Preview of upcoming GATK methods

Germline Copy Number Variations and Structural Variations

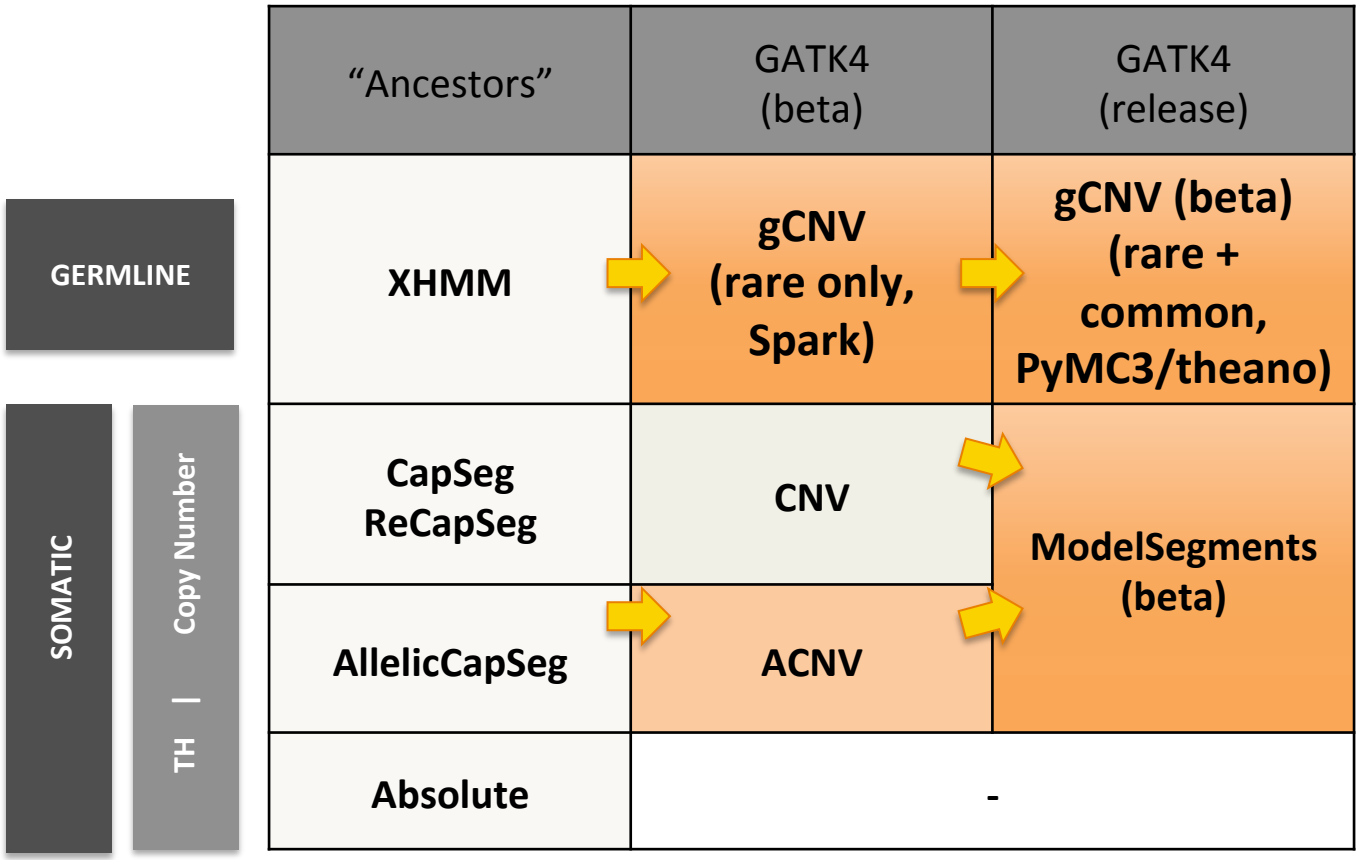
BROAD INSTITUTE

https://software.broadinstitute.org/gatk/

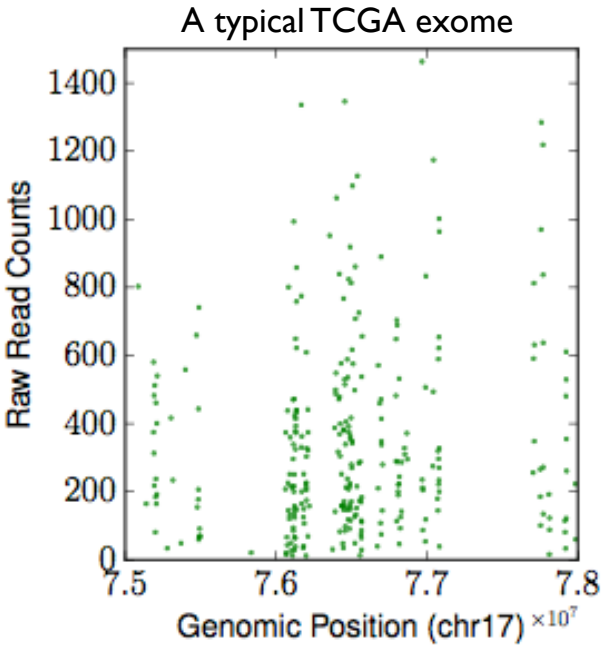
gatk

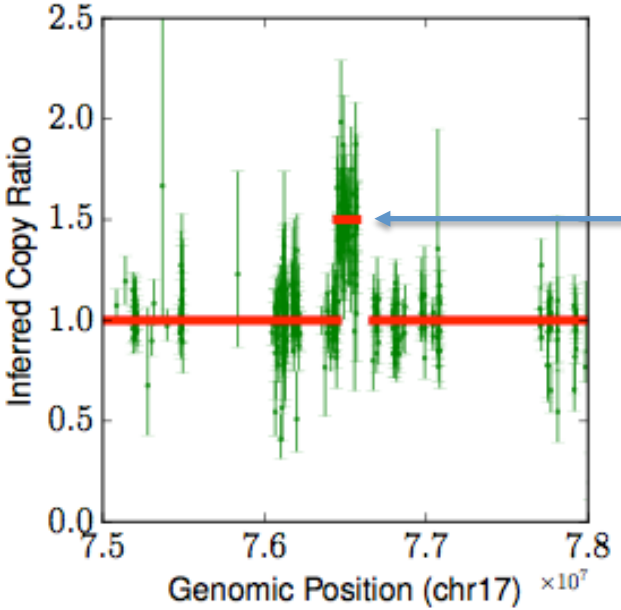
GATK-gCNV

History of GATK CNV tool development



Germline CNVs are short and difficult to call in WES



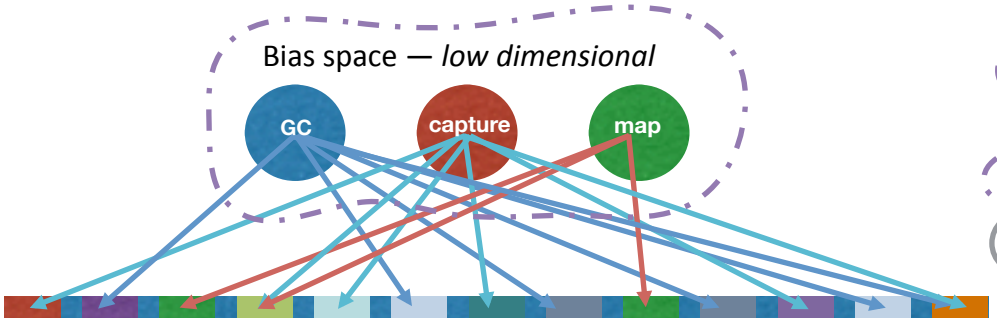


Germline CNV events typically short, but appear at regularly spaced copy-ratio states

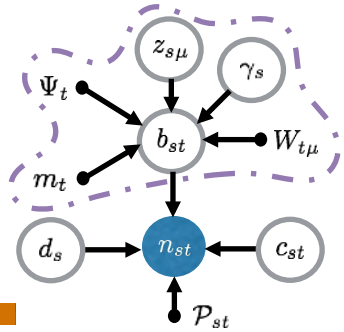
GATK GermlineCNVCaller (gCNV)

Denoising + segmentation + calling are all performed simultaneously!

gCNV Spark introduced a Bayesian coverage model



Target space — *high dimensional*



coverage model sub-unit

Generative model for target coverage

$$n_{st} \sim \text{Poisson}(\lambda_{st})$$

$$\lambda_{st} = d_s \times \mathcal{P}_{st} \times c_{st} \times e^{b_{st}}$$

- read depth
- germline ploidy
- copy ratio
- multiplicative bias

linear-Gaussian dimensional reduction

$$b_{st} \sim \mathcal{N}\left(\sum_{\mu=1}^{D} W_{t\mu} z_{s\mu} + m_t, \Psi_t + \gamma_s\right)$$

Model parameters

- Ψ_t target-specific variance
- $W_{t\mu}$ log-bias covariates
- m_t mean target-specific log-bias

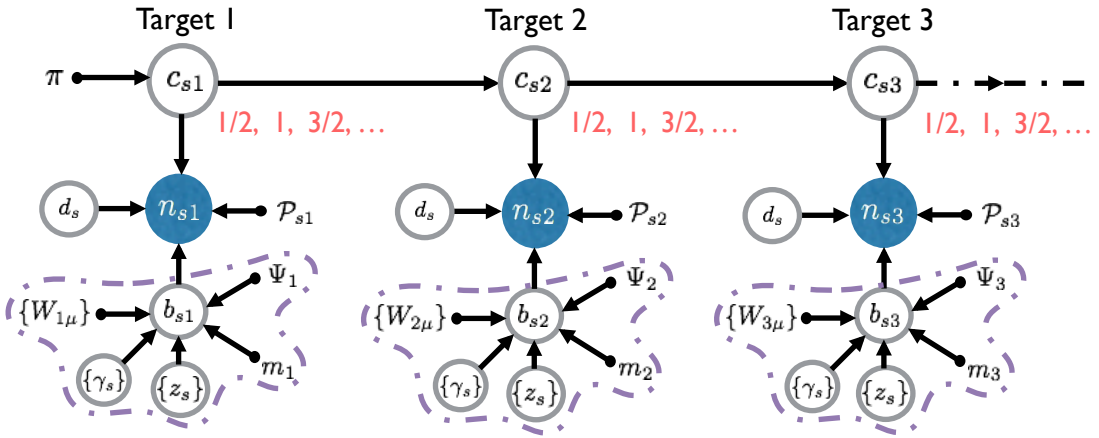
Sample-specific latent variables

- d_s average depth of coverage
- c_{st} copy ratio
- b_{st} log multiplicative bias
- γ_s sample-specific variance

Coverage model + HMM = denoise + segment + call

Given a panel of normal samples, gCNV simultaneously learns a model for denoising and calls germline CNVs.

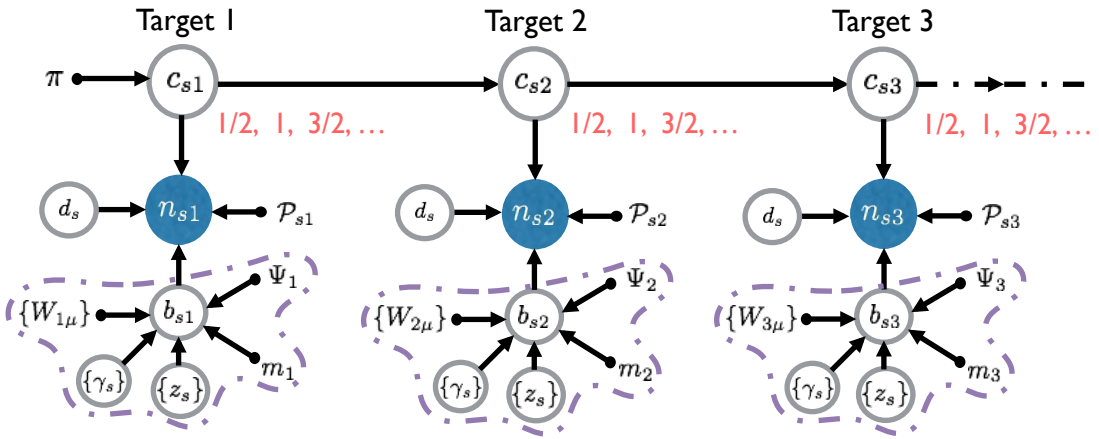
The learned model can then be used to denoise and call additional cases.



Coverage model + HMM = denoise + segment + call

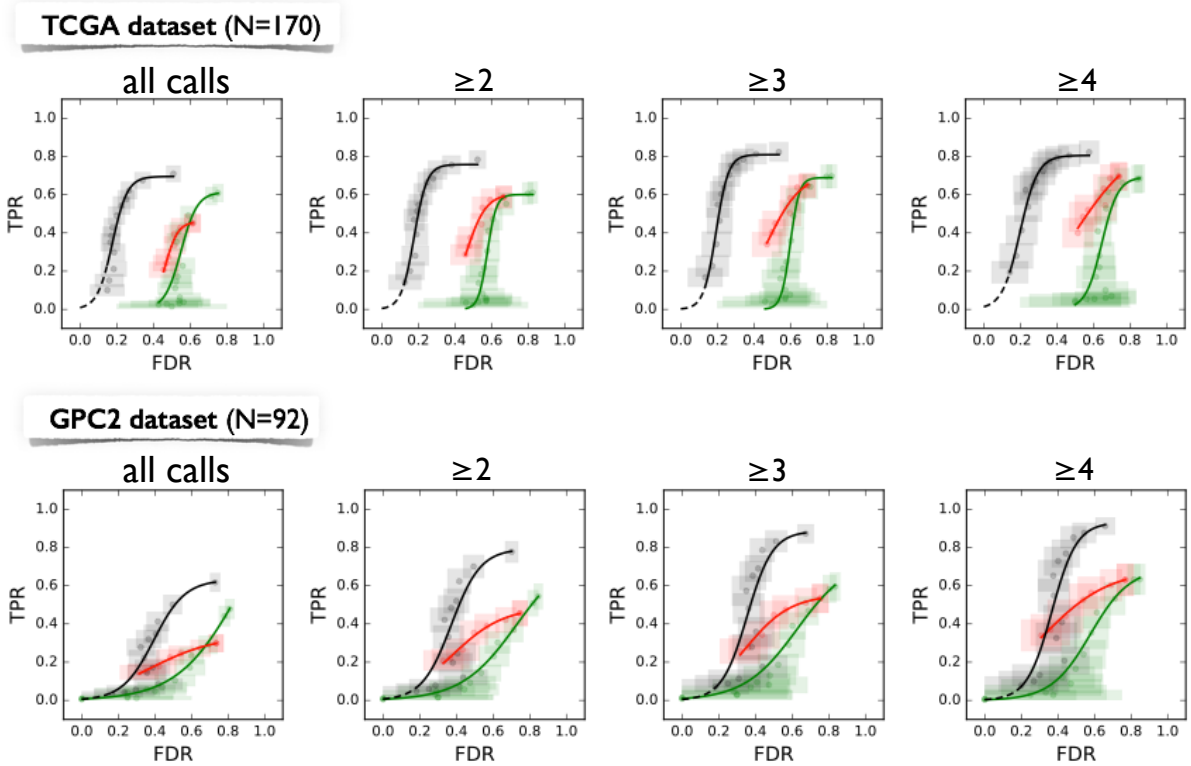
Given a panel of normal samples, gCNV simultaneously learns a model for denoising and calls germline CNVs.

The learned model can then be used to denoise and call additional cases.



Caveat: PCA denoising cannot handle ~~mixed-sex PoNs~~ or common CNVs in the PoN!

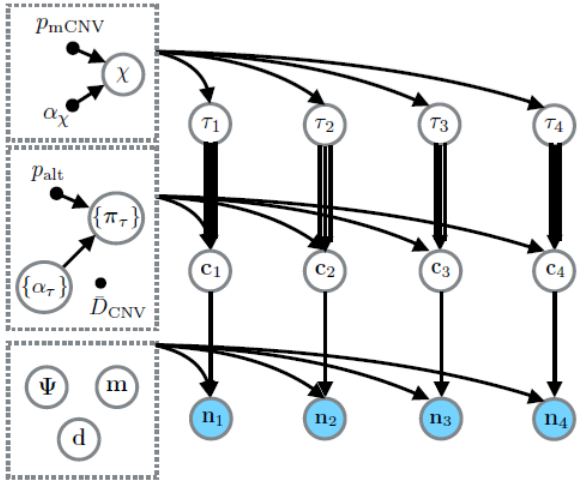
gCNV Spark greatly outperforms XHMM and CODEX

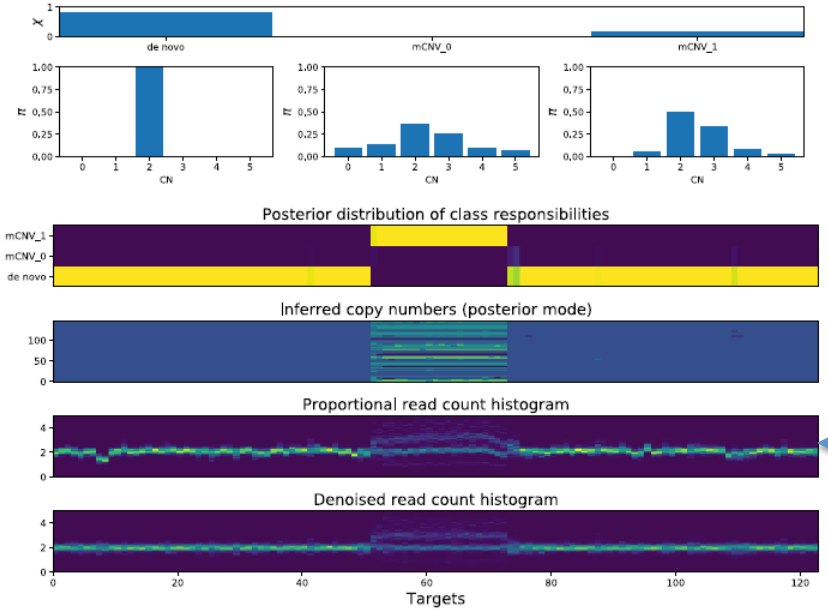


Compared WES calls from all tools against WGS Genome STRiP ground-truth calls

Additional runs by Monkol Lek on MYOSEQ cohort (N ~ 400) identified all previously known deletions and ~10 novel deletions (including on X) validated using MLPA

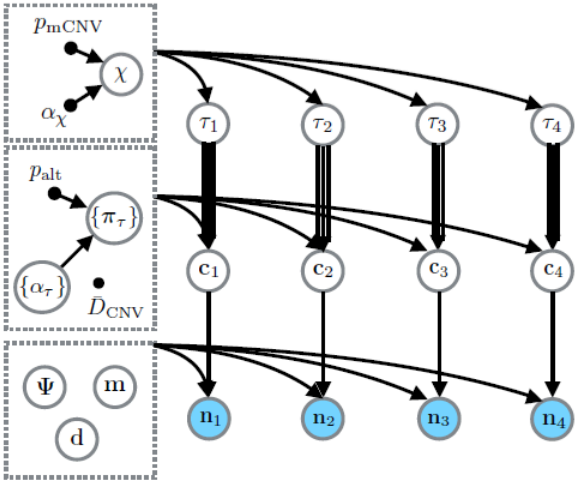
gCNV PyMC3/theano allows inference of common CNVs

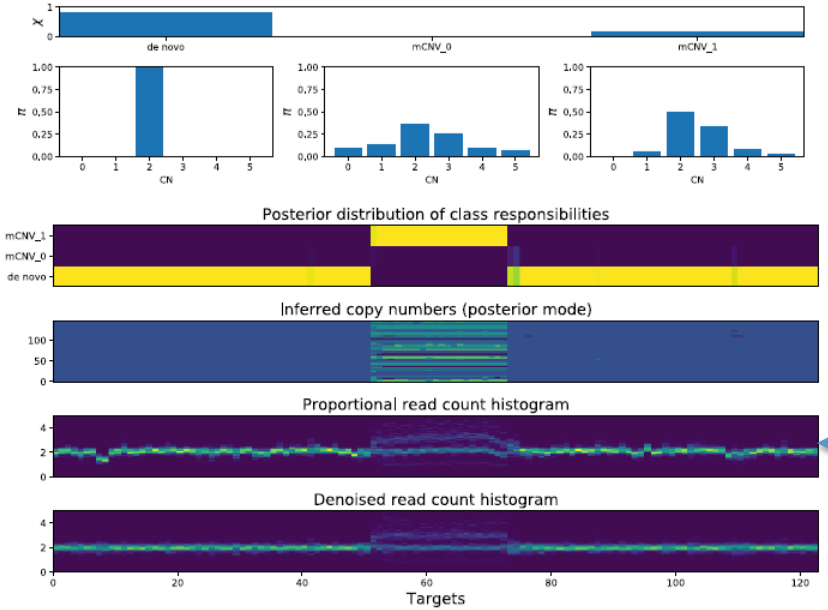




Possible multiallelic regions are identified in a hierarchical model, then a more flexible copy-number prior is used in these regions.

gCNV PyMC3/theano allows inference of common CNVs

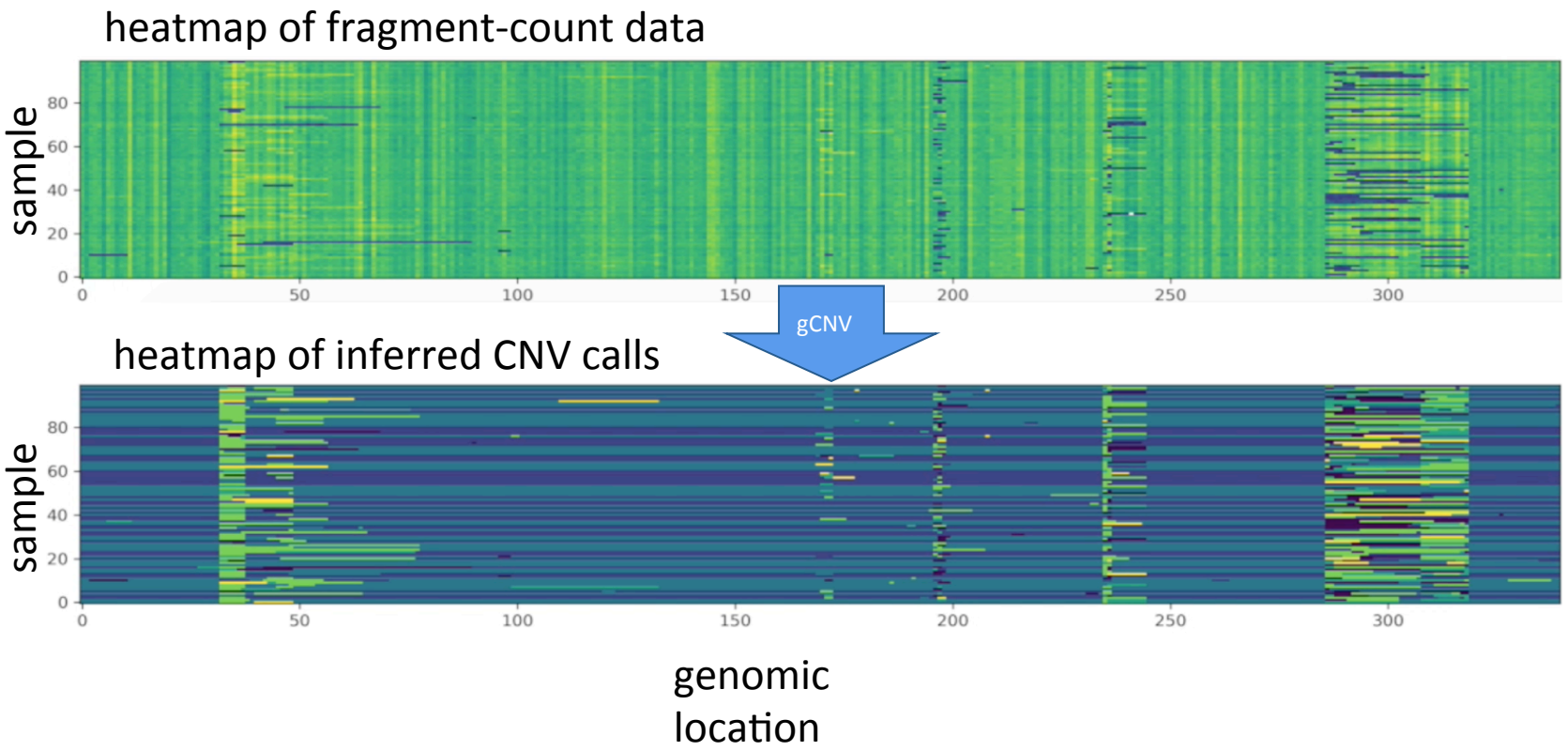




Possible multiallelic regions are identified in a hierarchical model, then a more flexible copy-number prior is used in these regions.

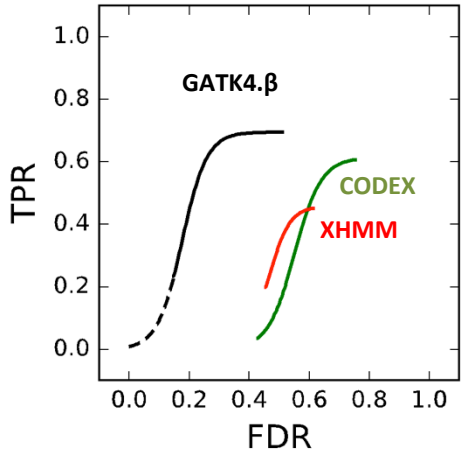
Caveat: PCA denoising cannot handle ~~mixed-sex PoNs~~ or ~~common CNVs in the PoN~~!

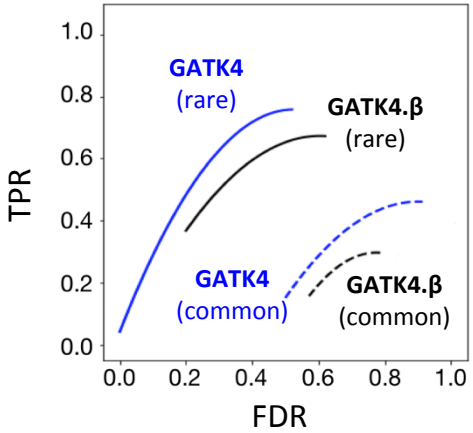
gCNV PyMC3/theano: inference in action



gCNV PyMC3/theano greatly outperforms gCNV Spark

TCGA (N=170) WES vs. WGS genomeSTRiP ground truth



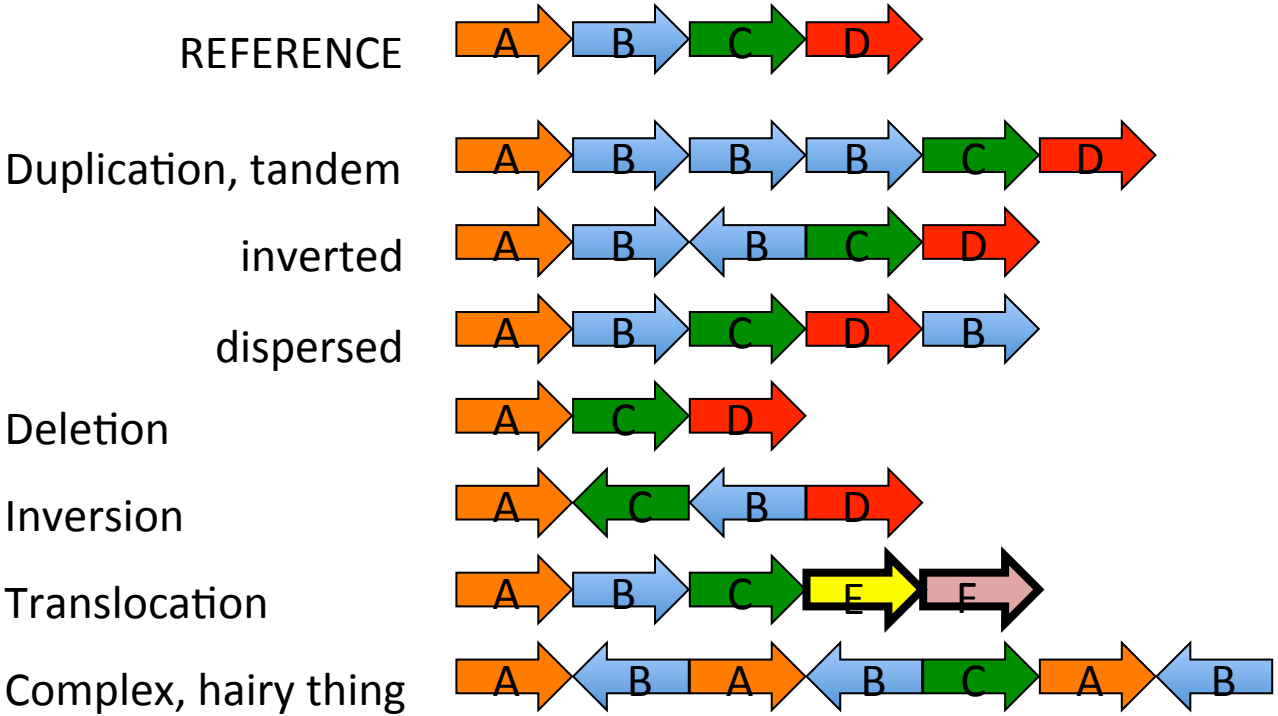


We are currently performing additional validations,
with a goal of running **automatic validations on public data** with every update!

GATK-SV

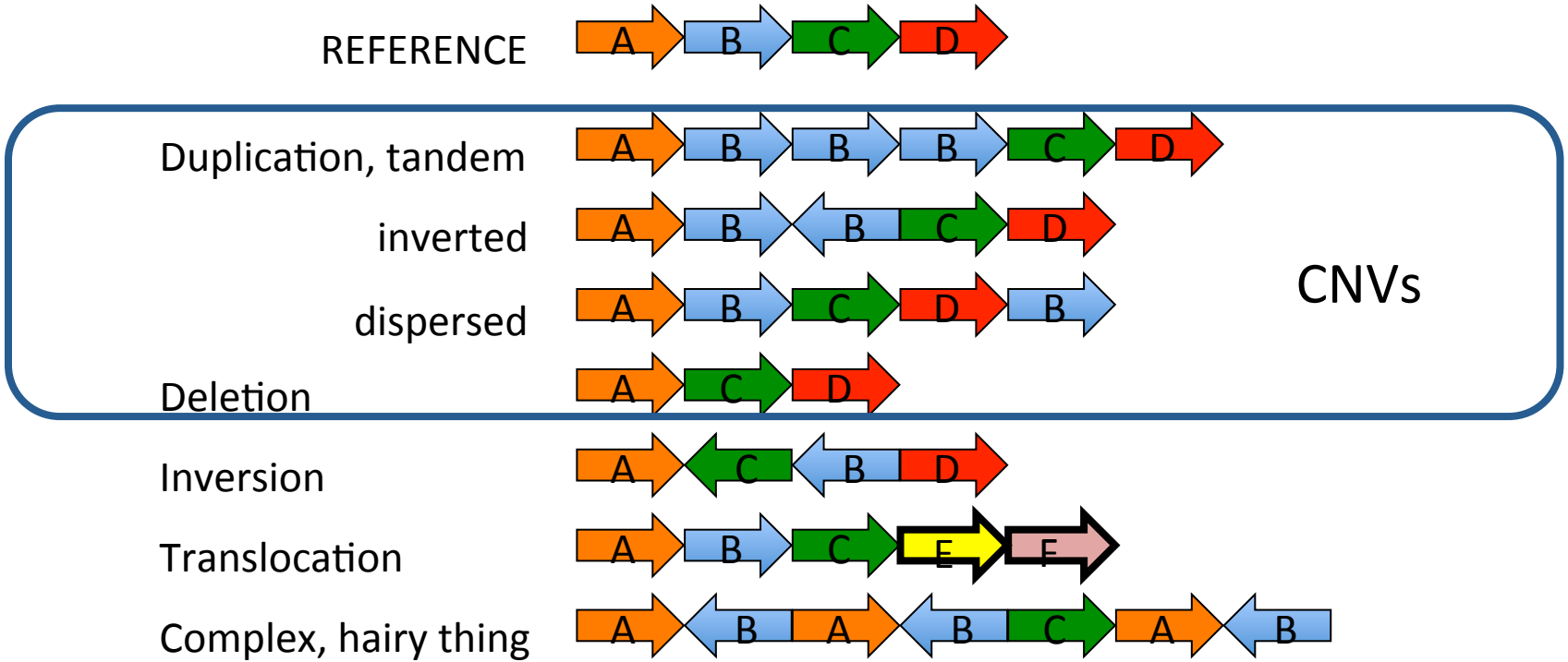
Structural variation

Definition: Any variant that affects 50bp or more of sequence



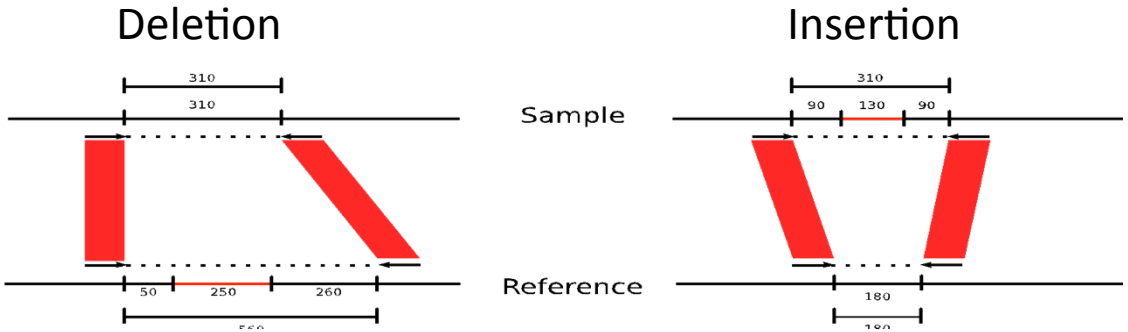
Structural variation

Definition: Any variant that affects 50bp or more of sequence

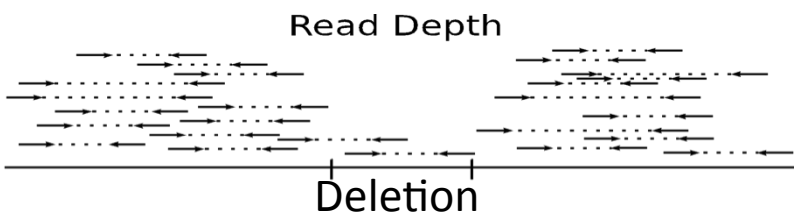


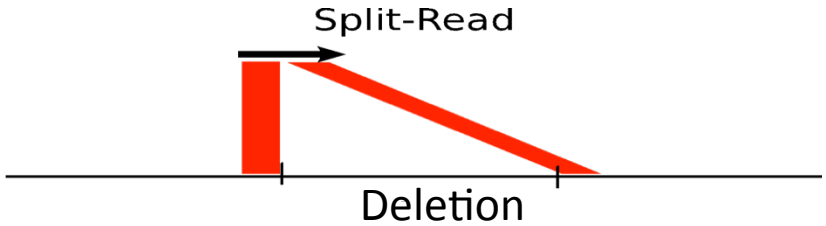
SVs signals in Short Reads

- Read Pair (RP): deviations in insert size and pair orientation of pairs



- Read Depth (RD): Fluctuations in coverage
- Split Read (SR): Look for reads clipped by the aligner at the breakpoint

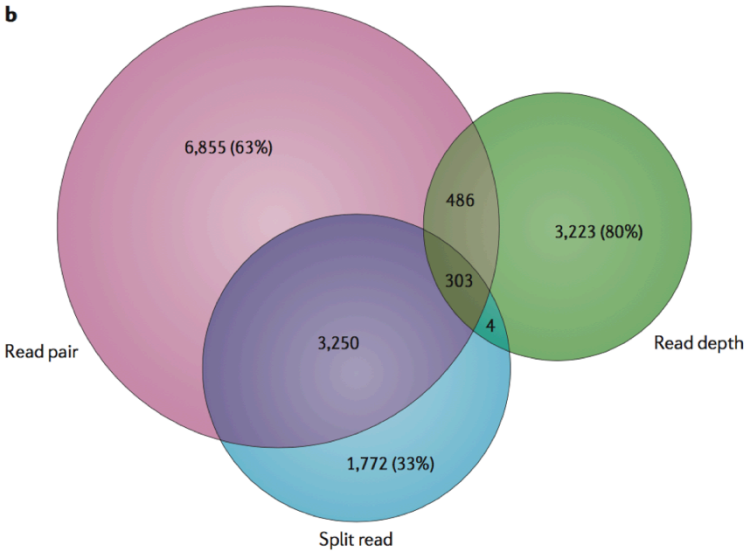




SV detection from short read data is hard

- Need to infer presence from short read mappings, not fully present in read sequences themselves
- SVs tend to occur in repetitive regions of the genome
- Successful pipelines require running multiple algorithms and determining consensus calls (lots of work)

Limited concordance between existing algorithmic strategies:

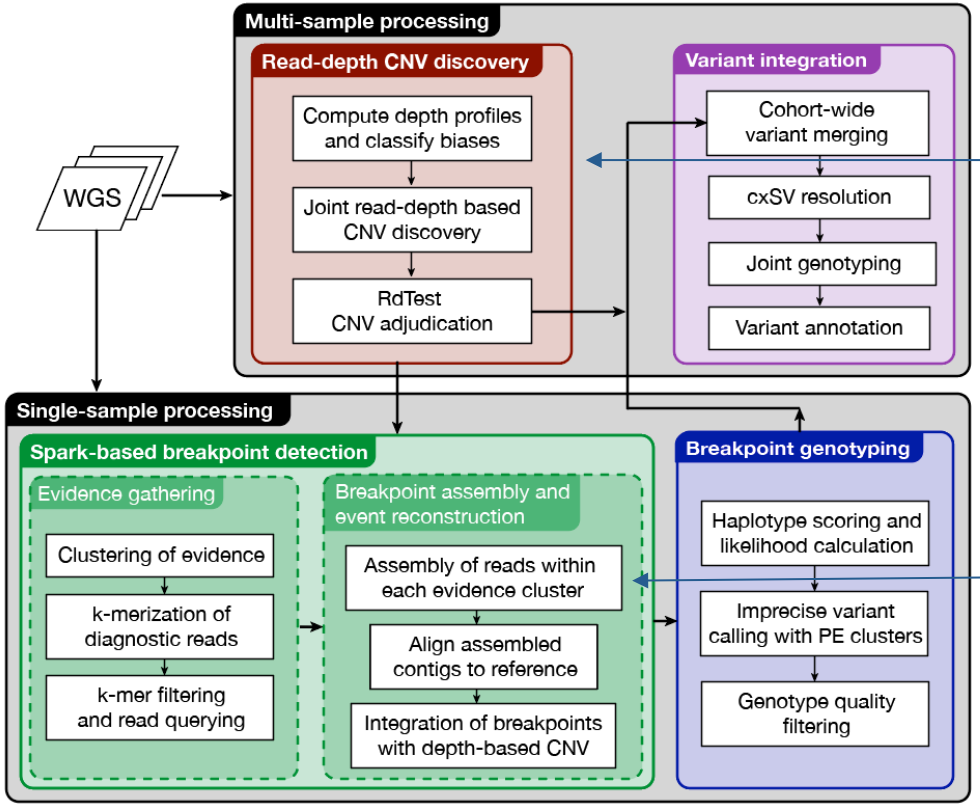


Alkan et al., *Nature Rev. Genetics*, 2011

GATK-SV (coming soon)

- Inspired by collaborations by diverse groups within Broad institute
- Based on local *de novo* assembly of candidate breakpoints
- Integrate read-depth, split-read, read-pair, and other mapping signals
- Work within a unified pipeline that integrates SV signals with CNV tools
- Build within GATK4 framework for robustness, supportability, interplay with other GATK pipelines
- Use Spark distributed computing framework to massively parallelize workflows and enable new approaches

The GATK-SV pipeline (WIP)



GATK CNV tools

SV breakpoint detection and assembly tools

Unique features of GATK-SV

- ***Spark-native:*** Seamlessly parallelizes to large Spark clusters, enables fast turnaround time
- ***Comprehensive assembly strategy:*** Overcomes reference aligner bias by searching the entire BAM file for reads that share k-mers with those in local assembly region; enables assembly of inserted sequence at breakpoints even if reads not mapped locally
- ***Designed to handle complex events:*** SV breakpoints are complex, we attempt to describe all rearrangements involved in an event simultaneously
- ***Integration of signals:*** designed to work with all SV signals in the data and produce a comprehensive call set (no need to run 12 tools and make consensus call set)
- ***Integrated, robust pipeline*****:** Integration with GATK CNV tools, support for standardized file formats, designed to work in GATK ecosystem.

Questions?