CNV detection

from targeted next-generation sequencing data

Anna Benet-Pagès

Johor 29.08.2018

Detection of structural variants and human disease

B/B^+

Calvin Bridges, 1936

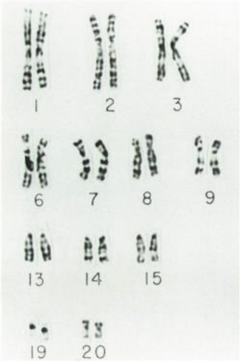
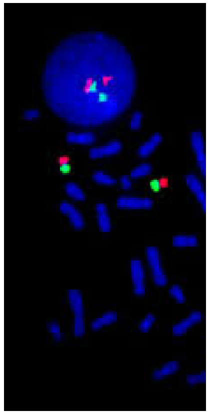
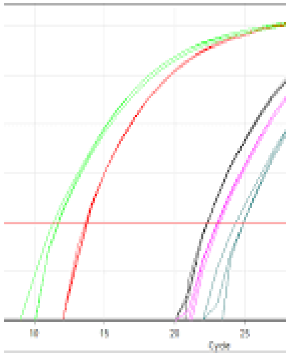


FIGURE 1B: G-banding karyotype of 46, XY individual.

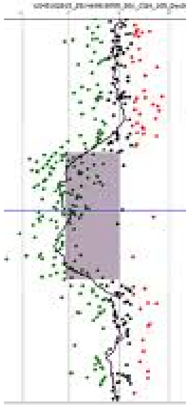
G-banding *Rowley et al.*, 1973



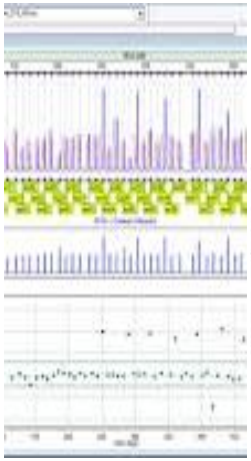
FISH *Langer et al.*, 1982



Quantitative PCR *Porcher et al.*, 1992



Array CGH *Pinkel et al.*, 1998c

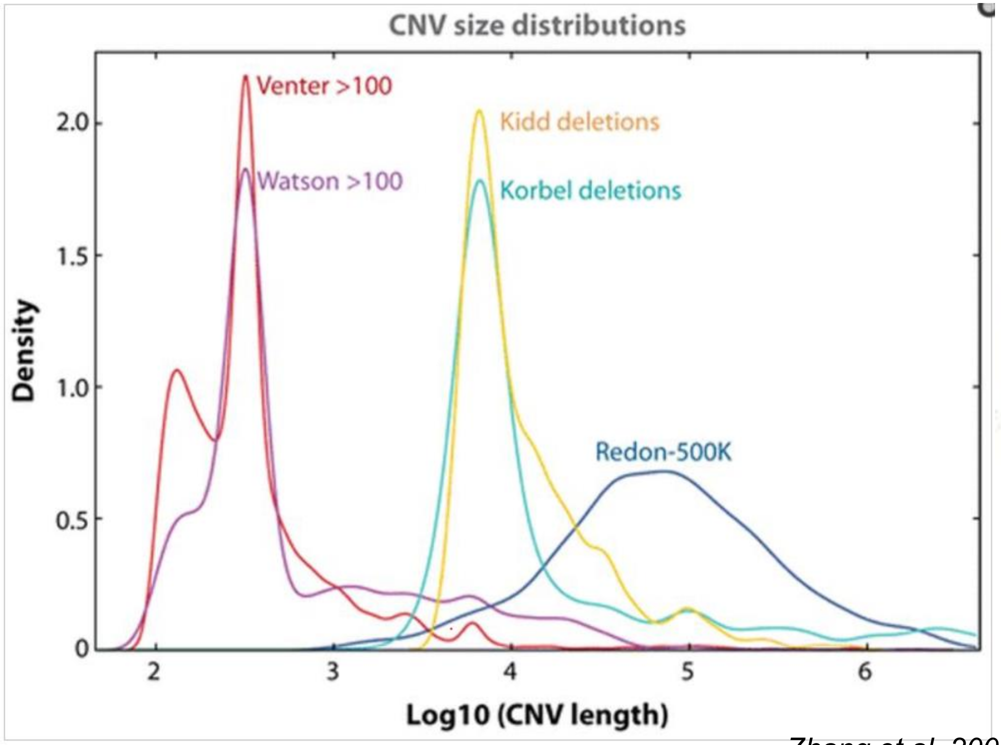


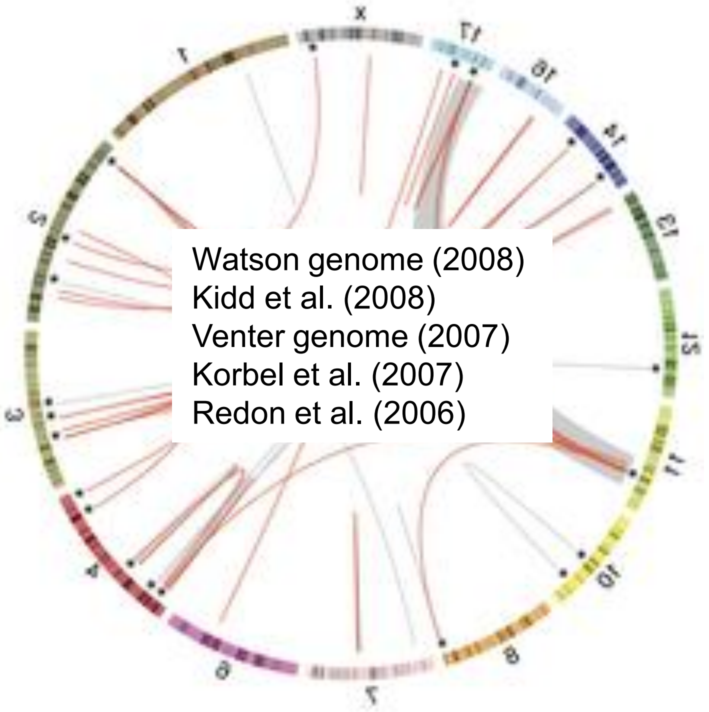
MLPA *Schouten et al.*, 2002

SNP array *Komura et al.*, 2006

High resolution technologies reveal small-size CNVs

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



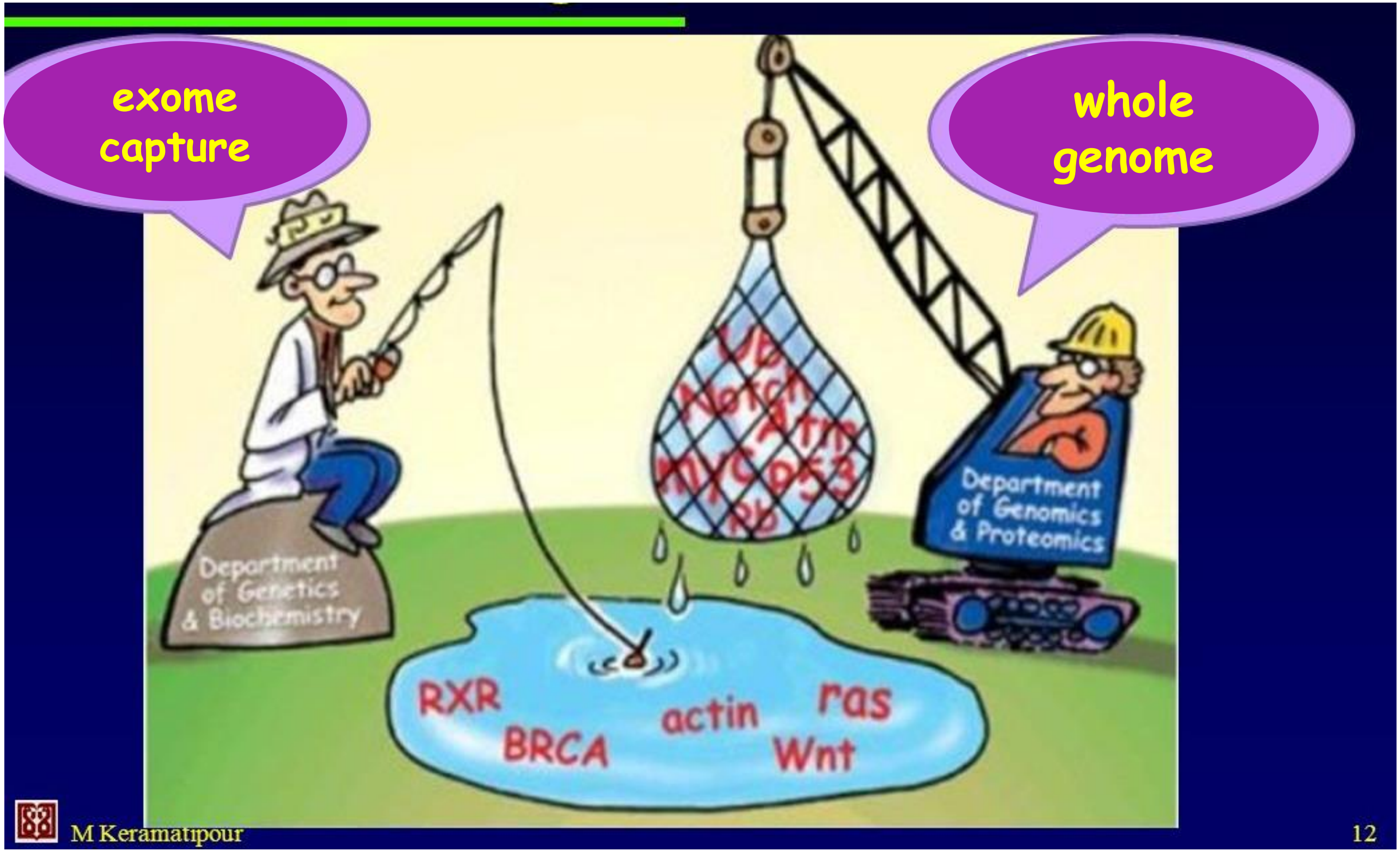


Watson genome (2008)
Kidd et al. (2008)
Venter genome (2007)
Korbel et al. (2007)
Redon et al. (2006)

Zhang et al. 2009

- Smaller structural variants are the most frequent

On the meanwhile…



CNV detection from targeted-capture data

Challenges:

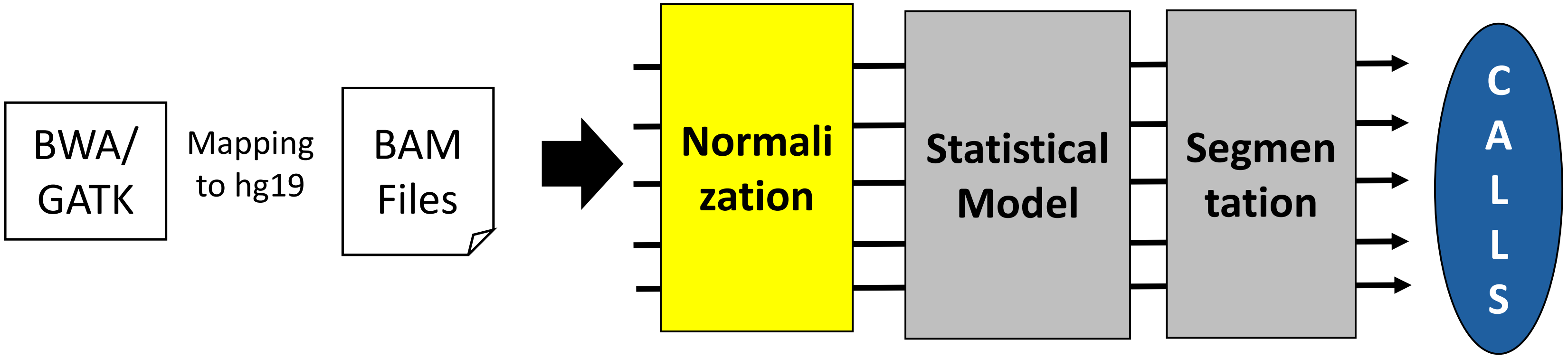
- CNV detection from exon capture approaches depends solely on read depth data
- Enrichment efficiency introduces a systematic noise in read depth data
- Coverage bias between sequencing runs and within samples of the same run
- Single exon events are extremely difficult to detect
- Control individuals are difficult to obtain (reference set / validation)
- Validation is expensive

CNV detection methods general considerations

➢ Which tool should I choose?

- applicable to capture data
- calling of rare CNVs
- 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 Pipeline Structure



BWA/
GATK

Mapping
to hg19

BAM
Files

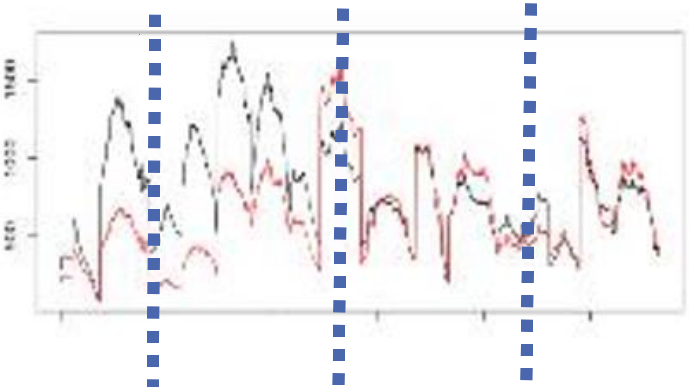
Normali
zation

Statistical
Model

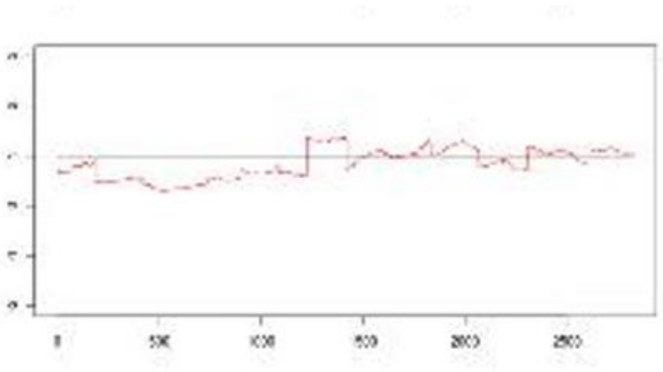
Segmen
tation

C
A
L
L
S

breakdown of the target
region exons/ windows

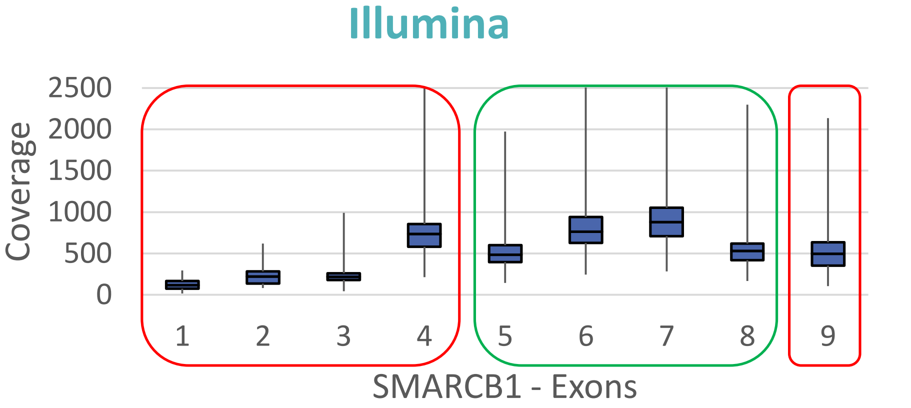


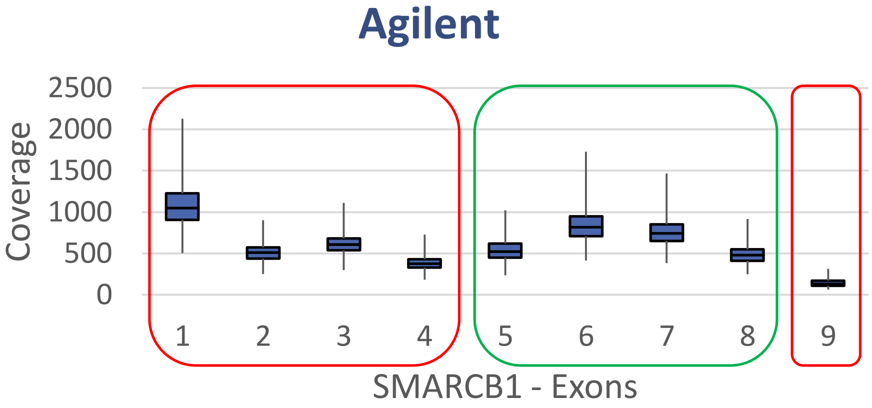
GC
Mapping qual
# reads
Ref. Set



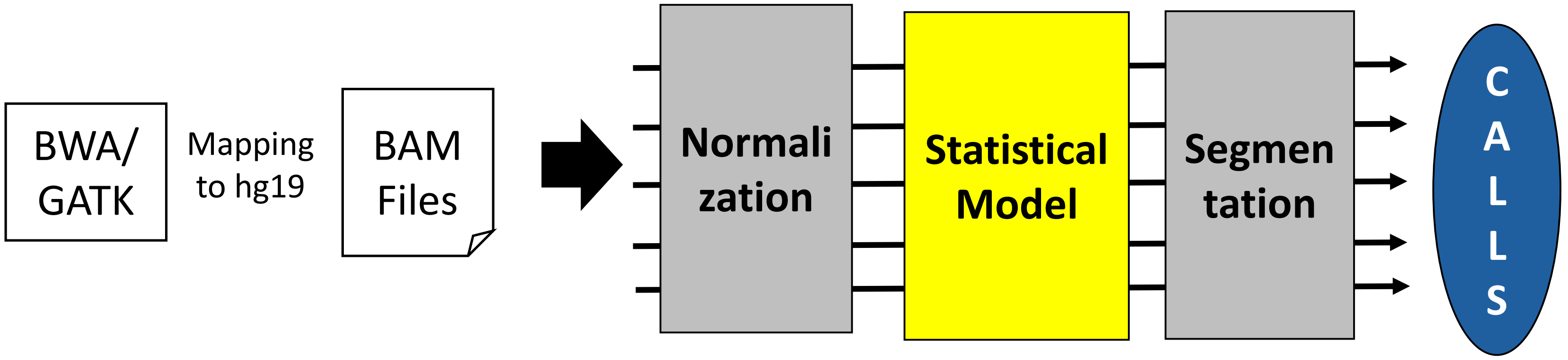
Reference Sets and data normalization

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

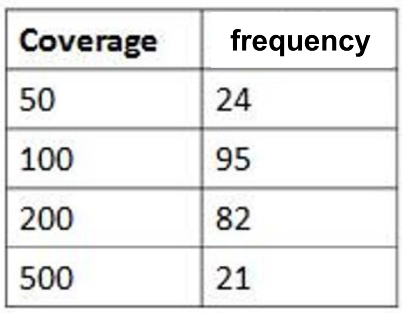




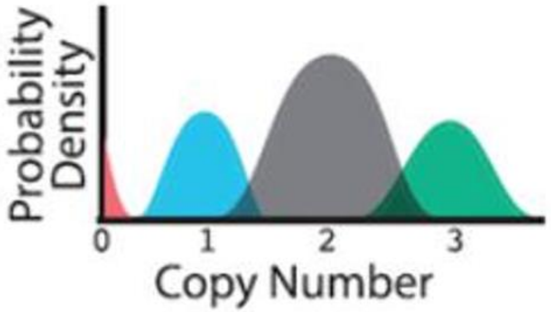
CNV Pipeline Structure



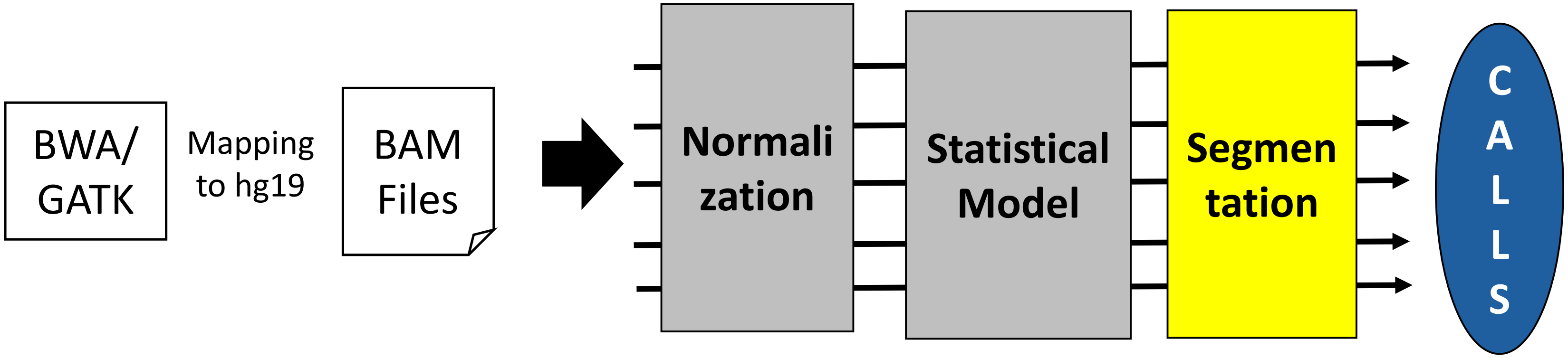
Poison distribution
Beta binomial
Negative binomial
Normal distribution



Coverage	frequency
50	24
100	95
200	82
500	21



CNV Pipeline Structure



BWA/
GATK

Mapping
to hg19

BAM
Files

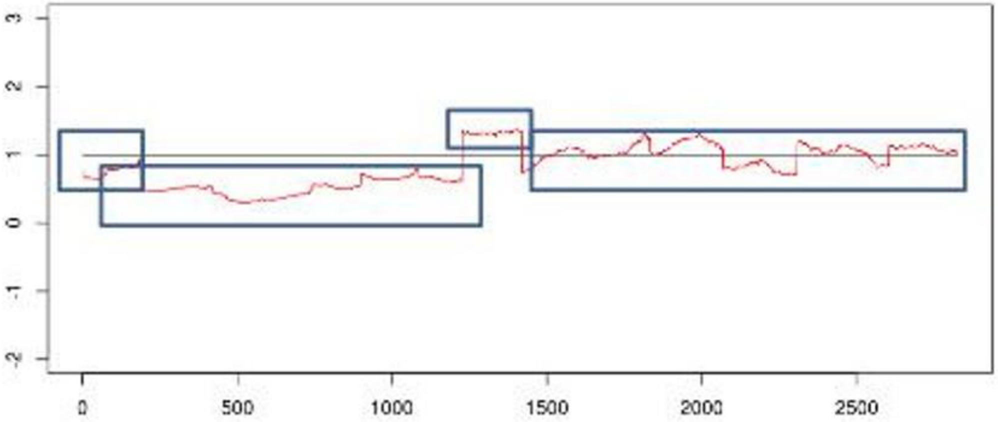
Normali
zation

Statistical
Model

Segmen
tation

C
A
L
L
S

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



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 **„meta-CNV-caller“** rcanavar, 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

Noll et al., Npjgenmed 2016

Meta-Tool CNV Detection Pipeline

- **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)

- **Inhouse method**

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

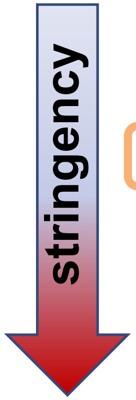
Performance of single tools

- Training set: true set of 146 CNV calls detected via MLPA.

	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%

Performance of tool combinations

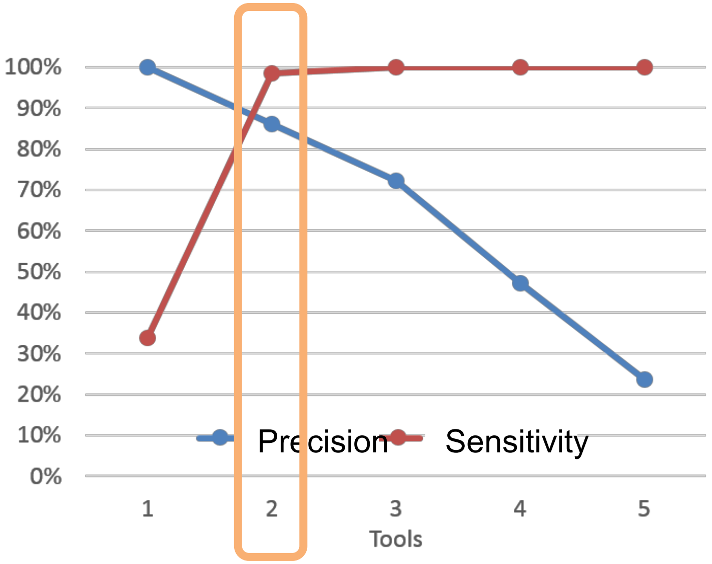
- What is the best number of tools required to call a variant?



stringency

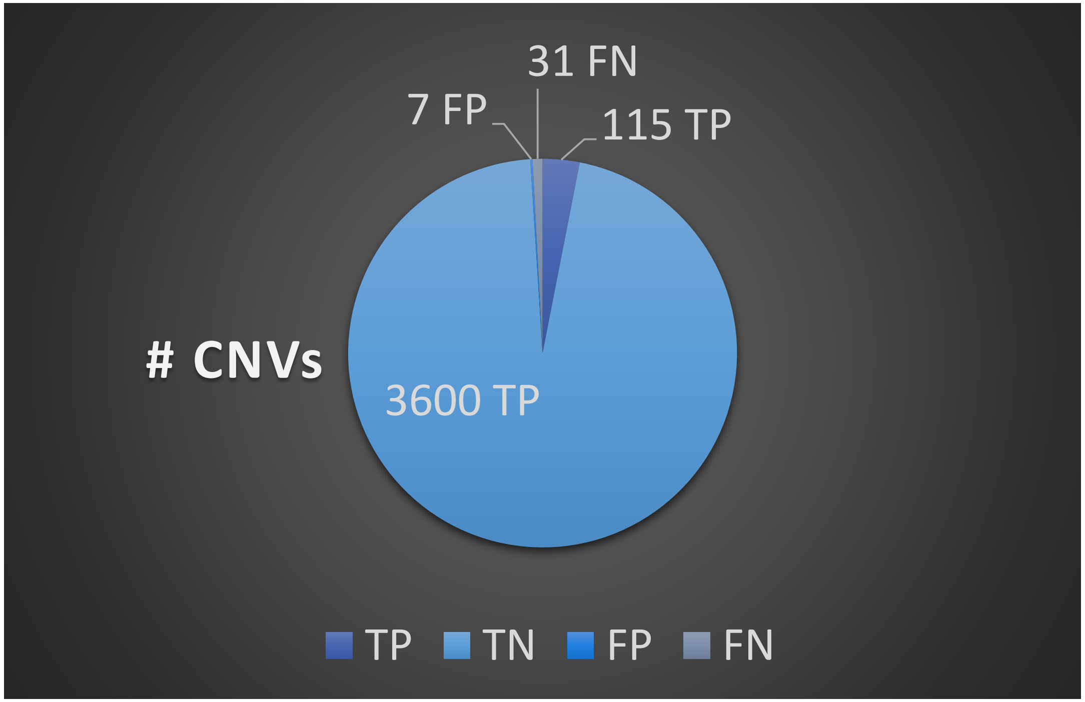
	TP	FN	FP	TN	Sensitivity (TPR)	Specificity (TNR)	Precision (PPV)	NPV
2 out of 5	115	7	8	3600	94.26%	99.78%	93.50%	99.81%
3 out of 5	72	50	1	3600	59.02%	99.97%	98.63%	98.63%
4 out of 5	28	94	1	3600	22.95%	99.97%	96.55%	97.46%
5 out of 5	4	118	0	3600	3.28%	100.00%	100.00%	96.83%

- Using two out of five concordant tool predictions shows the best balance between sensitivity and specificity



CNV Pipeline Evaluation

- >3700 MLPAs were performed in ~90 genes
- 146 CNVs (85 deletions / 61 gains)



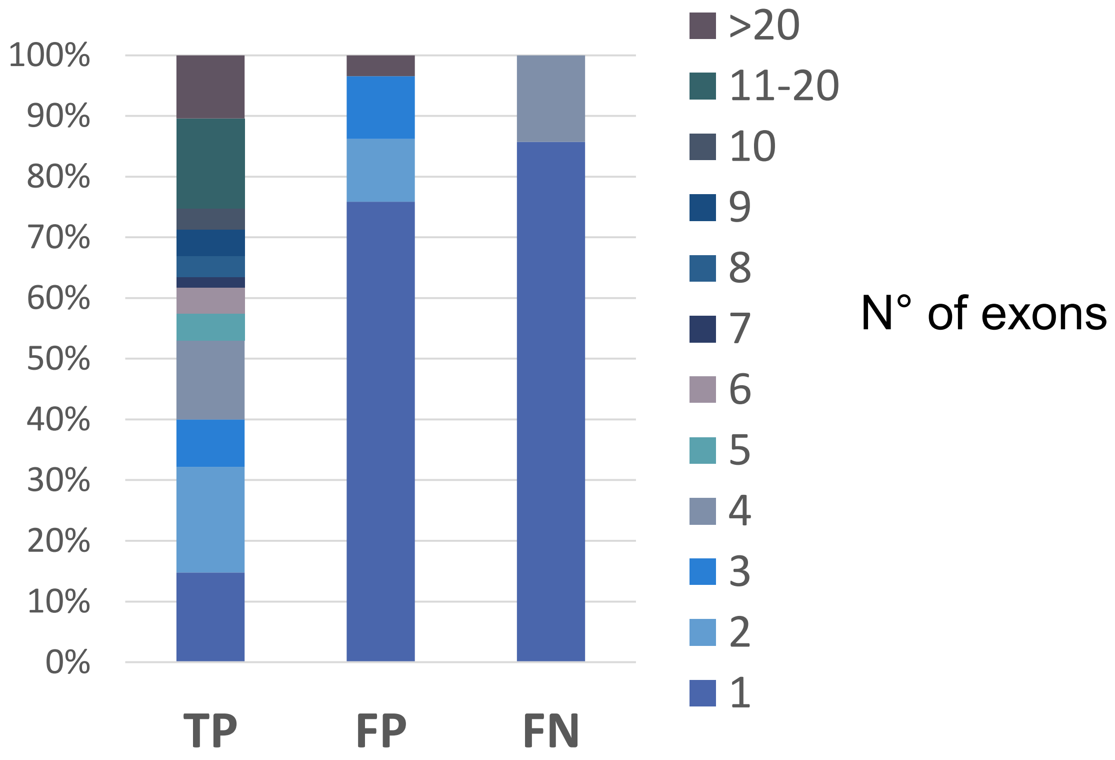
Sensitivity:	88.60%
Specificity:	98.88%
Precision:	71.40%

Pseudogenes excluded:

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

TP, FP , FN versus CNV size

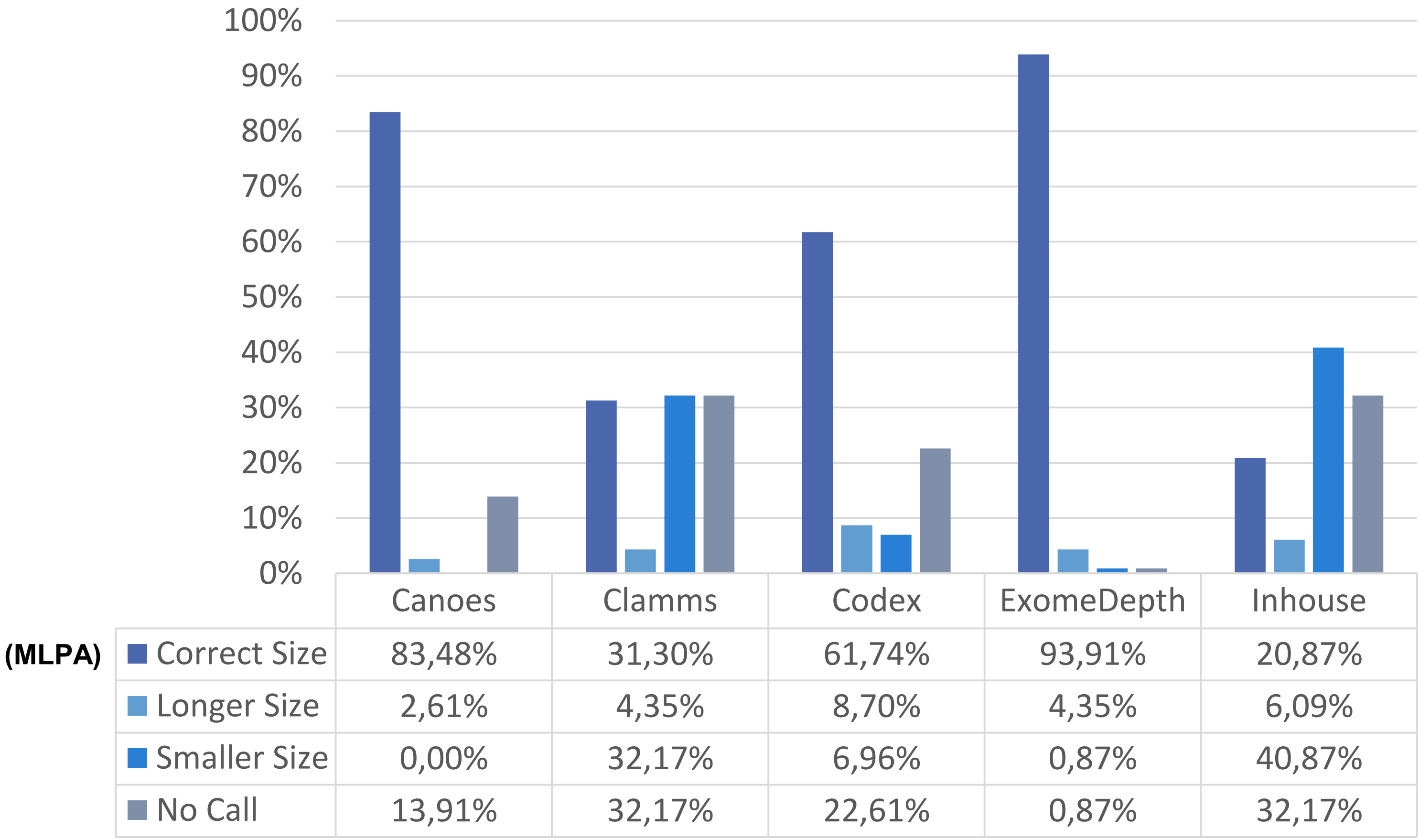
- Comparison of CNV sizes of TP, FP and FN calls detected by the combined CNV pipeline on the validation set.
- FP mainly cosiand FN calls consist mainly of single exon events.



definition of special quality thresholds for single exon events to minimize false negatives

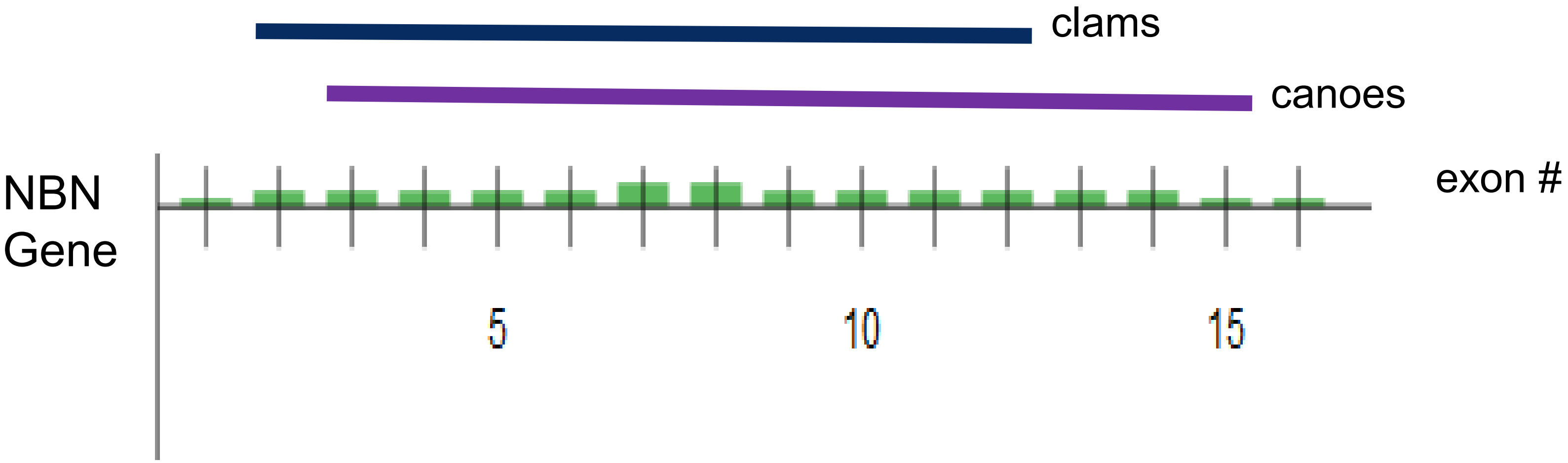
Discrepancies in CNV size detection between tools

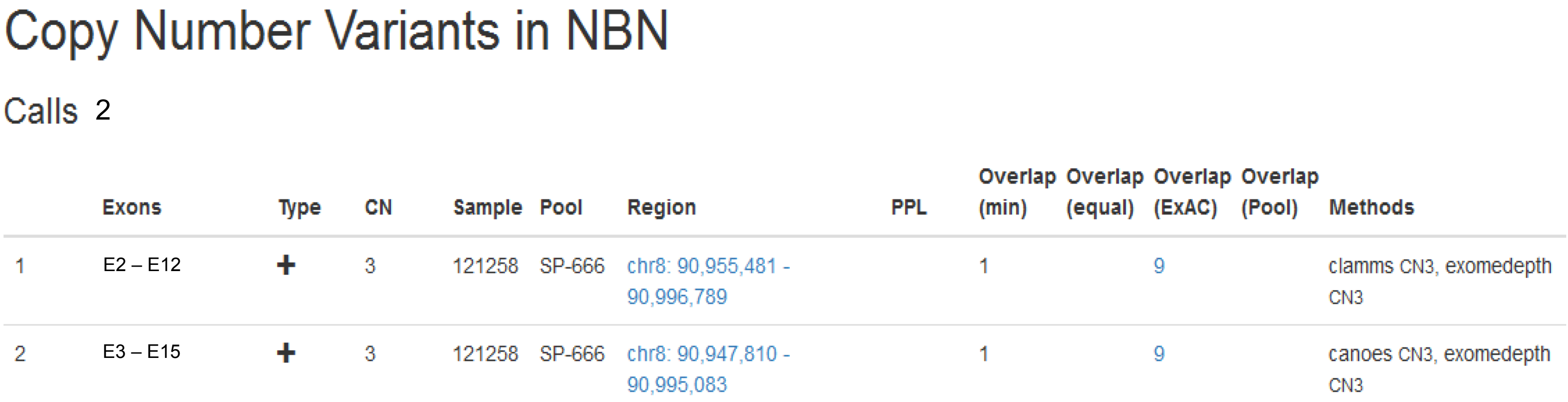
- Size of CNV calls were compared to MLPA
- Size is given as the number of exons within the call



		Canoes	Clamms	Codex	ExomeDepth	Inhouse
(MLPA)	■ Correct Size	83,48%	31,30%	61,74%	93,91%	20,87%
	■ Longer Size	2,61%	4,35%	8,70%	4,35%	6,09%
	■ Smaller Size	0,00%	32,17%	6,96%	0,87%	40,87%
	■ No Call	13,91%	32,17%	22,61%	0,87%	32,17%

Meta-CNV-caller: multi calls for one event





Copy Number Variants in NBN

Calls 2

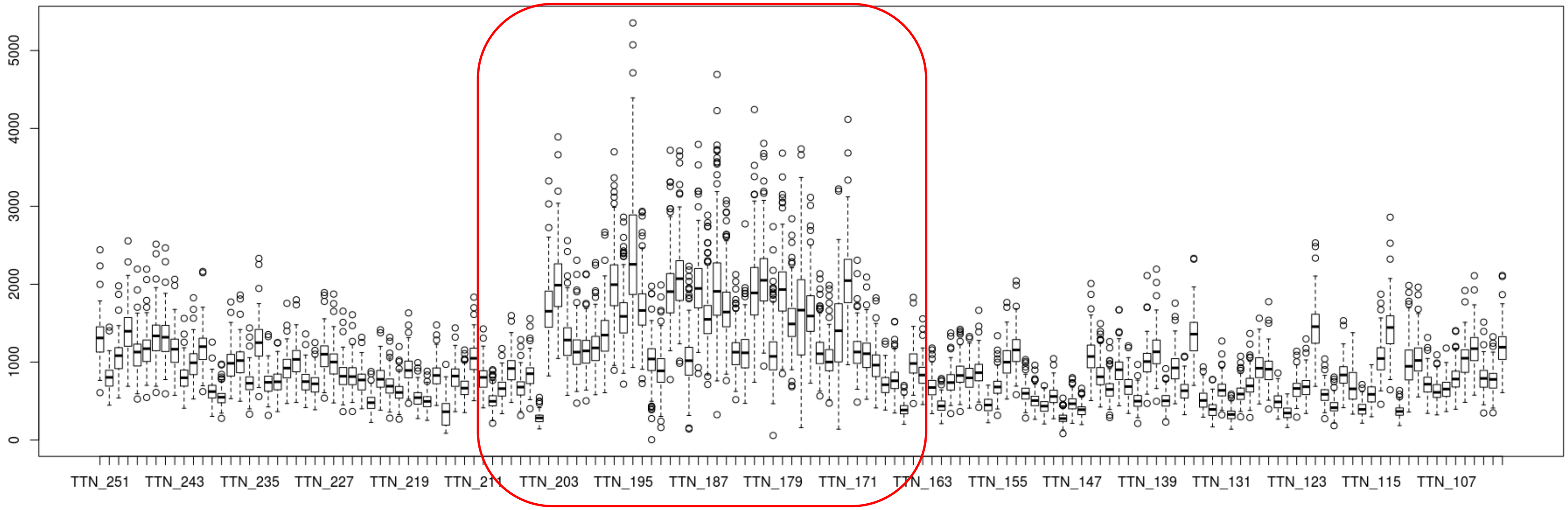
	Exons	Type	CN	Sample	Pool	Region	PPL	Overlap	Overlap	Overlap	Overlap	Methods
								(min)	(equal)	(ExAC)	(Pool)	
1	E2 – E12	+	3	121258	SP-666	chr8: 90,955,481 - 90,996,789		1		9		clamms CN3, exomedepth CN3
2	E3 – E15	+	3	121258	SP-666	chr8: 90,947,810 - 90,995,083		1		9		canoes CN3, exomedepth CN3

Challenges

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

non-uniform coverage = capture bias

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



can not be analyzed

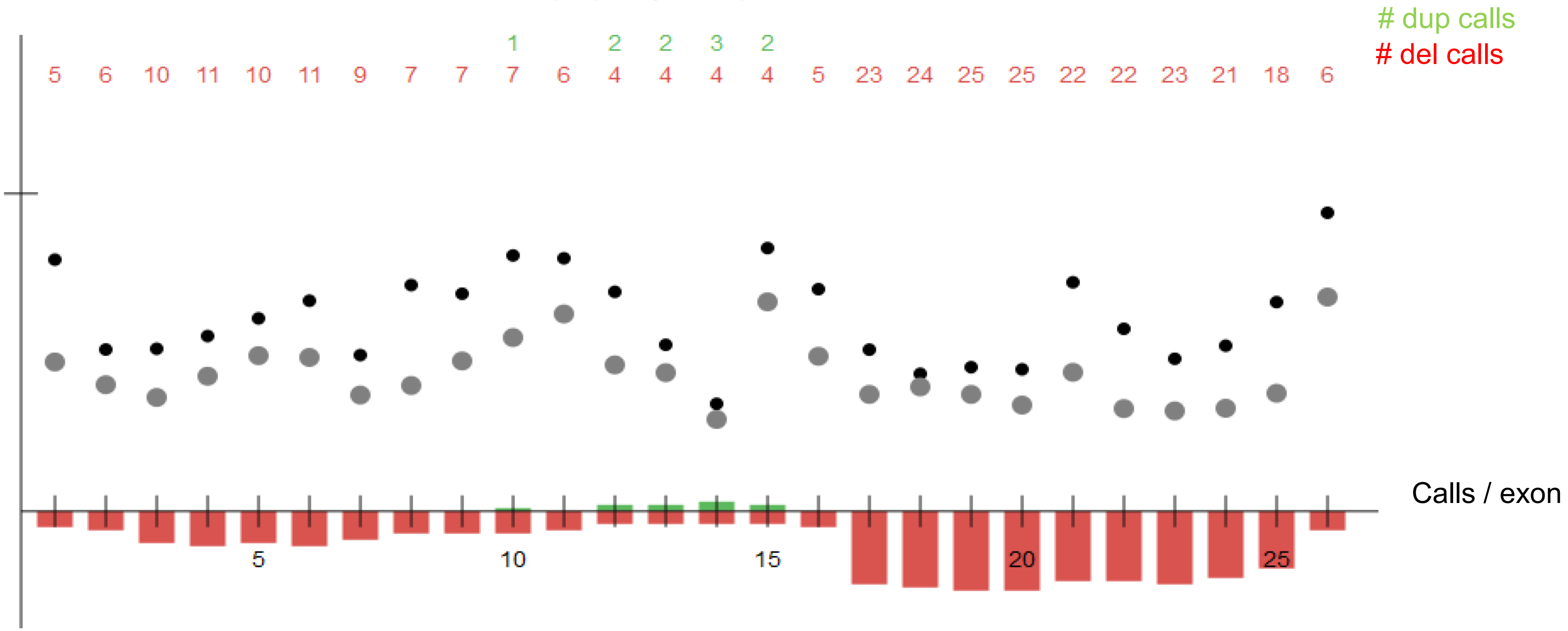
CNV calling artifacts

2. SCN1A

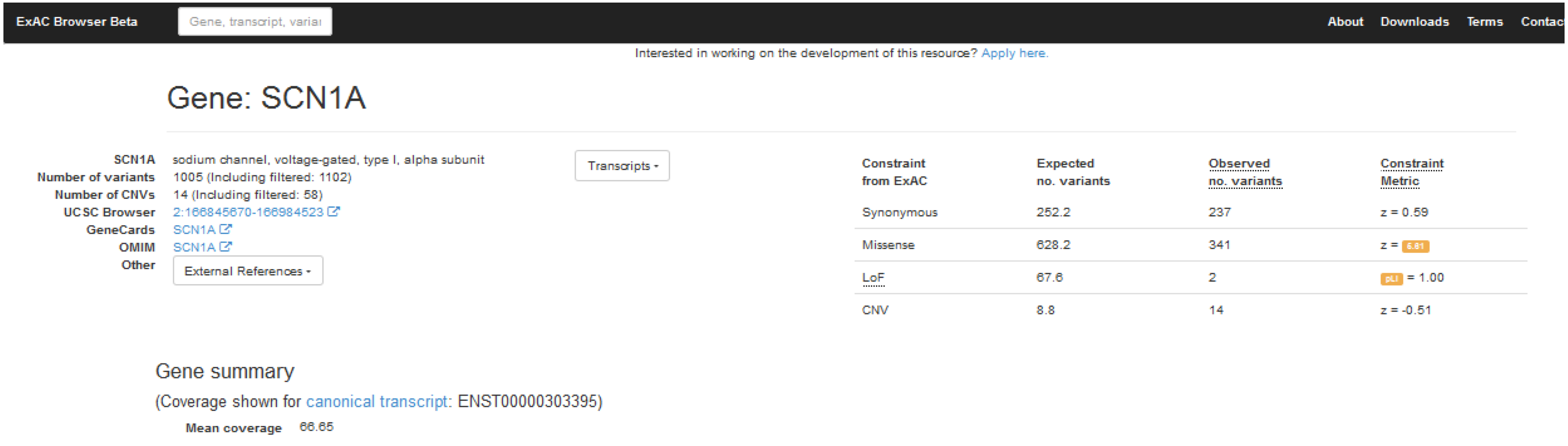
● Depth of coverage uniformity reference set

● Depth of coverage uniformity sample

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



CNV calling artifacts



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

Gene: SCN1A

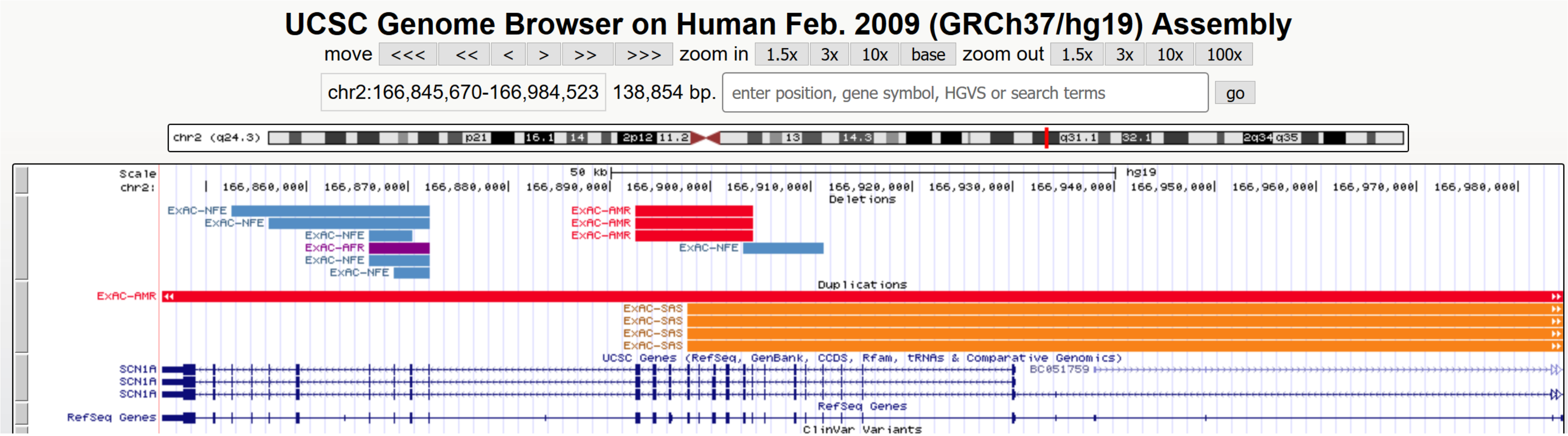
SCN1A	sodium channel, voltage-gated, type I, alpha subunit
Number of variants	1005 (Including filtered: 1102)
Number of CNVs	14 (Including filtered: 58)
UCSC Browser	2:166845670-166984523
GeneCards	SCN1A
OMIM	SCN1A
Other	External References ▾

Constraint from ExAC	Expected no. variants	Observed no. variants	Constraint Metric
Synonymous	252.2	237	z = 0.59
Missense	628.2	341	z = 6.81
LoF	67.6	2	pLI = 1.00
CNV	8.8	14	z = -0.51

Gene summary

(Coverage shown for canonical transcript: ENST00000303395)

Mean coverage 66.65



UCSC Genome Browser on Human Feb. 2009 (GRCh37/hg19) Assembly

chr2:166,845,670-166,984,523 138,854 bp.

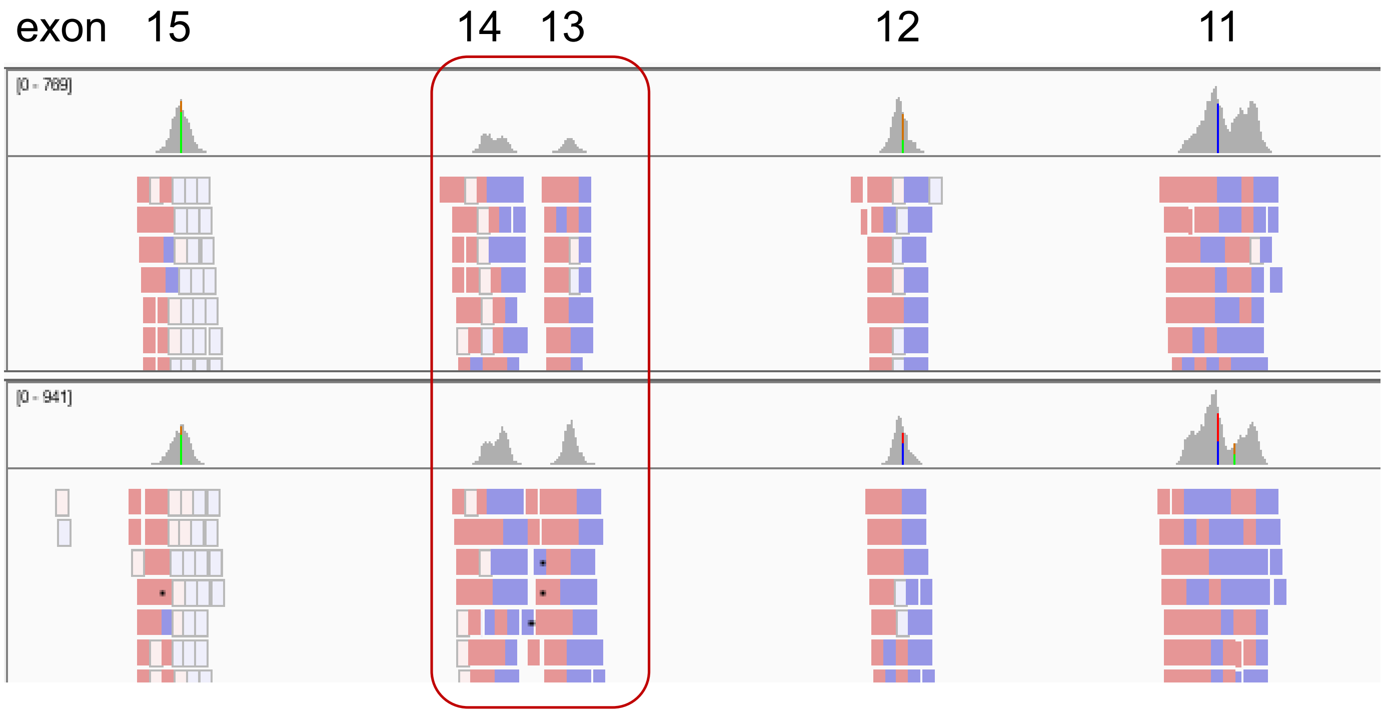
CNV calling in Pseudogenes

🟥 Forward read / unique mapping

🟦 Reverse read / unique mapping

□ Non-unique mapping

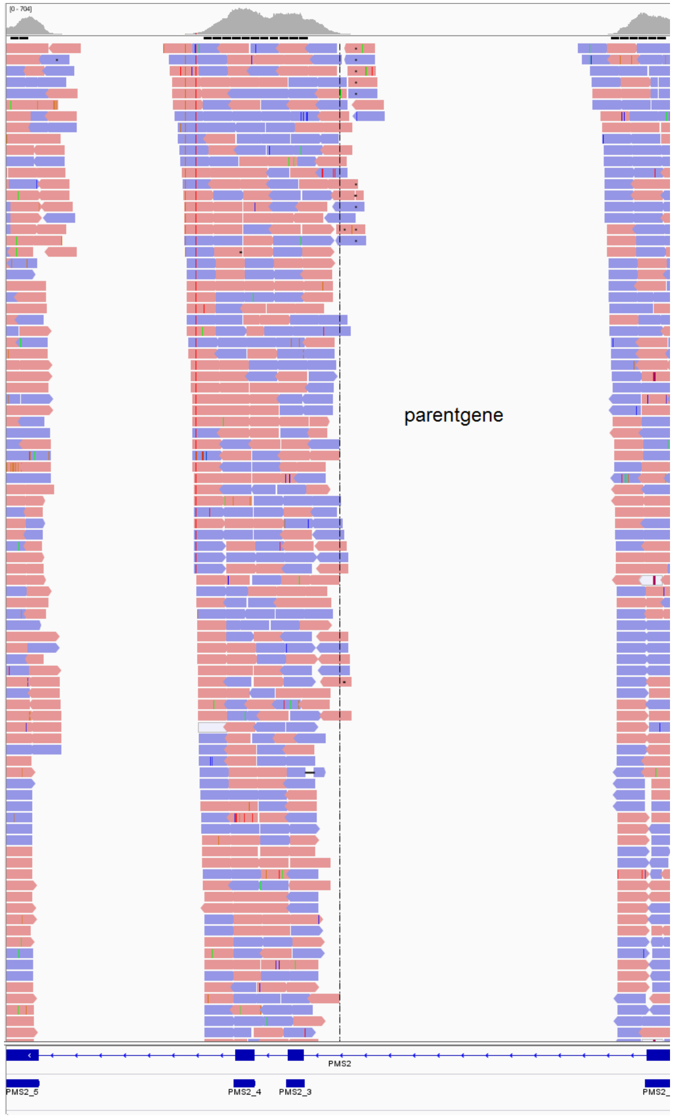
1. PMS2



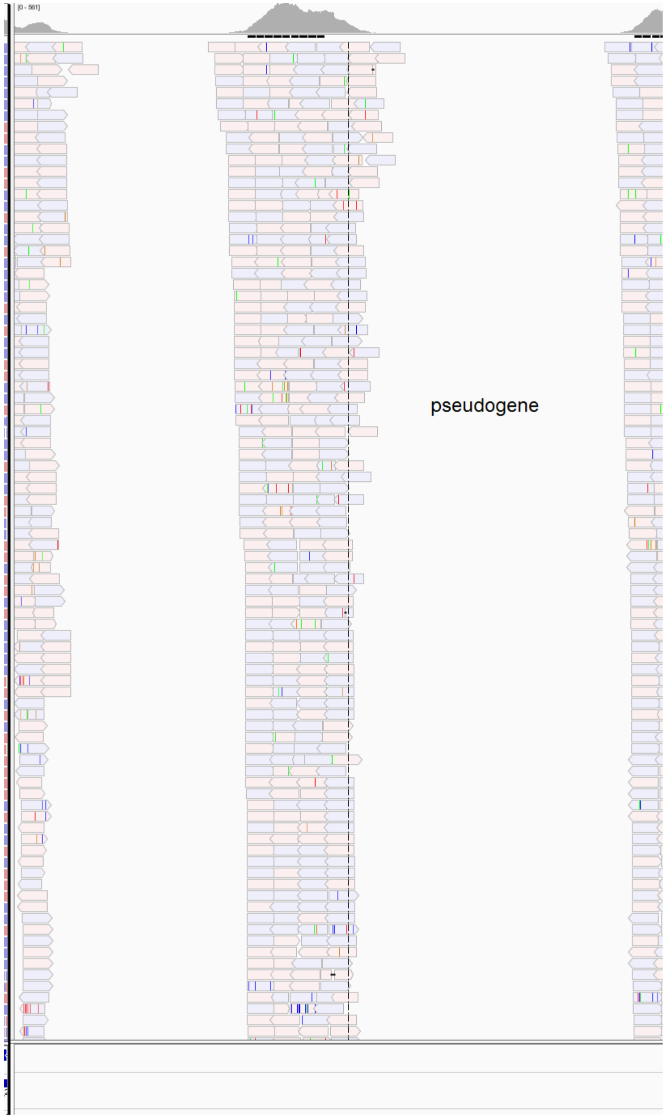
PMS2 exons 11 – 15 can not be analyzed

CNV calling in Pseudogenes

PMS2

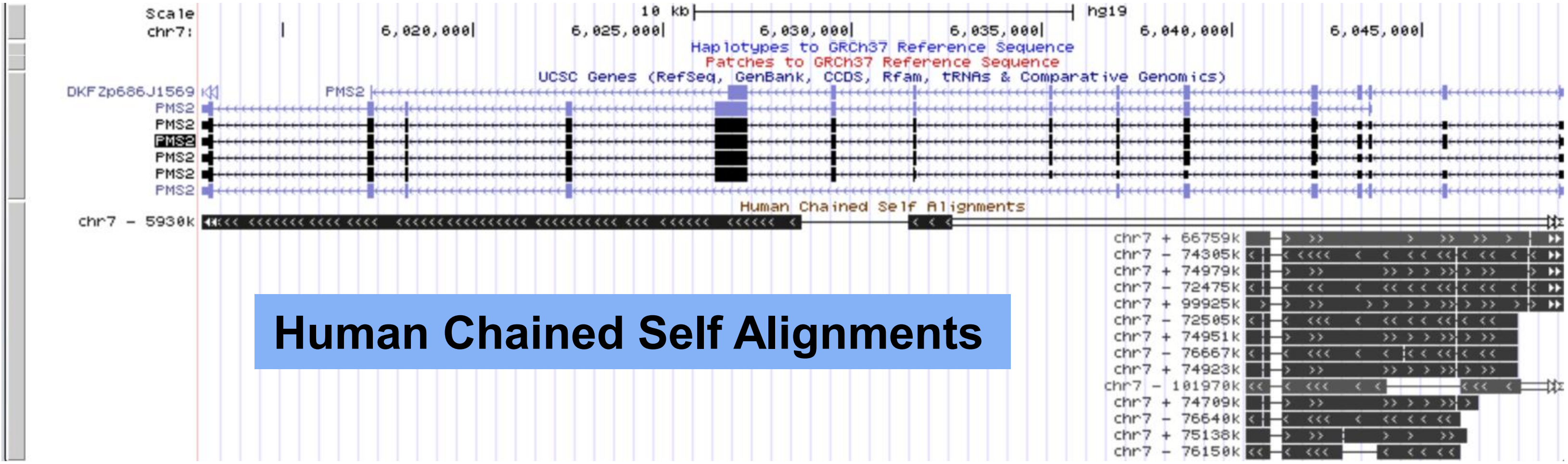


PMS2CL

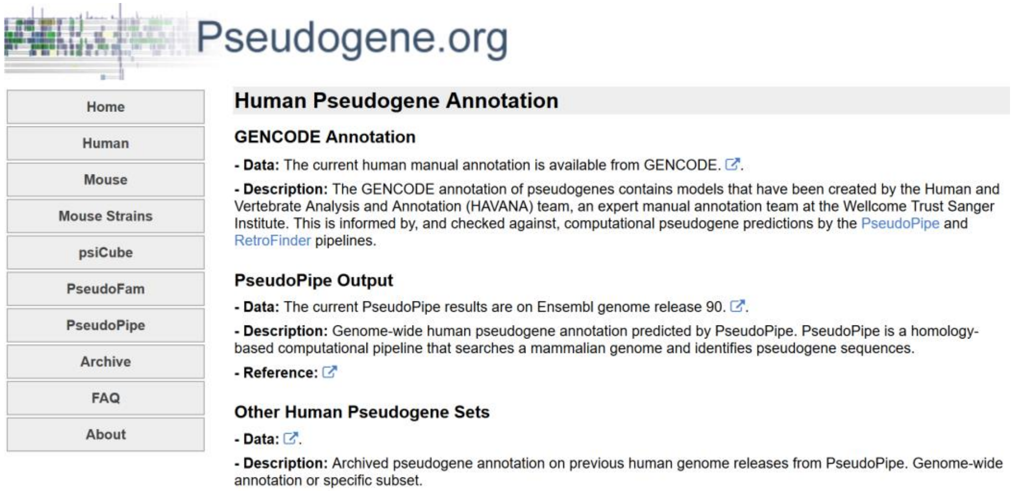


How to identify regions affected by pseudogenes

- alignments of the human genome with itself using blastz



Human Chained Self Alignments



Pseudogene.org

Home
Human
Mouse
Mouse Strains
psiCube
PseudoFam
PseudoPipe
Archive
FAQ
About

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:**

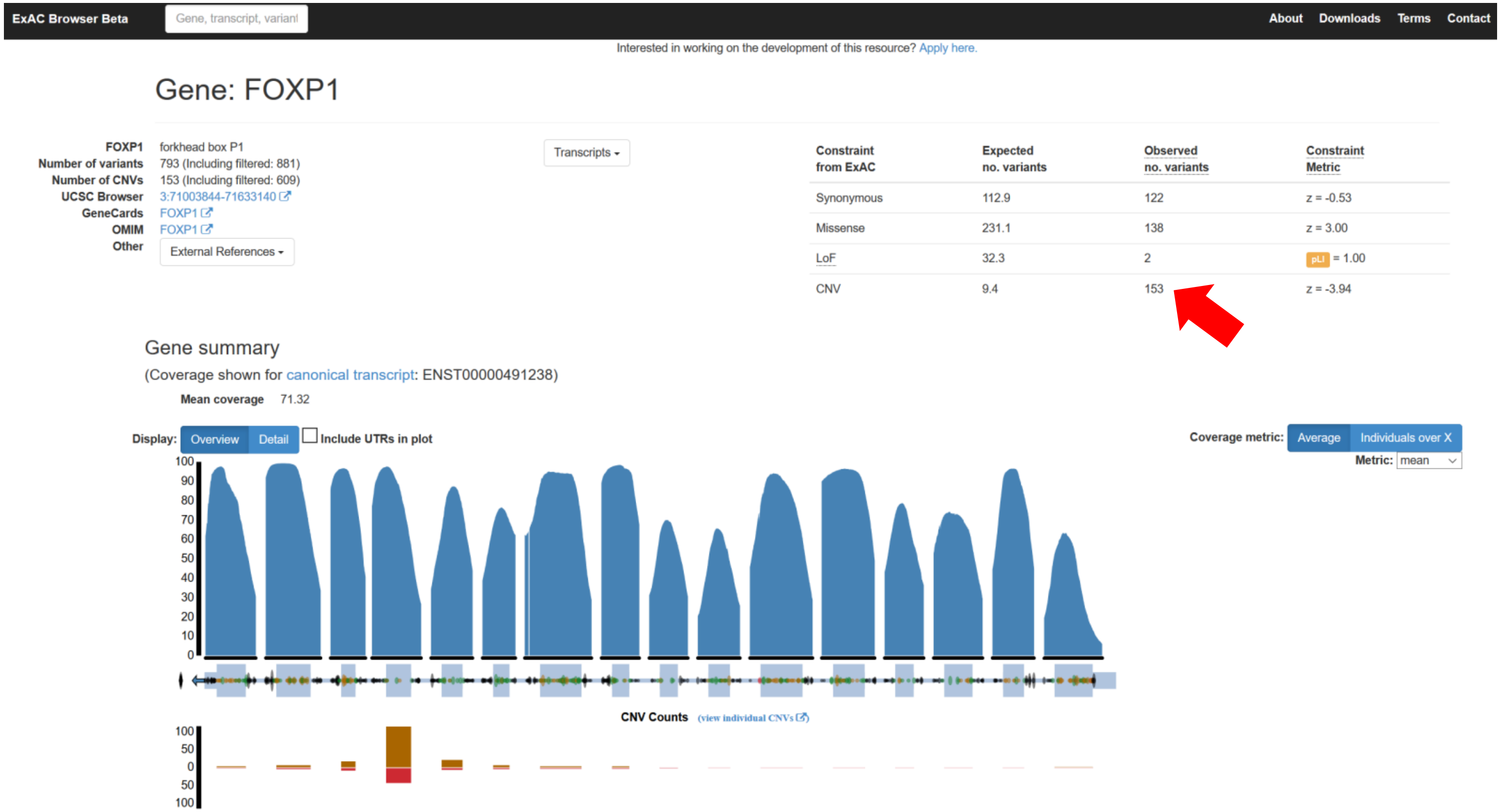
Other Human Pseudogene Sets

- **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



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

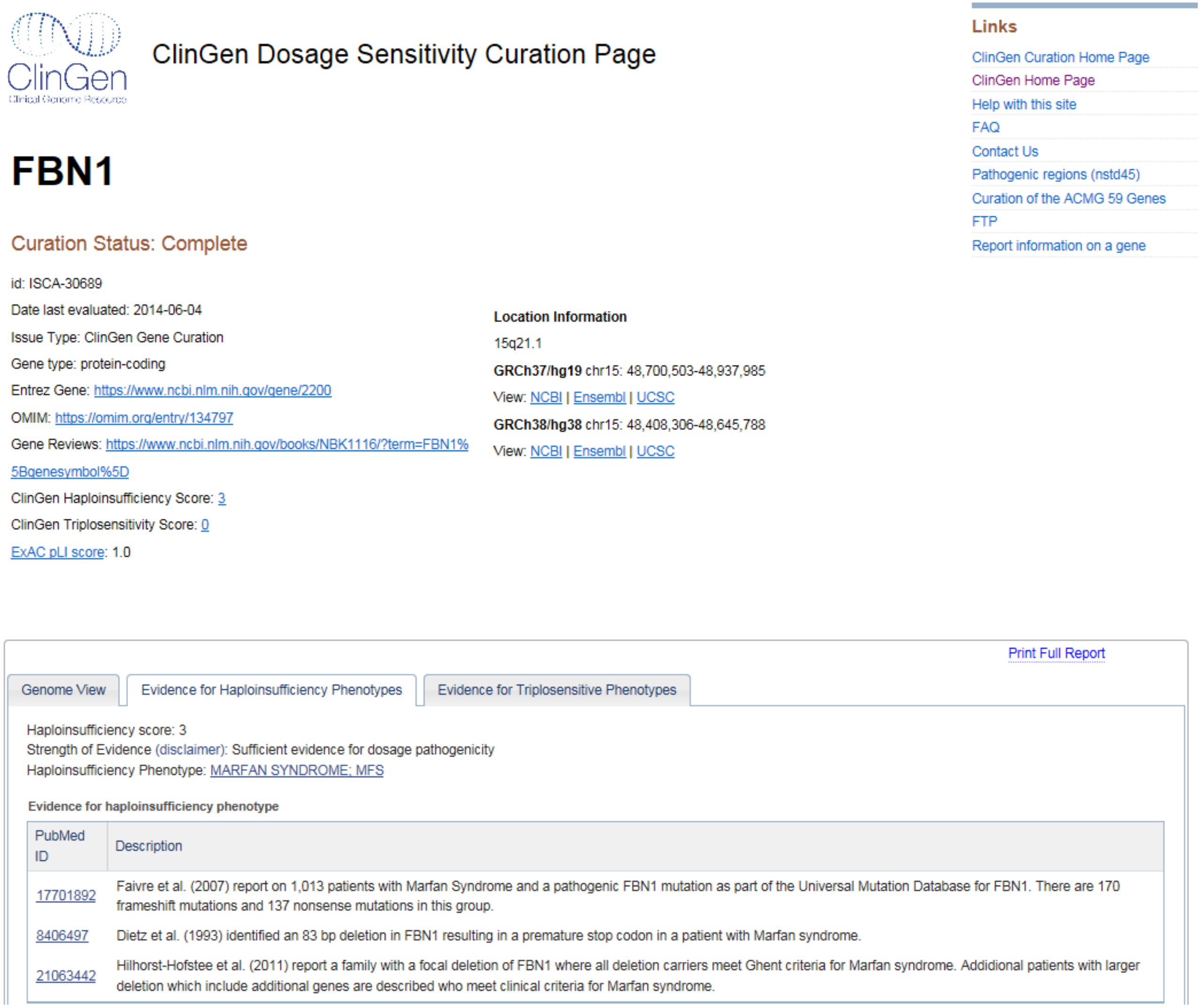
(Coverage shown for canonical transcript: ENST00000491238)

Mean coverage 71.32

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

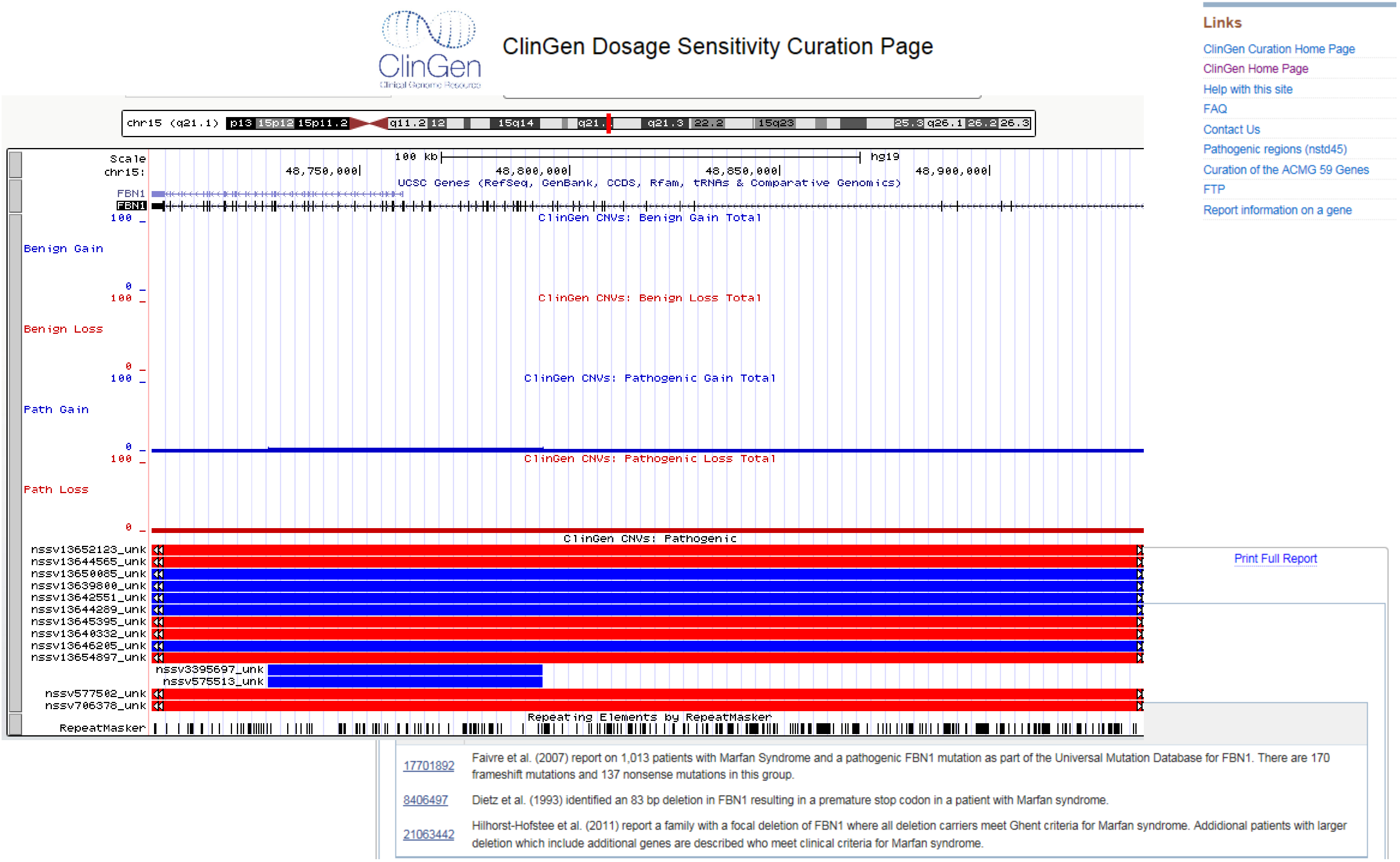
Genome View | Evidence for Haploinsufficiency Phenotypes | Evidence for Triplosensitive Phenotypes

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



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

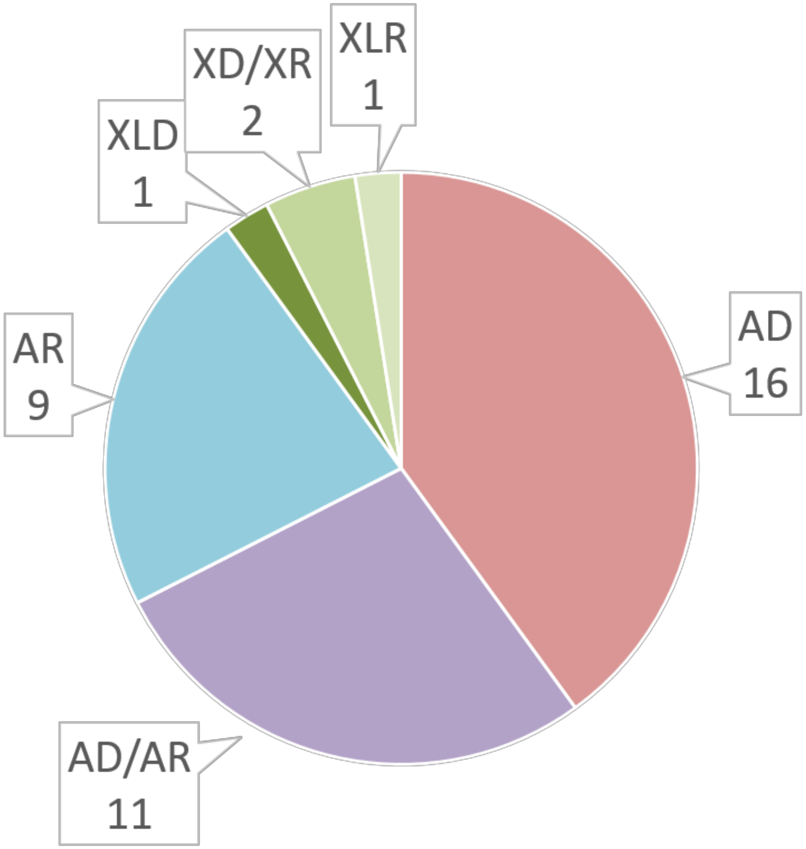
Print Full Report

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 **1600** individuals within the routine Dx

CNV clarified the underlying phenotype in 8 % of the cases

Increase of the diagnostic yield in 3%



benet-pages@mgz-muenchen.de