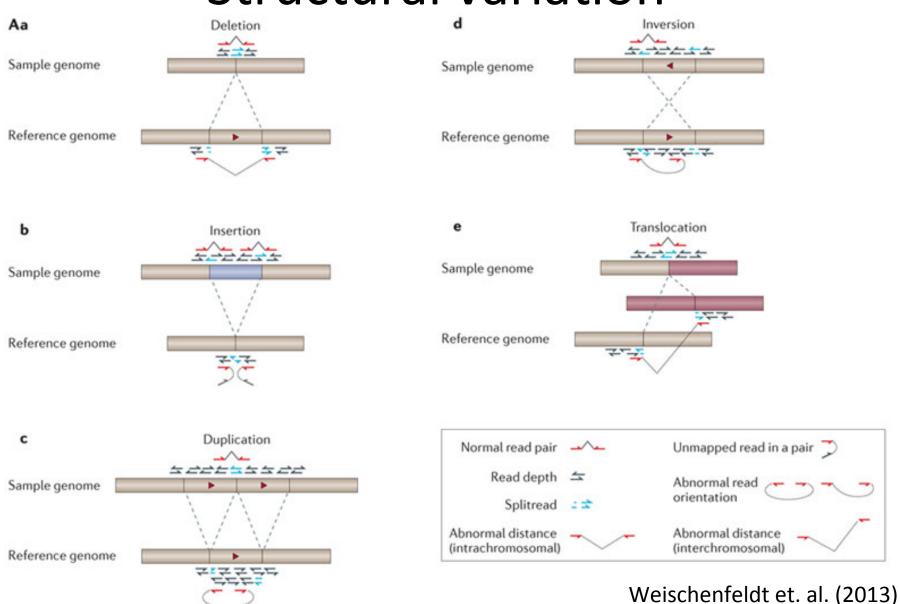
Accurate and fast detection of complex and nested structural variations using long read technologies

Fritz Sedlazeck

Friday, Oct 28 CSHL



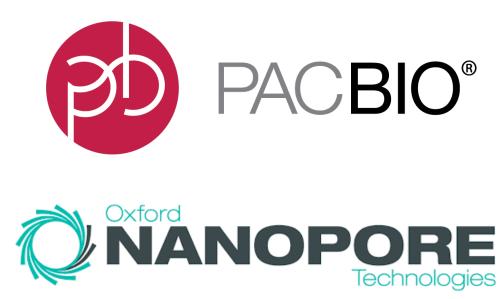
Structural variation



Long Read Technologies

- (+) SVs in repetitive regions
- (+) Can identify nested SVs

- (-) Higher error rate
- (-) Hard to align

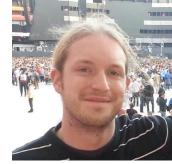


Hard to align

	w	5,091 bp												
	TYPE	14,826,000 bp	34,827,000 bp	34,828,000 bp	34,829,000 bp		34,830,000 bp	34,83						
	NAME DATA 1 DATA 5	14,620,000 bp	34,627,000 bp	34,828,000 bp	34,629,000 bp		34,630,000 bp	34,83						
	233													
		[0 - 87]		, AND ADDRAME IN THE DRIVER DRIVEN WITH A DRIVEN AND A DRIV	ALLENSEN, AND IN PRIME YOR A A DESIGN THE									
BWA-mem														
			· · · · · · · · · · · · · · ·			***								
				4 # 1 1 - 1 11 H 1 + 8 + 8 + 9 + 1 1 1 + 1 + 1 + 8 + 4 + 4 + 4 1 + 4 + 4 + 1 + 4 + 4 + 4 + 4 + 4 + 4 + 4 + 4 + 4 + 4		1+11+1+1+11								
						++++++++++++++++++++++++++++++++++++++								
			· · · · · · · · · · · · · · · · · · ·	******		***								
		· · · · · · · · · · · · · · · · · · ·	HIH HIH											
				· · · · · · · · · · · ·										
								· • • • • • •						
			· · · · · · · · · · · · · · · · · · ·	-		F BHHII								
BWA-MEN	VI													
	_	e e les les les les les les les les les	· · · · · · · · · · · · ·											
			· · · · · · · · · · · · · · · · · · ·											
			1											
			· · · · · · · · · · · · · · ·											
			· · · · · · · · · · · · · · · · · · ·											
			·· 4 //- · ···· 🗰											
			- +											
				-8-14 (8-86 (84) 18-11 - 1 (8-14) 18-14-10 (8-8) 18-41										
		· ·· · · ·	· · · · · · · · ·			**************************************								
simulated_sv.bed														
				IN	V_7									

Human genome: 1kb Inversion

Improving long read alignment



Philipp Rescheneder

- 1. Split the reads:
 - Translocations
 - Inversions
 - Duplications

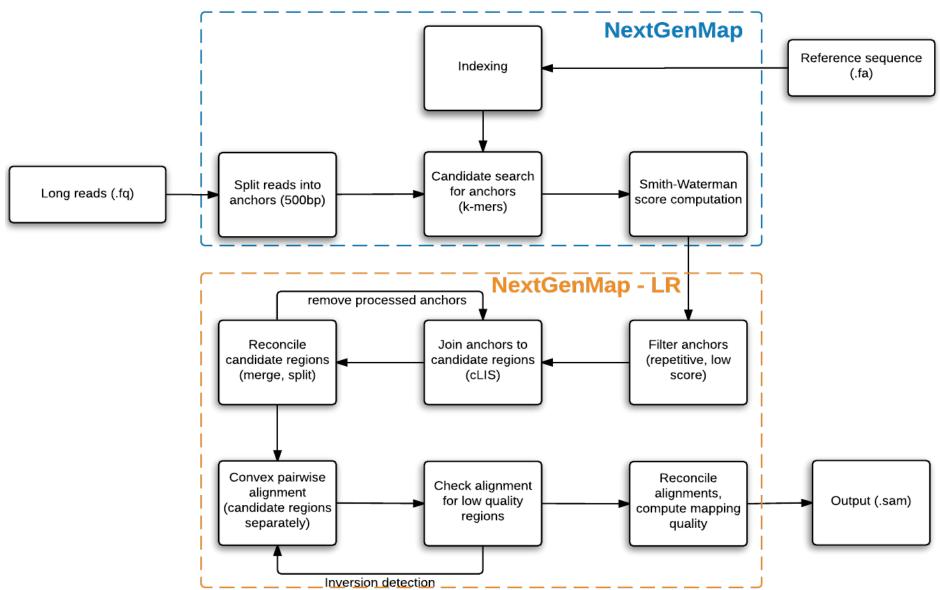


with the first particular and a star first a work of the first particular star for the star of the star of the

- 2. Improve alignment:
 - Insertions
 - Deletions

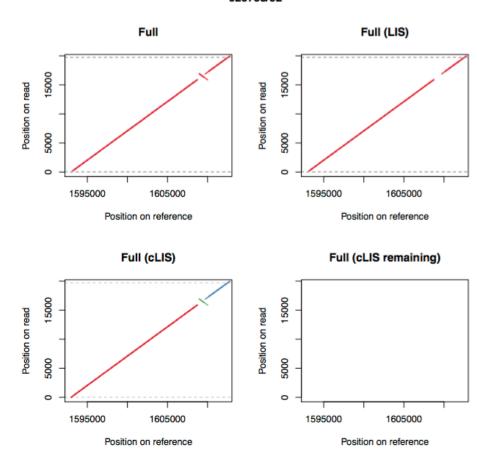
EI -F - I - UF FIHF F I F I - IF - FI FI - · I - FI FI - I - FI - F
I I -D HUE B HITELFE F +1 + FE + +-E F + I -E + F F F F H-H
HERRITER ALE CERTE ALE CERTE ALE ALE ALE ALE ALE ALE ALE ALE ALE AL
· · · · · · · · · · · · · · · · · · ·
FI FIFT I BEF FIB FFIF. IF FF FFFFFFFFFFFFFFFFFFFFFF
FFFTF-IFHI H-DEFT- FI FE IFTFT-IHFFTFTFTFTFTFTFT
ITTE B-BTEBB++FHIRE ITE IEI B B F I II
I BE E TERE E HERHHHHERE BILLERE
F-FFBFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF

NGM-LR workflow



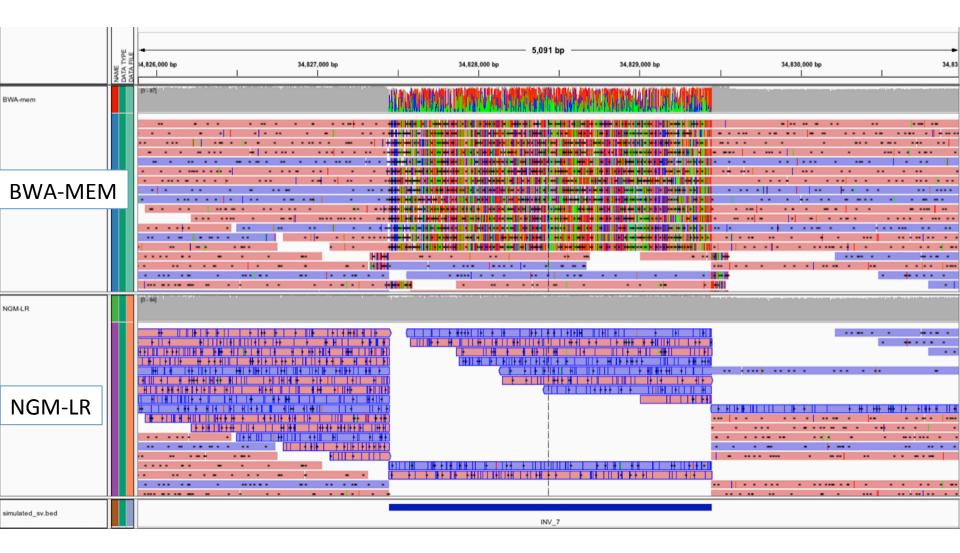
NGM-LR reconcile

Dot plot of 500bp segments from a read spanning a inversion



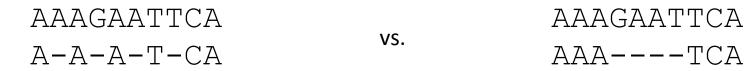
925758/52

NGM-LR inversion

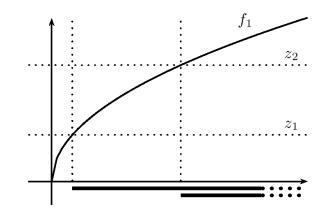


Convex Pairwise Alignment

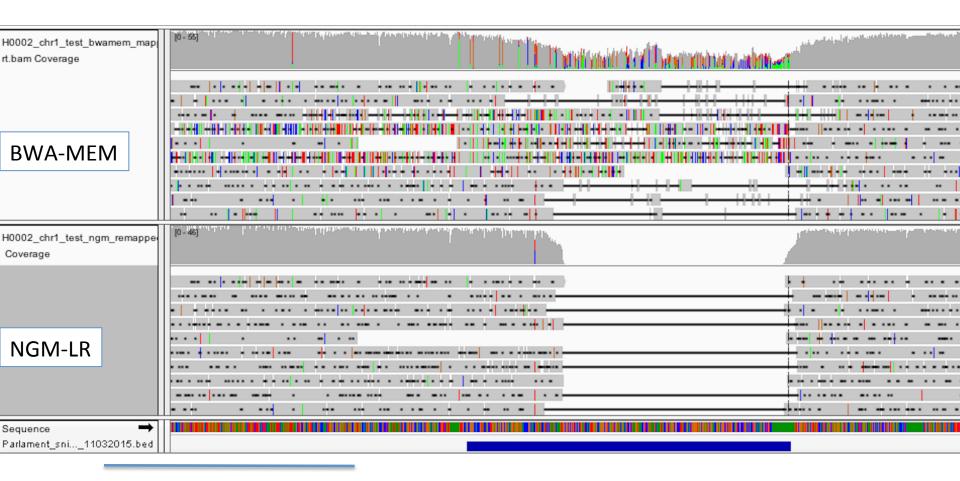
[0 - 65]	יאיי ו יאיי	цн., Гейн		an a		i hora	An	á hanin	dibi tadi yaan (talik	A	le posti se colo de terri		an arrest start for a first start		 · · · · · · · · · · · · · · · · · · ·	h litera en aparece	ال	a a da facel de la constana por	ա տարությունը։
							-+ + +												
• • •								HH						• •	 			· · · • • • · ·	
			┫╍═╁╍┫┥╎╼╾┥╸	4	┥┉║┉┫	H-		·	- 8 -8-8-8-8-8-8-8-8-8			+		• •• •	 				
			#	+ + - +	1[]-+[]++	HAN- MAR	┉╢┽┫┽╺┥┥┝┿┿												
			•		-	•	***		• •				·		 		· · ·	· · · · ···	
				· · · · · · · · · · · · · · · · · · ·															
••••••	• · · · · · ·		·			++ + 44				 - -					 · · • • · • · · ·			<mark>.</mark>	
				· · •••••		· · · · -	++++								 				
						-								•• •	 				
								-							 				-



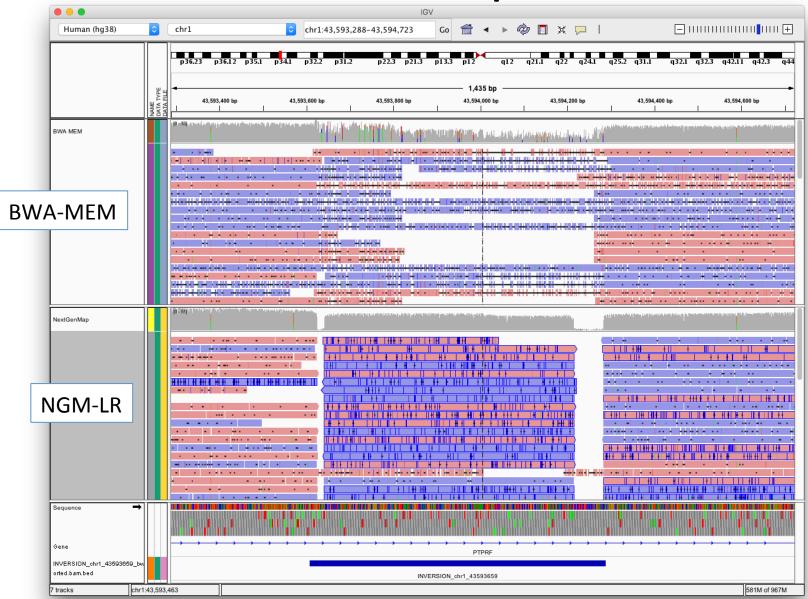
- Linear: gap cost always the same
- Affine: separate penalties for opening and extending a gap
- Convex: initially similar to affine, but becomes proportionally less costly for larger gaps



NGM-LR deletion



NGM-LR complex SV



Sniffles

- Analyzing split reads, alignment events and noisy regions.
- Parameter estimation
- Optional: Genotype estimation
- Optional: Clustering of SVs



Analyzing noisy regions

Extract the differences in the alignment

• Detect the noisy regions: Plane sweep algorithm

at the rest at a data of	and the second
+ + - + 	
	TE ERE ERE EN
	E + FI EI + FI + FI + FF I I + I + E + FF E + I + FI
	F F
I BE FILE F HURHARD BUILD	EL EL FIEINCE

 Store potential regions in a self balancing binary tree.

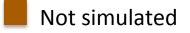
Simulation/Evaluation

- 1. Simulate 20 SVs of each type using SURVIVOR
- 2. Simulate Pacbio and illumina paired end reads
- 3. Evaluation using SURVIVOR

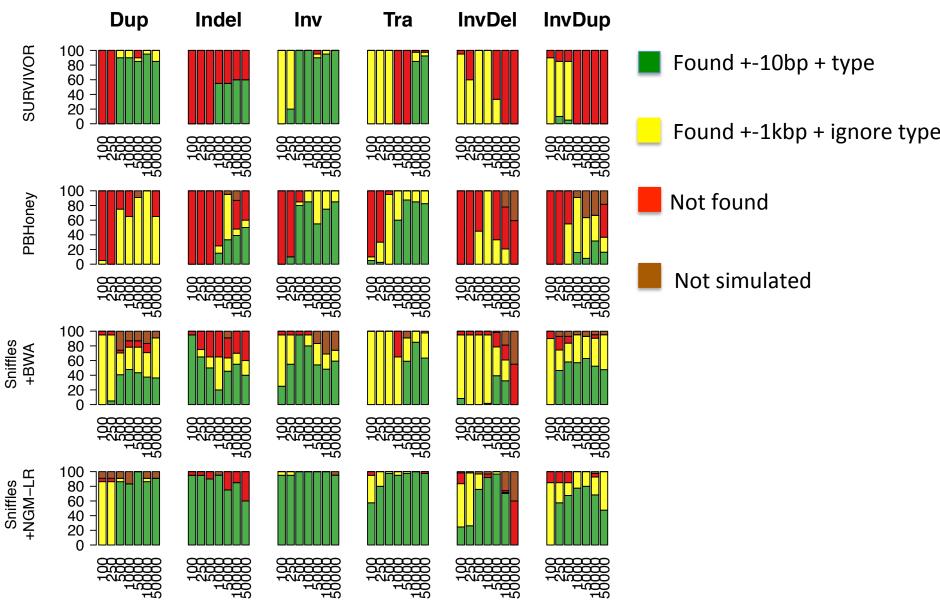
Found +-10bp + type

Found +-1kbp + ignore type

Not found



Evaluation of Sniffles



Summary

NextGenMap:

- Short read version: github.com/cibiv/NextGenMap
- Long read mapper: github.com/philres/nextgenmap-lr
 - Self detection of SVs
 - Manuscript in preparation

SURVIVOR:

- Toolkit for SV detection on short reads
- Simulation/Evaluation of current methods
- Consensus approach
- Accepted Nature Communications
- Available:

github.com/fritzsedlazeck/SURVIVOR

Sniffles:

- SVs detection for long reads
- Nested SV
- Manuscript in preparation
- Available: github.com/fritzsedlazeck/Sniffles

Future work:

- How much coverage is needed?
- Nanopore support
- Analysis of nested SVs
- Application to Cancer Genomes (See Maria Nattestad's Poster #79)

Acknowledgments



Cold Spring Harbor Laboratory

Maria Nattestad Han Fang Srividya Ramakrishnan



Daniel Jeffares Jürg Bähler Christophe Dessimoz



Philipp Rescheneder Moritz Smolka Arndt von Haeseler



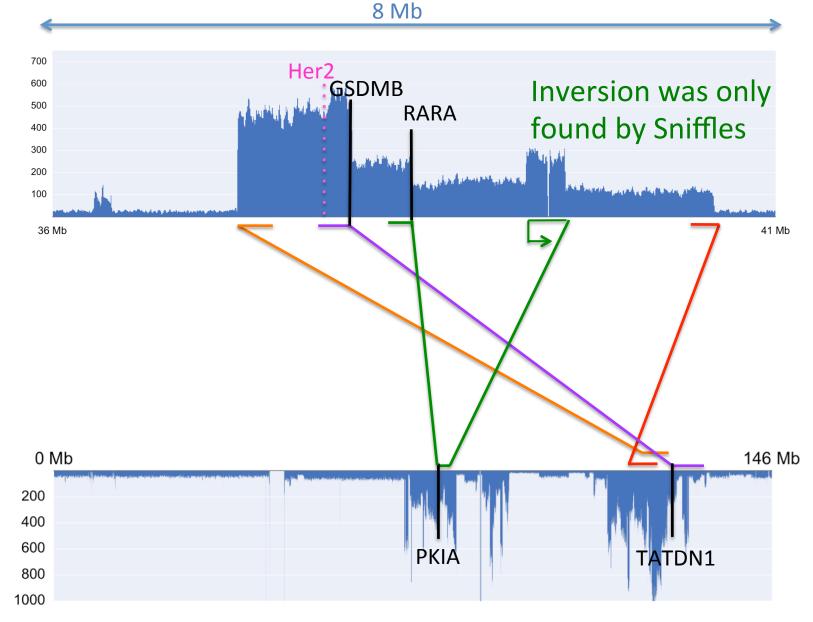
Michael Schatz

NGM-LR nanopore



Norris et al (2016)

Evaluation of Sniffles: SKBR3



Chromosome 8