Gerstein Lab slides for 2018 HGSVC meeting

M Gerstein
F Navarro, X Kong
S Kumar, S Li

other referenced contributors:
A Abyzov, A Harmanci, AE Urban, C Lee, DR Kim, DW Radke, E Khurana,
H Lam, J Bedford, J Du, J Korbel, M Snyder, X Mu, Y Fu, Y Zhang, Z Liu, Z Zhang

Overview

1. Past lab work on SVs as context
   a. Breakpoints
   b. Functional characterization

2. Relevance to future large-scale short-read sequencing

3. Current work on
   rapid, "generalized" breakpoint identifier

4. Current work on SV functional impact score

Previous work on finding breakpoints

- BreakSeq

(Lam et al; '09; Nature Biotech; Abyzov et al; '15; Nature Comm)

-Fast SV detection by alignment against a breakpoint library

- AGE (Abyzov & Gerstein; '11; Bioinfo)

-Optimal alignment over breakpoints

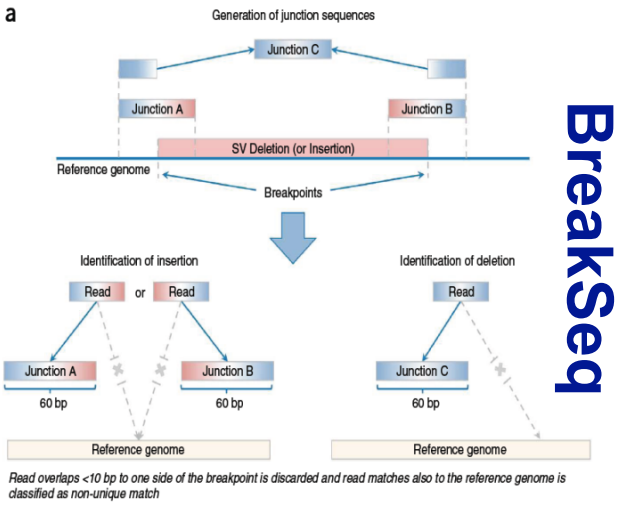
- SRiC (Zhang et al; '11; BMC Genomics)

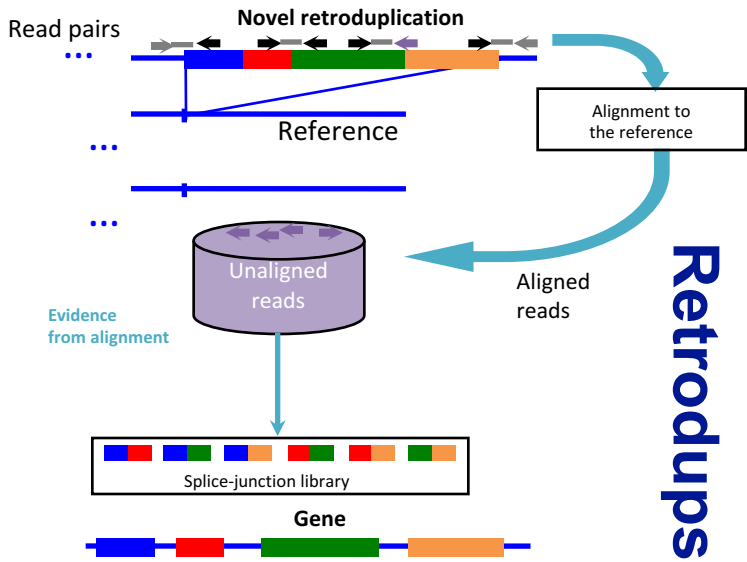
-Split read alignment, parameterized by simulation

- Retrodups

(Abyzov at al;'13; Genome Research; Zang at al; '17 Plos CompBio)

-Finding retroduplications & processed pseudogenes via alignment to junction lib.

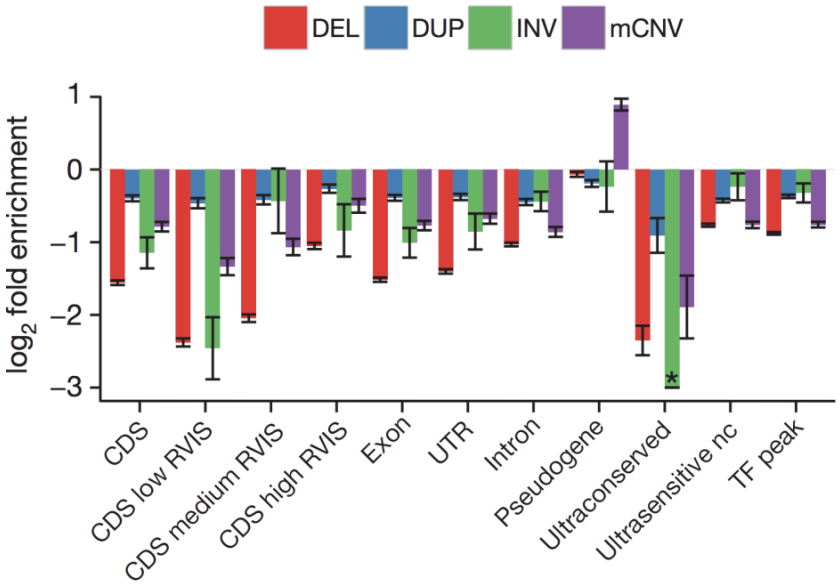


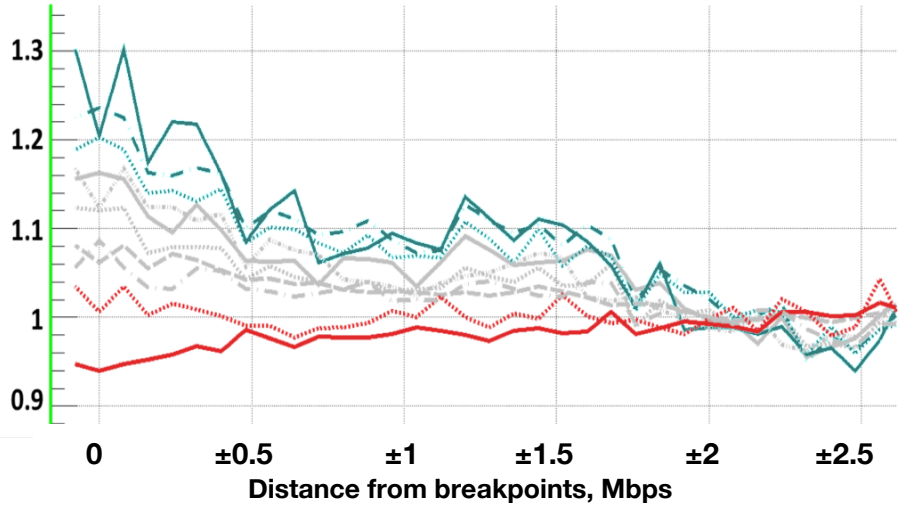


Previous work on functional characterization of SVs

- Overlaps enrichment of SVs [Sudmant et al. Nature '15. chaisson et al. submitted]
  - against functional annotations (CDS, TFBS, etc.)
  - using an appropriate null
- Mechanism classification of SV breakpoints
  -NAHR, NHR, etc. [Lam et al. Nat. Biotech. '10; Abyzov et al. Nat Commum '15]
- Patterns of epigenetic marks at SV breakpoints

[Khurana et al. Science '13]





Relevance to large-scale short read sequencing

- NHGRI GSP (CCDG+CMG) & TopMed will sequence 1000s of genomes w. short reads

- Useful to run these against a developed breakpoint library to genotype more SVs

- Useful to characterize the SVs in terms of functional impact to find those most important to disease
  -particularly for CMG

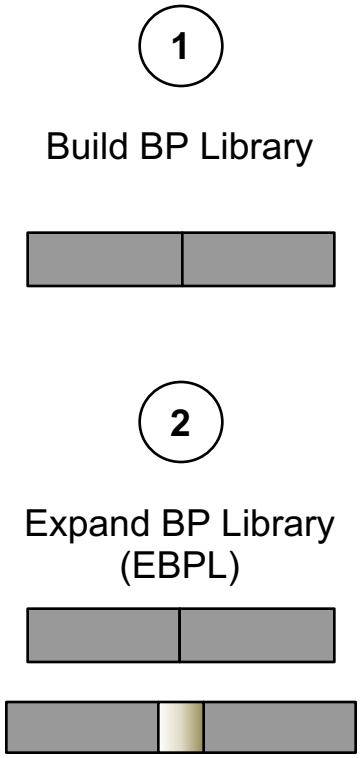
AIM A - Rapid Breakpoint Identification

Break point Library / Extended Break Point Library (1, 2)

- Many current and future resources are using short sequencing data
  - We are delopping rapid SV identification pipelines to characterize these samples
    - Genotype previously identified SVs
    - Detect new SVs

- Create an **integrated library** of Break Points in the Human Genome
  - Sources of Break Points
    - BreakSeq
    - Population-wise calls (i.e. 1000G Phase1-3)
    - Break points based on Long Reads
      - **HGSVC** (3 Trios)
        - PacBio
        - Nanopore
      - GiaB
        - NA12878 + Family
    - Mobile Element Polymorphisms (MEI)
    - Processed Pseudogenes

- Machine Learning to detect Breakpoint (abnormal reads)



- **Extend & Generalize** known Breakpoints to allow uncataloged gaps near **known BPs and other seq.**

AIM A - Rapid Breakpoint Identification

Indexing BPs and EBPs and "Fish" reads from FASTq pool (3, 4)

- Exhaustively index the human genome and BP/EBPs sequences
  - Compare and exclude BP/EBPs indexes already represented in the Ref Genome.

    3

    Index EBPL

  - Hash from unique BP/EBPs sequences
    - Alternatively, we could create k-mers from sequences
    - Multiple and Incremental BP library indexes releases

- Pool FastQ files and quickly run through the BP hash
  - Selecting only reads that are evidence for SVs

    4

    Efficiently Map reads to iEBPL

AIM A - Rapid Breakpoint Identification

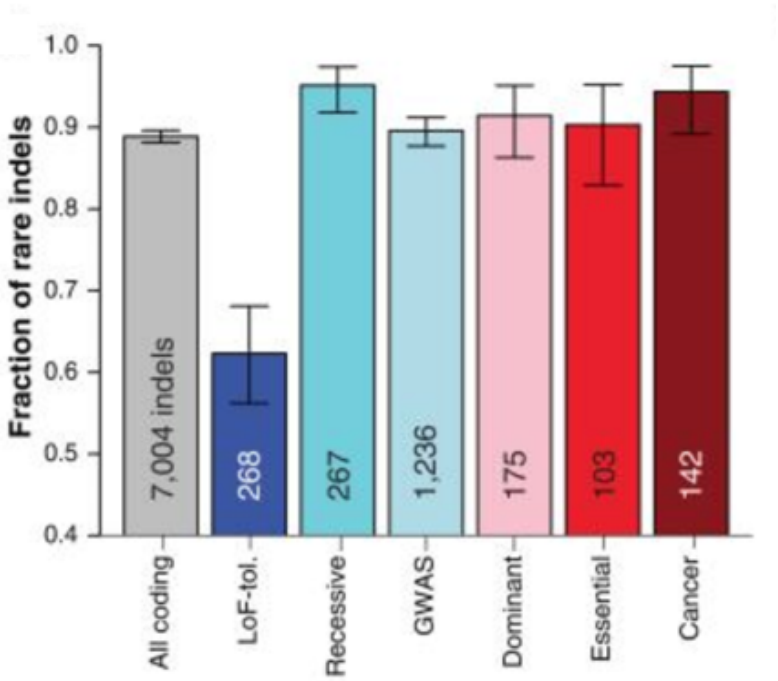
Software engineer features

- At scale / Fast
  - Able to quickly genotype/call SVs from thousands of genomes
- Cloud Aware (Docker, Spark, Hive)
- Biased (sensitive) calling in fragile(?) regions of the genome
  - Make up breakpoints
    - Homolog regions prone to recomb
    - TEs facilitating recomb

5

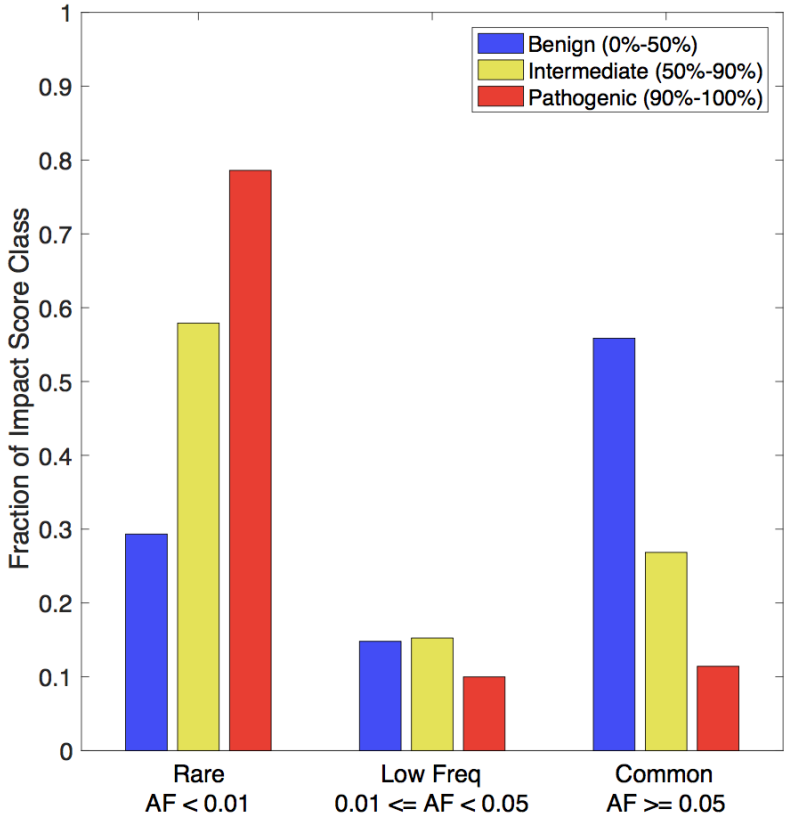
Call/Genotype SVs

Prioritizing SV by functional impact



[Khurana et al., *Science* ('13)]

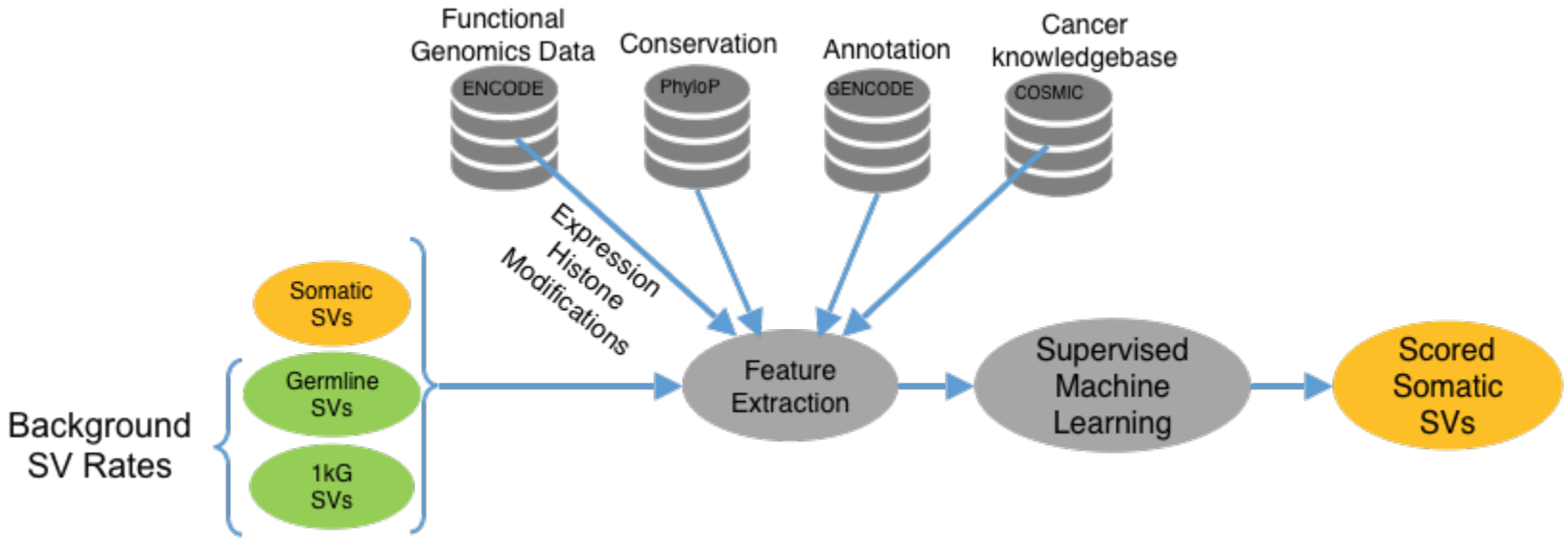
Fraction of Rare indels in coding-gene categories in 7 prostate cancer samples.



Ganel, et al.
https://www.biorxiv.org/content/early/2016/09/06/073833

Prioritization of SVs

- How do we prioritize SVs (somatic & germline)?
- SVs lengths vary widely: 50 base pairs – 100 megabase pairs
- Machine learning based scoring (Random Forest)

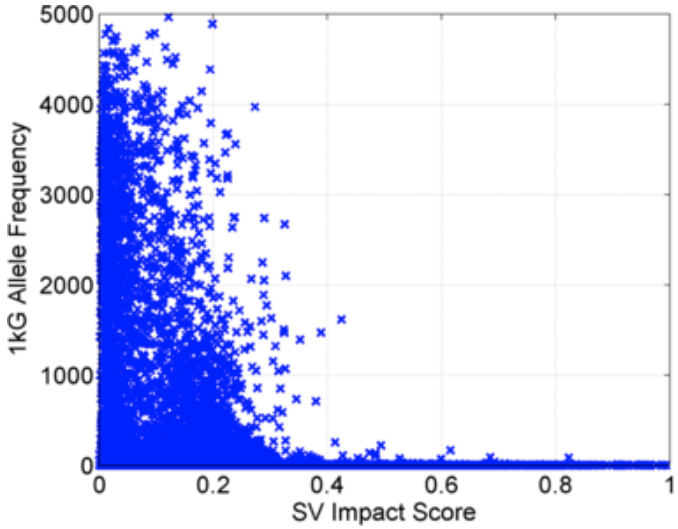


Kumar, et al, https://doi.org/10.1101/280446

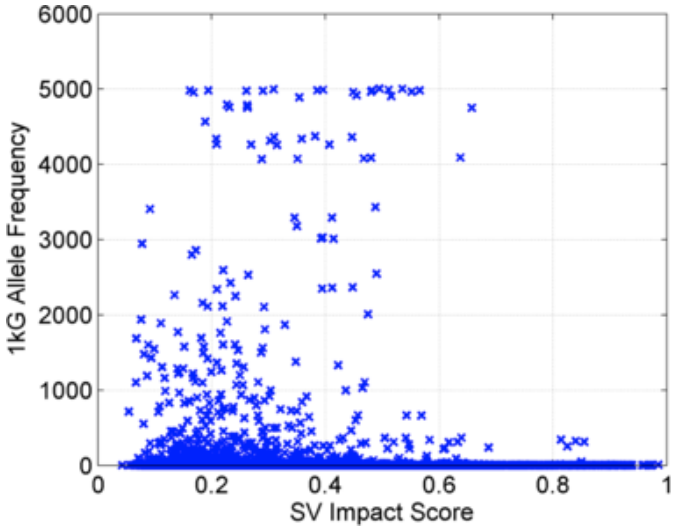
Prioritization of SVs

1KG phase3 data

Deletions



Duplication



1KG + Somatic

	Somatic	1kG
Somatic	566	113
1kG	243	458

	Somatic	**1kG**
Somatic	644	356
1kG	350	650

Overview

1. Past lab work on SVs as context
   a. Breakpoints
   b. Functional characterization

2. Relevance to future large-scale short-read sequencing

3. Current work on
   rapid, "generalized" breakpoint identifier

4. Current work on SV functional impact score

Gerstein Lab slides for
2018 HGSVC meeting

M Gerstein
F Navarro, X Kong
S Kumar, S Li

other referenced contributors:
A Abyzov, A Harmanci, AE Urban, C Lee,
DR Kim, DW Radke, E Khurana,
H Lam, J Bedford, J Du, J Korbel, M Snyder,
X Mu, Y Fu, Y Zhang, Z Liu, Z Zhang