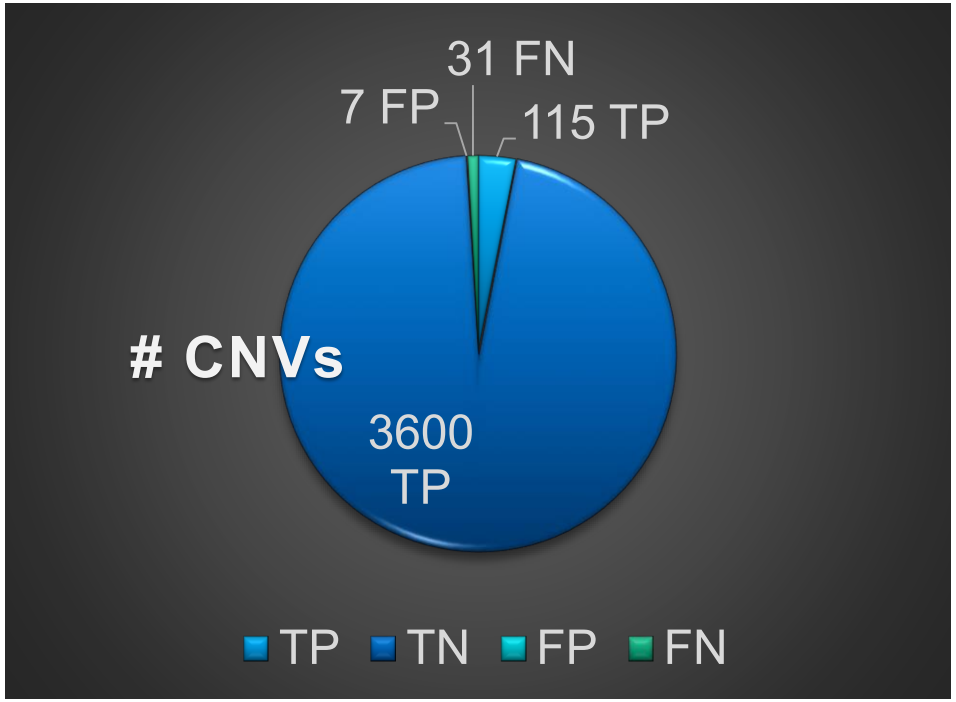


	Sensitivity (TPR)	Specificity (TNR)	Precision (PPV)	NPV
2 out of 5	94.26%	99.78%	93.50%	99.81%

➢ ***Use of two of five matching tools shows the best trade-off for sensitivity and specificity.***

CNV Pipeline Evaluation

- >3700 MLPAs performed in ~90 genes
- 146 CNVs (85 deletions / 61 gains)
- Minimal coverage per sample: 30X in >98% of the coding regions



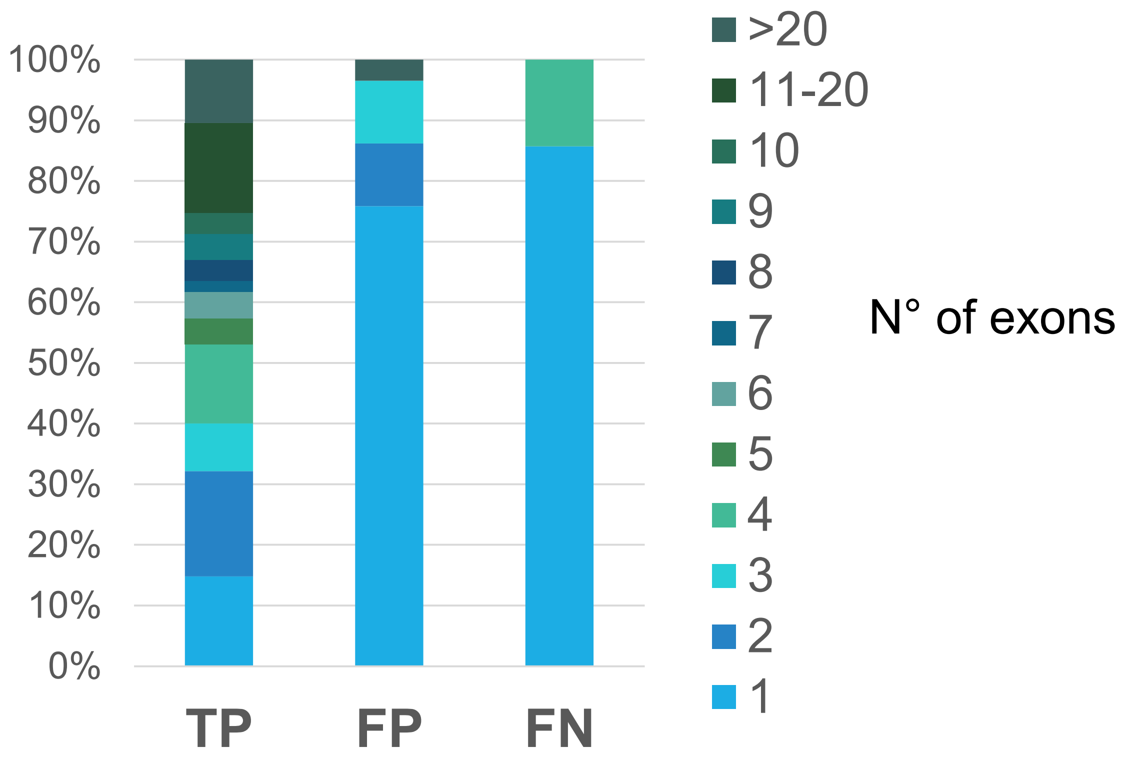
Sensitivity:	88.60%
Specificity:	98.88%
Precision:	71.40%

Performance increases considerably if homologous regions are excluded from the analysis (pseudogenes):

Sensitivity:	**94.26%**
Specificity:	**99.78%**
Precision:	**93.50%**

TP, FP , FN versus CNV size

- Comparison of CNV sizes of TP, FP and FN calls detected by the combined CNV pipeline on the validation set
- FP and FN calls consist mainly of single exon events

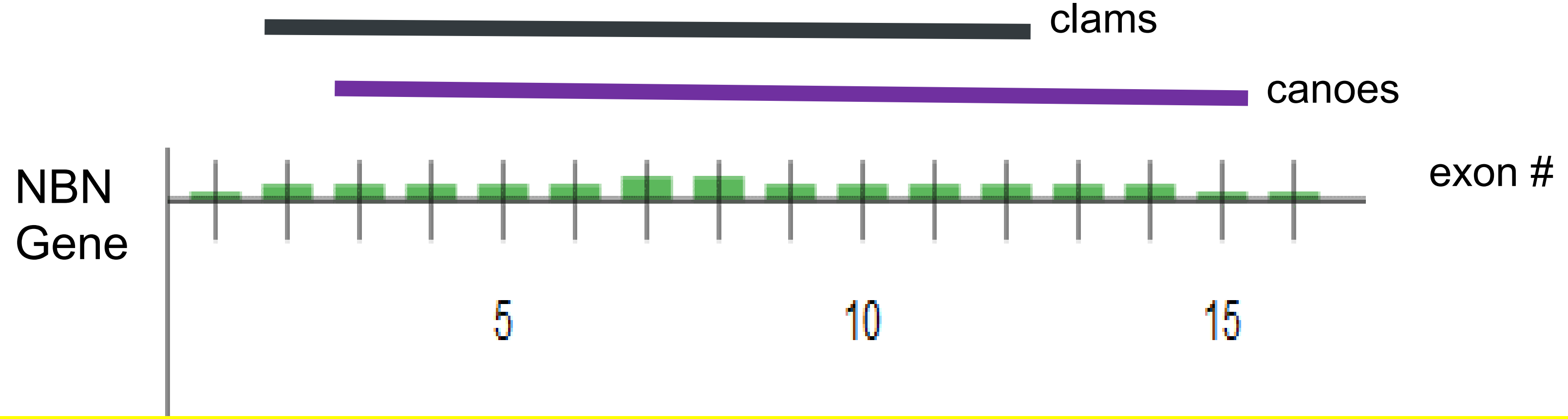


Challenges

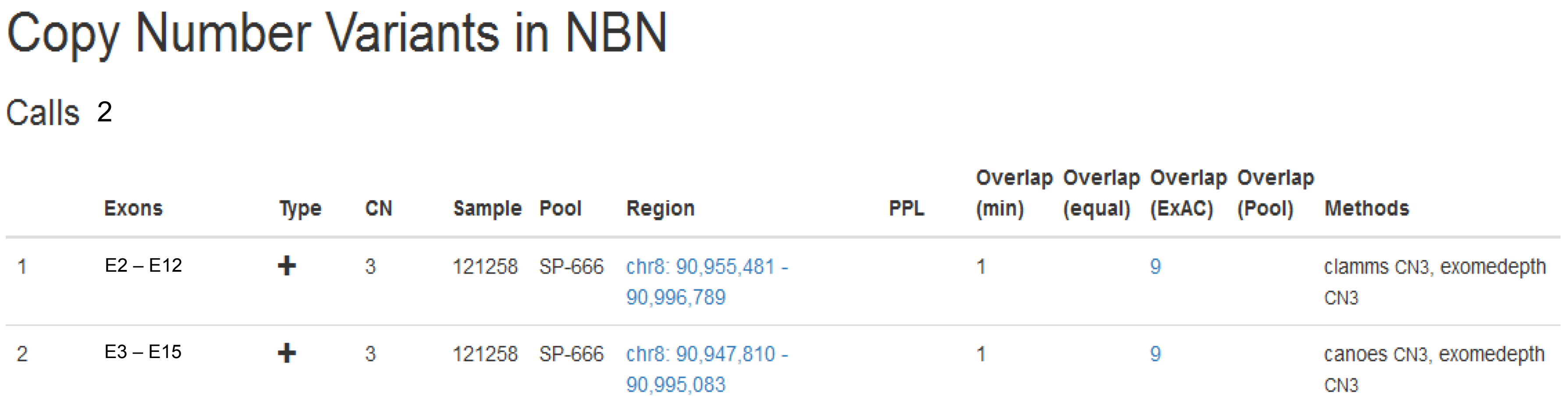
- Non-uniform of coverage
- CNV calling in homologous genomic regions (pseudogenes...)
- Clinical interpretation

Discrepancies in CNV size detection between tools

different tools = different calls for one event



Consider events separated, could be two copy numbers in two different alleles



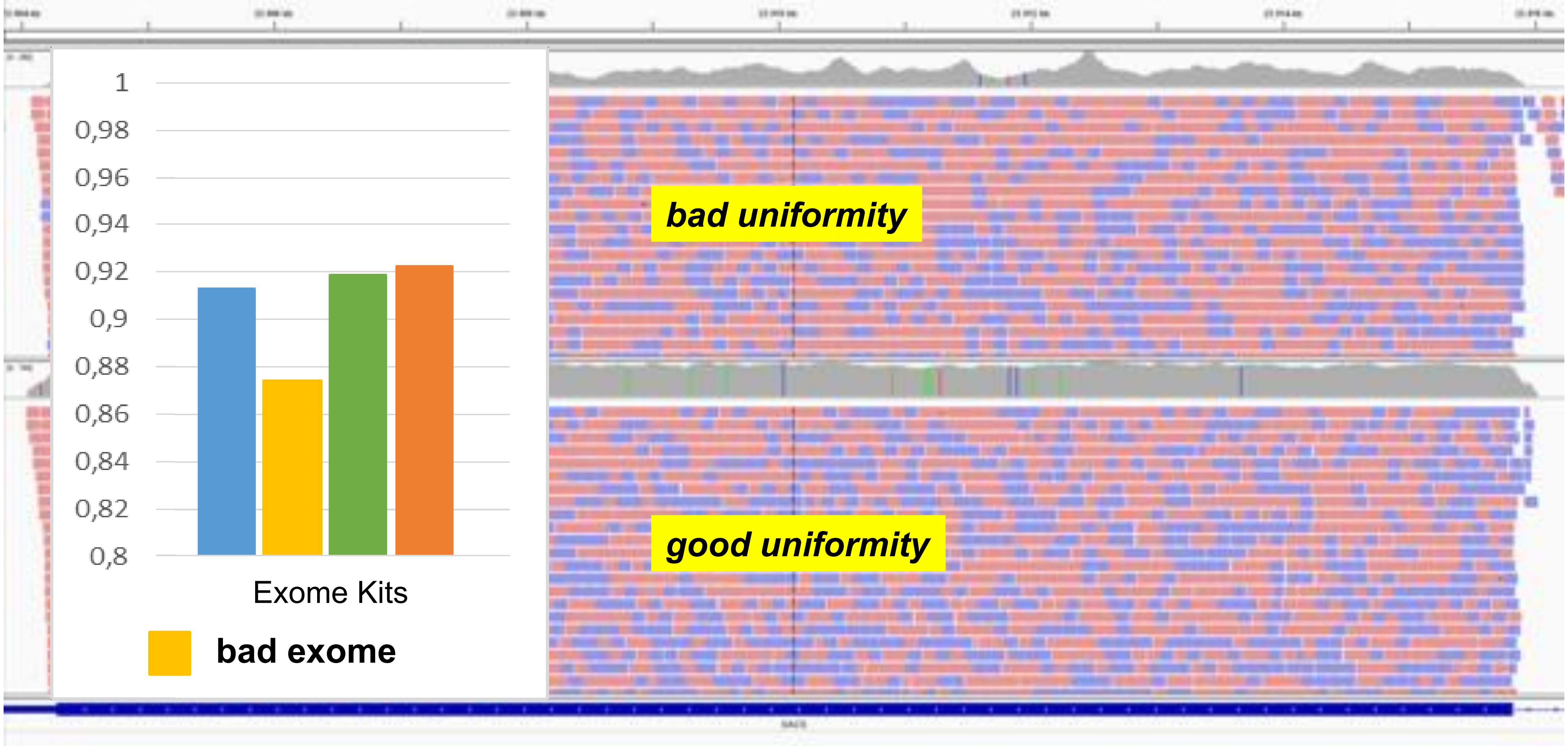
Copy Number Variants in NBN

Calls 2

								Overlap	Overlap	Overlap	Overlap	
	Exons	Type	CN	Sample	Pool	Region	PPL	(min)	(equal)	(ExAC)	(Pool)	Methods
1	E2 – E12	+	3	121258	SP-666	chr8: 90,955,481 - 90,996,789		1		9		clamms CN3, exomedepth CN3
2	E3 – E15	+	3	121258	SP-666	chr8: 90,947,810 - 90,995,083		1		9		canoes CN3, exomedepth CN3

non-uniform coverage = capture bias

- identification of reliable regions by assessment of capture efficiency to minimize false positives



bad uniformity

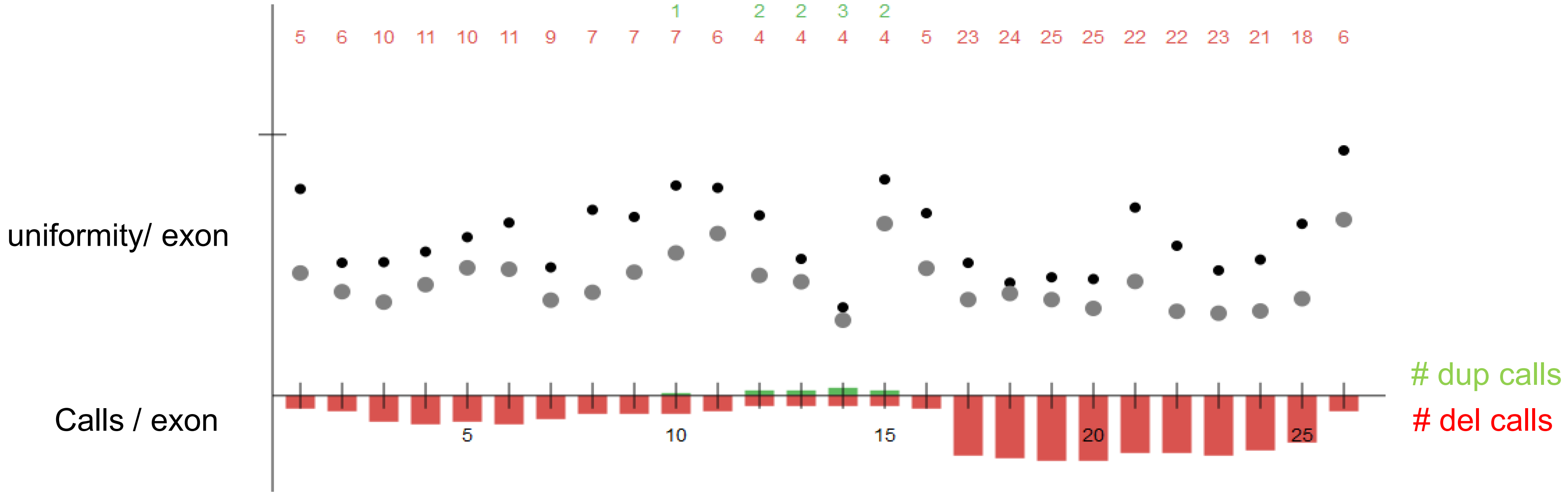
good uniformity

CNV calling artifacts

● Depth of coverage uniformity reference set
● Depth of coverage uniformity sample

2. SCN1A

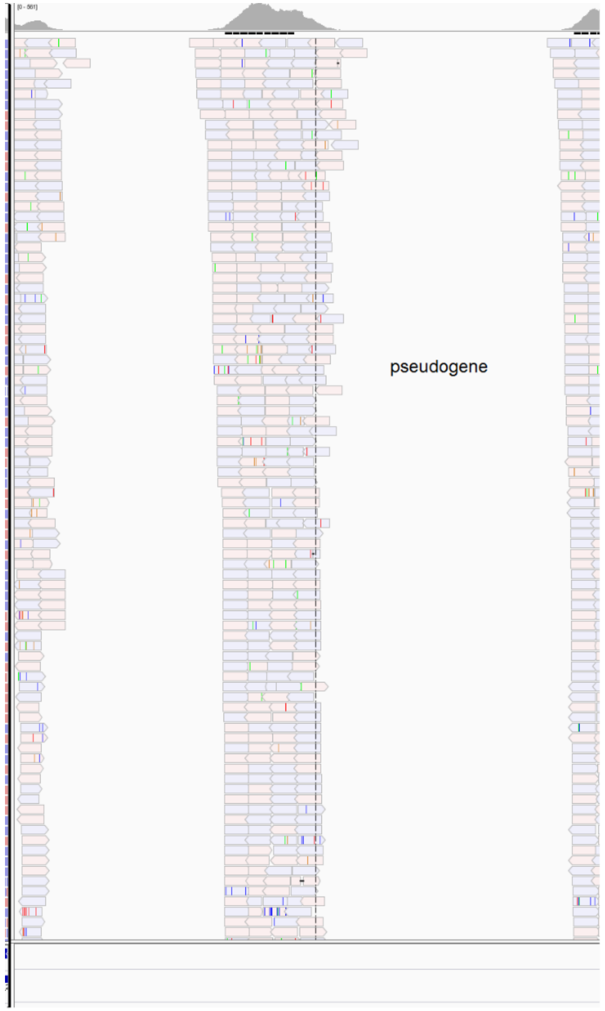
Number of exons	26
Duplications in	5 exons and 4 calls (frequency = 0.00)
Deletions in	26 exons and 39 calls (frequency = 0.01)



Non-unique mapping in homologous regions

- Non-unique mapping in pseudogenes increases the number of FP calls

PMS2



Blank reads represent reads with mapping quality equal to 0, reads can map to other regions

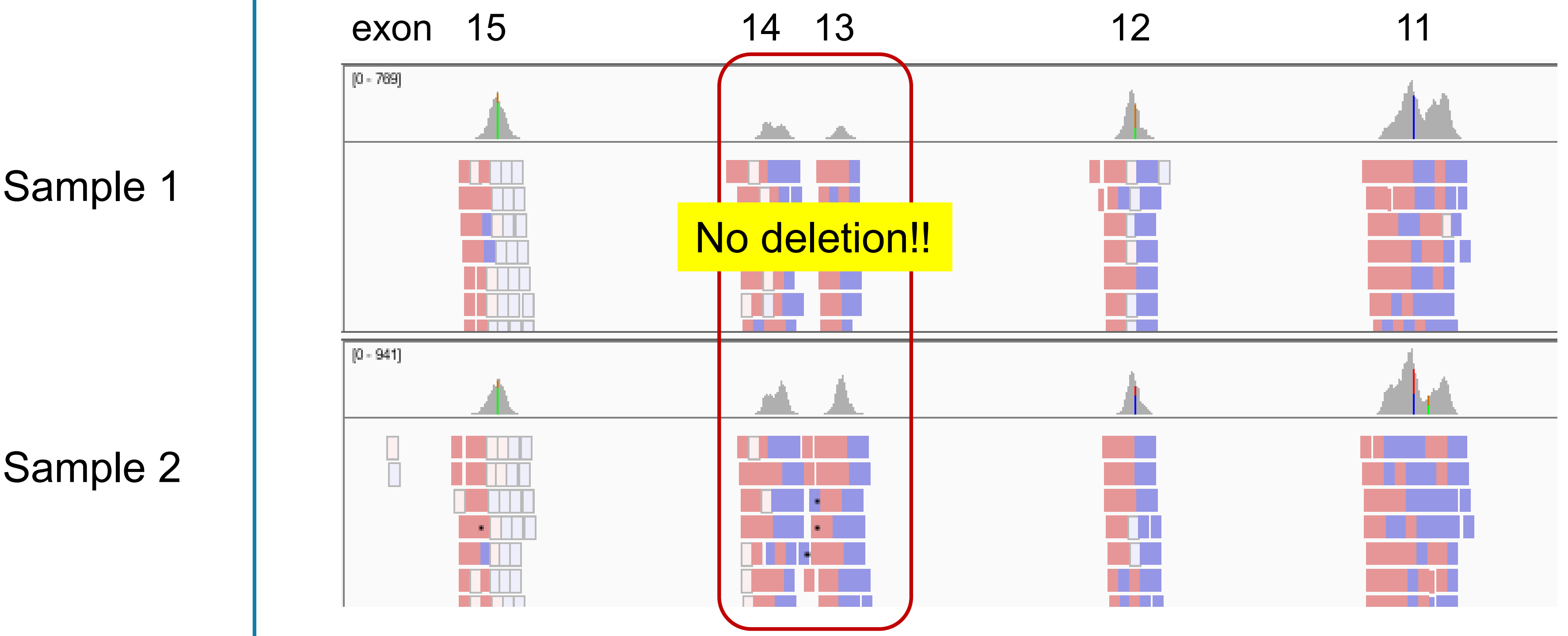
CNV calling in Pseudogenes

🟥 Forward read / unique mapping

🟦 Reverse read / unique mapping

□ Non-unique mapping

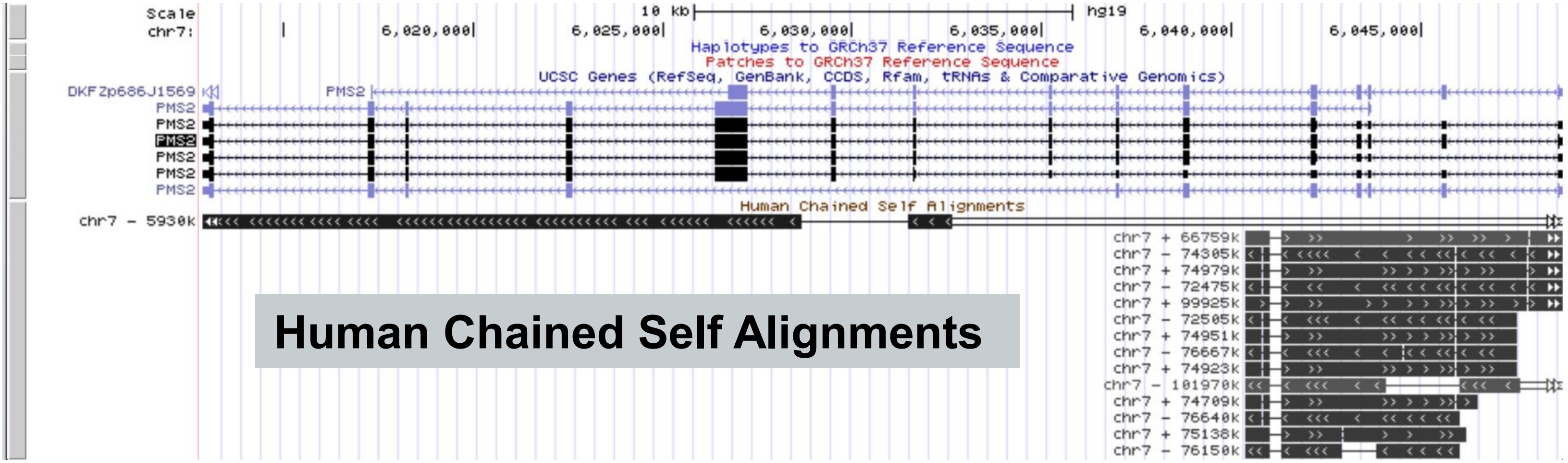
1. PMS2



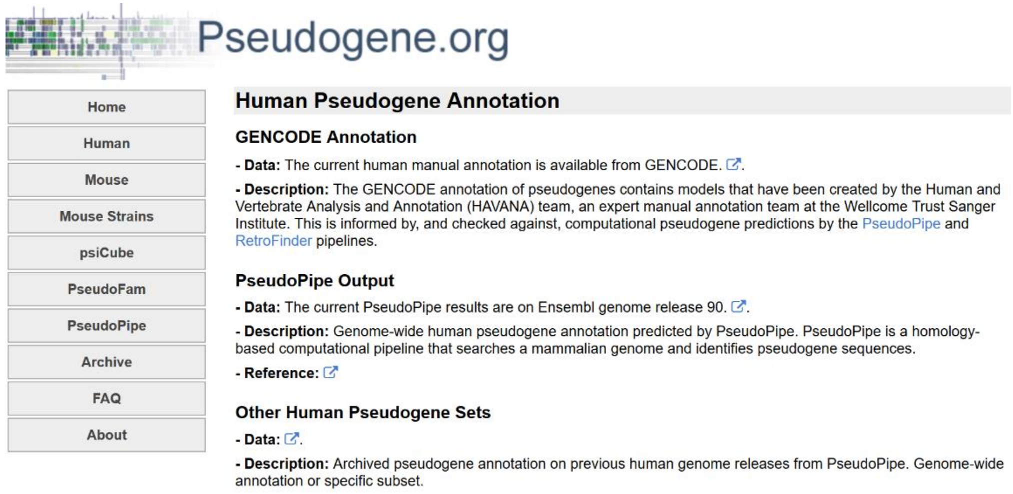
PMS2 exons 11 – 15 can not be analyzed

How to identify regions affected by pseudogenes

- alignments of the human genome with itself using blastz



Human Chained Self Alignments



Human Pseudogene Annotation

GENCODE Annotation

- **Data:** The current human manual annotation is available from GENCODE.

- **Description:** The GENCODE annotation of pseudogenes contains models that have been created by the Human and Vertebrate Analysis and Annotation (HAVANA) team, an expert manual annotation team at the Wellcome Trust Sanger Institute. This is informed by, and checked against, computational pseudogene predictions by the PseudoPipe and RetroFinder pipelines.

PseudoPipe Output

- **Data:** The current PseudoPipe results are on Ensembl genome release 90.

- **Description:** Genome-wide human pseudogene annotation predicted by PseudoPipe. PseudoPipe is a homology-based computational pipeline that searches a mammalian genome and identifies pseudogene sequences.

- **Reference:**

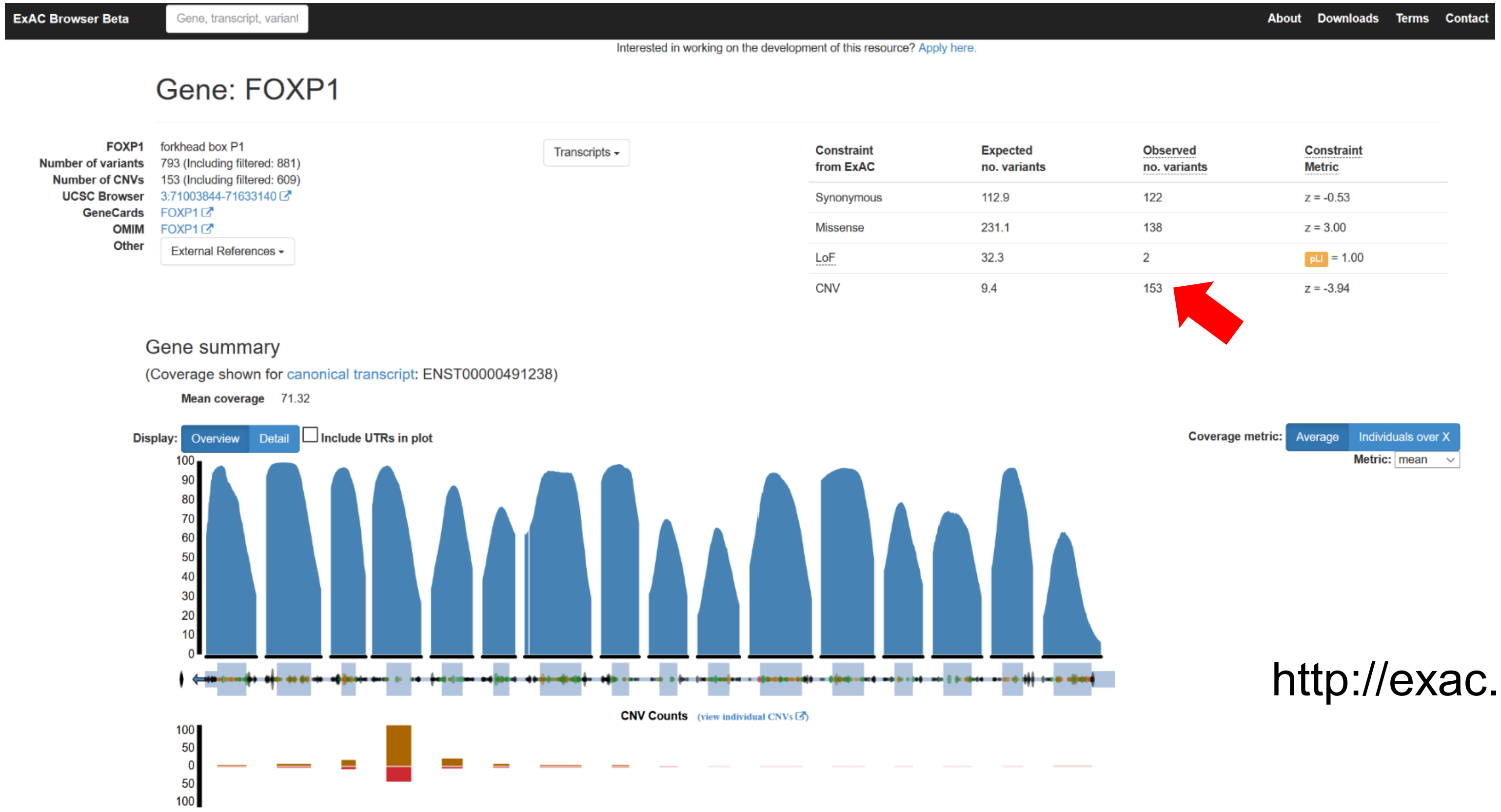
Other Human Pseudogene Sets

- **Data:**

- **Description:** Archived pseudogene annotation on previous human genome releases from PseudoPipe. Genome-wide annotation or specific subset.

Interpretation of CNV calls – population DB

- Deletions and duplications called based on read depth using XHMM; *Fromer et al.*
- Z score for the deviation of observed counts from the expected number



ExAC Browser Beta

Interested in working on the development of this resource? Apply here.

Gene: FOXP1

FOXP1	forkhead box P1
Number of variants	793 (Including filtered: 881)
Number of CNVs	153 (Including filtered: 609)
UCSC Browser	3:71003844-71633140
GeneCards	FOXP1
OMIM	FOXP1
Other	External References

Constraint from ExAC	**Expected no. variants**	**Observed no. variants**	**Constraint Metric**
Synonymous	112.9	122	z = -0.53
Missense	231.1	138	z = 3.00
LoF	32.3	2	pLI = 1.00
CNV	9.4	153	z = -3.94

Gene summary

(Coverage shown for canonical transcript: ENST00000491238)

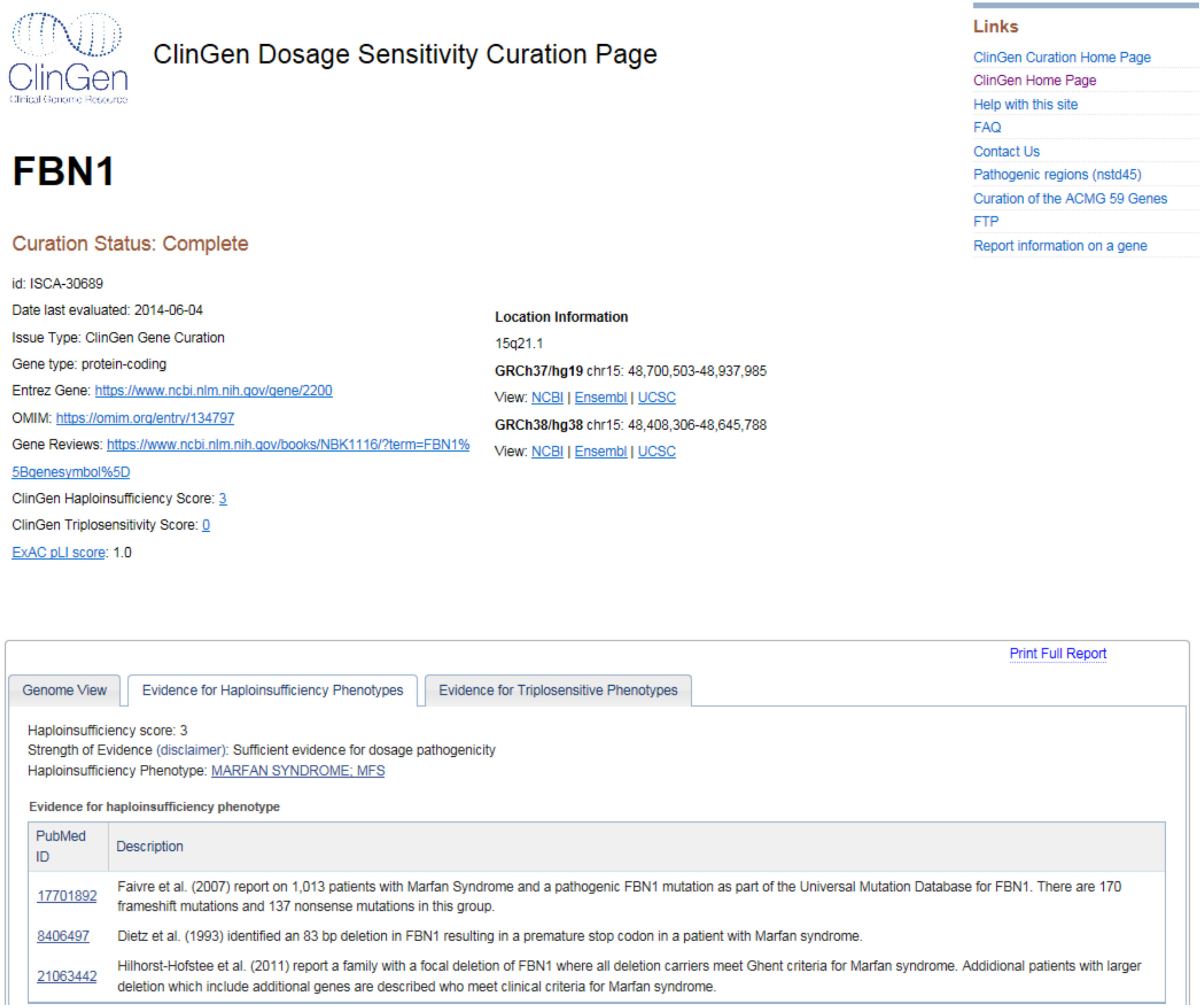
Mean coverage 71.32

http://exac.broadinstitute.org/

Positive Z scores indicate that the gene had fewer variants than expected.
Negative Z scores are given to genes that had a more variants than expected.

Interpretation of CNV calls – clinical DB

- DGV
- DECIPHER
- ClinVar
- ClinGen



ClinGen Dosage Sensitivity Curation Page

Links

- ClinGen Curation Home Page
- ClinGen Home Page
- Help with this site
- FAQ
- Contact Us
- Pathogenic regions (nstd45)
- Curation of the ACMG 59 Genes
- FTP
- Report information on a gene

FBN1

Curation Status: Complete

id: ISCA-30689
Date last evaluated: 2014-06-04
Issue Type: ClinGen Gene Curation
Gene type: protein-coding
Entrez Gene: https://www.ncbi.nlm.nih.gov/gene/2200
OMIM: https://omim.org/entry/134797
Gene Reviews: https://www.ncbi.nlm.nih.gov/books/NBK1116/?term=FBN1%5Bgenesymbol%5D
ClinGen Haploinsufficiency Score: 3
ClinGen Triplosensitivity Score: 0
ExAC pLI score: 1.0

Location Information
15q21.1
GRCh37/hg19 chr15: 48,700,503-48,937,985
View: NCBI | Ensembl | UCSC
GRCh38/hg38 chr15: 48,408,306-48,645,788
View: NCBI | Ensembl | UCSC

Print Full Report

Genome View | Evidence for Haploinsufficiency Phenotypes | Evidence for Triplosensitive Phenotypes

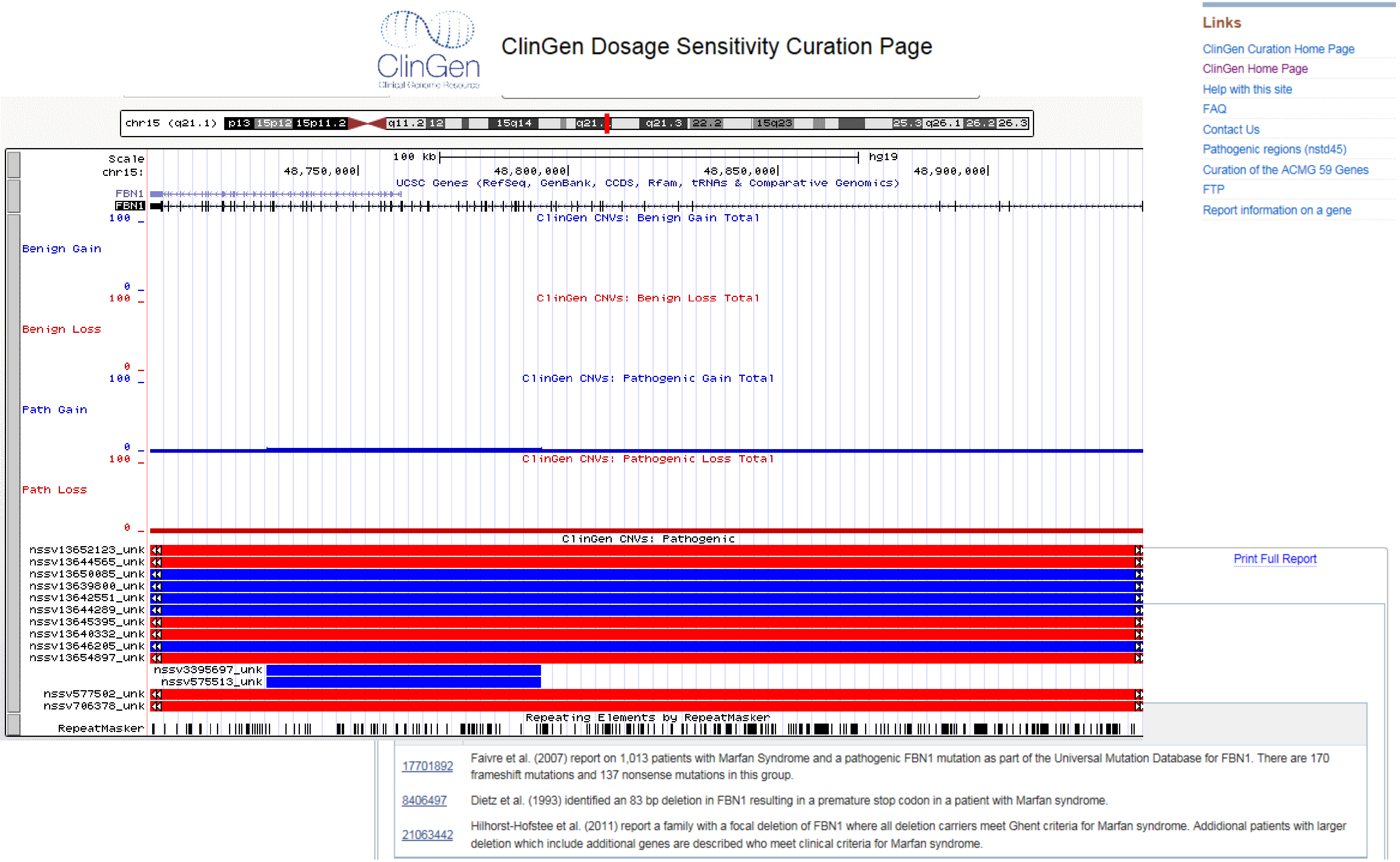
Haploinsufficiency score: 3
Strength of Evidence (disclaimer): Sufficient evidence for dosage pathogenicity
Haploinsufficiency Phenotype: MARFAN SYNDROME; MFS

Evidence for haploinsufficiency phenotype

PubMed ID	Description
17701892	Faivre et al. (2007) report on 1,013 patients with Marfan Syndrome and a pathogenic FBN1 mutation as part of the Universal Mutation Database for FBN1. There are 170 frameshift mutations and 137 nonsense mutations in this group.
8406497	Dietz et al. (1993) identified an 83 bp deletion in FBN1 resulting in a premature stop codon in a patient with Marfan syndrome.
21063442	Hilhorst-Hofstee et al. (2011) report a family with a focal deletion of FBN1 where all deletion carriers meet Ghent criteria for Marfan syndrome. Addidional patients with larger deletion which include additional genes are described who meet clinical criteria for Marfan syndrome.

Interpretation of CNV calls – clinical DB

ClinVar track



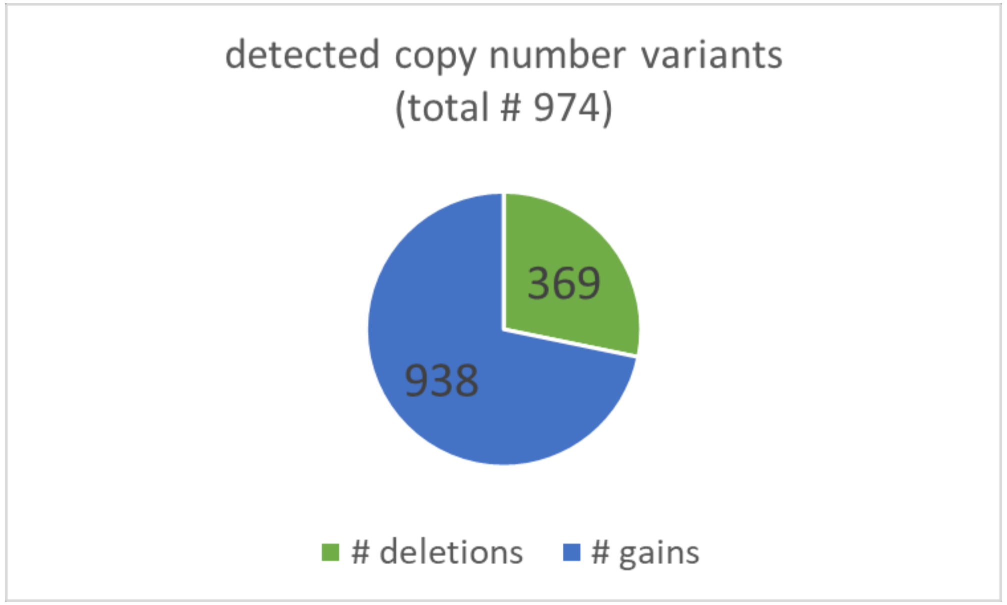
17701892 Faivre et al. (2007) report on 1,013 patients with Marfan Syndrome and a pathogenic FBN1 mutation as part of the Universal Mutation Database for FBN1. There are 170 frameshift mutations and 137 nonsense mutations in this group.

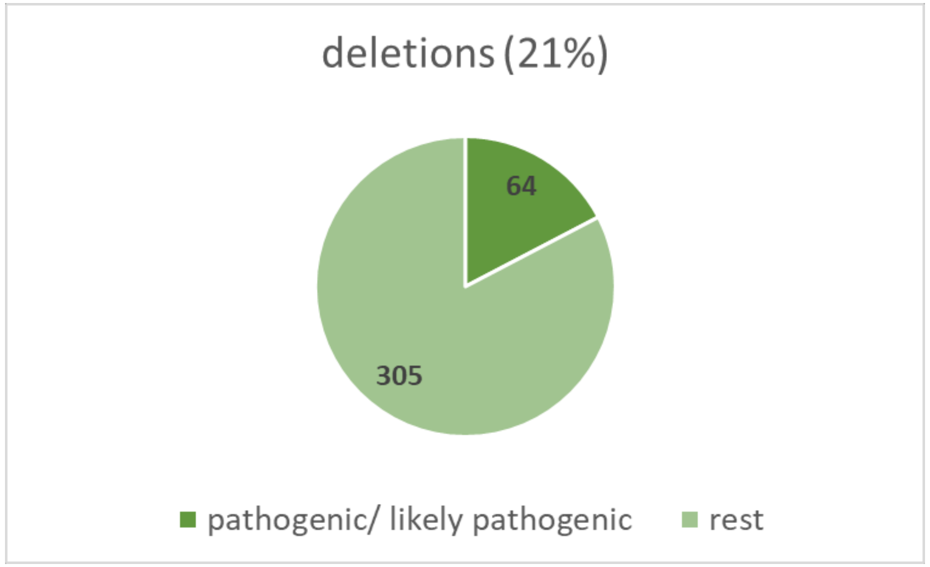
8406497 Dietz et al. (1993) identified an 83 bp deletion in FBN1 resulting in a premature stop codon in a patient with Marfan syndrome.

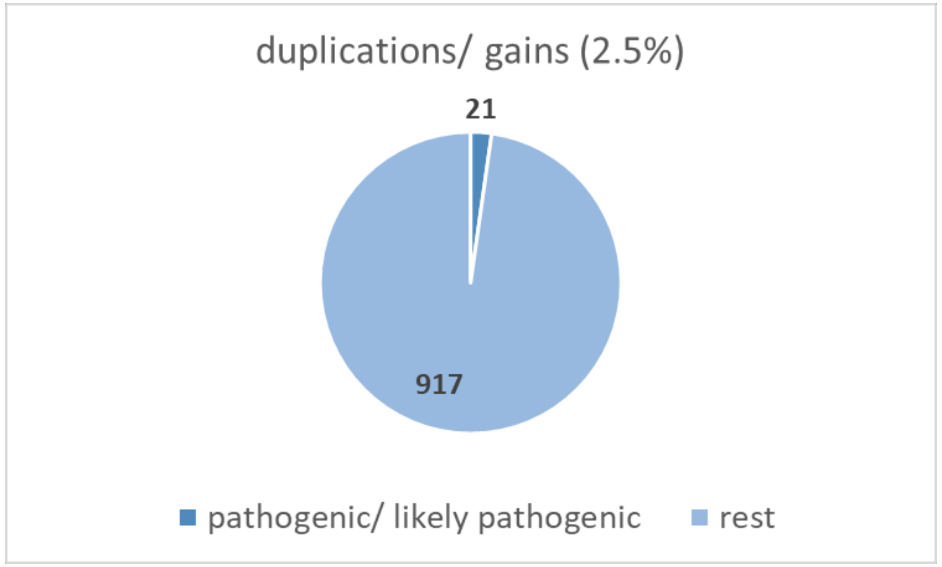
21063442 Hilhorst-Hofstee et al. (2011) report a family with a focal deletion of FBN1 where all deletion carriers meet Ghent criteria for Marfan syndrome. Addidional patients with larger deletion which include additional genes are described who meet clinical criteria for Marfan syndrome.

CNV analysis on ~3700 individuals within the routine Dx

- Rare diseases
- Hereditary cancer


detected copy number variants
(total # 974)


deletions (21%)

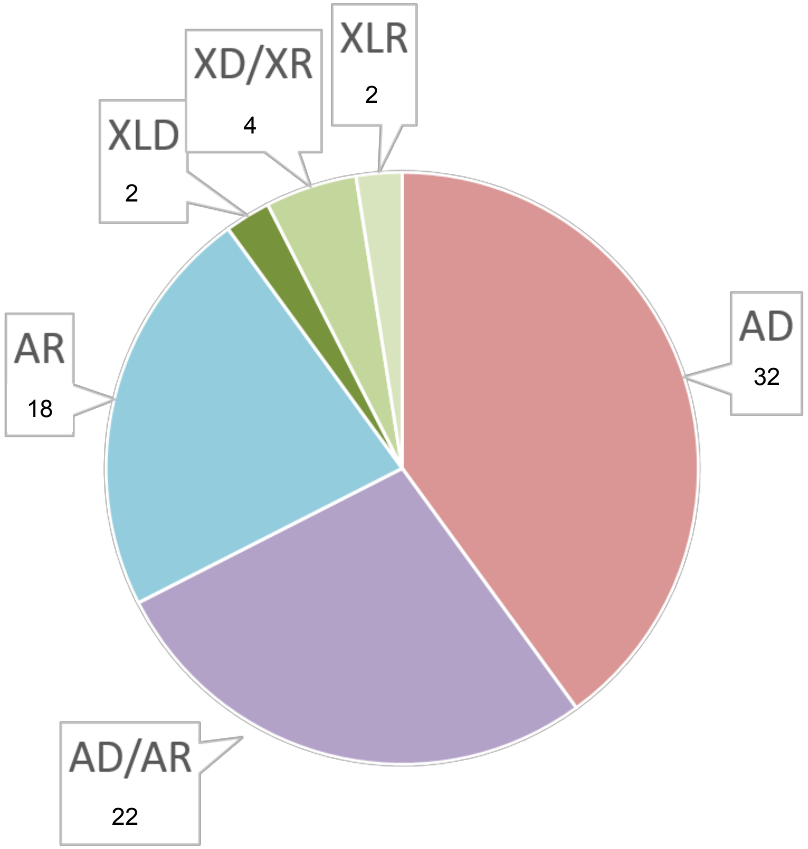

duplications/ gains (2.5%)

CNV analysis diagnostic yield

- CNV clarified the underlying phenotype in 8 % of the cases

- Increase the overall diagnostic yield in ~5% compared to MLPA approach

- Despite of the challenges CNV detection based on WES data may give a quick insight into CNV patterns for a specific disease or phenotype



MGZ

benet-pages@mgz-muenchen.de

Anke Arnold | Bioinformatics Data Scientist

Florentine Scharf | Bioinformatics Data Scientist

Glen Currier | Software Engineer

Tobias Wohlfrom | Bioinformatics Software Engineer

Madalina Giurgiu | Software Engineer, MsC Bioinformatics Student

https://www.mgz-muenchen.de/startseite.html