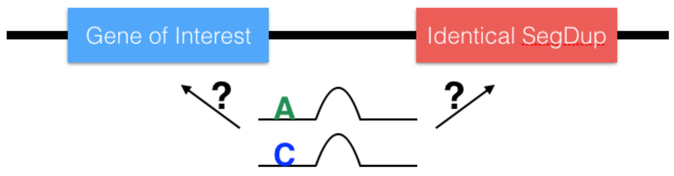
PseudoSlayer: Clinical-grade variant identification within segmental duplications using NGS data

Daniel J Kvitek, Leah Matzat, Sylvia Lin, Yang Li, Jason Funt, Erik Gafni
Invitae Corporation, San Francisco, CA

Disclosure statement: All authors are employees and stockholders of Invitae Corporation.

The Problem

- Segmental duplications (SegDups) complicate variant calling from short-read next-generation sequencing (NGS) data.
- High levels of sequence identity between SegDup regions prevents unique alignment of NGS reads.
- Therefore, clinical-grade variant calls in these regions are difficult or impossible with NGS data using standard read alignment and variant identification approaches.

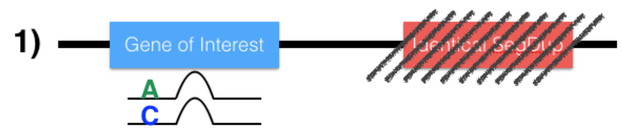


Aligned location of sequence reads is ambiguous owing to sequence identity.

Definitions:
- Segmental duplication (SegDup): Long stretches of nearly identical DNA sequence that occur two or more times in the genome that arose from duplication events that appear fixed in the population
- Paralog: SegDup of a gene region
- PSV (**P**aralogous **S**equence **V**ariant): A sequence difference between SegDups

The Solution

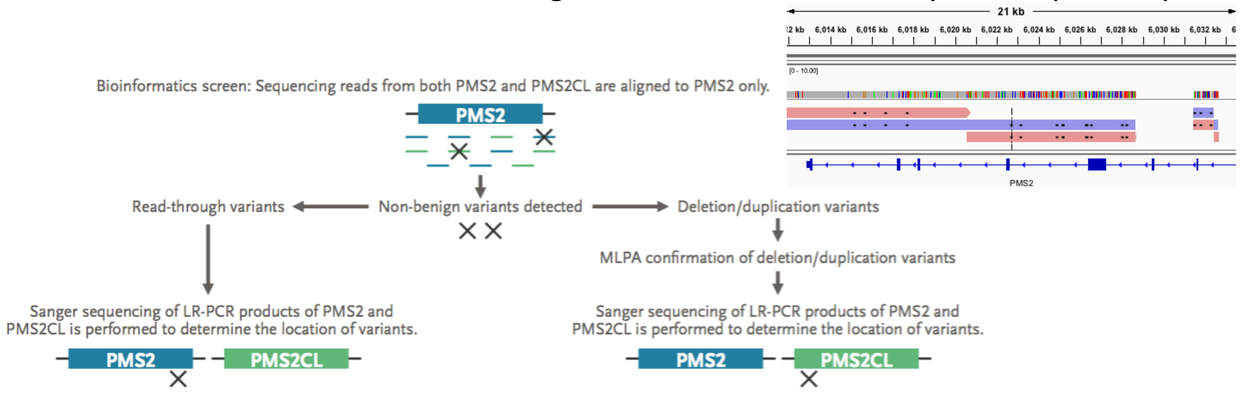
- We developed **PseudoSlayer,** a general two-step method for identifying variants within SegDups: (1) a bioinformatics (BFX) screen, where NGS data are aligned to a SegDup-masked reference; (2) if variants of interest are identified, the true variant location/ identity is disambiguated with additional lab assays (eg LR-PCR + sequencing, MLPA).
- The majority of samples are screen negative and require no additional assays, which keeps turnaround time and cost low while maintaining the outstanding quality Invitae is committed to deliver.



Masking SegDups enables read alignment, and additional lab assays reveal the true location and identity of the variant.

PMS2

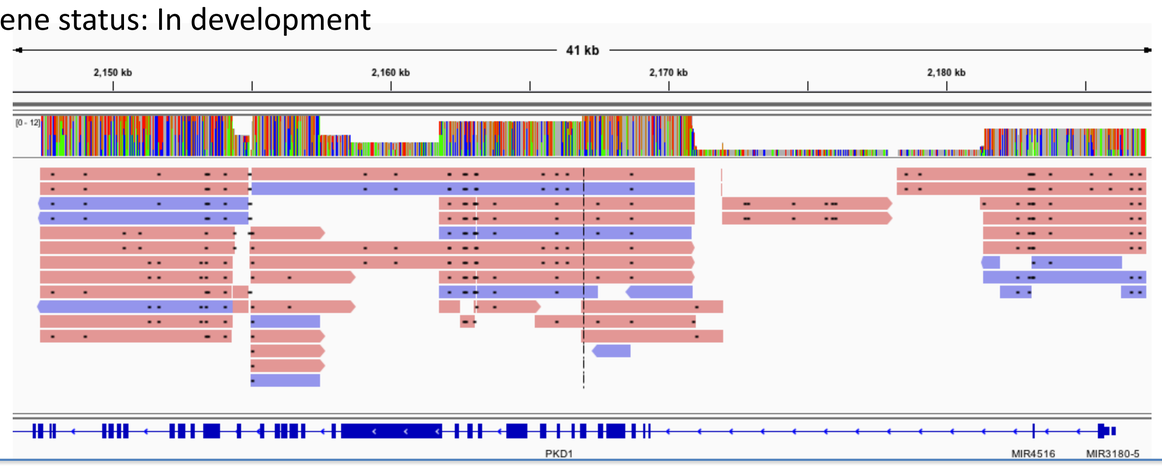
- Related syndrome: Lynch syndrome (HNPCC)
- # of SegDups: 1 (PMS2CL), covering PMS2 exons 9, 11-15
- Avg % SegDup identity over exons: 99.8% (exons 12-15)
- Length of exonic sequence in SegDup region: 1.2 kb
- Unique feature: Gene conversion causes polymorphic PSVs.
- Specific solution: Mask PMS2CL, BFX screen, disambiguate variants with LR-PCR/MLPA.
- Gene status: Full commercial offering, validated, 100% sensitivity, 100% specificity



- https://www.invitae.com/en/webinars/
- http://blog.invitae.com/full-pms2-testing-at-invitae-weve-got-you-covered/
- http://marketing.invitae.com/acton/attachment/7098/f-0139/1/-/-/-/WP103-1_PMS2%20Sequencing%20NGS%20Validation%20Summary.pdf

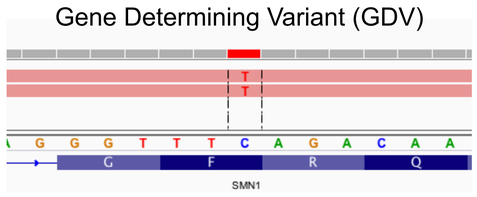
PKD1

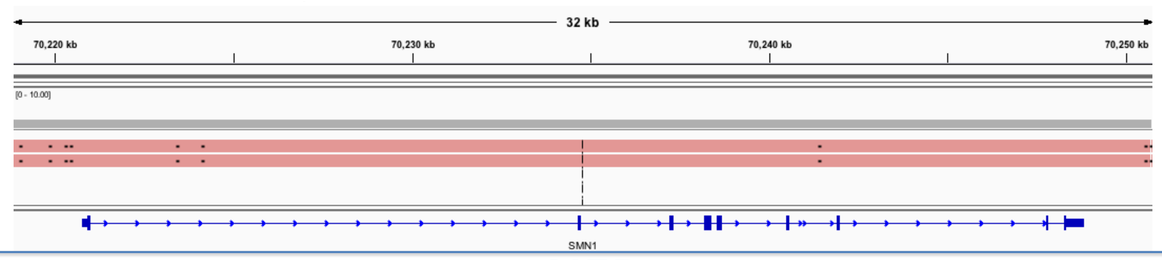
- Related syndrome: Autosomal dominant polycystic kidney disease (ADPKD)
- # of SegDups: 6, covering PKD1 exons 1-33
- Avg % SegDup identity over exons: 98.4%
- Length of exonic sequence in SegDup region: 54.5 kb
- Unique features: Uncharacterized polymorphic PSVs, rare gene conversion reported.
- Specific solution: Mask polymorphic PSV sites, unique read alignment, disambiguate (del/dup|gene conversion) with LR-PCR.
- Gene status: In development



SMN1

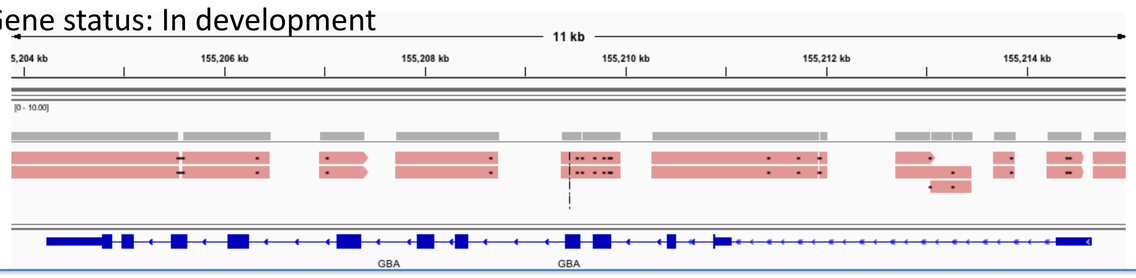
- Related syndrome: Spinal muscular atrophy (SMA)
- # of SegDups: 1 (SMN2), covering full SMN1 gene
- Avg % SegDup identity over exons: 99.9%
- Length of exonic sequence in SegDup region: 1.8 kb
- Unique feature: Extreme copy number polymorphism/fluidity; SMN2 is 15% functional.
- Specific solution: Mask SMN2, BFX screen, call specific exon 7 copy number using the GDV allele balance.
- Gene status: In development





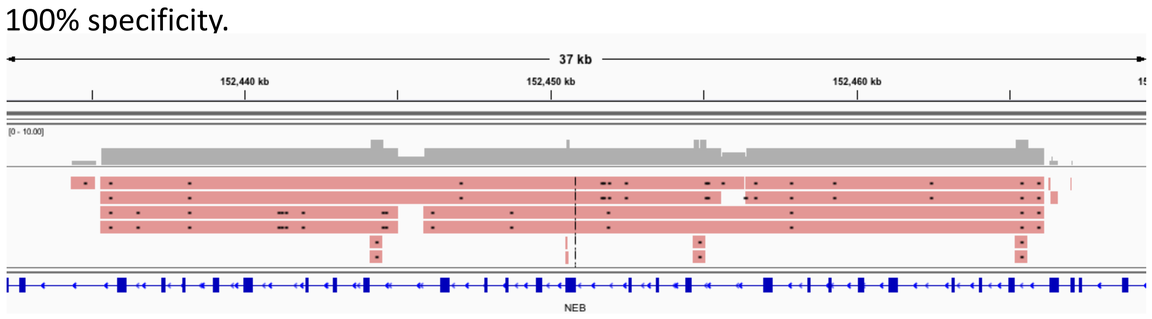
GBA

- Related syndrome: Gaucher's disease (GD)
- # of SegDups: 1 (GBAP1), covering full GBA gene, with interruptions
- Avg % SegDup identity over exons: 97.5%
- Length of exonic sequence in SegDup region: 3.2 kb
- Unique features: Uncharacterized polymorphic PSVs; gene conversion and fusion genes can be pathogenic.
- Specific solution: Mask polymorphic PSV sites, unique read alignment, disambiguate (del/dup|gene conversion|fusion gene) with LR-PCR.
- Gene status: In development



NEB

- Related syndrome: Autosomal recessive nemaline myopathy 2 (NEM2)
- # of SegDups: 2, covering exons 82-89
- Avg % SegDup identity over exons: 99.8%
- Length of exonic sequence in SegDup region: 4.4 kb
- Unique feature: Intra-genic triplication of eight exons
- Specific solution: Mask two of three triplications, BFX screen.
- Gene status: Commercial offering for sequence variants, validated, 100% sensitivity, 100% specificity.



SBDS

- Related syndrome: Shwachman-Diamond syndrome (SDS)
- # of SegDups: 1 (SBDSP1), covering full SBDS gene
- Avg % SegDup identity over exons: 98.2%
- Length of exonic sequence in SegDup region: 1.5 kb
- Unique feature: 60% of pathogenic variants due to gene conversion (PMID: 12496757)
- Specific solution: Unique alignment, disambiguate (del/dup|gene conversion) with LR-PCR.
- Gene status: In development