

# **Detection and visualization of complex** structural variants from NGS data

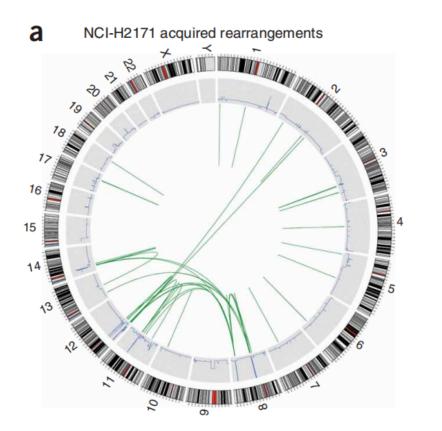
Valerio Vitali Institute for Evolution and Biodiversity University of Münster, Germany SESSION ON VARIANT CALLING

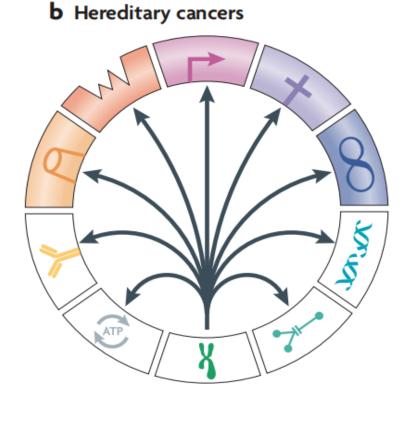
BIOINFORMATICS



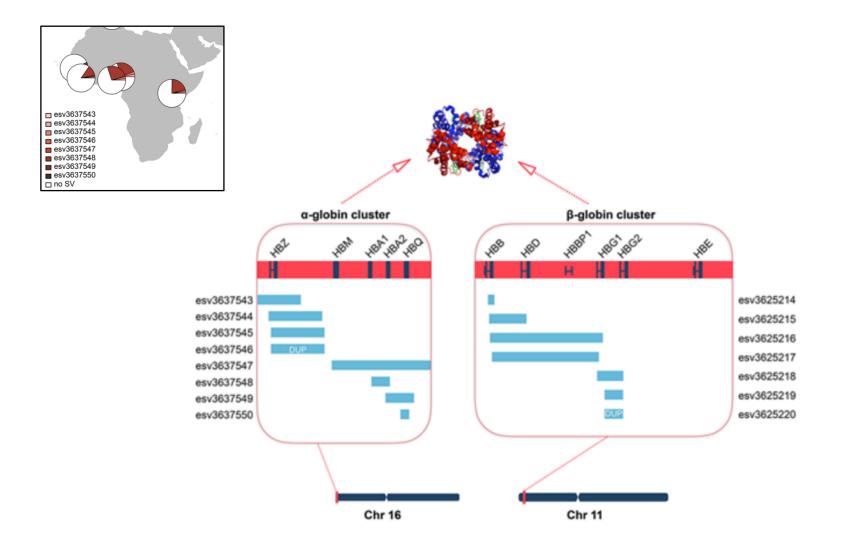
Summer School 2019

## Structural Variants: initiation of cancer

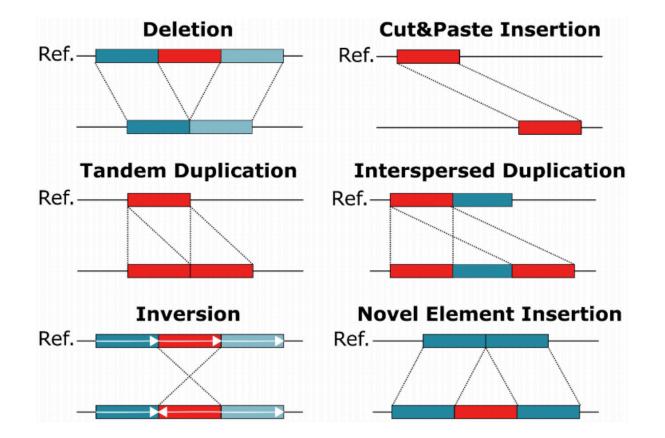




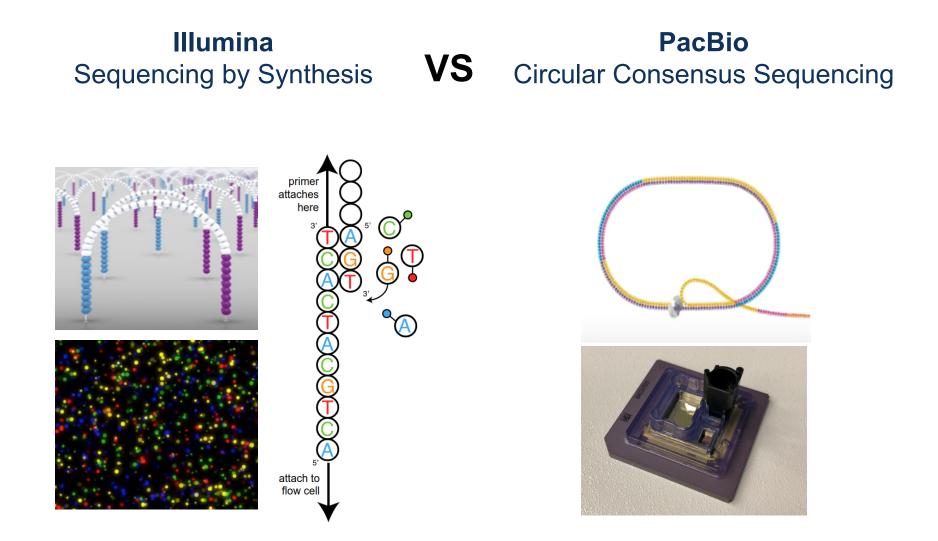
#### Structural Variants: balancing selection



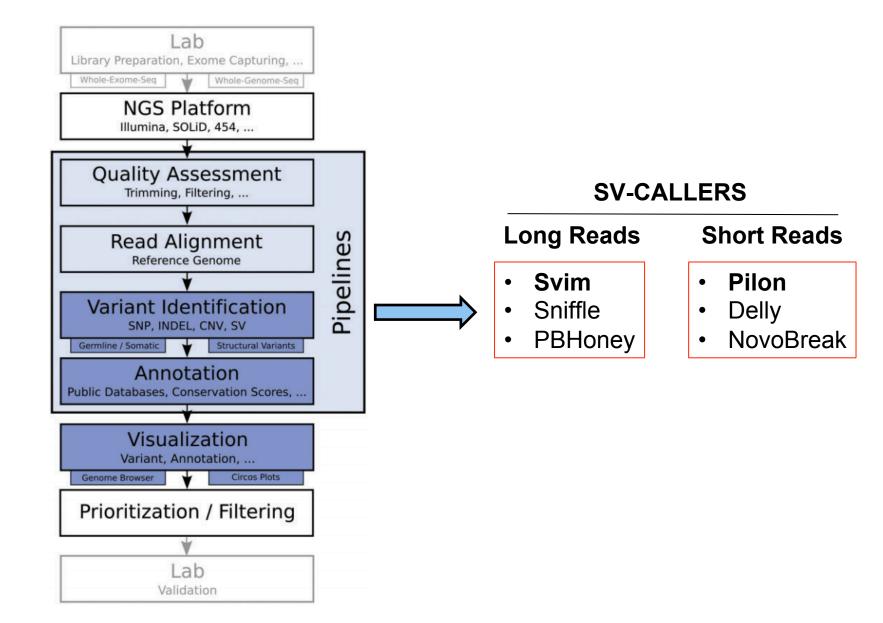
#### Structural Variants: types



#### Illumina SBS vs PacBio CCS



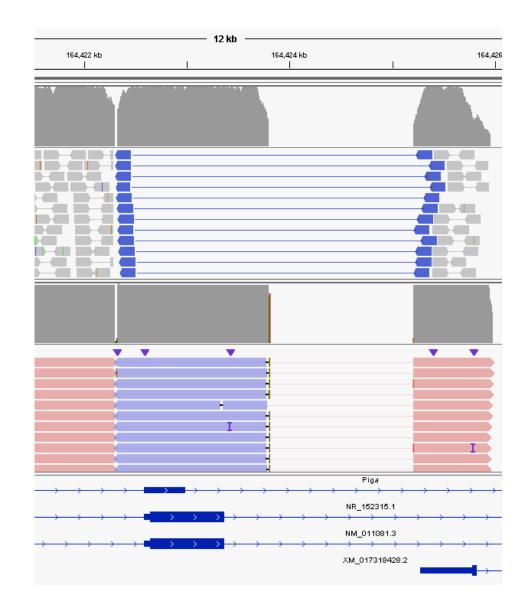
#### Standard NGS workflow



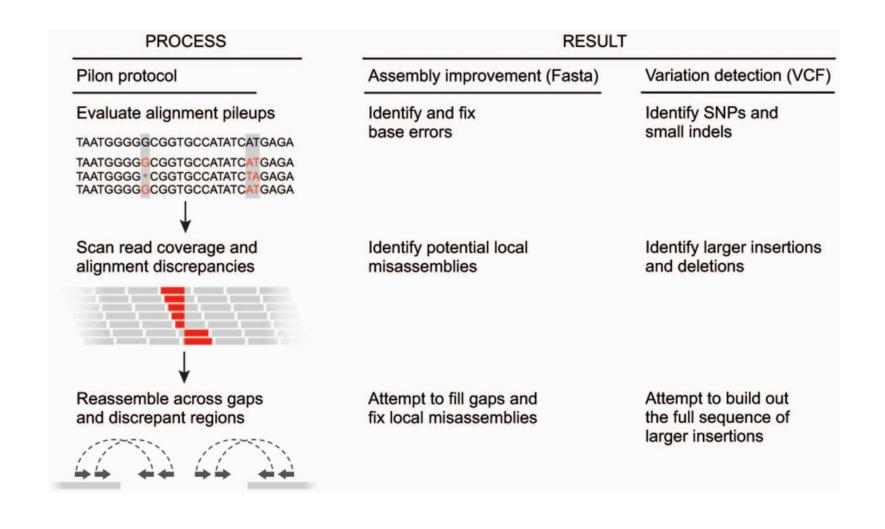
# Short vs Long reads

- + Higher base quality
- + Lower costs
- Indirect evidence for SVs
- Mapping on repeated regions

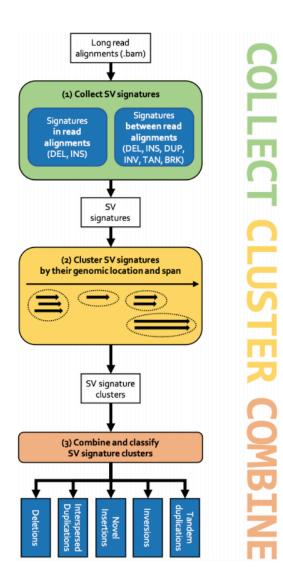
- + Accurate mapping on repeated and low complexity regions
- + Often span the whole SV
- Higher sequencing costs
- Elevated error rate (5-15%)



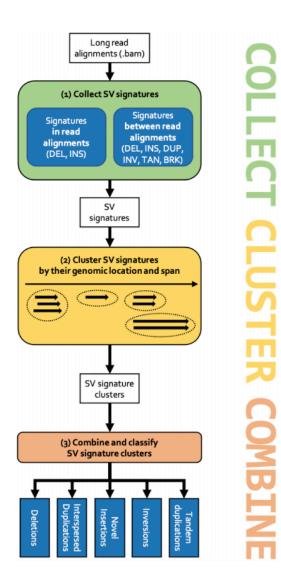
# SV calling from PE-reads: Pilon

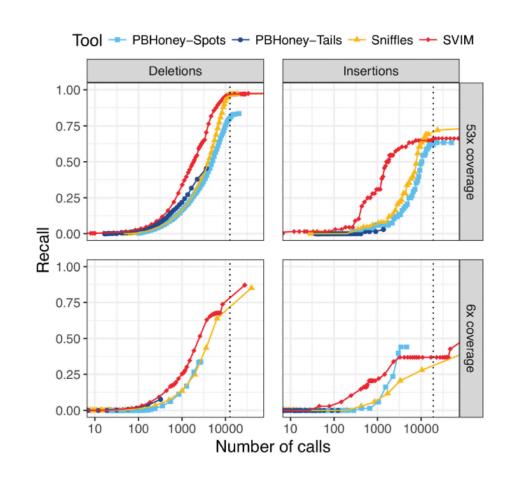


# SV calling from long reads: svim



# SV calling from long reads: svim





# Variant Call Format (VCF)

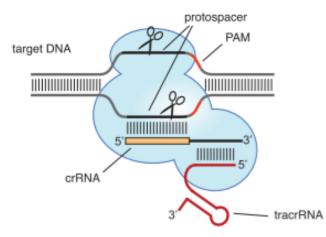
##fileformat=VCFv4.1									
##fileDate=20190909 Men ed. Sign in [Slack 🗁 20/ per Maile Dear El Data Flood El 4) How to Fix Inter El Relutonals de Unux compatibility en la Contra de Co									
##source="Pilon version 1.23 Mon Nov 26 16:04:05 2018 -0500"									
##PILON="genome GRCm38.p6 ChrX.fafrags PIGA15ILL003.trmd.sorted.bamoutput PIGA15ILL003outdir ./PIGA15ILL003.Pilon outputchangesvcftargets									
NC_000086.7:164420377-164425982fix all, breaksmindepth 3"									
##reference=file:/home/waltermint/Desktop/PIPE1_test/GRCm38.p6_ChrX.fa									
##contig= <id=nc_000086.7,length=5606></id=nc_000086.7,length=5606>									
##FILTER= <id=lowcov,description="low at="" coverage="" good="" location"="" of="" reads=""></id=lowcov,description="low>									
##FILTER= <id=amb,description="ambiguous evidence="" genome"="" haploid="" in=""></id=amb,description="ambiguous>									
##FILTER= <id=del,description="this a="" another="" base="" change="" deletion="" event="" from="" in="" is="" or="" record"=""></id=del,description="this>									
##INF0= <id=dp.number=1,type=integer,description="valid been="" depth;="" filtered"="" have="" may="" read="" reads="" some=""></id=dp.number=1,type=integer,description="valid>									
##INF0= <id=td,number=1,type=integer,description="total bad="" depth="" including="" pairs"="" read=""></id=td,number=1,type=integer,description="total>									
##INF0= <id=pc,number=1,type=integer,description="physical across="" coverage="" inserts="" locus"="" of="" valid=""></id=pc,number=1,type=integer,description="physical>									
##INFO= <id=bq,number=1,type=integer,description="mean at="" base="" locus"="" quality=""></id=bq,number=1,type=integer,description="mean>									
##INF0= <id=mq,number=1,type=integer,description="mean at="" locus"="" mapping="" quality="" read=""></id=mq,number=1,type=integer,description="mean>									
##INFO= <id=qd_number=1,type=integer_description="variant by="" confidence="" depth"="" quality=""></id=qd_number=1,type=integer_description="variant>									
##INF0= <id=bc,number=4,type=integer,description="count as,="" at="" cs,="" gs,="" locus"="" of="" ts=""></id=bc,number=4,type=integer,description="count>									
##INF0= <id=qp,number=4,type=integer,description="percentage &="" as,="" at="" by="" cs,="" gs,="" locus"="" mq="" of="" q="" ts="" weighted=""></id=qp,number=4,type=integer,description="percentage>									
##INF0= <id=ic,number=1,type=integer,description="number here"="" insertion="" of="" reads="" with=""></id=ic,number=1,type=integer,description="number>									
##INF0= <id=dc,number=1,type=integer,description="number deletion="" here"="" of="" reads="" with=""></id=dc,number=1,type=integer,description="number>									
##INF0= <id=xc,number=1,type=integer,description="number clipped="" here"="" of="" reads=""></id=xc,number=1,type=integer,description="number>									
##INF0= <id=ac,number=a,type=integer,description="allele allele,="" alt="" as="" count="" each="" for="" genotypes,="" in="" listed"="" order="" same="" the=""></id=ac,number=a,type=integer,description="allele>									
##INF0= <id=af,number=a,type=float,description="fraction allele(s)"="" alternate="" evidence="" in="" of="" support=""></id=af,number=a,type=float,description="fraction>									
##INF0= <id=svtype,number=1,type=string,description="type of="" structural="" variant"=""></id=svtype,number=1,type=string,description="type>									
##INF0= <id=svlen,number=.,type=string,description="difference alleles"="" alt="" and="" between="" in="" length="" ref=""></id=svlen,number=.,type=string,description="difference>									
##INF0= <id=end,number=1,type=integer,description="end described="" in="" of="" position="" record"="" the="" this="" variant=""></id=end,number=1,type=integer,description="end>									
##INF0= <id=imprecise,number=0,type=flag,description="imprecise (alt="" change="" contains="" from="" local="" ns)"="" reassembly=""></id=imprecise,number=0,type=flag,description="imprecise>									
##FORMAT= <id=gt,number=1,type=string,description="genotype"></id=gt,number=1,type=string,description="genotype">									
##FORMAT= <id=ad,number=.,type=string,description="allelic alleles="" alt="" and="" depths="" for="" in="" listed"="" order="" ref="" the=""></id=ad,number=.,type=string,description="allelic>									
##FORMAT= <id=dp,number=1,type=string,description="approximate been="" depth;="" filtered"="" have="" may="" read="" reads="" some=""></id=dp,number=1,type=string,description="approximate>									
##ALT= <id=dup.description="possible duplication"="" seamental=""></id=dup.description="possible>									
#CHROM POS ID REF ALT CONQUAL FILTER INFO FORMAT SAMPLE									
NC_000086.7 164420377 . C . 0 LowCov DP=0;TD=0;BQ=0;MQ=0;QD=0;BC=0,0,0;QP=0,0,0;QP=0,0,0;PC=1;IC=0;AC=0;AF=0.00 GT									
0/0									
NC_000086.7 164420378 . A . 0 LowCov DP=0;TD=0;BQ=0;MQ=0;QD=0;BC=0,0,0;QP=0,0,0;QP=0,0,0;PC=2;IC=0;AC=0;AF=0.00 GT									
NC_000086.7 164420379 . T . 0 LowCov DP=0;TD=0;BQ=0;MQ=0;QD=0;BC=0,0,0;QP=0,0,0;QP=0;DC=0;CC=0;AC=0;AF=0.00 GT									
0/0									
NC 000086.7 164420380 . G . 0 LowCov DP=0:TD=0:B0=0:M0=0:0D=0:BC=0.0.0:0P=0.0.0:0P=3:IC=0:AC=0:AC=0:AF=0.00 GT									

# Variant Call Format (VCF)

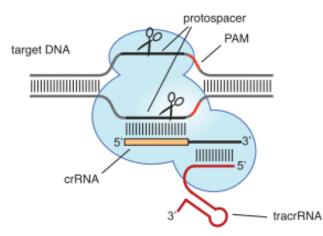
##fileformat=VCFv4.1										
##fileDate=20190909										
##fileDate=20190909 ##source="Pilon version 1.23 Mon Nov 26 16:04:05 2018 -0500"										
##Source="Piton version 1.23 Mon Nov 20 10:04:05 2018 -0500" ##PILON="genome GRCm38.p6 ChrX.fafrags PIGA15ILL003.trmd.sorted.bamoutput PIGA15ILL003outdir ./PIGA15ILL003.Pilon outputchangesvcftargets										
NC_000086.7:164420377-164425982fix all,breaksmindepth 3" ##reference_file:/beme/upltermint/Deckten/DTPE1_test/CPCm28_n6_CbrX_fa										
##reference=file:/home/waltermint/Desktop/PIPE1_test/GRCm38.p6_ChrX.fa										
##contig= <id=nc_000086.7,length=5606></id=nc_000086.7,length=5606>										
##FILTER= <id=lowcov,description="low at="" coverage="" good="" location"="" of="" reads=""></id=lowcov,description="low>										
##FILTER= <id=amb,description="ambiguous evidence="" genome"="" haploid="" in=""></id=amb,description="ambiguous>										
##FILTER= <id=del,description="this a="" another="" base="" change="" deletion="" event="" from="" in="" is="" or="" record"=""></id=del,description="this>										
##INF0= <id=dp,number=1,type=integer,description="valid been="" depth;="" filtered"="" have="" may="" read="" reads="" some=""></id=dp,number=1,type=integer,description="valid>										
##INF0= <id=td,number=1,type=integer,description="total bad="" depth="" including="" pairs"="" read=""></id=td,number=1,type=integer,description="total>										
##INFO= <id=pc,number=1,type=integer,description="physical across="" coverage="" inserts="" locus"="" of="" valid=""></id=pc,number=1,type=integer,description="physical>										
##INF0= <id=bq,number=1,type=integer,description="mean at="" base="" locus"="" quality=""></id=bq,number=1,type=integer,description="mean>										
##INFO= <id=mq,number=1,type=integer,description="mean at="" locus"="" mapping="" quality="" read=""></id=mq,number=1,type=integer,description="mean>										
##INFO= <id=qd,number=1,type=integer,description="variant by="" confidence="" depth"="" quality=""></id=qd,number=1,type=integer,description="variant>										
##INFO= <id=bc,number=4,type=integer,description="count as,="" at="" cs,="" gs,="" locus"="" of="" ts=""></id=bc,number=4,type=integer,description="count>										
##INFO= <id=qp,number=4,type=integer,description="percentage &="" as,="" at="" by="" cs,="" gs,="" locus"="" mq="" of="" q="" ts="" weighted=""></id=qp,number=4,type=integer,description="percentage>										
##INFO= <id=ic,number=1,type=integer,description="number here"="" insertion="" of="" reads="" with=""> Dion="dbSNP membership" build 129"&gt;</id=ic,number=1,type=integer,description="number>										
##INFO= <id=dc,number=1,type=integer,description="number deletion="" here"="" of="" reads="" with=""></id=dc,number=1,type=integer,description="number>										
##INFO= <id=xc,number=1,type=integer,description="number clipped="" here"="" of="" reads=""></id=xc,number=1,type=integer,description="number>										
##INFO= <id=ac,number=a,type=integer,description="allele allele,="" alt="" as="" count="" each="" for="" genotypes,="" in="" listed"="" order="" same="" the=""></id=ac,number=a,type=integer,description="allele>										
##INF0= <id=af,number=a,type=float,description="fraction allele(s)"="" alternate="" evidence="" in="" of="" support=""></id=af,number=a,type=float,description="fraction>										
##INF0= <id=svtype,number=1,type=string,description="type of="" structural="" variant"=""></id=svtype,number=1,type=string,description="type>										
##INF0= <id=svlen,number=.,type=string,description="difference alleles"="" alt="" and="" between="" in="" length="" ref=""></id=svlen,number=.,type=string,description="difference>										
##INFO= <id=end,number=1,type=integer,description="end described="" in="" of="" position="" record"="" the="" this="" variant=""></id=end,number=1,type=integer,description="end>										
##INF0= <id=imprecise,number=0,type=flag,description="imprecise (alt="" change="" contains="" from="" local="" ns)"="" reassembly=""></id=imprecise,number=0,type=flag,description="imprecise>										
##FORMAT= <id=gt,number=1,type=string,description="genotype"></id=gt,number=1,type=string,description="genotype">										
##FORMAT= <id=ad,number=.,type=string,description="allelic alleles="" alt="" and="" depths="" for="" in="" listed"="" order="" ref="" the=""> Wallo Columns</id=ad,number=.,type=string,description="allelic>										
##FORMAT= <td=dp_number=1_type=string_description="approximate_read_depth; some_reads_may_have_been_filtered"=""></td=dp_number=1_type=string_description="approximate_read_depth;>										
##ALT= <id=dup.description="possible duplication"="" seamental=""></id=dup.description="possible>										
#CHROM POS ID REF ALT TOT QUAL FILTER INFO FORMAT SAMPLE V2										
NC_000086.7 164420377 . C . 0 LowCov DP=0;TD=0;BQ=0;MQ=0;QD=0;BC=0,0,0,0;QP=0,0,0,0;PC=1;IC=0;DC=0;XC=0;AF=0.00 GT										
NC_000086.7 164420378 . A . 0 LowCov DP=0;TD=0;BQ=0;MQ=0;QD=0;BC=0,0,0,0;QP=0,0,0,0;PC=2;IC=0;DC=0;XC=0;AC=0;AF=0.00 GT										
<b>0/0</b> 20 17330										
NC_000086.7 164420379 . T . 0 LowCov DP=0;TD=0;BQ=0;MQ=0;QD=0;BC=0,0,0,0;QP=0,0,0,0;PC=2;IC=0;AC=0;AC=0;AF=0.00 GT										
NC 000086.7 164420380 . G . 0 LowCov DP=0:TD=0:B0=0:M0=0:DD=0:BC=0.0.0.0:OP=0.0.0.0:PC=3:IC=0:DC=0:AC=0:AF=0.00 GT										

# Variant Call Format (VCF)

#### ##fileformat=VCFv4.1 ##fileDate=20190909 ##source="Pilon version 1.23 Mon Nov 26 16:04:05 2018 -0500" ##PILON="--genome GRCm38.p6 ChrX.fa --frags PIGA15ILL003.trmd.sorted.bam --output PIGA15ILL003 --outdir ./PIGA15ILL003.Pilon output --changes --vcf --targets NC 000086.7:164420377-164425982 --fix all,breaks --mindepth 3" ##reference=file:/home/waltermint/Desktop/PIPE1 test/GRCm38.p6 ChrX.fa ##contig=<ID=NC 000086.7,length=5606> ##FILTER=<ID=LowCov.Description="Low Coverage of good reads at location"> ##FILTER=<ID=Amb.Description="Ambiguous evidence in haploid genome"> ##FILTER=<ID=Del.Description="This base is in a deletion or change event from another record"> ##INF0=<ID=DP.Number=1.Type=Integer.Description="Valid read depth: some reads may have been filtered"> ##INF0=<ID=TD.Number=1.Type=Integer.Description="Total read depth including bad pairs"> ##INF0=<ID=PC,Number=1,Type=Integer,Description="Physical coverage of valid inserts across locus"> ##INF0=<ID=BQ,Number=1,Type=Integer,Description="Mean base quality at locus"> ##INF0=<ID=MQ,Number=1,Type=Integer,Description="Mean read mapping quality at locus">, ##INF0=<ID=QD,Number=1,Type=Integer,Description="Variant confidence/quality by depth"> ##INF0=<ID=BC,Number=4,Type=Integer,Description="Count of As, Cs, Gs, Ts at locus"> ##INFO=<ID=QP,Number=4,Type=Integer,Description="Percentage of As, Cs, Gs, Ts weighted by Q & MQ at locus"> ##INF0=<ID=IC,Number=1,Type=Integer,Description="Number of reads with insertion here"> ##INF0=<ID=DC,Number=1,Type=Integer,Description="Number of reads<u>with deletion here"></u> ##INF0=<ID=XC,Number=1,Type=Integer,Description="Number of reads clipped here"> ##INFO=<ID=AC,Number=A,Type=Integer,Description="Allele count in genotypes, for each ALT allele, in the same order as listed"> ##INF0=<ID=AF,Number=A,Type=Float,Description="Fraction of evidence in support of alternate allele(s)"> ##INF0=<ID=SVTYPE,Number=1,Type=String,Description="Type of structural variant"> ##INF0=<ID=SVLEN,Number=.,Type=String,Description="Difference in length between REF and ALT alleles"> ##INFO=<ID=END,Number=1,Type=Integer,Description="End position of the variant described in this record"> ##INF0=<ID=IMPRECISE,Number=0,Type=Flag,Description="Imprecise change from local re<u>assembly (ALT contains Ns)"></u> ##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype"> ##FORMAT=<ID=AD,Number=.,Type=String,Description="Allelic depths for the ref and alt alleles in the order listed"> ##FORMAT=<ID=DP.Number=1.Type=String.Description="Approximate read depth; some reads may have been filtered"> ##ALT=<ID=DUP,Description="Possible segmental duplication"> #CHROM POS REF ID ALT OUAL FILTER INFO FORMAT SAMPLE NC 000086.7 164420377 0 LowCov DP=0:TD=0:B0=0:M0=0:0D=0:BC=0.0.0.0:0P=0.0.0.0:PC=1:IC=0:DC=0:XC=0:AC=0:AF=0.00 GT 0/0 NC 000086.7 164420378 0 LowCov DP=0;TD=0;BQ=0;MQ=0;QD=0;BC=0,0,0,0;QP=0,0,0;PC=2;IC=0;DC=0;XC=0;AC=0;AF=0.00 GT 0/0 NC 000086.7 164420379 LowCov DP=0;TD=0;BQ=0;QD=0;BC=0,0,0,0;QP=0,0,0,0;PC=2;IC=0;DC=0;XC=0;AC=0;AF=0.00 GT 0 0/0 NC 000086.7 DP=0:TD=0:B0=0:M0=0:0D=0:BC=0.0.0.0:0P=0.0.0.0:PC=3:TC=0:DC=0:XC=0:AF=0.00.0 164420380 LOWCOV

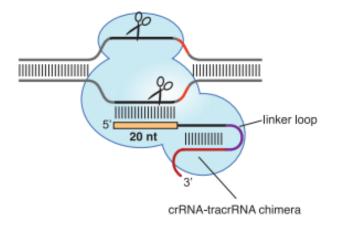


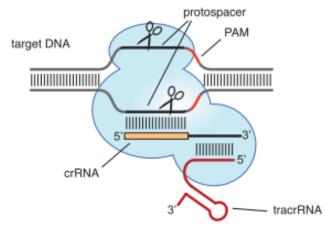
Cas9 programmed by crRNA:tracrRNA duplex



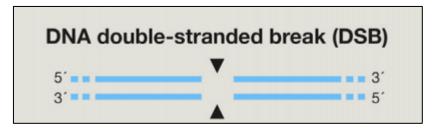
Cas9 programmed by crRNA:tracrRNA duplex

Cas9 programmed by single chimeric RNA

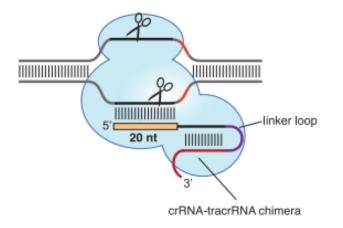


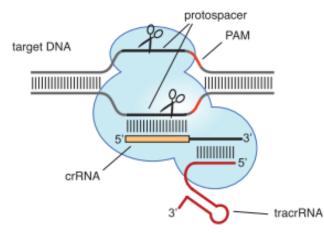


Cas9 programmed by crRNA:tracrRNA duplex

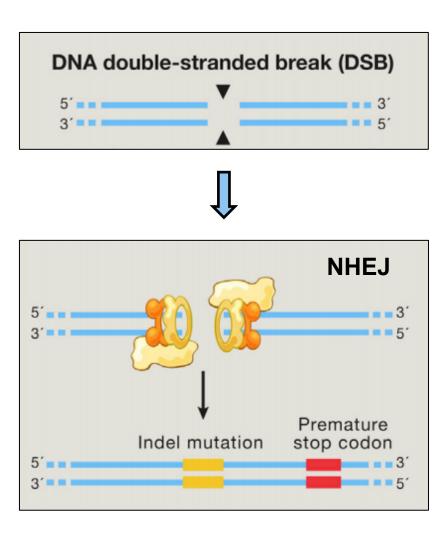


Cas9 programmed by single chimeric RNA

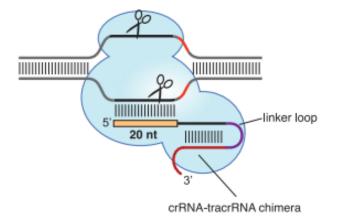




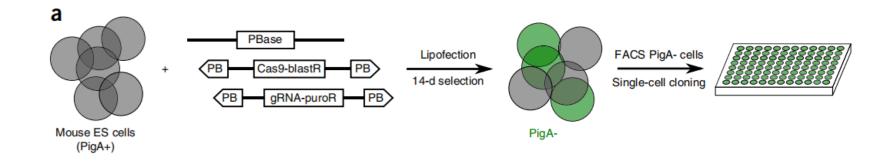
Cas9 programmed by crRNA:tracrRNA duplex



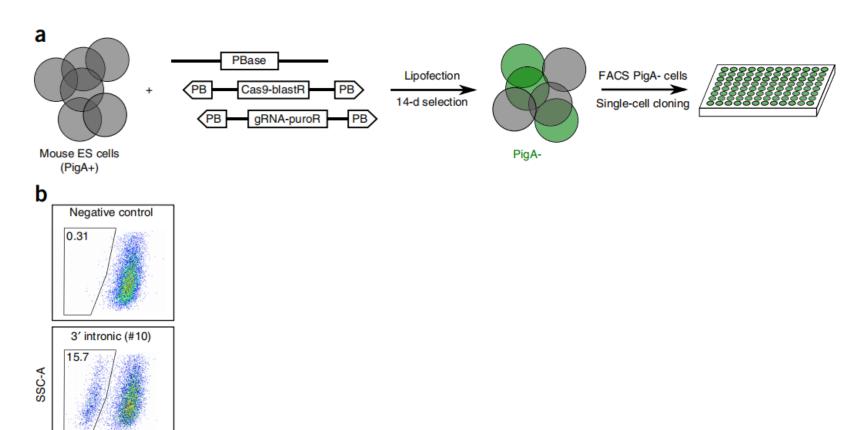
Cas9 programmed by single chimeric RNA

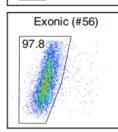


## CRISPR-Cas9 leads to complex SVs



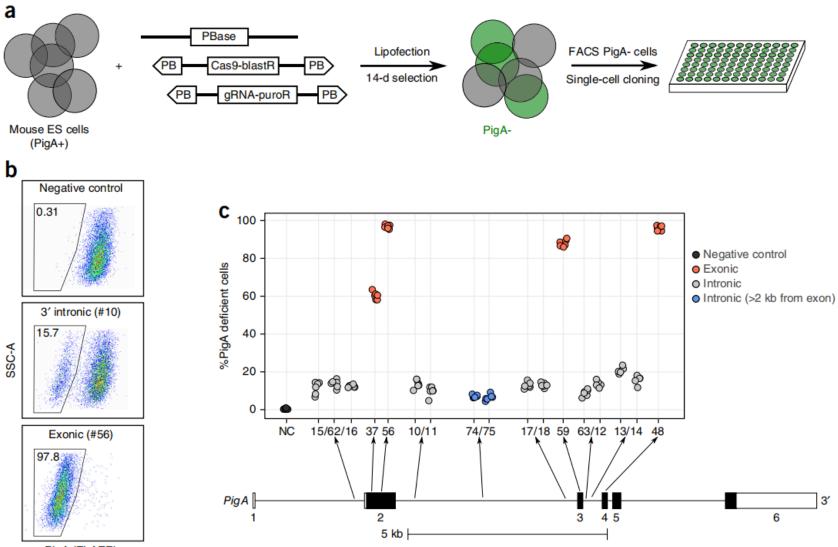
## CRISPR-Cas9 leads to complex SVs





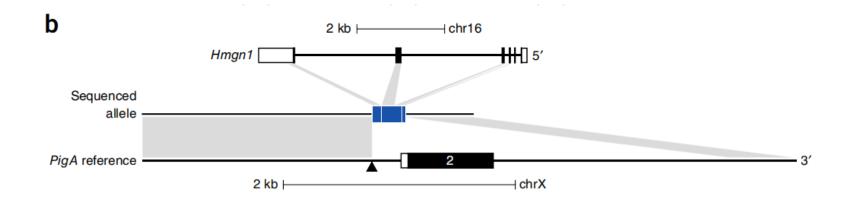
PigA (FLAER)

### **CRISPR-Cas9** leads to complex SVs

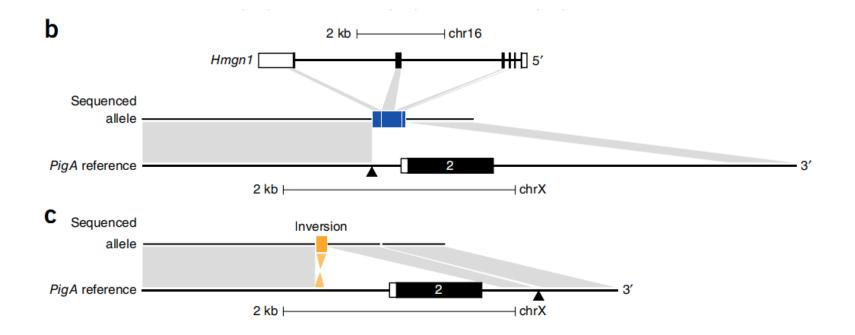


PigA (FLAER)

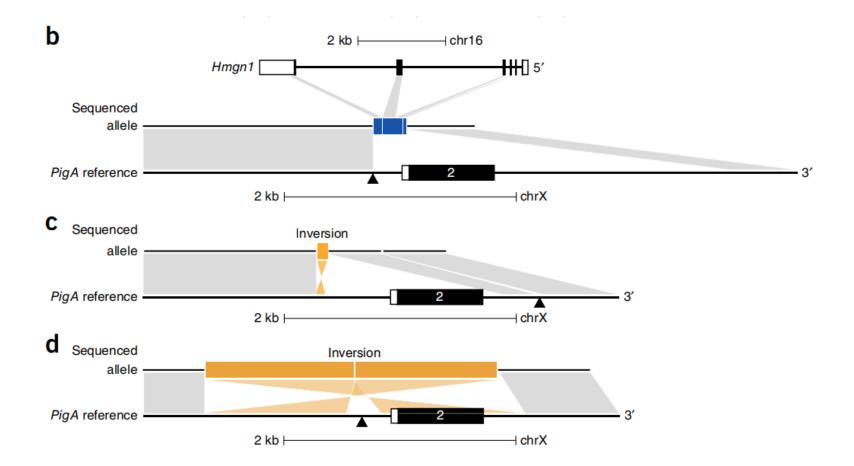
#### CRISPR-Cas9 induced SVs



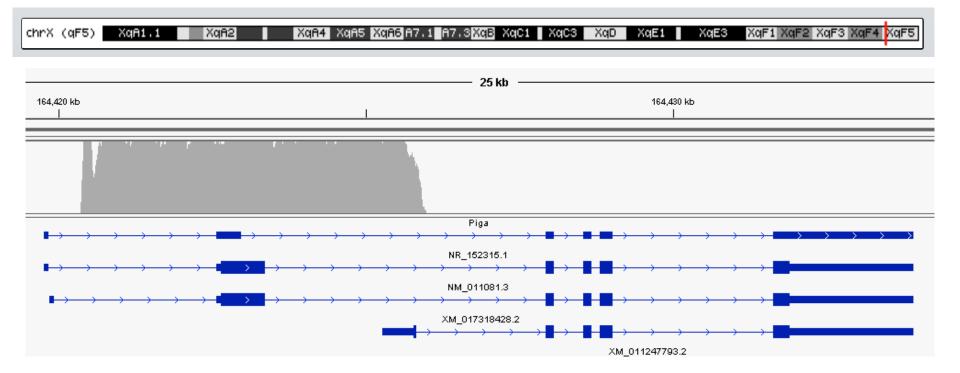
#### CRISPR-Cas9 induced SVs



#### CRISPR-Cas9 induced SVs



# Targeted sequencing of PigA gene



# SV\_Dataset

#### 🕨 🚞 Instructions

- Ground\_Truth.txt
   ILL\_sample.txt
   PAC\_sample.txt
   PIPE1\_tutorial.html
   PIPE1\_workflow.html
- PIPE2\_tutorial.html
- PIPE2\_workflow.html
- PIPE3\_workflow.html

#### 🕨 🚞 PIPE1

for\_the\_braves
 adapters.fa
 PIPE1\_fastq.tar.gz

#### 🗎 🗎 PIPE2

For\_the\_braves
PIPE2\_fastq.tar.gz

#### 🕨 💼 Ref\_genome

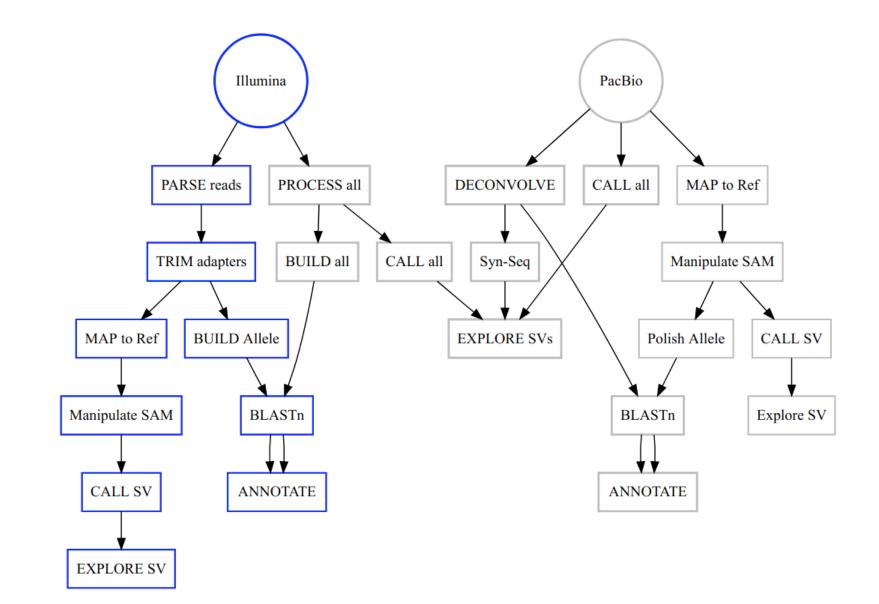
GRCm38.p6\_ChrX.fa

ref\_GRCm38.p6\_top\_level.gff3.gz

# PIPE1: Illumina PE-reads

sample_ID	target	expr	gRNA	library	reads	COV barcode
PIGA00ILL001	NA	POS	sham	SC	1866	52.1 CTGCGTGCTCTACGAC
PIGA56ILL070	EXON	NEG	g56	SC	1930	54.5 CATAGCGACTATCGTG
PIGA10ILL066	INTRON	POS	g10	SC	2070	58.4 GCTCGACTGTGAGAGA
PIGA15ILL071	INTRON	POS	g15	SC	1740	48.7 ACTCTCGCTCTGTAGA
PIGA15ILL010	INTRON	NEG	g15	SC	1638	65.2 CAGTGAGAGCGCGATA
PIGA15ILL003	INTRON	NEG	g15	SC	1826	70.2 GCAGACTCTCACACGC
PIGA15ILL012	INTRON	NEG	g15	SC	1832	80.7 GTGTGAGATATATATC
PIGA15ILL017	INTRON	NEG	g15	SC	2050	76.6 GACAGCATCTGCGCTC
PIGA15ILL031	INTRON	NEG	g15	SC	1822	63.4 TACTAGAGTAGCACTC
PIGA15ILL050	INTRON	NEG	g15	SC	960	52.8 TGTGTATCAGTACATG
PIGA15ILL059	INTRON	NEG	g15	SC	1500	53.7 ACACGCATGACACACT
PIGA15ILL062	INTRON	NEG	g15	SC	1424	47.0 GATCTCTACTATATGC
PIGA15ILL067	INTRON	NEG	g15	SC	2356	70.3 ACAGTCTATACTGCTG
PIGA15ILL069	INTRON	NEG	g15	SC	2334	72.4 ATGATGTGCTACATCT
PIGA15ILLPOL	INTRON	NEG	g15	poly	16416	442.6 TGCTCGCAGTATCACA

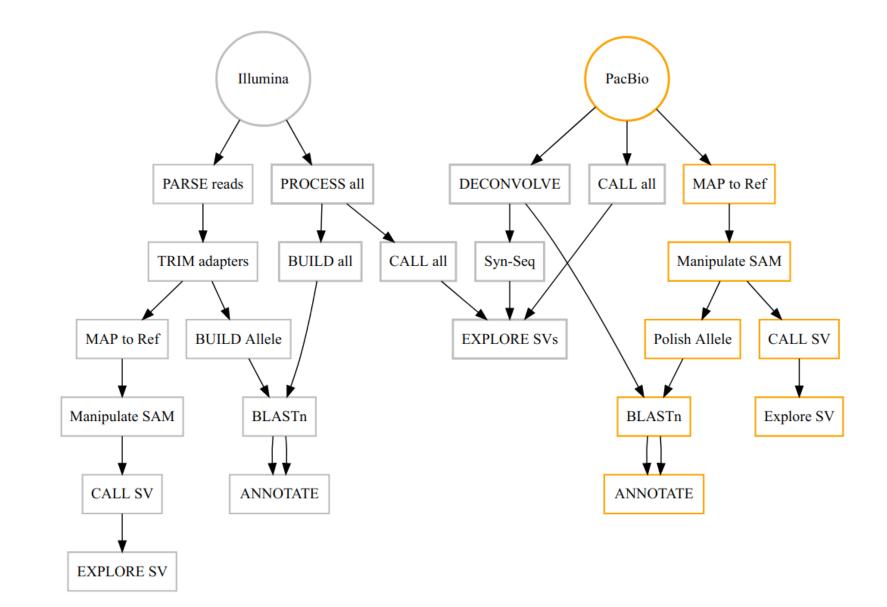
## **PIPE1: workflow**



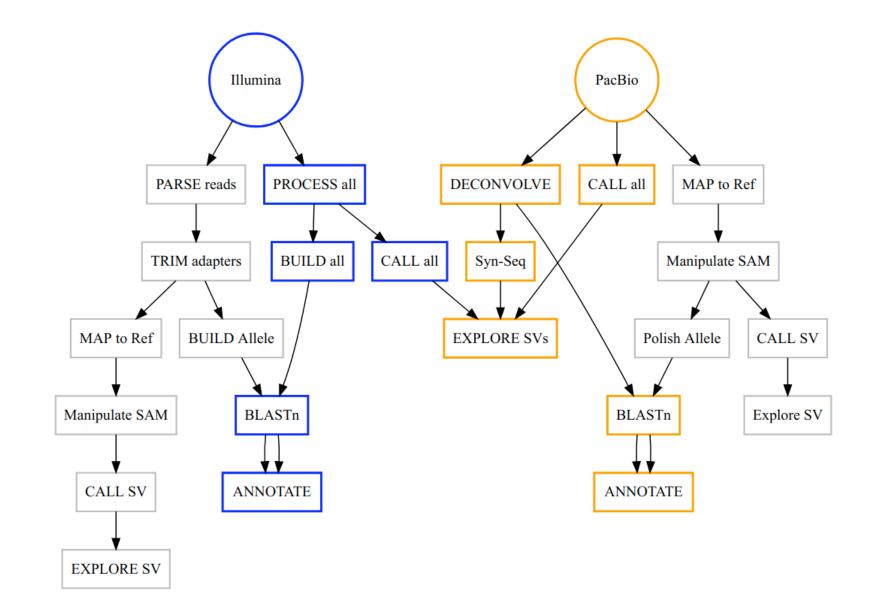
# PIPE2: PacBio FLNC reads

sample_ID	target	expr	gRNA	library	reads
PIGA00PAC001	NA	POS	sham	SC	21
PIGA56PAC070	EXON	NEG	g56	SC	21
PIGA10PAC066	INTRON	POS	g10	SC	21
PIGA15PAC071	INTRON	POS	g15	SC	21
PIGA15PAC010	INTRON	NEG	g15	SC	21
PIGA15PAC003	INTRON	NEG	g15	SC	21
PIGA15PAC012	INTRON	NEG	g15	SC	21
PIGA15PAC017	INTRON	NEG	g15	SC	21
PIGA15PAC031	INTRON	NEG	g15	SC	21
PIGA15PAC050	INTRON	NEG	g15	SC	21
PIGA15PAC059	INTRON	NEG	g15	SC	21
PIGA15PAC062	INTRON	NEG	g15	SC	21
PIGA15PAC067	INTRON	NEG	g15	SC	21
PIGA15PAC069	INTRON	NEG	g15	SC	21
PIGA15PACPOL	INTRON	NEG	g15	poly	210

## **PIPE2: workflow**



### **PIPE3: workflow**



#### Variants Deconvolution

