Personalized Phased Diploid Genomes of the EN-TEx Samples

Michael Schatz, Fritz Sedlazeck, Han Fang, Maria Nattestad, Ruibang Luo, Srividya Ramakrishnan, Charlotte Darby, Philipp Rescheneder, Alex Dobin, Carrie Davis, Ashwin Prakash, Anna Vlasova, Alessandra Breschi, Roderic Guigo, Tom Gingeras

Oct 2, 2017

ENTex Update

CSH

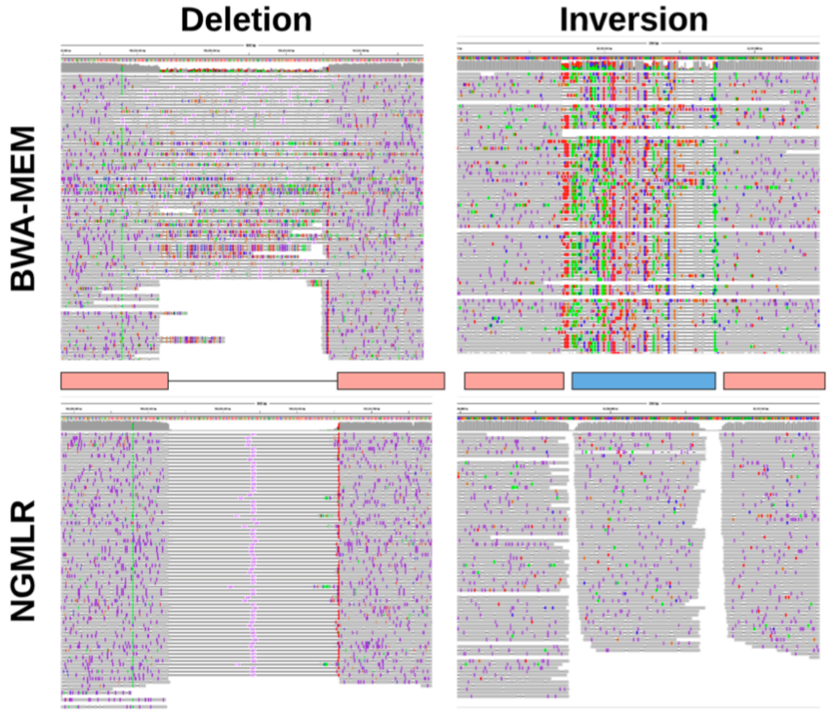
CRG
Centre
de Regulació
Genòmica

NGMLR and Sniffles

"NGMLR + Sniffles paper ": Accurate detection of complex structural variations using single molecule sequencing
Sedlazeck, FJ *et al.* (2017) *bioRxiv* https://doi.org/10.1101/169557

"SKBR3 paper": Complex rearrangements and oncogene amplifications revealed by long-read DNA and RNA sequencing of a breast cancer cell line
Nattestad, M *et al.* (2017) bioRxiv https://doi.org/10.1101/174938

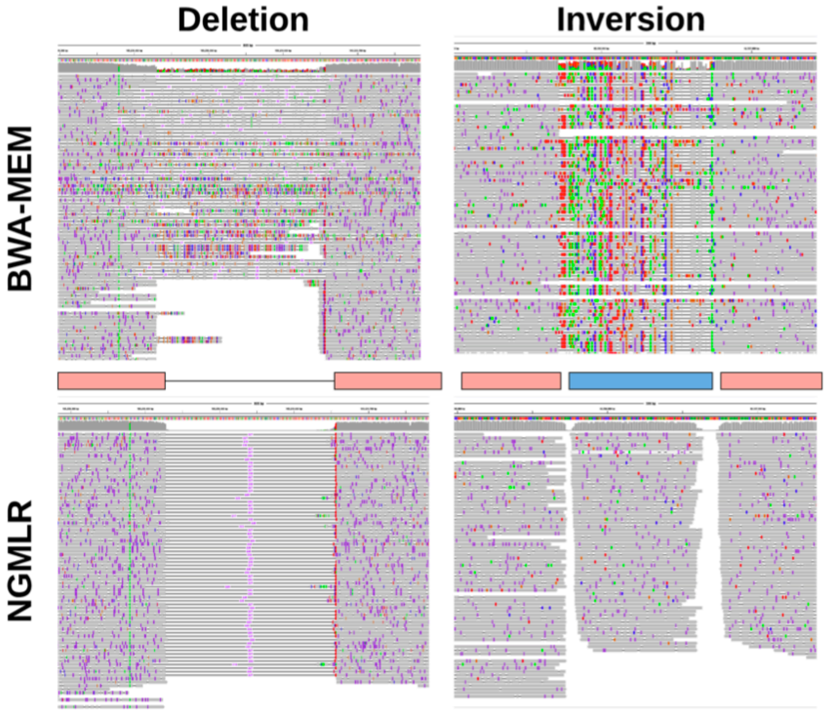
NGMLR and Sniffles

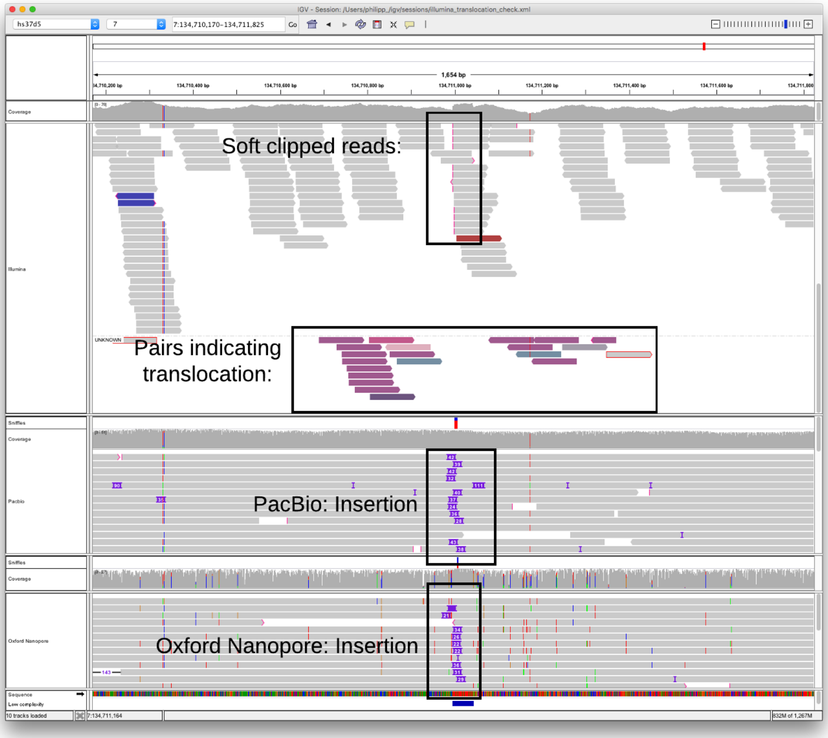


“NGMLR + Sniffles paper ”: Accurate detection of complex structural variations using single molecule sequencing
Sedlazeck, FJ *et al.* (2017) *bioRxiv* https://doi.org/10.1101/169557

”SKBR3 paper”: Complex rearrangements and oncogene amplifications revealed by long-read DNA and RNA sequencing of a breast cancer cell line
Nattestad, M *et al.* (2017) bioRxiv https://doi.org/10.1101/174938

NGMLR and Sniffles





Soft clipped reads:

Pairs indicating translocation:

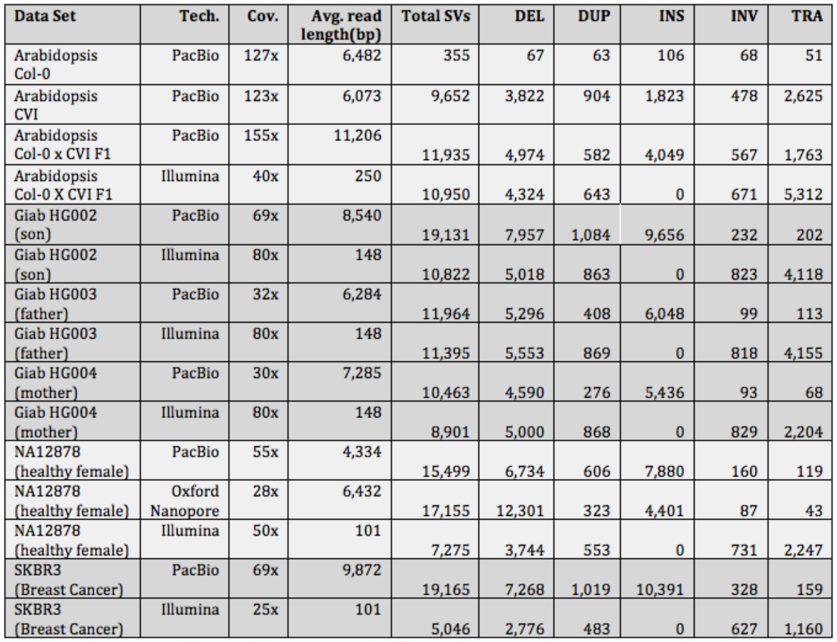
PacBio: Insertion

Oxford Nanopore: Insertion

"NGMLR + Sniffles paper ": Accurate detection of complex structural variations using single molecule sequencing
Sedlazeck, FJ *et al.* (2017) *bioRxiv* https://doi.org/10.1101/169557

"SKBR3 paper": Complex rearrangements and oncogene amplifications revealed by long-read DNA and RNA sequencing of a breast cancer cell line
Nattestad, M *et al.* (2017) bioRxiv https://doi.org/10.1101/174938

NGMLR and Sniffles



Data Set	Tech.	Cov.	Avg. read length(bp)	Total SVs	DEL	DUP	INS	INV	TRA
Arabidopsis Col-0	PacBio	127x	6,482	355	67	63	106	68	51
Arabidopsis CVI	PacBio	123x	6,073	9,652	3,822	904	1,823	478	2,625
Arabidopsis Col-0 x CVI F1	PacBio	155x	11,206	11,935	4,974	582	4,049	567	1,763
Arabidopsis Col-0 X CVI F1	Illumina	40x	250	10,950	4,324	643	0	671	5,312
Giab HG002 (son)	PacBio	69x	8,540	19,131	7,957	1,084	9,656	232	202
Giab HG002 (son)	Illumina	80x	148	10,822	5,018	863	0	823	4,118
Giab HG003 (father)	PacBio	32x	6,284	11,964	5,296	408	6,048	99	113
Giab HG003 (father)	Illumina	80x	148	11,395	5,553	869	0	818	4,155
Giab HG004 (mother)	PacBio	30x	7,285	10,463	4,590	276	5,436	93	68
Giab HG004 (mother)	Illumina	80x	148	8,901	5,000	868	0	829	2,204
NA12878 (healthy female)	PacBio	55x	4,334	15,499	6,734	606	7,880	160	119
NA12878 (healthy female)	Oxford Nanopore	28x	6,432	17,155	12,301	323	4,401	87	43
NA12878 (healthy female)	Illumina	50x	101	7,275	3,744	553	0	731	2,247
SKBR3 (Breast Cancer)	PacBio	69x	9,872	19,165	7,268	1,019	10,391	328	159
SKBR3 (Breast Cancer)	Illumina	25x	101	5,046	2,776	483	0	627	1,160

Table 1: Summary of detected SVs across 15 different data sets. SVs were reported with a min. size of 50bp using SURVIVOR based on Delly, Lumpy and Manta for Illumina or Sniffles for PacBio or Oxford Nanopore requiring at least 10 reads. Supplementary Table 5 shows all the data sets used.

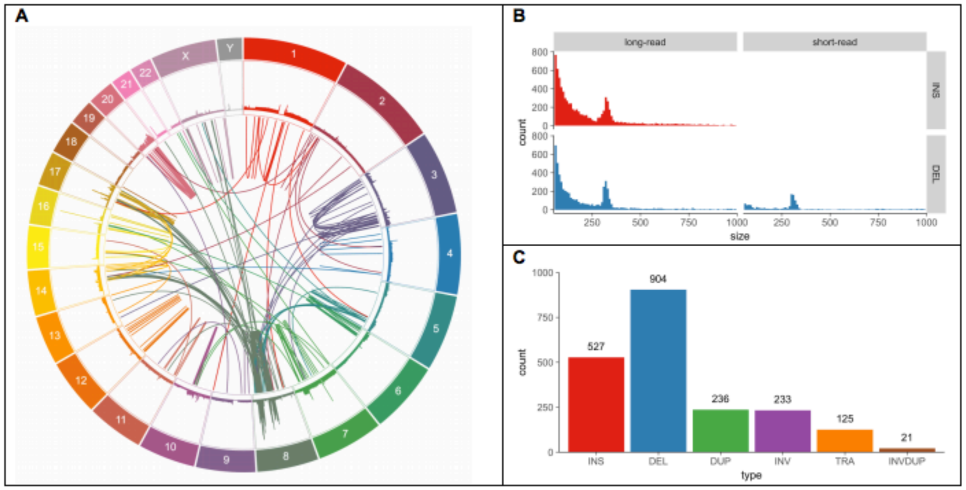


Figure 1 | Variants found in SK-BR-3 with PacBio long-read sequencing. (A) Circos plot showing long-range (larger than 10 kbp or interchromosomal) variants found by Sniffles from split-read alignments, with read coverage shown in the outer track. (B) Variant size histogram of deletions and insertions from size 50 bp up to 1 kbp found by log-read (Sniffles) and short-read (Survivor 2-caller consensus) variant-calling, showing similar size distributions for insertions and deletions from long reads but not for short reads where insertions are entirely missing. (C) Sniffles variant counts by type for variants above 1 kbp in size, including translocations and inverted duplications.

"NGMLR + Sniffles paper ": Accurate detection of complex structural variations using single molecule sequencing
Sedlazeck, FJ *et al.* (2017) *bioRxiv* https://doi.org/10.1101/169557

"SKBR3 paper": Complex rearrangements and oncogene amplifications revealed by long-read DNA and RNA sequencing of a breast cancer cell line
Nattestad, M *et al.* (2017) bioRxiv https://doi.org/10.1101/174938

Structural Variations Concordance (ENC-002)

	Sniffles	Falcon	LongRanger	SuperNova	SURVIVOR2	MegaHit
Sniffles	17,107					
Falcon	7,857	12,241				
LongRanger	2,823	1,946	3,785			
SuperNova	3,394	2,837	1,486	18,862		
SURVIVOR2	3,291	2,163	2,274	1,646	6,631	
MegaHit	1,858	1,529	569	1,378	687	3,855

PacBio

10X Genomics

Illumina

Main Diagonal

- Calls per tool

Outer triplets

- Concordance by Technology

Inner triplets

- Concordance by Assembly
- Concordance by Mappers

Overall:

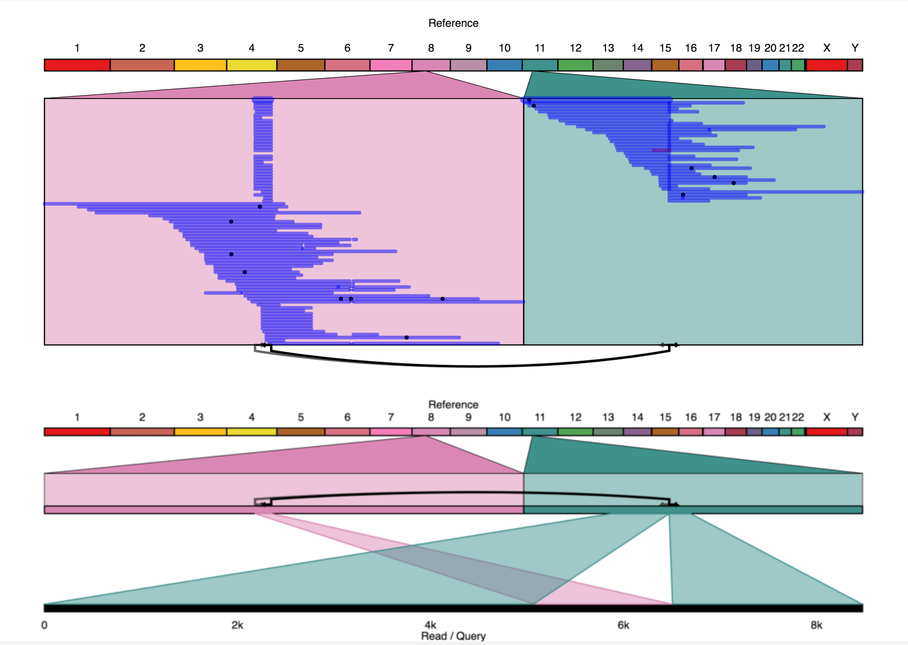
- We need multiple technologies and approaches

Sniffles PacBio Variant Calls (ENC-002)

Sniffles calls

	All SVs (50bp+)	Large SVs (10kbp+)
Deletions	7,124	163
Duplications	1,642	153
Insertions	7,904	0
Inversions	275	144
Translocations	162	162
All	***17,107***	***622***

Translocation in Ribbon



Ribbon: Visualizing complex genome alignments and structural variation
Nattestad et al. (2016) *bioRxiv* doi: http://dx.doi.org/10.1101/082123

Sniffles PacBio Variant Calls

ENC-002

	All SVs (50bp+)	Large SVs (10kbp+)
Deletions	7,124	163
Duplications	1,642	153
Insertions	7,904	0
Inversions	275	144
Translocations	162	162
All	***17,107***	***622***

ENC-003

	All SVs (50bp+)	Large SVs (10kbp+)
Deletions	7,747	128
Duplications	1,511	116
Insertions	9,528	0
Inversions	224	106
Translocations	101	101
All	***19,111***	***451***

Current Calls: 50bp+ & supported by 10+ reads (High Confidence)
Sensitive Analysis: 10bp+ event & supported by 5+ reads

Variants Per Chromosome

ENC-002

Chr	DEL	DUP	INV	INS	TRA
1	459	106	13	606	19
2	550	116	10	576	21
3	376	50	7	453	16
4	526	50	9	451	7
5	400	69	13	383	4
6	443	88	5	502	7
7	441	103	10	445	16
8	372	75	2	388	7
9	242	50	15	340	13
10	372	84	10	408	10
11	285	72	16	397	4
12	315	64	16	405	7
13	252	56	3	319	8
14	179	26	1	210	1
15	168	27	1	212	1
16	358	226	115	250	2
17	271	63	2	273	2
18	204	46	3	219	0
19	215	42	6	255	0
20	199	88	7	244	8
21	161	44	3	167	7
22	124	44	5	156	0
X	197	41	2	201	2
Y	15	11	0	29	0
M	0	1	1	15	0

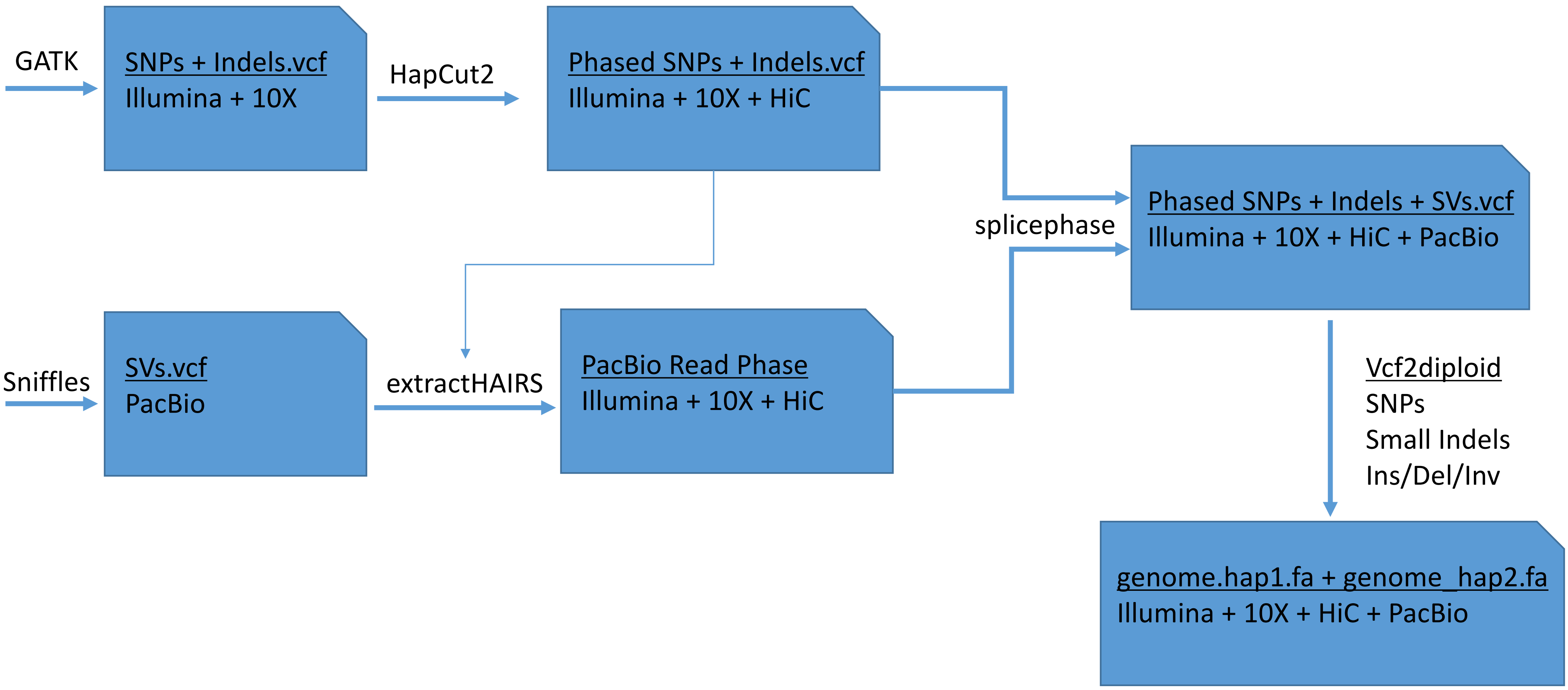
Raw Coverage: 165,114,151,916bp / 3Gb = 55.0x
Raw Length: 7538 +/- 5610 bp
Mean alignment length: 5257 +/- 5457 bp

ENC-003

Chr	DEL	DUP	INV	INS	TRA
1	480	91	11	727	10
2	609	125	9	686	7
3	390	46	6	560	9
4	525	47	6	557	2
5	439	62	13	461	4
6	490	75	7	594	3
7	507	105	9	546	9
8	372	73	7	482	4
9	286	49	19	397	12
10	409	78	9	450	1
11	316	71	11	439	4
12	344	49	9	451	1
13	294	57	3	389	11
14	182	31	3	244	1
15	189	21	0	235	1
16	347	163	67	266	2
17	261	71	5	352	1
18	232	41	2	283	0
19	262	54	3	302	0
20	206	78	12	306	8
21	157	45	4	188	8
22	168	37	1	188	1
X	273	39	8	403	2
Y	9	2	0	18	0
M	0	1	0	4	0

Raw coverage: 171,860,416,260bp / 3Gb = 57.2x
Raw length: 6974 +/- 5854 bp
Mean alignment length: 5682 +/- 4673 bp

Diploid Construction



GATK

SNPs + Indels.vcf
Illumina + 10X

HapCut2

Phased SNPs + Indels.vcf
Illumina + 10X + HiC

Sniffles

SVs.vcf
PacBio

extractHAIRS

PacBio Read Phase
Illumina + 10X + HiC

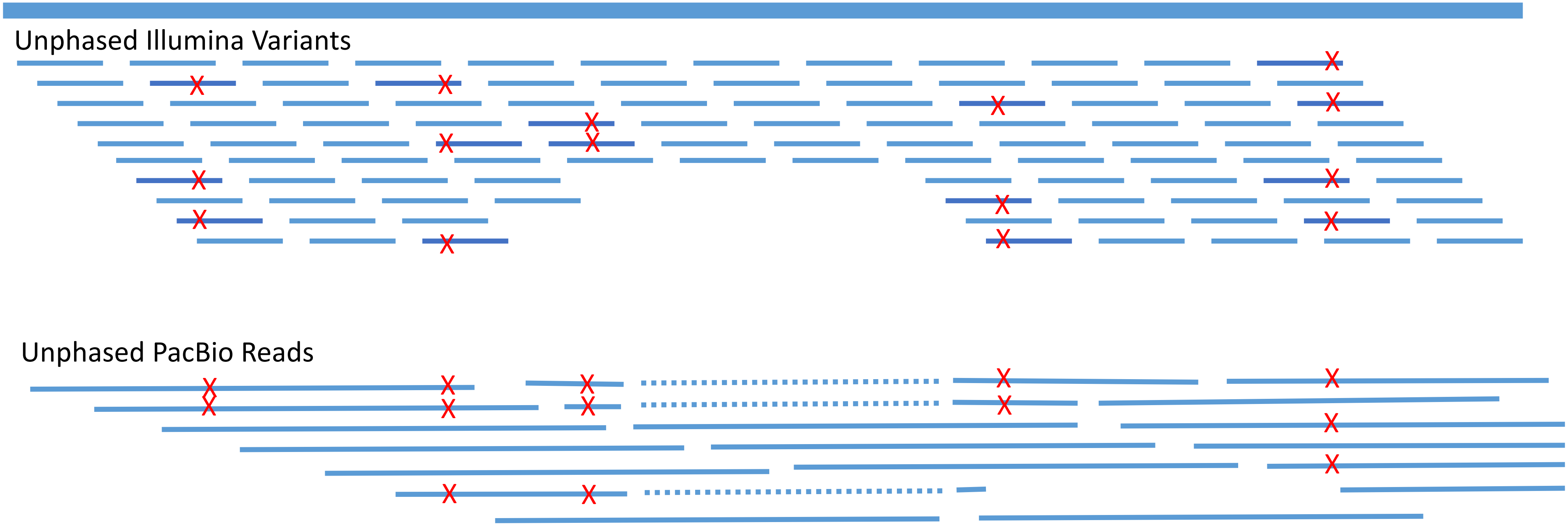
splicephase

Phased SNPs + Indels + SVs.vcf
Illumina + 10X + HiC + PacBio

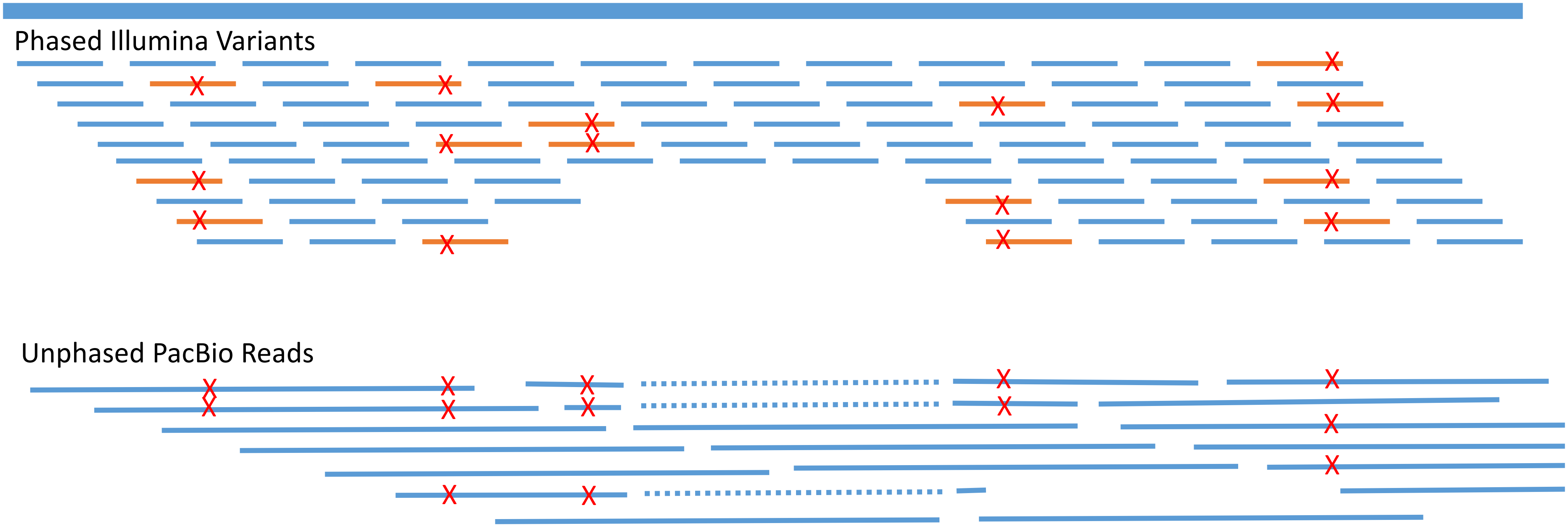
Vcf2diploid
SNPs
Small Indels
Ins/Del/Inv

genome.hap1.fa + genome_hap2.fa
Illumina + 10X + HiC + PacBio

SVPhaser



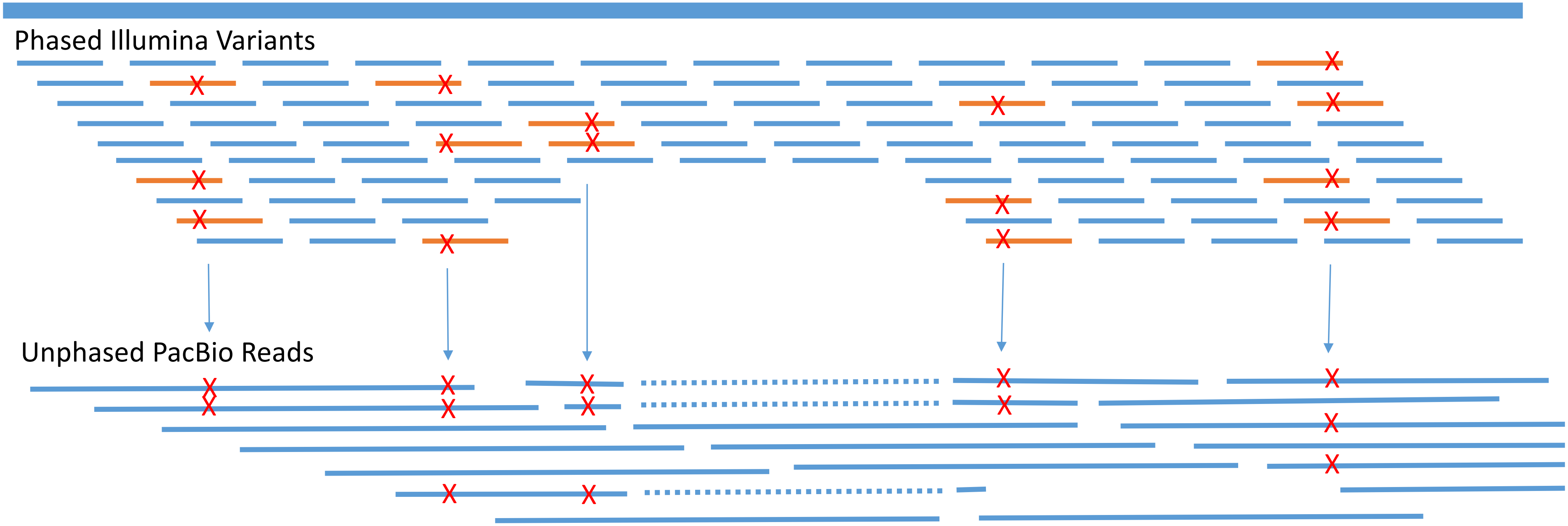
SVPhaser



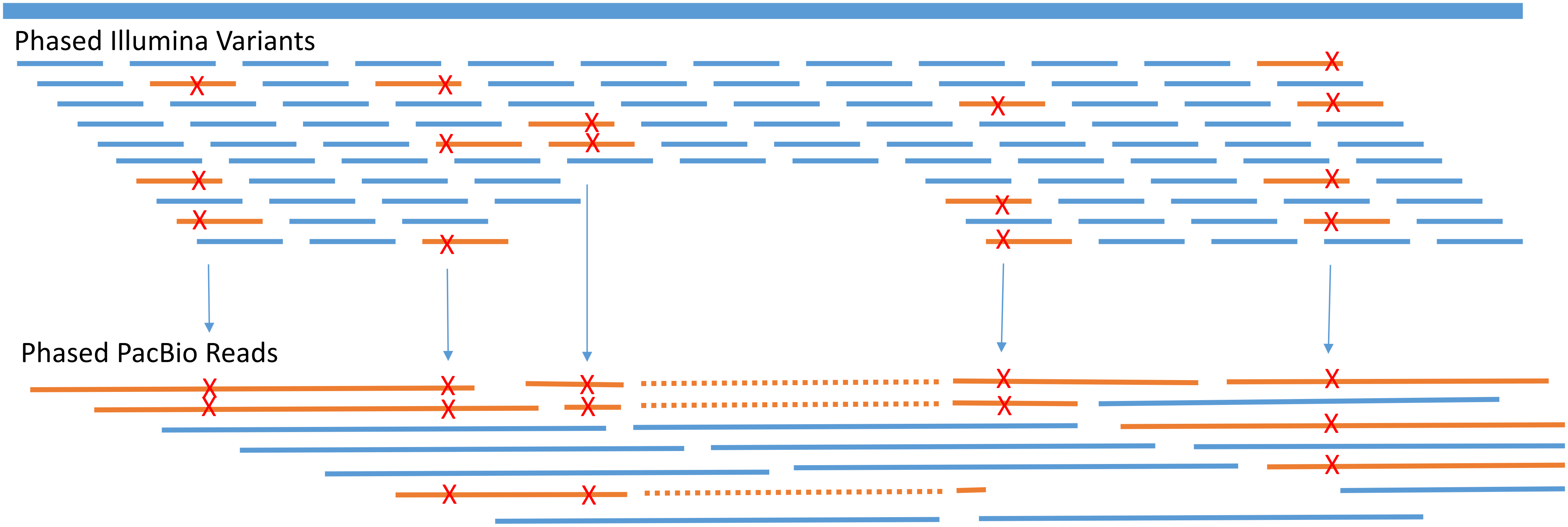
Phased Illumina Variants

Unphased PacBio Reads

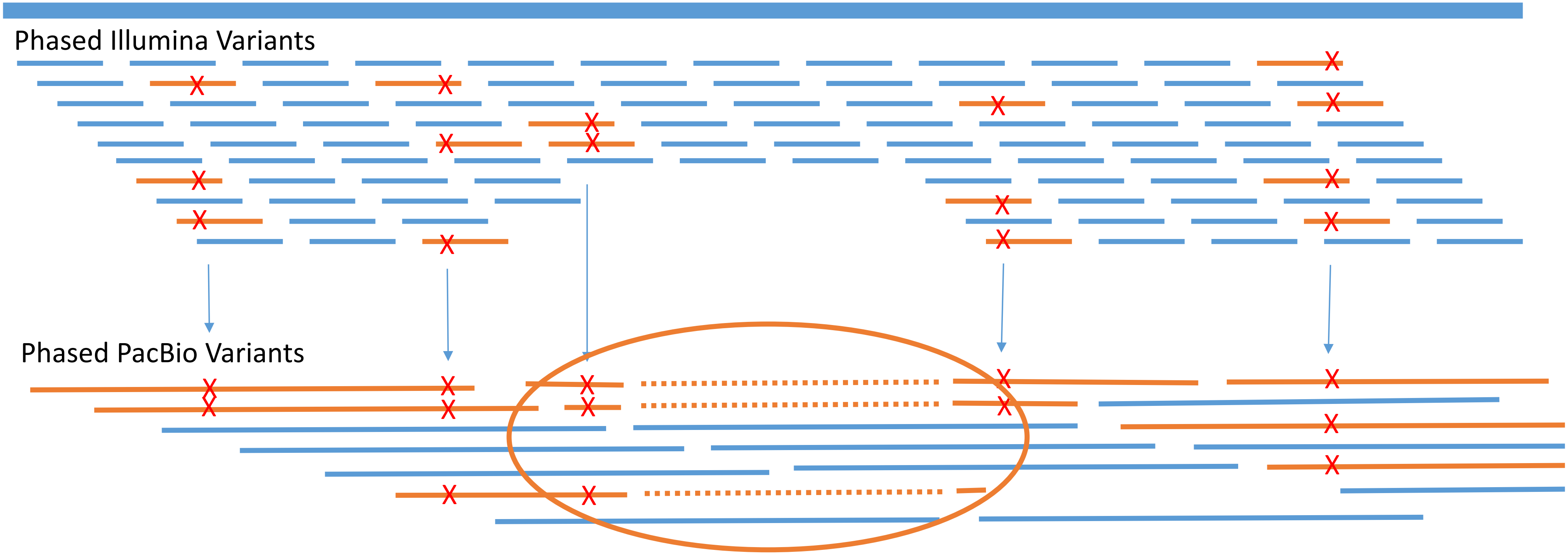
SVPhaser



SVPhaser

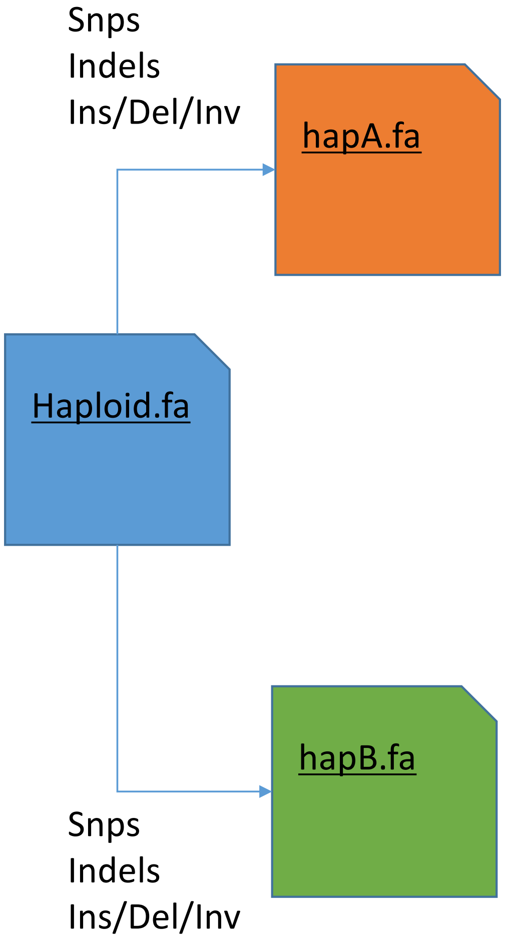


SVPhaser



Deletion must be on the orange haplotype!

SVPhaser Testing



Snps
Indels
Ins/Del/Inv

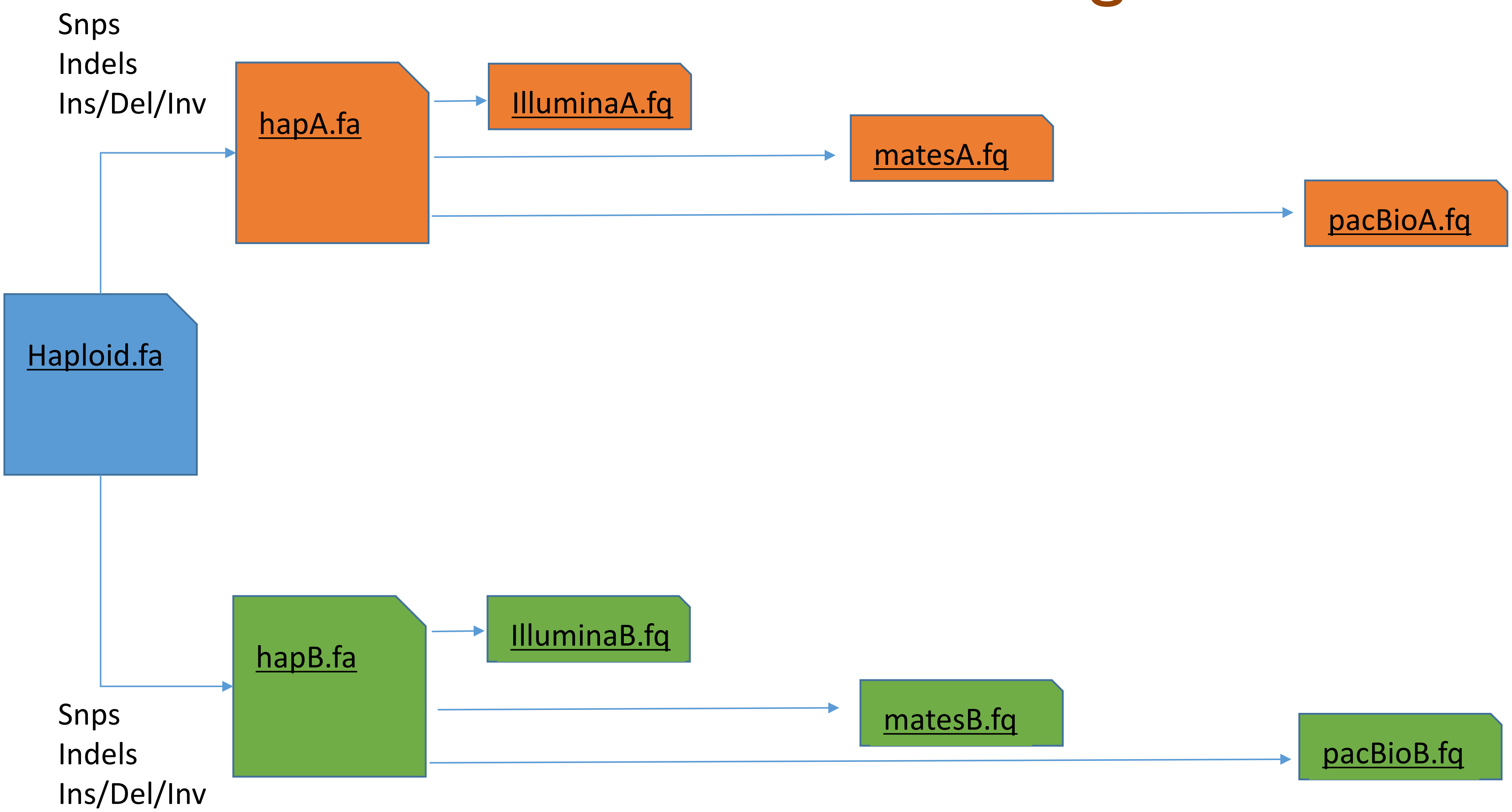
hapA.fa

Haploid.fa

hapB.fa

Snps
Indels
Ins/Del/Inv

SVPhaser Testing



Snps
Indels
Ins/Del/Inv

hapA.fa

IlluminaA.fq

matesA.fq

pacBioA.fq

Haploid.fa

hapB.fa

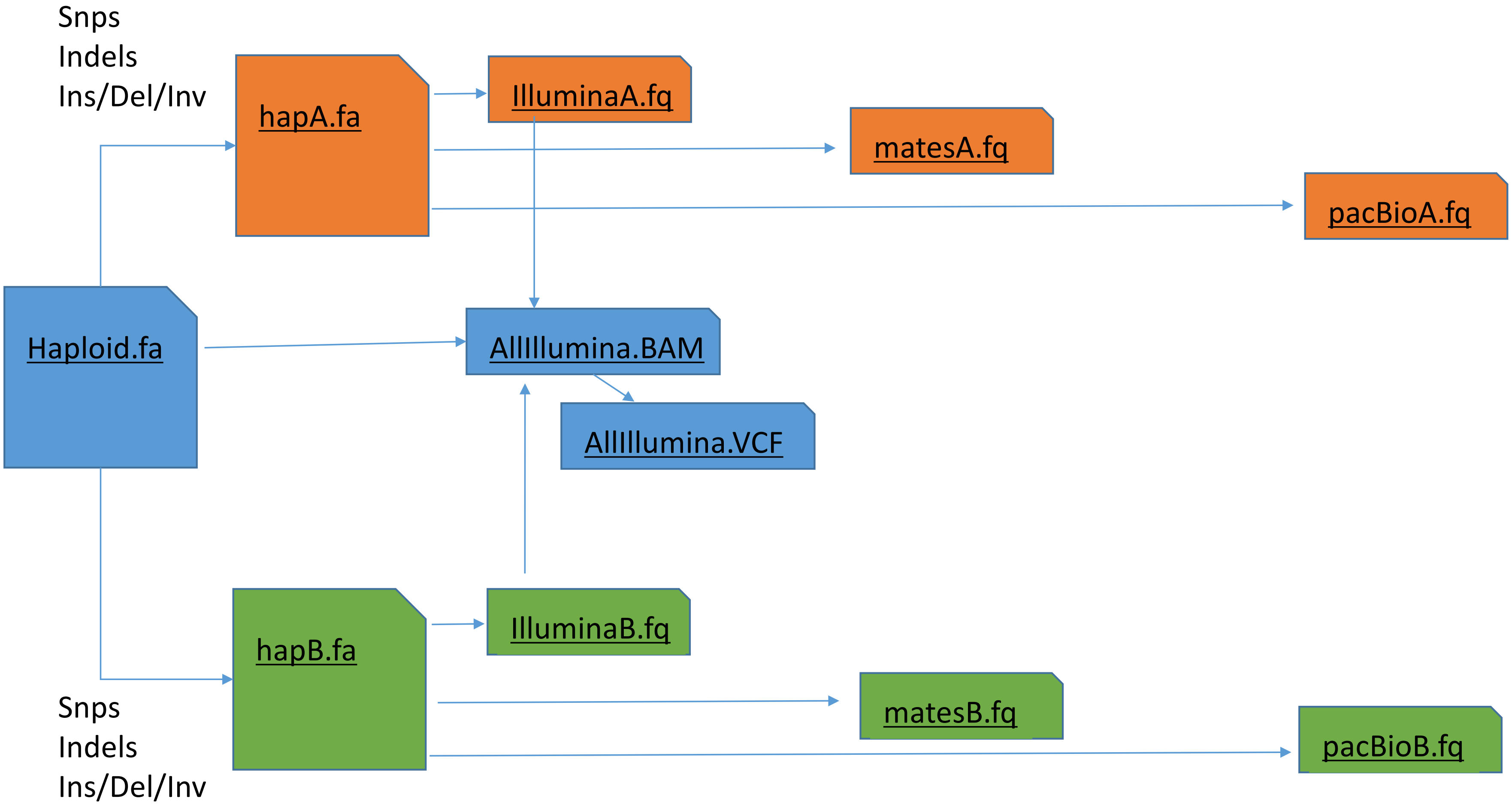
IlluminaB.fq

matesB.fq

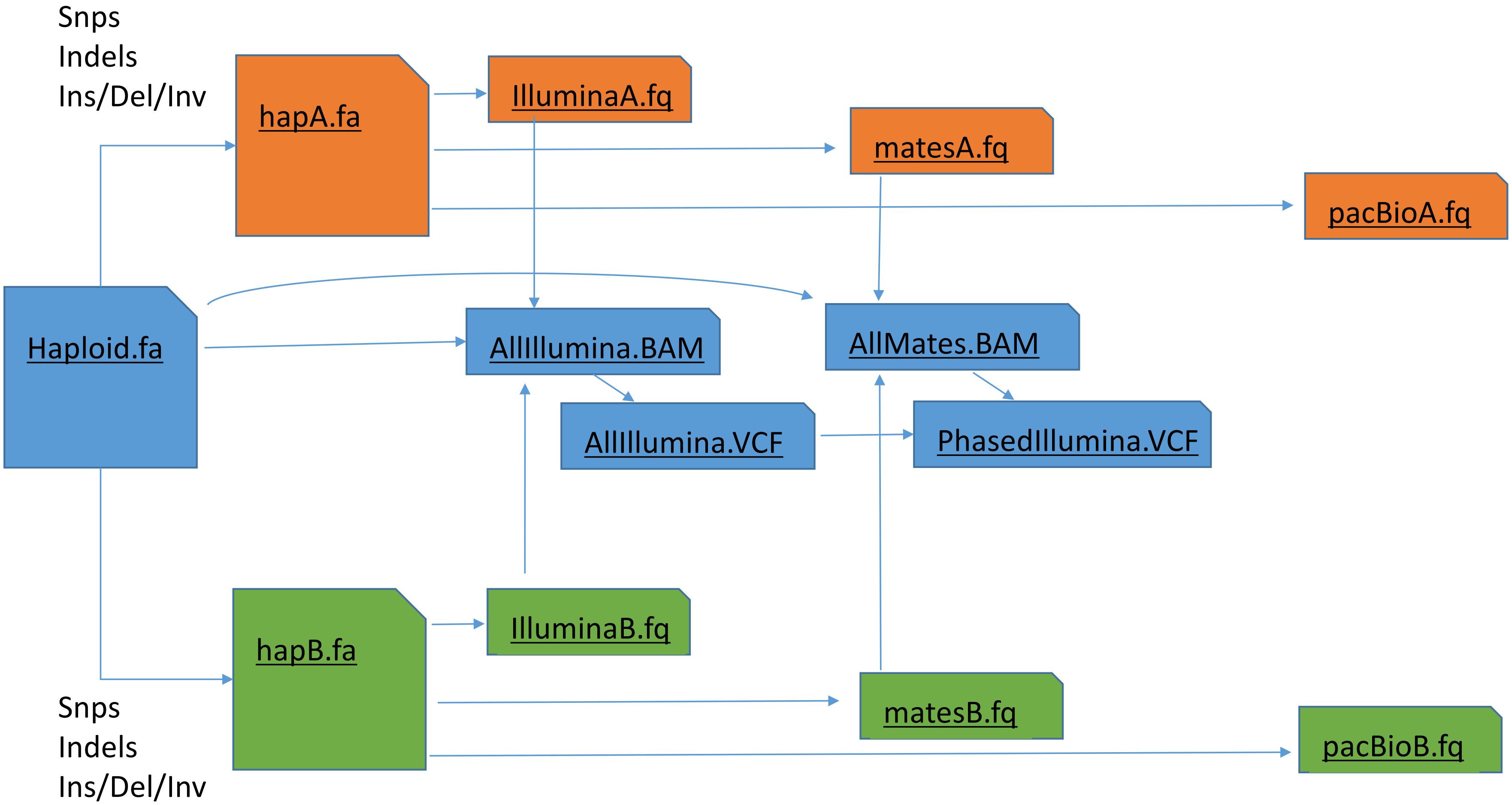
pacBioB.fq

Snps
Indels
Ins/Del/Inv

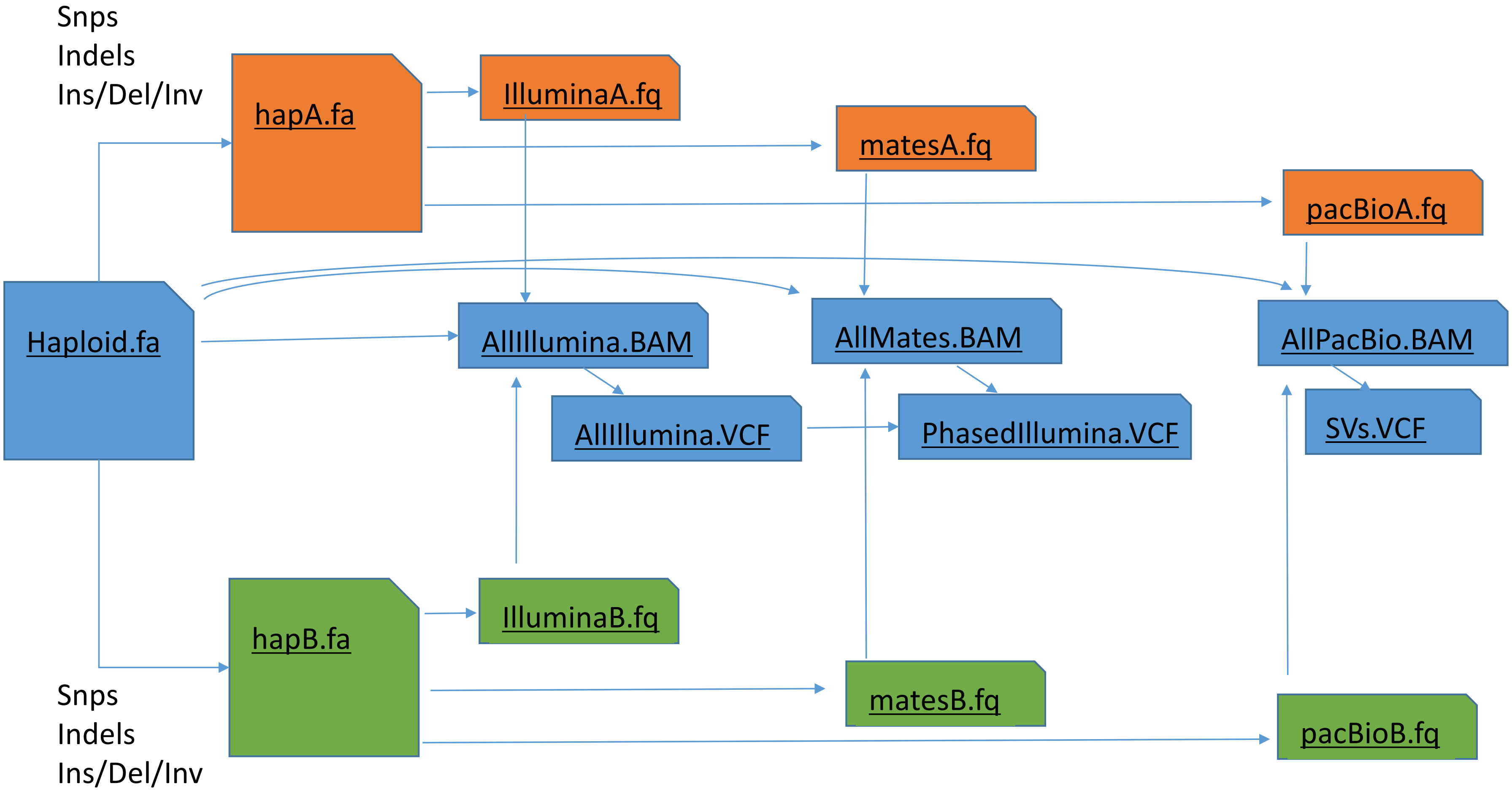
SVPhaser Testing



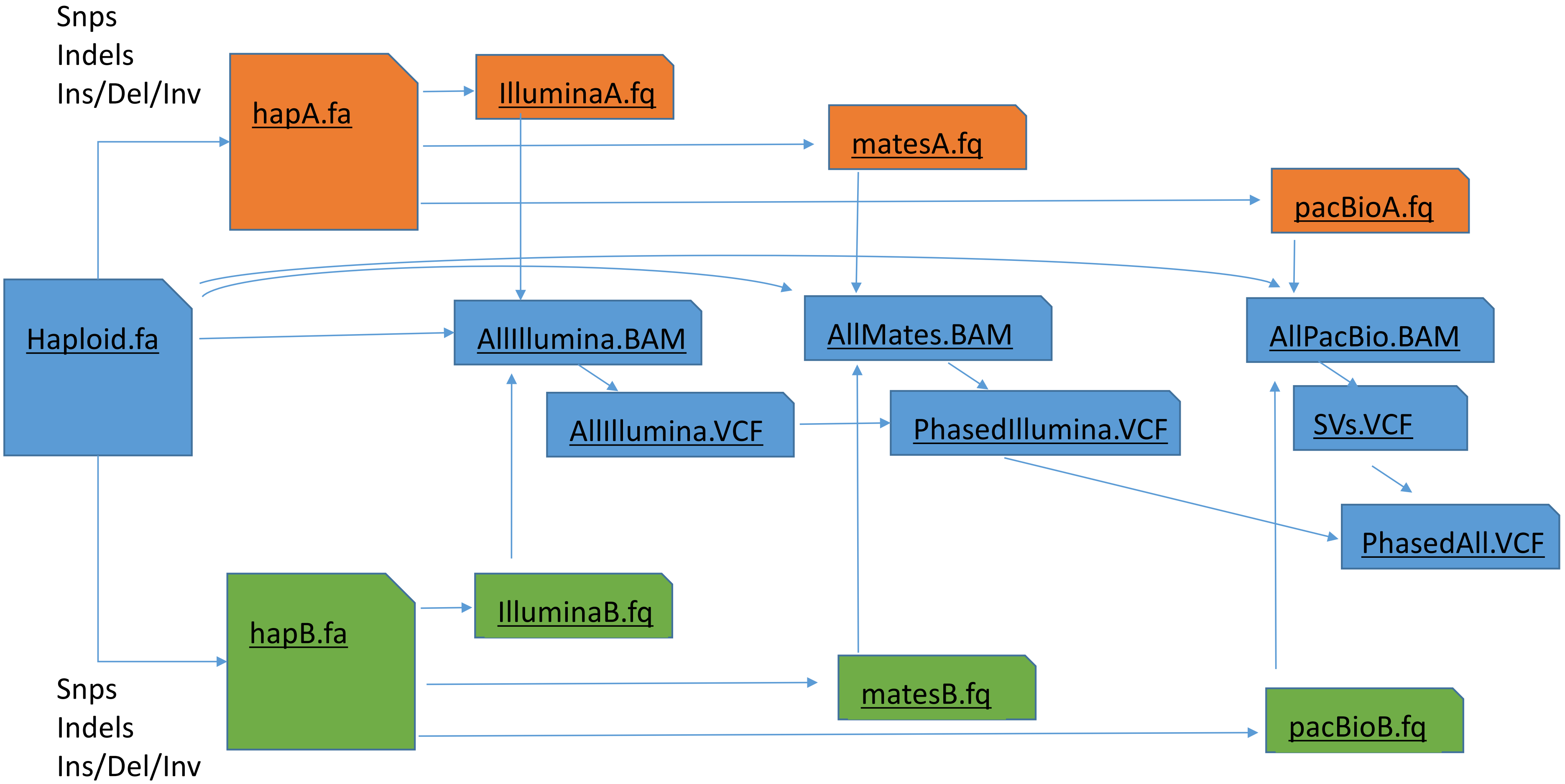
SVPhaser Testing



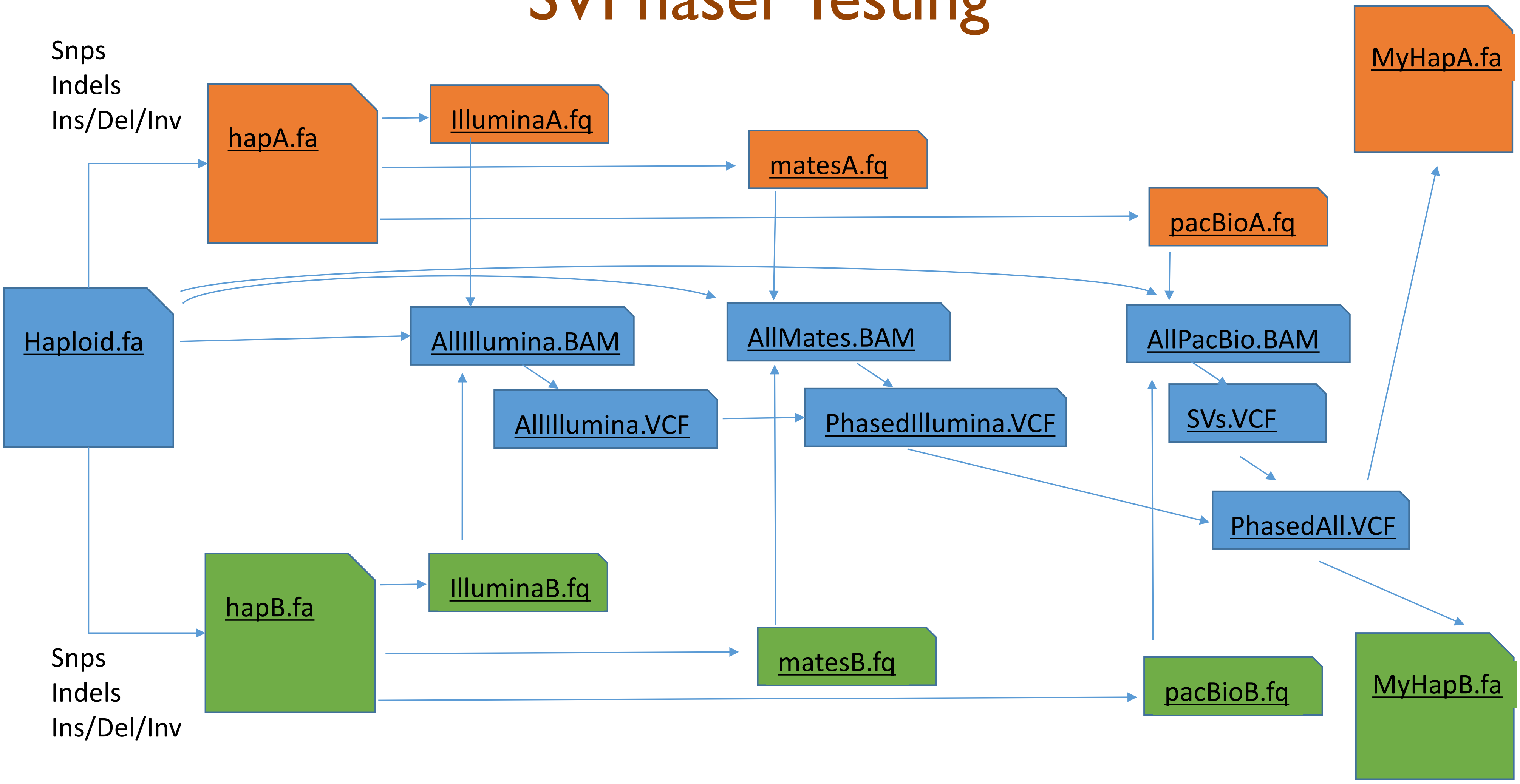
SVPhaser Testing



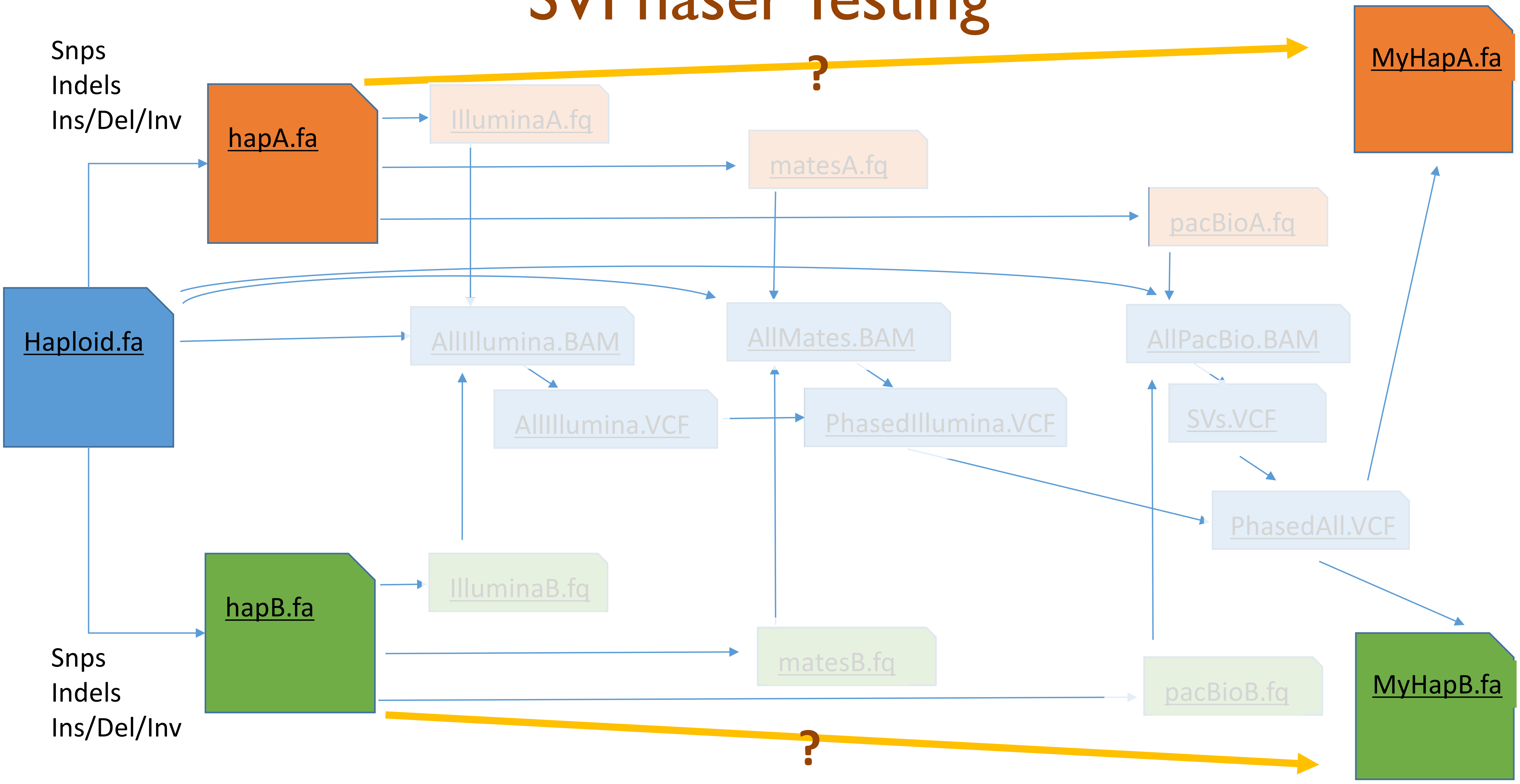
SVPhaser Testing



SVPhaser Testing



SVPhaser Testing



Snps
Indels
Ins/Del/Inv

hapA.fa

IlluminaA.fq

matesA.fq

pacBioA.fq

?

MyHapA.fa

Haploid.fa

AllIllumina.BAM

AllMates.BAM

AllPacBio.BAM

AllIllumina.VCF

PhasedIllumina.VCF

SVs.VCF

PhasedAll.VCF

hapB.fa

IlluminaB.fq

matesB.fq

pacBioB.fq

MyHapB.fa

Snps
Indels
Ins/Del/Inv

?

SVPhaser Testing (standard AlleleSeq)

```
$ nucmer mutA.fasta chr1_hapA.fa

    [S1]     [E1]  |     [S2]     [E2]  |  [LEN 1]  [LEN 2]  |  [% IDY]  |  [LEN R]  [LEN Q]  |  [COV R]  [COV Q]  | [TAGS]
=========================================================================================================================
       1     6093  |        1     6093  |     6093     6093  |   100.00  |   705447   699930  |     0.86     0.87  | chr1             chr1_maternal
    7339    21137  |     6094    19892  |    13799    13799  |    99.99  |   705447   699930  |     1.96     1.97  | chr1             chr1_maternal
   22158    28972  |    19892    26706  |     6815     6815  |   100.00  |   705447   699930  |     0.97     0.97  | chr1             chr1_maternal
   28973    66028  |    28648    65703  |    37056    37056  |   100.00  |   705447   699930  |     5.25     5.29  | chr1             chr1_maternal
   66833    71776  |    65699    70641  |     4944     4943  |    99.94  |   705447   699930  |     0.70     0.71  | chr1             chr1_maternal
   71768    75348  |    71610    75190  |     3581     3581  |   100.00  |   705447   699930  |     0.51     0.51  | chr1             chr1_maternal
   75348   103011  |    76044   103707  |    27664    27664  |    99.98  |   705447   699930  |     3.92     3.95  | chr1             chr1_maternal
  103011   123468  |   104645   125102  |    20458    20458  |   100.00  |   705447   699930  |     2.90     2.92  | chr1             chr1_maternal
  124039   141409  |   125103   142473  |    17371    17371  |   100.00  |   705447   699930  |     2.46     2.48  | chr1             chr1_maternal
  142112   145387  |   142474   145749  |     3276     3276  |   100.00  |   705447   699930  |     0.46     0.47  | chr1             chr1_maternal
  146107   178148  |   145750   177791  |    32042    32042  |    99.99  |   705447   699930  |     4.54     4.58  | chr1             chr1_maternal
  178826   189042  |   177792   188008  |    10217    10217  |    99.97  |   705447   699930  |     1.45     1.46  | chr1             chr1_maternal
  189665   195023  |   188009   193367  |     5359     5359  |   100.00  |   705447   699930  |     0.76     0.77  | chr1             chr1_maternal
  195024   200386  |   195238   200600  |     5363     5363  |   100.00  |   705447   699930  |     0.76     0.77  | chr1             chr1_maternal
  201520   231233  |   200601   230315  |    29714    29715  |    99.99  |   705447   699930  |     4.21     4.25  | chr1             chr1_maternal
  231222   268940  |   232037   269755  |    37719    37719  |   100.00  |   705447   699930  |     5.35     5.39  | chr1             chr1_maternal
  268941   273282  |   271210   275551  |     4342     4342  |    99.98  |   705447   699930  |     0.62     0.62  | chr1             chr1_maternal
  273282   275274  |   276293   278285  |     1993     1993  |   100.00  |   705447   699930  |     0.28     0.28  | chr1             chr1_maternal
  276698   292589  |   278286   294177  |    15892    15892  |   100.00  |   705447   699930  |     2.25     2.27  | chr1             chr1_maternal
  294512   306412  |   294178   306078  |    11901    11901  |   100.00  |   705447   699930  |     1.69     1.70  | chr1             chr1_maternal
  307636   313944  |   306078   312386  |     6309     6309  |   100.00  |   705447   699930  |     0.89     0.90  | chr1             chr1_maternal
  313942   323954  |   313941   323953  |    10013    10013  |   100.00  |   705447   699930  |     1.42     1.43  | chr1             chr1_maternal
  325156   327294  |   323952   326090  |     2139     2139  |   100.00  |   705447   699930  |     0.30     0.31  | chr1             chr1_maternal
  327987   340849  |   326090   338952  |    12863    12863  |   100.00  |   705447   699930  |     1.82     1.84  | chr1             chr1_maternal
  340843   345937  |   340468   345562  |     5095     5095  |   100.00  |   705447   699930  |     0.72     0.73  | chr1             chr1_maternal
  347851   361244  |   345563   358956  |    13394    13394  |   100.00  |   705447   699930  |     1.90     1.91  | chr1             chr1_maternal
  361838   378128  |   358957   375247  |    16291    16291  |   100.00  |   705447   699930  |     2.31     2.33  | chr1             chr1_maternal
  379230   427511  |   375248   423529  |    48282    48282  |    99.99  |   705447   699930  |     6.84     6.90  | chr1             chr1_maternal
  427511   447306  |   424063   443858  |    19796    19796  |   100.00  |   705447   699930  |     2.81     2.83  | chr1             chr1_maternal
  447307   448221  |   445585   446499  |      915      915  |   100.00  |   705447   699930  |     0.13     0.13  | chr1             chr1_maternal
  449845   480347  |   446500   477002  |    30503    30503  |    99.99  |   705447   699930  |     4.32     4.36  | chr1             chr1_maternal
  480348   506343  |   478764   504759  |    25996    25996  |   100.00  |   705447   699930  |     3.69     3.71  | chr1             chr1_maternal
  507867   509684  |   504760   506577  |     1818     1818  |   100.00  |   705447   699930  |     0.26     0.26  | chr1             chr1_maternal
  509684   528583  |   507296   526195  |    18900    18900  |   100.00  |   705447   699930  |     2.68     2.70  | chr1             chr1_maternal
```

SVPhaser Testing (enhanced AlleleSeq)

```
$ nucmer mutA.fasta chr1_hapA.fa

    [S1]     [E1]  |     [S2]     [E2]  |  [LEN 1]  [LEN 2]  |  [% IDY]  |  [LEN R]  [LEN Q]  |  [COV R]  [COV Q]  | [TAGS]
===============================================================================================================================
       1   705447  |        1   706650  |   705447   706650  |    99.52  |   705447   706650  |   100.00   100.00  | chr1      chr1_maternal
```

```
$ nucmer mutB.fasta chr1_hapB.fa

    [S1]     [E1]  |     [S2]     [E2]  |  [LEN 1]  [LEN 2]  |  [% IDY]  |  [LEN R]  [LEN Q]  |  [COV R]  [COV Q]  | [TAGS]
===============================================================================================================================
       1   706496  |        1   707791  |   706496   707791  |    99.49  |   706496   707791  |   100.00   100.00  | chr1      chr1_paternal
```

SVPhaser Testing (after)

```
$ nucmer mutA.fasta chr1_hapA.fa

    [S1]     [E1]  |     [S2]     [E2]  |  [LEN 1]  [LEN 2]  |  [% IDY]  |  [LEN R]  [LEN Q]  |  [COV R]  [COV Q]  | [TAGS]
=================================================================================================================================
       1   705447  |        1   706650  |   705447   706650  |    99.52  |   705447   706650  |   100.00   100.00  | chr1	chr1_maternal
```

```
$ nucmer mutB.fasta chr1_hapB.fa

    [S1]     [E1]  |     [S2]     [E2]  |  [LEN 1]  [LEN 2]  |  [% IDY]  |  [LEN R]  [LEN Q]  |  [COV R]  [COV Q]  | [TAGS]
=================================================================================================================================
       1   706496  |        1   707791  |   706496   707791  |    99.49  |   706496   707791  |   100.00   100.00  | chr1	chr1_paternal
```

One end-to-end alignment per haplotype
All SVs correctly phased and inserted
☺

SVPhaser Testing (after)

```
$ nucmer mutA.fasta chr1_hapA.fa

    [S1]     [E1]  |     [S2]     [E2]  |  [LEN 1]  [LEN 2]  |  [% IDY]  |  [LEN R]  [LEN Q]  |  [COV R]  [COV Q]  | [TAGS]
===============================================================================================================================
       1   705447  |        1   706650  |   705447   706650  |    99.52  |   705447   706650  |   100.00   100.00  | chr1      chr1_maternal
```

```
$ nucmer mutB.fasta chr1_hapB.fa

    [S1]     [E1]  |     [S2]     [E2]  |  [LEN 1]  [LEN 2]  |  [% IDY]  |  [LEN R]  [LEN Q]  |  [COV R]  [COV Q]  | [TAGS]
===============================================================================================================================
       1   706496  |        1   707791  |   706496   707791  |    99.49  |   706496   707791  |   100.00   100.00  | chr1      chr1_paternal
```

Small amount of residual differences (0.5%)

- Nearly impossible to have 100% perfect SNP calling
- A small percent of insertions have misreported sequence
  - -> Fix is in progress

One end-to-end alignment per haplotype
All SVs correctly phased and inserted
☺