Integrating SV calls across pipelines

Rameen Beroukhim & Peter Campbell, on behalf of PCAWG-6

Challenge

- 2000+ whole cancer genomes
- (At least) 3 different pipelines
  - 3 different sets of SV calls (outputting genomic breakpoints)
  - 3 different sets of copy number segmentations
- How do we generate a ‘consensus’ set of SV & CN calls for each genome?

1,000 Genomes experience

- Multiple different platforms
  - Read depth; aberrant read-pairs; split read mapping; *de novo* assembly etc

- Relied heavily on:
  - Orthogonal data sets (PacBio, trios, validation by PCR/capillary sequencing, copy number arrays)
  - Genotype concordance across individuals

Courtesy of Jan Korbel & Mark Gerstein

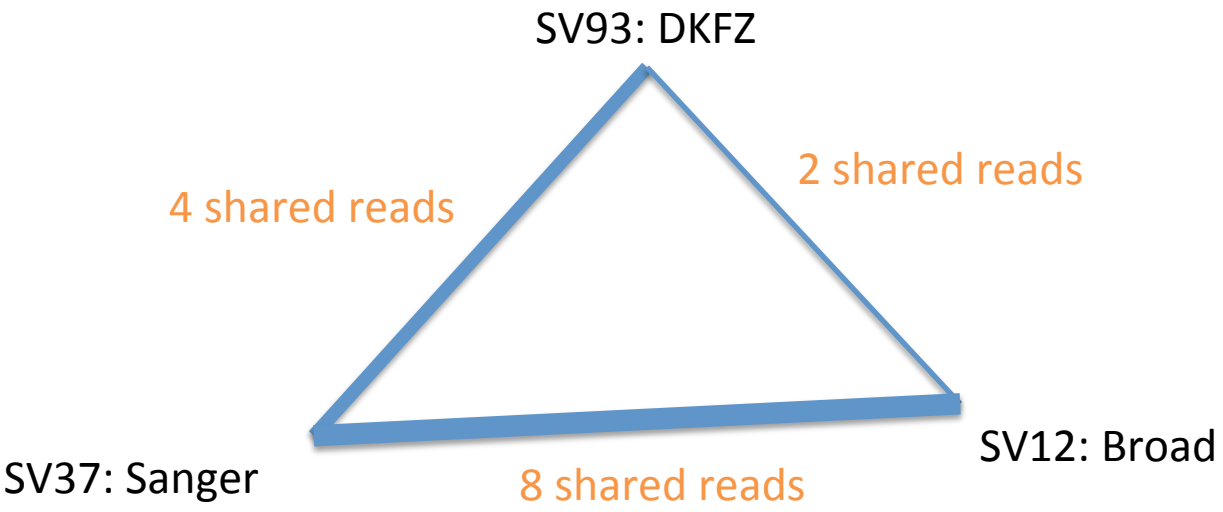
Scenarios for integrating SV calls

- Concordance
  - All pipelines call same variant at same location and with same microhomology / NTS
- Semi-concordance
  - Same basic variant but details of exact location or microhom/ NTS differ
- Differential calls
  - Variant called by some pipelines but not others
- Split calls
  - Some pipelines call 2 different variants, where others lump together as a single variant

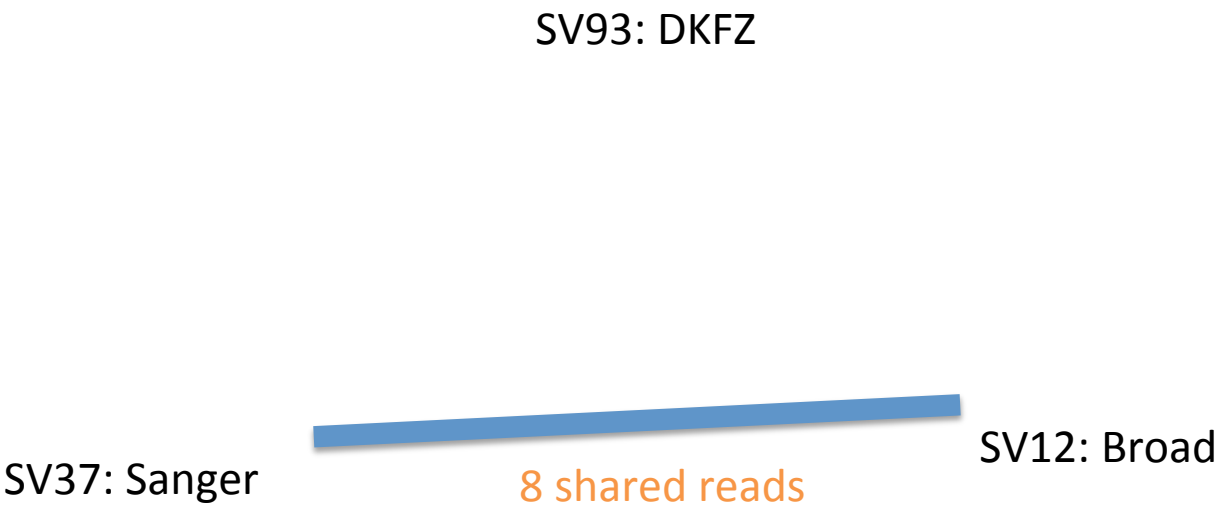
Proposal: Cliques based on supporting reads

- Each SV call in each pipeline is made on the basis of supporting reads
- Where the pipelines use the same supporting reads in SV calls, probably reporting the same variant
- One can quantify concordance between pipelines in clique membership

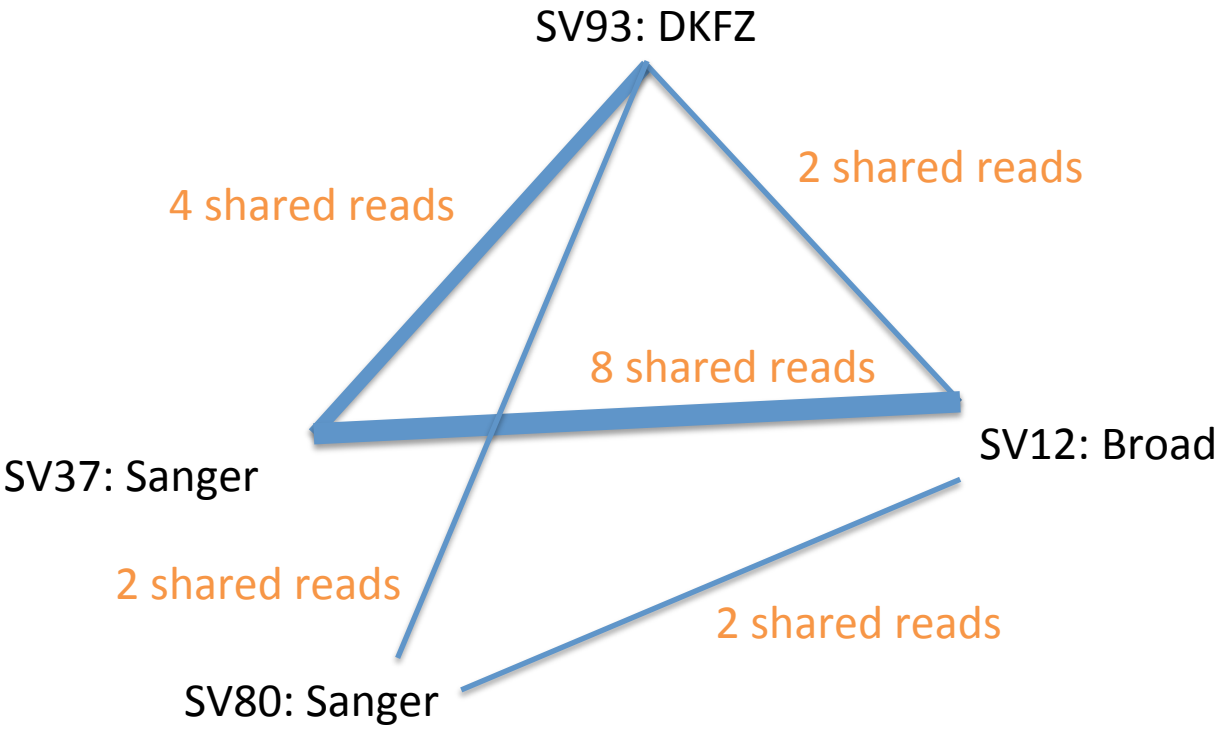
Cliques of SV calls in PCAWG – Concordance & semi-concordance



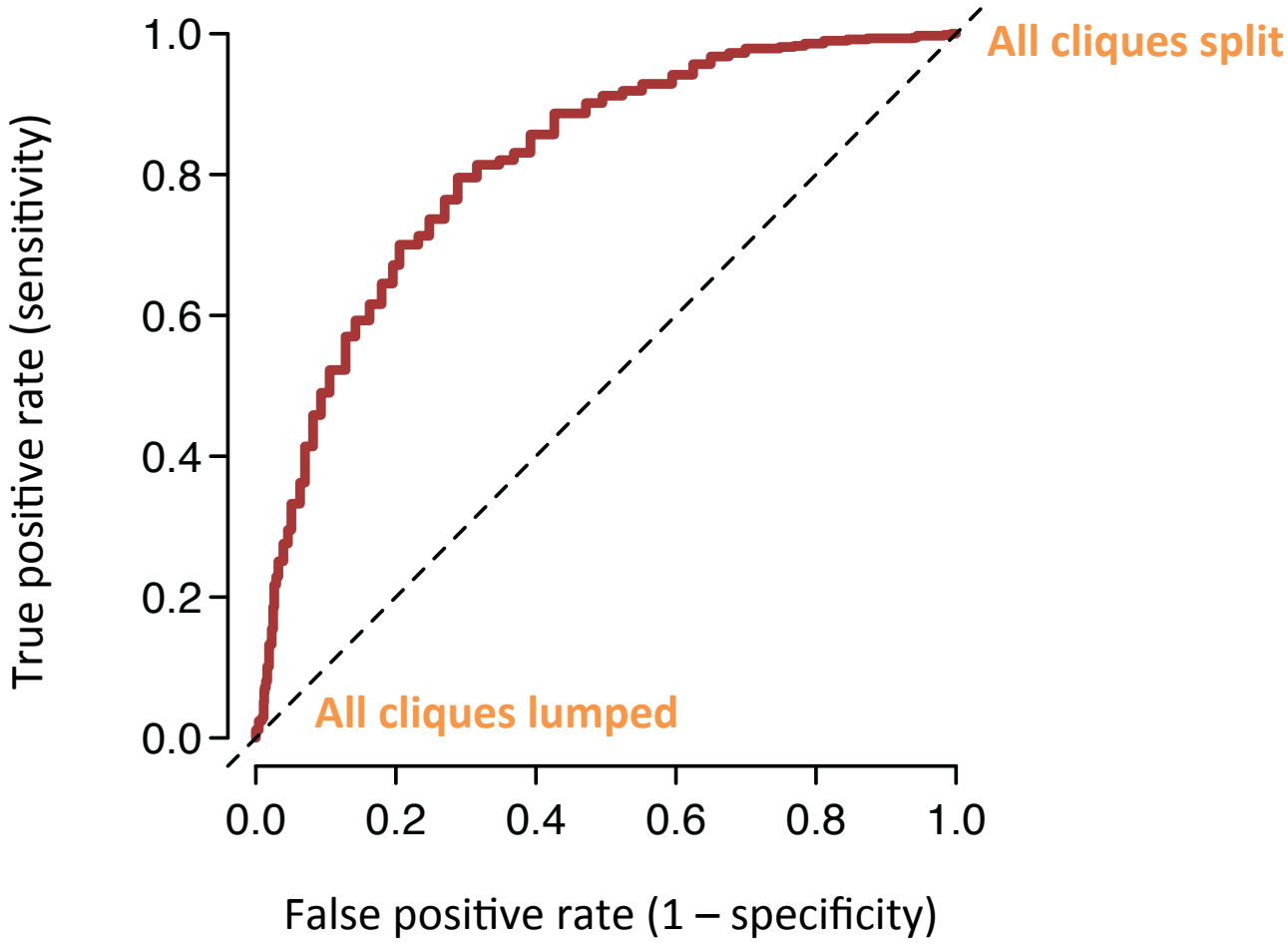
Cliques of SV calls in PCAWG – Differential calls



Cliques of SV calls in PCAWG – Split calls



Lumping and splitting cliques



Potential ‘gold standard’ data sets to train cut-offs for lumping vs splitting

- A set of visually inspected copy number changes from core 50 genomes
- 3 cell lines with 10% BAC libraries sequenced to finishing standard
  - Bignell et al, Genome Research 2007
- Samples with several lesions sequenced
  - Eg primary & met; multiple lesions of primary
- PCR validation of novel calls
- Ultra-high depth medulloblastoma genome