panelcn.MOPS and SavvyCNV together detect the most exon-level copy number variants from targeted-capture

Sharing software instead of data is useful for clinical collaboration

A Multi-Site Comparison of Copy Number Variant Callers for Targeted-Capture Sequencing

S. Piatek[1], E. Sollars[2], P. Lombard[1], A. C. Davies[3], H. Ahlfors[1] and L. Jenkins[1]

@stefanpiatek

stefan.piatek@nhs.net

[1] *London North Genomic Laboratory Hub, Great Ormond Street Hospital, United Kingdom*
[2] *Sheffield Diagnostic Genetics Service, Sheffield Children's Hospital, United Kingdom*
[3] *The University of Manchester, United Kingdom*

Introduction

- Exon-level Copy Number Variants (CNVs) are clinically important but gold standard methods of detection (MLPA and ddPCR) are not scalable
- Clinical laboratories routinely use targeted-capture Next Generation Sequencing (NGS) for single nucleotide variant detection
- CNVs can be detected from NGS data, it is unknown which CNV caller is best at the exon-level
- Using NGS data for CNV calling would:
  - Decrease costs and increase capacity for genes that are routinely tested for CNVs
  - Improve diagnostic yield in gene panels that are not routinely tested for CNVs
- Each laboratory has few positive CNVs
- We created **CNV-patissier to automate the CNV caller bake off** across multiple NHS sites

Methods

Only data with confirmed CNV-status used (by MLPA)

146 positive CNVs, 191 negative samples

- **The Institute of Cancer Research (ICR) 96-Exon Validation series:**
  Enzymatic shearing & TruSight Cancer Panel
- **Great Ormond Street Hospital (GOSH):**
  Physical shearing & Agilent SureSelect capture
- **Sheffield Children's Hospital (SCH):**
  Physical shearing & Agilent SureSelect capture
- **Three other NHS sites** awaiting local approval

Automation of analysis, using CNV-patissier

- Reproducible: Docker images for the same configuration, versions and dependencies
- Default parameters for each CNV caller as suggested by documentation for targeted-capture
- Only relevant data saved to SQLite database for simple pooling at lead site

GitHub python docker

Results

- High sensitivity callers:
  DECoN, ExomeDepth, GATK, panelcn.MOPS and SavvyCNV (Figure 1)
- DECoN detected 2 true positive CNVs that ExomeDepth was not able to
- All true positive calls were in the intersection of SavvyCNV and panelcn.MOPS (Figure 2)
  - Sensitivity: 0.97 (95% CI: 0.93 – 0.99)
  - Specificity: 0.91 (95% CI: 0.85 – 0.94)

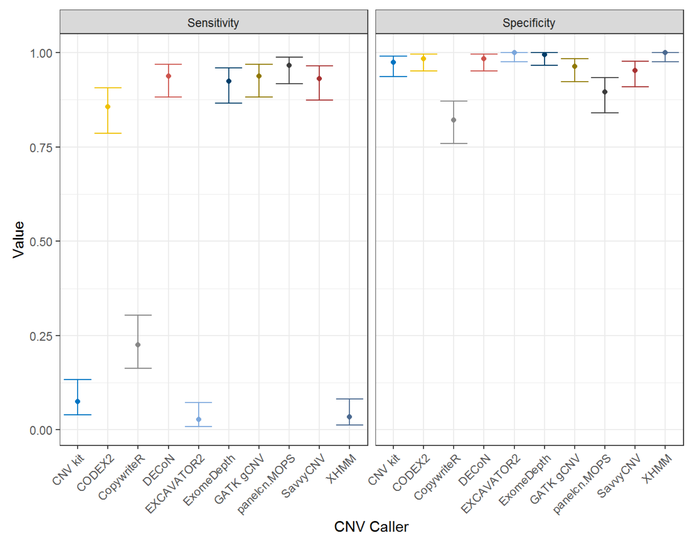


***Figure 1**: CNV caller sensitivity and specificity (with 95% confidence intervals)*

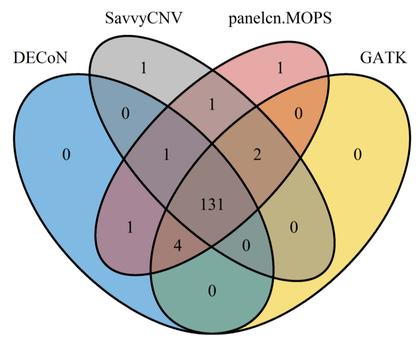


***Figure 2**: Venn Diagram of true positive calls from top 4 most-sensitive CNV callers. Four positive CNVs of the 146 were not detected*

Conclusions

- Final analysis with all NHS sites should reach 300 positive CNVs for a higher certainty comparison
- An intersection of CNV callers should be used for exon-level detection using targeted capture NGS
- As patient sequencing data was never shared, each NHS site could easily agree to collaborate

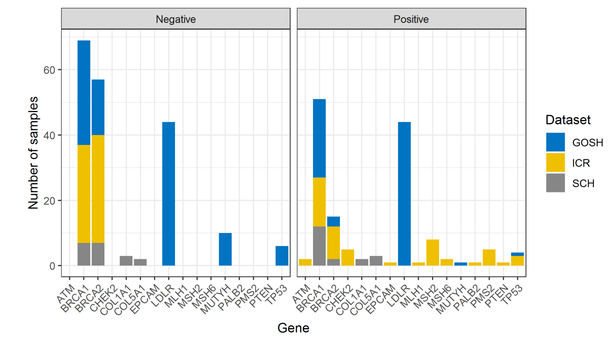
Extra information:

Targeted-capture callers considered:

All freely available, released 2012 - 2018

CNV caller	Read depth normalisation or modelling	Panel of normals?	Calling method	On-target or off-target reads
CANOES	Negative binomial distribution model with GC content correction	Yes	Hidden Markov model and maximum likelihood Viterbi algorithm	On-target
Canvas	Weighted average, Outlier removal and GC content correction	Yes	Unbalanced Haar wavelet transform then deviation model	On-target
CLAMMS	Finite mixture model, GC content and low mappability	Yes	Hidden Markov model	On-target
CNV-CH	GC content correction	No	Convex hull algorithm	On-target
CNVkit	Rolling median depth with GC content & repetitive region normalisation	Yes	Segmentation and copy number determination	Both
CNVPanelizer	Trimmed mean of M	No	Bootstrapped subsampling	On-target
CODEX2	Latent factor model	Yes	Recursive Poisson-likelihood segmentation algorithm	On-target
CoNIFER	Singular vector decomposition	Yes	z-score	On-target
CONTRA	Log ratio, GC content correction	Yes	Binning and linear interpolation	On-target
CopywriteR	Loess-based correction for mappability and GC content	Matched normal	Circular binary segmentation	Off-target
DECoN	Binomial model, GC content correction	Closest correlation from all	Hidden Markov model and maximum likelihood Viterbi algorithm	On-target
EXCAVATOR2	Mean read count, median normalisation for GC content, mappability and region size	Yes	Shifting level model segmentation and Gaussian model	Both
ExomeDepth	Binomial model, GC content correction	Yes, closest correlation	Hidden Markov model and maximum likelihood Viterbi algorithm	On-target
FishingCNV	Singular vector decomposition	Yes	Circular binary segmentation	On-target
GATK gCNV	GC content correction, Bayesian read depth model	Yes	Hierarchical graphical model	On-target
panelcn.MOPS	Upper quartile normalisation	Yes, closest correlation	Mixture of Poissons	On-target
SavvyCNV (unpublished)	Singular vector decomposition	No	Hidden Markov model with crossover in integer copy number	Both
VisCap	Fraction of coverage, median-normalised	No	Log2 ratio thresholds and outlier in batch	On-target
WISExome	Normalised to total counts over whole genome	Within-sample reference	z-score	On-target
XHMM	Singular vector decomposition, filter highly variable regions	No	z-score and merging	On-target

Samples with confirmed CNV-status used:



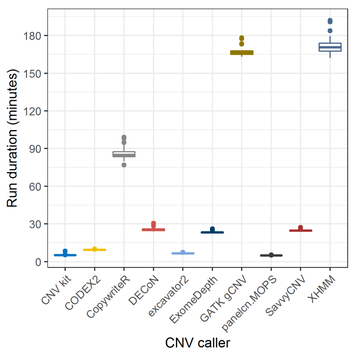
Docker Hub images and tool versions:

CNV Caller	Base docker image and tag	Tool version	Supporting tools (version)	Docker image used and tag
Canvas	mono:5.14	1.11.0	Python (3.7.2)	stefpiatek/canvas: 1.11.0
CNV kit	ubuntu:18.04	0.9.6	Python (2.7.15rc) R (3.4.4)	etal/cnvkit:0.9.6
CODEX2	r-base:3.5.1	26e796c	R (3.5.1) CODEX (bioconductor 3.8) BSgenome (bioconductor 3.8)	stefpiatek/codex2: 26e796c
CopywriteR	r-base:3.5.1	2.14.1		stefpiatek/copywriter: 2.14.1
DECoN	r-base:3.1.2	1.0.2	R (3.1.2)	stefpiatek/decon:1.0.2
EXCAVATOR2	r-base:3.5.1	1.1.2	Bedtools (2.26.0) Samtools (1.9)	stefpiatek/excavator2: 1.1.2
ExomeDepth	r-base:3.5.1	1.1.10	R (3.5.1)	stefpiatek/ exomedepth:1.1.10
GATK gCNV	broadinstitute/gatk: gatkbase-2.0.3	4.1.0.0	openJDK (1.8.0_191) Python (3.6.2)	broadinstitute/gatk: 4.1.0.0
panelcn.MOPS	r-base:3.5.1	1.4.0	R (3.5.1)	stefpiatek/ panelcn_mops:1.4.0
SavvyCNV	java:8-jdk	f996a83	openJDK (1.8.0_111) GATK 3.8.1.0	stefpiatek/savvycnv: f996a83
WISExome	continuumio/ miniconda:4.5.4	-	Python (2.7.15 Anaconda)	stefpiatek/wiseexome:latest
XHMM	broadinstitute/ gatk3:3.8-1	1.0	GATK 3.8.1.0 plinkseq (0.10)	stefpiatek/xhmm:1.0

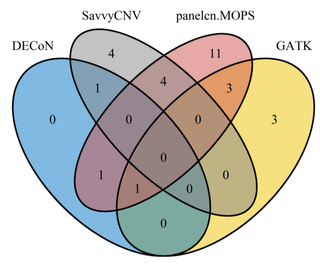
Callers removed from comparison:

- Canvas: no CNVs from targeted-capture
- WISExome: not functional

Run duration using ICR data



False positives



Take a picture for link to project (and poster) or
github.com/stefpiatek/cnv-patissier

NHS London North NHS Genomic Laboratory Hub