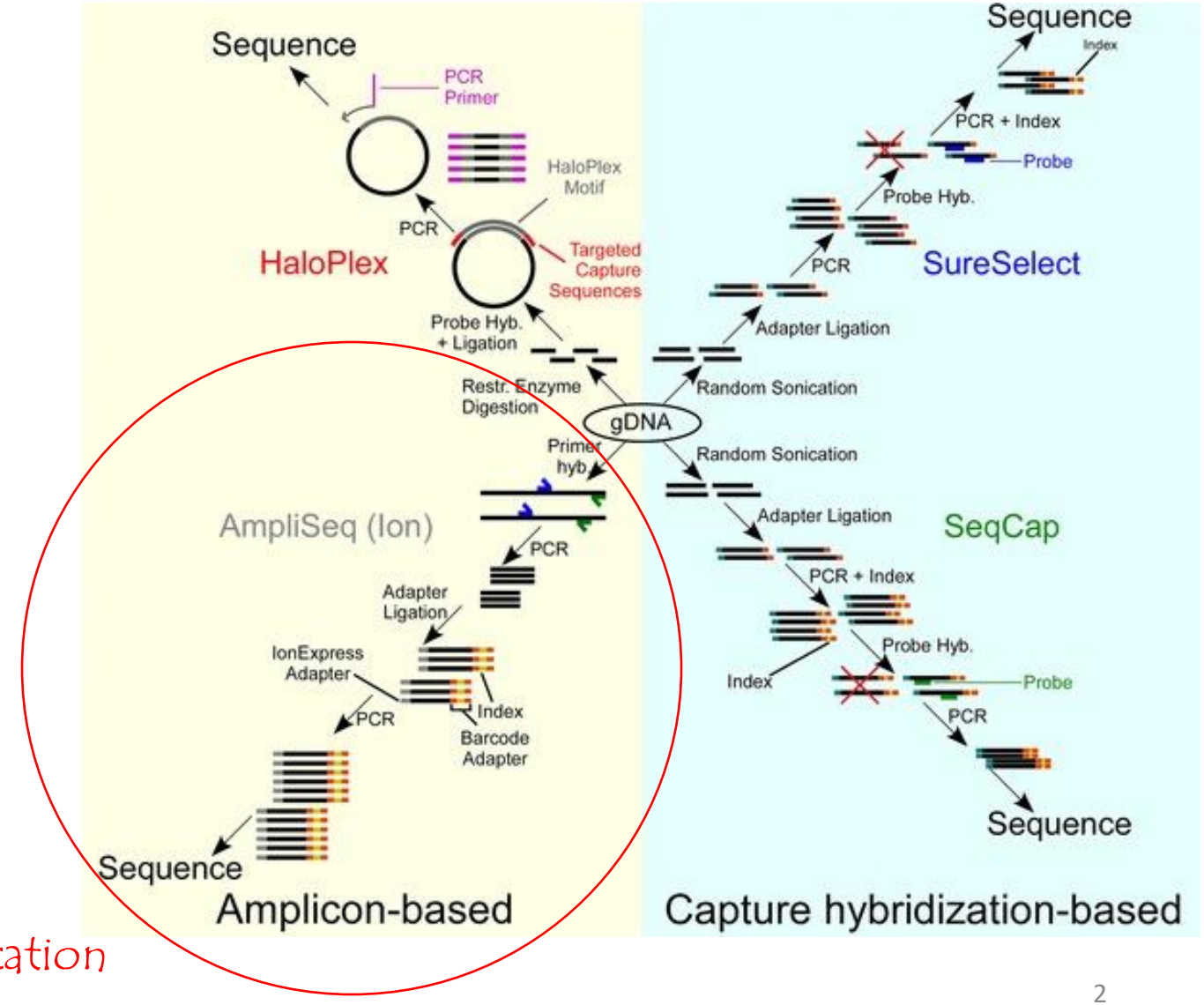
Pseudogene identification

Sandy Teng (Dec. 1st, 2022)

Motivation

- Cancer panel design
  - Amplicon-based capture method
  - # of amplicon sequences: 4,107
- Why do we need to check the specificity of amplicon sequences?
  - Noise information
- Strategy
  - Screening whole genome to identify all possible candidates



Sequence amplification

Alignment result

- bwa, blat, megablast, blastn (Total = 4,107 query IDs)

	bwa	blat	megablast	blastn
Computation time	5 secs	9 mins	16 mins	140 mins (94 mins/ thread number = 8)
# of IDs with multiple hits (i.e., pseudogene)	117	105	105	118
# of IDs with a single hit (i.e., self sequence)	3,990	4,002	4,002	3,989
# of alignments	4,402	6,627	391,457	20,569,682

Tools comparison

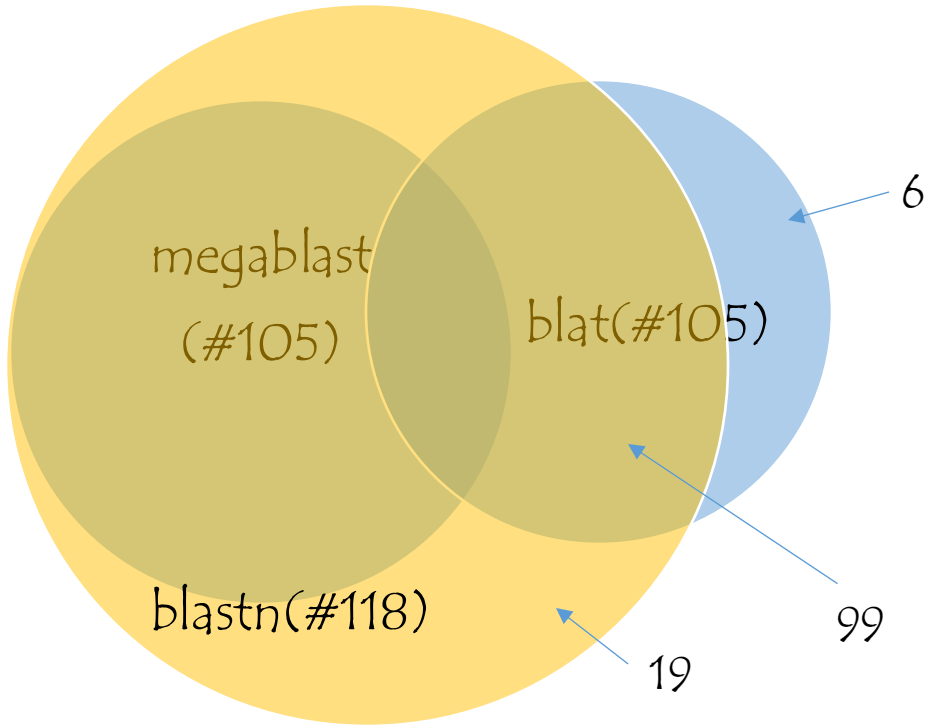
- Alignment parameters

	bwa	blat	megablast	blastn
Number of matching nucleotides	19 (-k)	11 (-tileSize)	28 (-word_size)	11 (-word_size)
Penalty for 5'- and 3'-end clipping	[5,5]	NA	NA	NA
Penalty for gap	Affine-gap penalty: gap open + gap extension	Maximum gap size = 2	Gap extension cost = None Gap opening cost = 0	Gap extension cost = 2 Gap opening cost = 5
Penalty for mismatch	Mismatch penalty	Minimum identity = 90%	Nucleic mismatch = -2 Nucleic match = 1	Nucleic mismatch = -3 Nucleic match = 2

Tools comparison

- All query sequences were aligned by these tools
- blastn identified more alignments

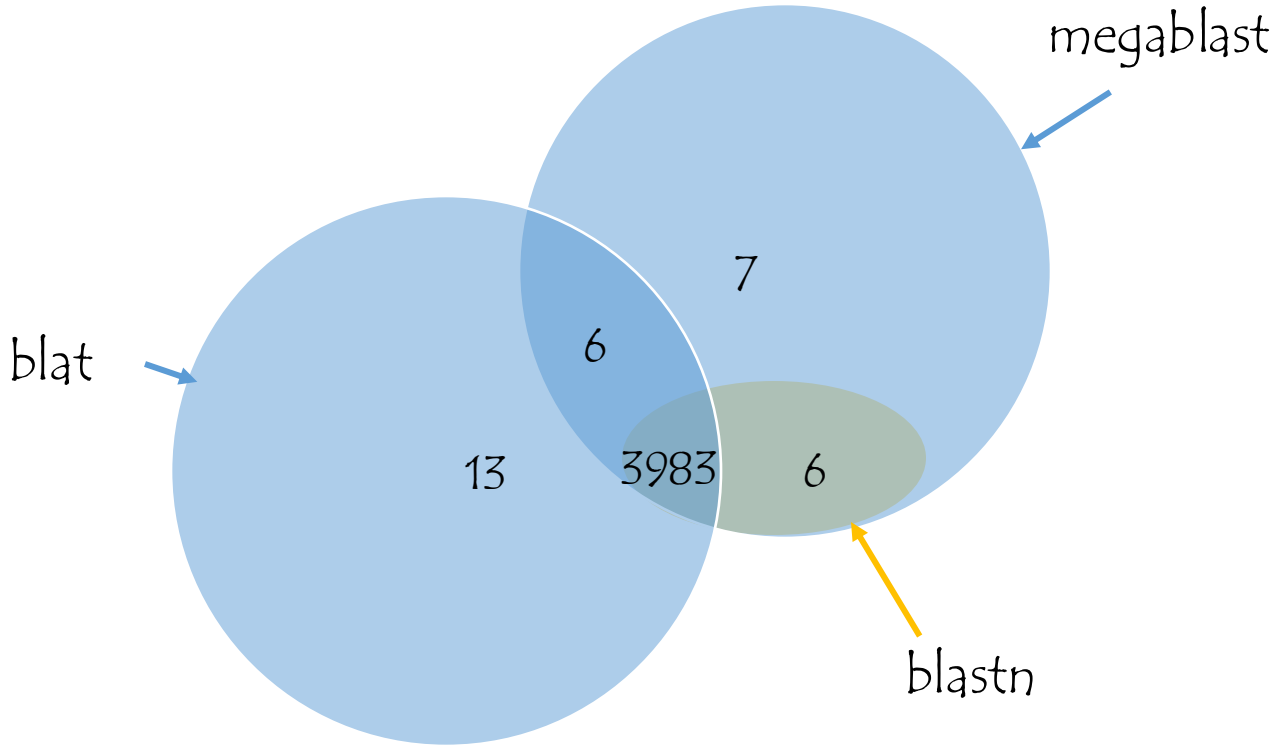
	bwa	blat	megablast	blastn
Computation time	5 secs	9 mins	16 mins	140 mins (94 mins/ thread number = 8)
# of IDs with multiple hits (i.e., pseudogene)	117	105	105	118
# of IDs with a single hit (i.e., self sequence)	3,990	4,002	4,002	3,989
# of alignments	4,402	6,627	391,457	20,569,682



Tools comparison

- Single hits

	bwa	blat	megablast	blastn
Computation time	5 secs	9 mins	16 mins	140 mins (94 mins/ thread number = 8)
# of IDs with multiple hits (i.e., pseudogene)	117	105	105	118
# of IDs with a single hit (i.e., self sequence)	3,990	4,002	4,002	3,989
# of alignments	4,402	6,627	391,457	20,569,682



Observations

- Amplicons
  - GENEID_PIK3CA_POOL_1_ID_PGD651-PIK3CA-CDS-09-1
  - GENEID_NRAS_POOL_2_ID_PGD651-NRAS-SNP-17-1
  - GENEID_NF1_POOL_2_ID_PGD651-NF1-CDS-33-2

Inconsistent sequence (example 1)

- blastn, blat (GENEID_PIK3CA_POOL_1_ID_PGD651-PIK3CA-CDS-09-1)

Alignment (tool)	Alignment length	# of mismatch	# of gap opening	# of gap	Identity	Chr.	s.start	s.end
Self	158	0	0	0	100	3	179218196	179218353
Alignment 1 (blat)	132	1	0	0	99.24	22	16572014	16572145
Alignment 2 (blat)	24	0	0	0	100	22	16572147	16572170
Alignment 3 (blastn/megablast)	158	2	1	1	98.101	22	16572014	16572170

```
Query  1         ATTTTATTTTACAGAGTAACAGACTAGCTAGAGACAATGAATTAAGGGAAAATGACAAAG  60
                 ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  16572014  ATTTTATTTTACAGAGTAACAGACTAGCTAGAGACAATGAATTAAGGGAAAATGACAAAG  16572073

Query  61        AACAGCTCAAAGCAATTTCTACACGAGATCCTCTCTCTGAAATCACTGAGCAGGAGAAAG  120
                 |||||||||||||||||||||||||||||||||||||||||||||| ||||||||||||
Sbjct  16572074  AACAGCTCAAAGCAATTTCTACACGAGATCCTCTCTCTGAAATCACTGCGCAGGAGAAAG  16572133

Query  121       ATTTTCTATGGAGTCACAGGTAAGTGCTAAAATGGAGA  158
                 ||||||||||||  ||||||||||||||||||||||||
Sbjct  16572134  ATTTTCTATGGA-CCACAGGTAAGTGCTAAAATGGAGA  16572170
```

Inconsistent sequence (example 2)

- GENEID_NRAS_POOL_2_ID_PGD651-NRAS-SNP-17-1

Alignment (tool)	Alignment length	# of mismatch	# of gap opening	# of gap	Identity	Chr.	s.start	s.end
Self	138	0	0	0	100	1	114718029	114718166
Alignment 1 (blat)	139	6	1	1	94.96	5	179082532	179082670
Alignment 2 (blastn/megablast)	137	5	1	1	95.62	5	179082534	179082670

Range 1: 3 to 139 Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Identities	Gaps	Strand
217 bits(240)	5e-62	131/137(96%)	1/137(0%)	Plus/Plus

```
GG  Query  3    TGTTAAAACTGGTGTAATAGCTCAATAGAA-TAAGTATTCCAGATTTCGGGAGGGATGAA  61
                ||||||||||||||||||||||||||| |||| ||||||||||||||||||||||||||||
GC  Sbjct  3    TGTTAAAACTGGTGTAATAGCTCAAAAGAACTAAGTATTCCAGATTTCGGGAGGGATGAA  62

    Query  62   GAGGGAGATATTCAGAACCCTTCACCAGATTCCCCCCAACTTGATCATAGTGGATTAATG  121
                |||||||||||||||| |||||||||||||| ||||||||||||||||||||||||||||
    Sbjct  63   GAGGGAGATATTCAGAAACCTTCACCAGATTCCTCCCAACTTGATCATAGTGGATTAATG  122

    Query  122  GTGTGCTTTGTGGATGT  138
                |||||||||||||||
    Sbjct  123  ACGTGCTTTGTGGATGT  139
```

Inconsistent sequence (example 3)

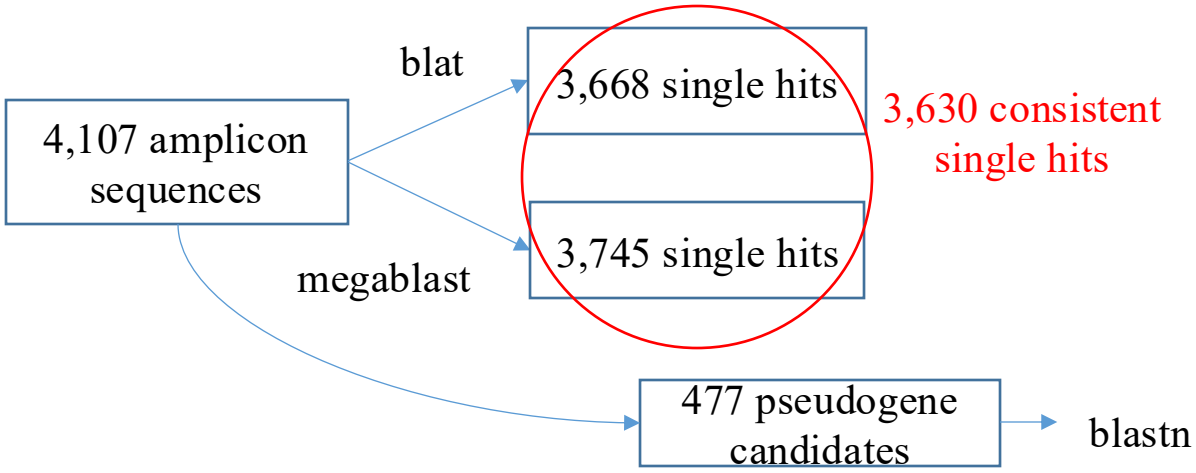
- GENEID_NF1_POOL_2_ID_PGD651-NF1-CDS-33-2

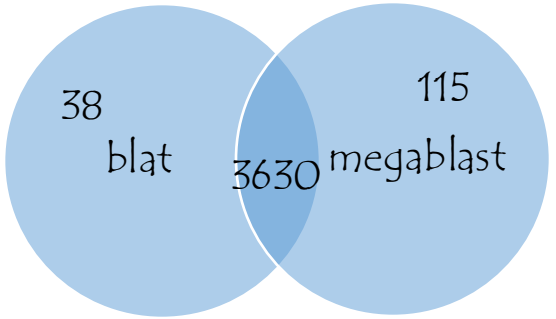
Alignment (tool)	Alignment length	# of mismatch	#of gap opening	# of gap	Identity	Chr.	s.start	s.end
Self	144	0	0	0	100	17	31259014	31259157
Alignment 1 (blat)	135	9	0	0	93.33	15	20923283	20923149
Alignment 2 (blastn/megablast)	144	11	0	0	92.361	15	20923283	20923140

```
Query  1         CCCTGTTTTATTGTGTAGATACTTCAGAGTATTGCCAATCATGTTCTCTTCACAAAAGAA  60
                 |||||||||||||||| ||||||||||||||||||||||||||||||||||||||||||
Sbjct  20923283  CCCTGTTTTATTGTGTAAATACTTCAGAGTATTGCCAATCATGTTCTCTTCACAAAAGAA  20923224

Query  61        GAACATATGCGGCCTTTCAATGATTTTGTGAAAAGCAACTTTGATGCAGCACGCAGGTAA  120
                 || ||||||||||||||||||||||||||||||||||| |||||||||||   | ||||||
Sbjct  20923223  GAGCATATGCGGCCTTTCAATGATTTTGTGAAAAGCAGCTTTGATGCAGCTTGAAGGTAA  20923164

Query  121       TTTTCTTGCCACTTACTCAGTTGC  144
                    | |||||||||| ||| ||||
Sbjct  20923163  GCTACTTGCCACTTATTCACTTGC  20923140
```

2-stage identification

- Stage 1: Identify single hits (megablast, blat)
  - 3,630 amplicons identified
- Stage 2: Identify pseudogenes (blastn)
  - 477 pseudogene candidates





(Example 1) pseudogene candidate

- GENEID_NF1_POOL_1_ID_PGD651-NF1-SNP-09-1

Alignment (tool)	Aligned Ratio	# of mismatch	# of gap opening	# of gap	Identity	Chr.	s.start	s.end
Self	1	0	0	0	100	17	31232522	31232638
Alignment 1 (blat)	0.846	10	0	0	89.90	15	21503257	21503141
Alignment 2 (blastn)	1	14	0	0	88.034	15	21503239	21503141
Alignment 3 (blastn)	1	14	0	0	88.034	15	20929556	20929440
Alignment 4 (blastn)	1	14	0	0	88.034	15	21940538	21940422

(Example 2) pseudogene candidate

- GENEID_AKT2_POOL_2_ID_PGD651-AKT2-CDS-11-3

Alignment (tool)	Aligned Ratio	# of mismatch	# of gap opening	# of gap	Identity	Chr.	s.start	s.end
Self	1	0	0	0	100	19	40235953	40236127
Alignment 2 (blat/blastn)	0.8857	12	0	0	92.260	14	104772941	104773095
Alignment 3 (blastn)	0.811	33	0	0	76.761	1	243552804	243552945

```
Query  13         ATCTCTTCCATGAGGATGAGCTCGAAGAGGCGCTCGTGGTCCTGGTTGTAGAAGGGCAGG  72
                  || |||||||| |  || |  || || ||   ||| ||||||||||||||||| || |
Sbjct  243552804  ATGTCTTCCATTAATATTAATTCAAAAAGTTTCTCATGGTCCTGGTTGTAGAAAGGTAAC  243552863

Query  73         CGGCCGCACATCATCTCGTACATGACCACACCCAGCCCCCACCAGTCCACGGCCCGGCCA  132
                  |  || ||||||||| || |||||||| || || || |||||||||||| || || ||||||
Sbjct  243552864  CTCCCACACATCATTTCATACATGACAACCCCTAGGCCCCACCAGTCTACTGCTCGGCCA  243552923

Query  133        TAGTCATTGTCCTCCAGCACCT  154
                  |||||||| || || | |||||
Sbjct  243552924  TAGTCATTATCTTCTAACACCT  243552945
```

(Example 3) pseudogene with many gaps (02/25)

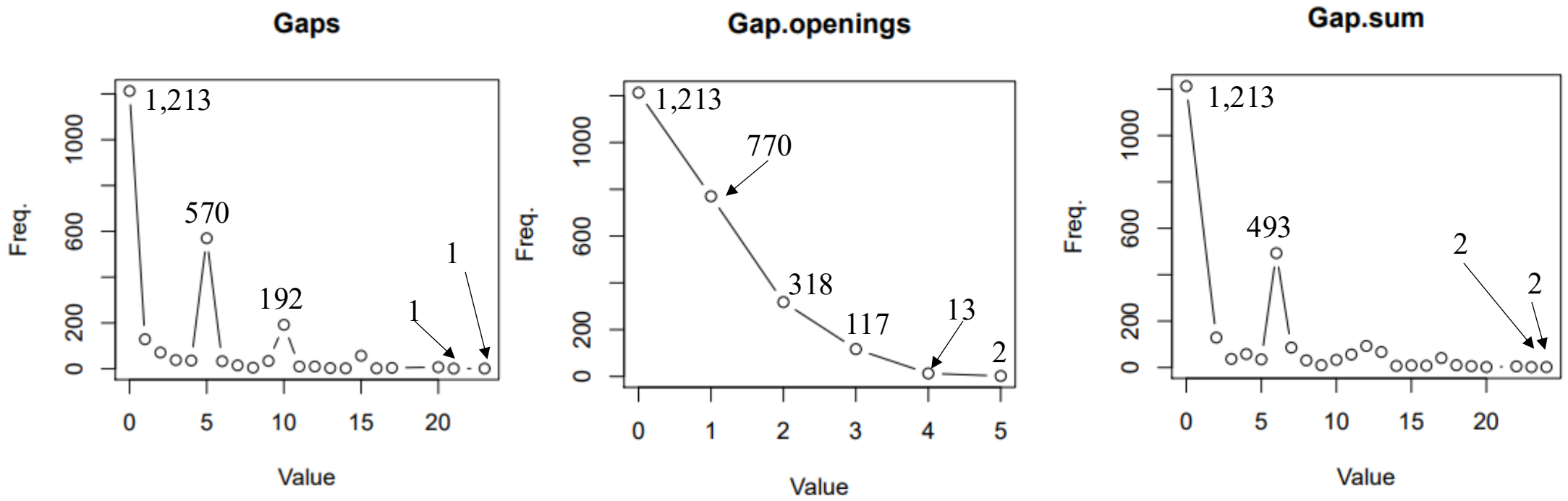
- GENEID_NF1_POOL_2_ID_PGD651-NF1-CDS-23-2

Alignment (tool)	Aligned Ratio	# of mismatch	# of gap opening	# of gap	Identity	Chr.	s.start	s.end
Self	1	0	0	0	100	17	31230265	31230439
Alignment 1 (blastn)	1	7	0	0	96.00	15	20931895	20931721
Alignment 2 (blastn)	1	8	1	23	84.34	2	131190529	131190332

```
Query  1          TTCGTGTGCTTGGGAATATGGTCCATGCAATTCAAATAAAAACGAAACTGTGTCAATTAG  60
                  || |||||||||||||||||||||||||||||||||||||||||||||||||||| || |
Sbjct  131190529  TTTGTGTGCTTGGGAATATGGTCCATGCAATTCAAATAAAAACGAAACTGTGTCAGTTGG  131190470

Query  61         TTGAAGTAATGATGGCAAGGAGAGATGACCTCTCATTT----------------------  98
                  ||||||||| ||||||||||||||||||||||||||
Sbjct  131190469  TTGAAGTAACGATGGCAAGGAGAGATGACCTCTCATTTTGCCGAGAGATGACCTCTCATT  131190410

Query  99         -TGCCAAGAGATGAAATTTAGGTGAGTTCTCAAAAGAGCAATGTAGGGTCTTGTAAATCT  157
                   |||||||||| ||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  131190409  TTGCCAAGAGATAAAATTTAGGTGAGTTCTCAAAAGAGCAATGTAGGGTCTTGTAAATCT  131190350

Query  158        TAATATGTCCAATGAAGT  175
                  || | ||| |||||||||
Sbjct  131190349  TAGTTTGTTCAATGAAGT  131190332
```

Plots (blastn -- 2,433 alignments)

- Gap.sum = Gaps + Gap.openings



Pseudogene identification

- Criteria for identical hit:

① Aligned ratio = aligned length / amplicon length = 1, gap + gap open == 0
  - Single variant noise
  - # of pseudogenes: 87

② Aligned ratio = aligned length / amplicon length = 1, gap + gap open <= 5 and identity >= 90%
  - Insertion/deletion variant noise
  - # of pseudogenes: 112

Discarded amplicon 1

- GENEID_FGFR1_POOL_2_ID_PGD651-FGFR1-SNP-10-1

Alignment (tool)	Aligned Ratio	# of mismatch	# of gap opening	# of gap	Identity	Chr.	s.start	s.end
Self	1	0	0	0	100	8	38434576	38434693
Alignment 1 (blastn)	1	13	1	1	88.136	1	101256434	101256318
Alignment 2 (blastn)	1	14	2	2	86.441	18	11928995	11928880

```
Query  1          CATTCAAACTGGTCCCTTCACTTAGACATTCTTTTCCTTTCCTCCTCTGAATCAAGTCAG  60
                  ||||| ||||||||| ||||||| || |||||||||||| ||||||||| |||||||||
Sbjct  101256434  CATTCGAACTGGTCCTTTCACTTTGAGATTCTTTTCCTTTGCTCCTCTGA-TCAAGTCAG  101256376

Query  61         CACACACCTTCTCCAGGGATTTTACGCTGCGGATCATTAGAGGGATTCGAATTTGGTG  118
                  |||||||||||||||||||||||||  ||||| |  ||| || |||||||||| ||||
Sbjct  101256375  CACACACCTTCTCCAGGGATTTTACTTTGCGGCTTGTTACAGTGATTCGAATTCGGTG  101256318
```

Chr. 1

```
Query  1          CATTCAAACTGGTCCCTTCACTTAGACATTCTTTTCCTTTCCTCCTCTGAATCAAGTCAG  60
                  ||||  ||||||||| ||||||| || |||||||||||| | ||||||| |||||||||
Sbjct  11928995   CATTTGAACTGGTCCTTTCACTTTGAGATTCTTTTCCTTTGCCCCTCTGA-TCAAGTCAG  11928937

Query  61         CACACACCTTCTCCAGGGATTTTACGCTGCGGATCATTAGAGGGATTCGAATTTGGTG  118
                  |||||||||||||||||||||||||  ||| | || |||||| ||||| || ||||||
Sbjct  11928936   CACACACCTTCTCCAGGGATTTTACCTTGCTGCTCCTTAGAGTGATTCTAA-TTGGTG  11928880
```

Chr. 18

Discarded amplicon 2

- GENEID_NF1_POOL_1_ID_PGD651-NF1-SNP-09-1

Alignment (tool)	Aligned Ratio	# of mismatch	# of gap opening	# of gap	Identity	Chr.	s.start	s.end
Self	1	0	0	0	100	17	31232522	31232638
Alignment 1 (blastn)	1	14	0	0	88.034	15	20929556	20929440
Alignment 2 (blastn)	1	14	0	0	88.034	15	21503257	21503141
Alignment 3 (blastn)	1	14	0	0	88.034	15	21940538	21940422

Alignment 1

```
Query  1         CATGTCCAACATAGCACACTTCATAATAAGCCACCCTGGCTGATTATCGCGAGAGAGGAG  60
                 |||| |||||| ||||  ||||||||| || ||||||||||||||||||  ||||||||||
Sbjct  20929556  CATGCCCAACACAGCATGCTTCATAATGAGTCACCCTGGCTGATTATCCTGAGAGAGGAG  20929497

Query  61        AGAAACAGTTAACCCAGGGCCATTCACACCATGCACATATGATTGTTTTGGAATGTC  117
                 |||| ||||||| ||||||||| ||||||| ||||||| |||| ||||||||||||||
Sbjct  20929496  AGAAGCAGTTAATCCAGGGCCAGTCACACCGTGCACATGTGATAGTTTTGGAATGTC  20929440
```

Alignment 2

```
Query  1         CATGTCCAACATAGCACACTTCATAATAAGCCACCCTGGCTGATTATCGCGAGAGAGGAG  60
                 |||| |||||| ||||  ||||||||| || ||||||||||||||||||  ||||||||||
Sbjct  21503257  CATGCCCAACACAGCATGCTTCATAATGAGTCACCCTGGCTGATTATCCTGAGAGAGGAG  21503198

Query  61        AGAAACAGTTAACCCAGGGCCATTCACACCATGCACATATGATTGTTTTGGAATGTC  117
                 |||| ||||||| ||||||||| ||||||| ||||||| |||| ||||||||||||||
Sbjct  21503197  AGAAGCAGTTAATCCAGGGCCAGTCACACCGTGCACATGTGATAGTTTTGGAATGTC  21503141
```

Alignment 3

```
Query  1         CATGTCCAACATAGCACACTTCATAATAAGCCACCCTGGCTGATTATCGCGAGAGAGGAG  60
                 |||| |||||| ||||  ||||||||| || ||||||||||||||||||  ||||||||||
Sbjct  21940538  CATGCCCAACACAGCATGCTTCATAATGAGTCACCCTGGCTGATTATCCTGAGAGAGGAG  21940479

Query  61        AGAAACAGTTAACCCAGGGCCATTCACACCATGCACATATGATTGTTTTGGAATGTC  117
                 |||| ||||||| ||||||||| ||||||| ||||||| |||| ||||||||||||||
Sbjct  21940478  AGAAGCAGTTAATCCAGGGCCAGTCACACCGTGCACATGTGATAGTTTTGGAATGTC  21940422
```

Discarded amplicon 3

- GENEID_NF1_POOL_2_ID_PGD651-NF1-CDS-10-3

Alignment (tool)	Aligned Ratio	# of mismatch	# of gap opening	# of gap	Identity	Chr.	s.start	s.end
Self	1	0	0	0	100	17	31201109	31201253
Alignment 1 (blastn)	1	14	1	1	89.655	18	14156655	14156512
Alignment 2 (blastn)	1	14	2	3	88.276	21	14001614	14001755

```
Query  1         TGCCTTGTTTCTTGCTTTCGTATAAGCCCTCACAACAACCAACACTTTAAGGTGAGAGCA  60
                 |||||||||||||||| ||||| |||||||||||||||||||||| |||||||||||| |||
Sbjct  14156655  TGCCTTGTTTCTTGCCTTCGTGTAAGCCCTCACAACAACCAACAGTTTAAGGTGAGGGCA  14156596

Query  61        TTGGTTTTTATCTAACTATATTTACTGATGCTGTTATCCTTTATAAACAAAAAGACTATA  120
                 |||||||||||||||||||| |||||||||||  ||||||||||||||||  |||||||| |
Sbjct  14156595  TTGGTTTTTATCTAACTATGTTTACTGATGCCATTATCCTTTATAAACGGAAAGACTAGA  14156536

Query  121       GAGATTAATAGGTTCACTTTTATCG  145
                 | || ||| ||||||||| | |||||
Sbjct  14156535  GGGA-TAACAGGTTCACCTCTATCG  14156512
```

Chr. 18

```
Query  1         TGCCTTGTTTCTTGCTTTCGTATAAGCCCTCACAACAACCAACACTTTAAGGTGAGAGCA  60
                 |||||||||||||||| ||||| |||||||||||||||||||||   |||||||||||| |||
Sbjct  14001614  TGCCTTGTTTCTTGCCTTCGTGTAAGCCCTCACAACAACCAAG--TTTAAGGTGAGGGCA  14001671

Query  61        TTGGTTTTTATCTAACTATATTTACTGATGCTGTTATCCTTTATAAACAAAAAGACTATA  120
                  |||||||||||||||||||| |||||||||||| ||||||||||||||||  |||||||| |
Sbjct  14001672  CTGGTTTTTATCTAACTATGTTTACTGATGCCGTTATCCTTTATAAACGGAAAGACTAGA  14001731

Query  121       GAGATTAATAGGTTCACTTTTATCG  145
                 | || ||| ||||||||| | |||||
Sbjct  14001732  GGGA-TAACAGGTTCACCTCTATCG  14001755
```

Chr. 21

Summary

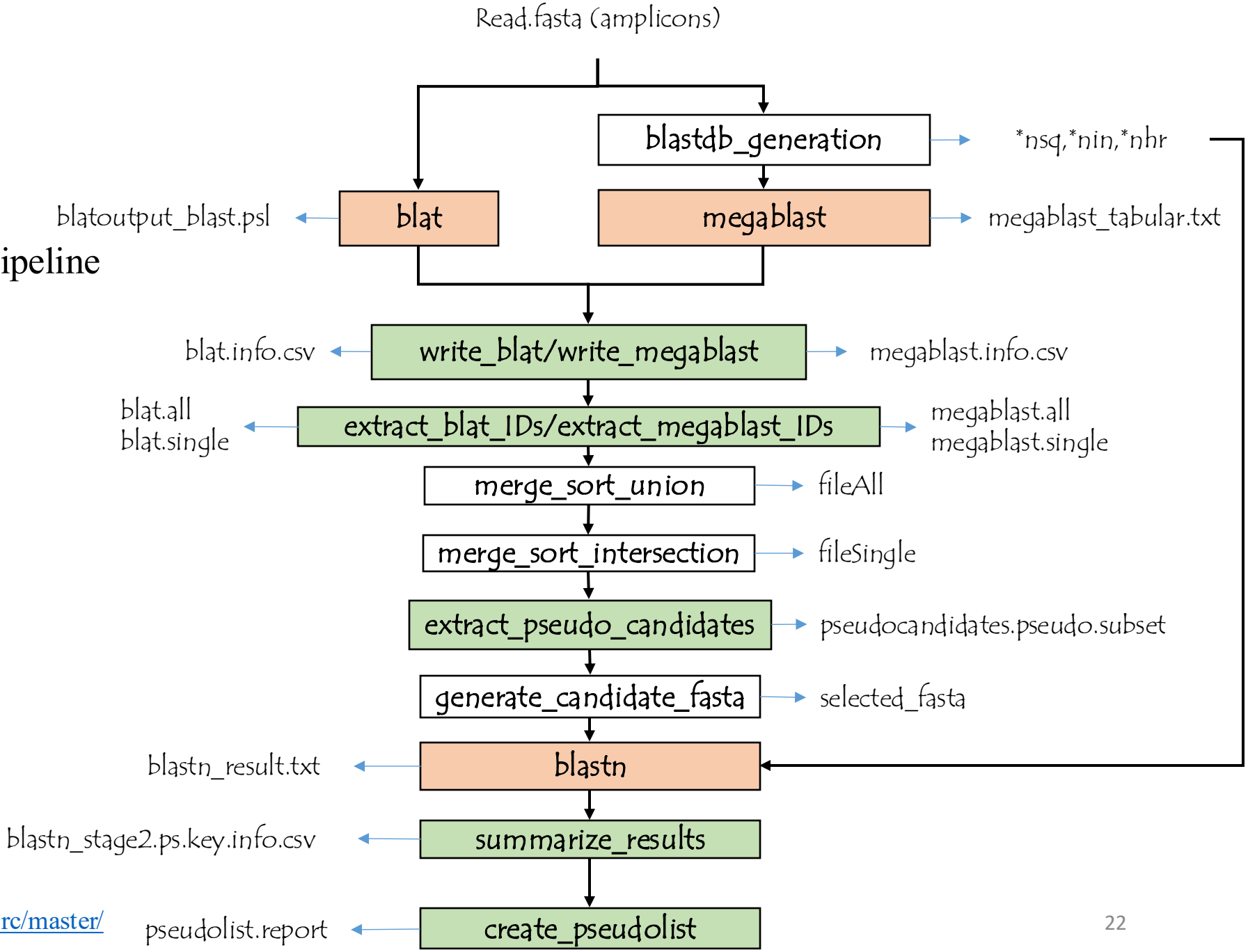
- Since different scoring functions are implemented in bwa, blat, megablast and blastn, the alignments for the same amplicon generated by different aligners may vary in terms of length and genomic location.
- If an amplicon can be aligned to other sequences, the corresponding sequences produced using PCR may be possibly generated from other sequences. The amplicon is thereby considered as a pseudogene candidate.
- 2-stage identification process utilizes blat and megablast to identify amplicons that have consistent single hits among the two tools, and confirms whether the rest of the amplicons are pseudogenes using blastn.
- If an amplicon can be perfectly aligned to a sequence other than itself, the aligned sequence can be regarded as a possible source of location noise (single variant noise).
- If an amplicon can be completely aligned (i.e., aligned length = amplicon length) to other sequence with gap(s) or gap opening(s), the aligned sequence can be considered as a possible source of insertion/ deletion noise.

Pipeline implementation

Pseudogene detection (nextflow) pipeline

Workflow

- Pseudogene detection pipeline
- Aligners
- Python codes



Repo: https://bitbucket.org/sandyteng/psf_2/src/master/

(Input) File formats

- amplicon
  - amplicon fasta file
- refgenome
  - reference genome/database fasta file
- pycode_list
  - the directory for “extract_single_hits_from_infofiles.py”
- pycode_info: the directory for “blat_info_arg.py”
- pycode_union: the directory for “extract_ids_from_pseudolists.py”
- pycode_all: the directory for “extract_all_hits_from_infofiles.py”
- pycode_summary: the directory for “blat_keyinfo_positions_arg.py”
- pycode_pseudolist: the directory for “blastn_stage2_pseudogenelist_arg.py”
- dbasestr: reference genome/database string
- publish_dir: result directory

psf_2 / params / PA039_GRCh38.json

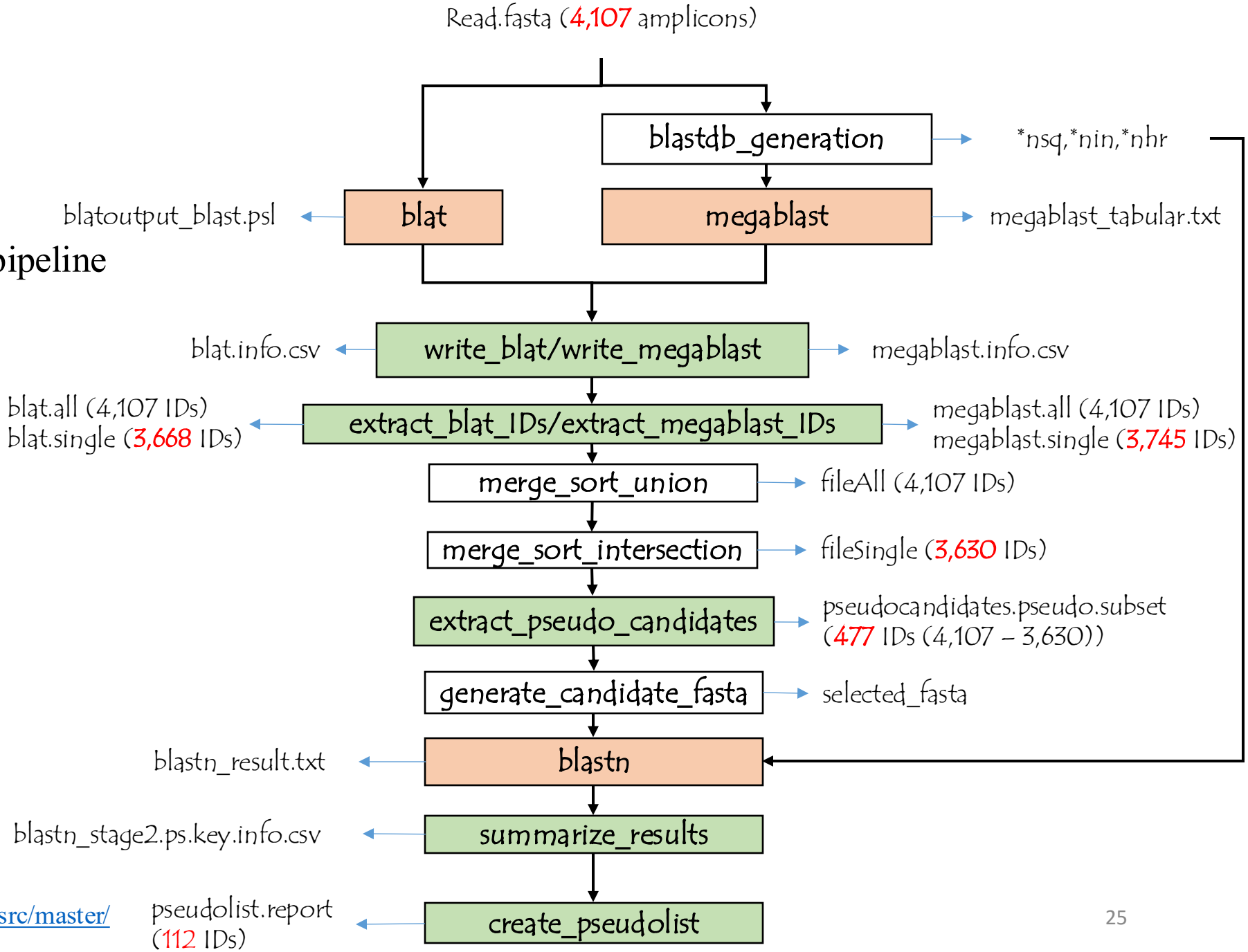
```
{
    "amplicon":"/mnt/BI3/Team_workdir/sandyteng_workdir/PseudoGene_pipeline/data/PA039_amplicon_v20210824-38.read.fasta",
    "publish_dir":"/mnt/BI3/Team_workdir/sandyteng_workdir/PseudoGene_pipeline_results/PA039/",
    "refgenome":"/mnt/BI3/Team_workdir/sandyteng_workdir/PseudoGene_pipeline/refgenome/GRCh38.p2.mask1.fasta",
    "pycode_list":"/mnt/BI3/Team_workdir/sandyteng_workdir/PseudoGene_pipeline/repo_dev/psf_2/python_code/extract_single_hits_from_infofiles.py",
    "pycode_info":"/mnt/BI3/Team_workdir/sandyteng_workdir/PseudoGene_pipeline/repo_dev/psf_2/python_code/blat_info_arg.py",
    "pycode_union":"/mnt/BI3/Team_workdir/sandyteng_workdir/PseudoGene_pipeline/repo_dev/psf_2/python_code/extract_ids_from_pseudolists.py",
    "pycode_all":"/mnt/BI3/Team_workdir/sandyteng_workdir/PseudoGene_pipeline/repo_dev/psf_2/python_code/extract_all_hits_from_infofiles.py",
    "pycode_summary":"/mnt/BI3/Team_workdir/sandyteng_workdir/PseudoGene_pipeline/repo_dev/psf_2/python_code/blat_keyinfo_positions_arg.py",
    "pycode_pseudolist":"/mnt/BI3/Team_workdir/sandyteng_workdir/PseudoGene_pipeline/repo_dev/psf_2/python_code/blastn_stage2_pseudogenelist_arg.py",
    "dbasestr":"GRCh38.p2.mask1"
}
```

(Output) File formats

- Blastn database files: database files for blastn & megablast
  - *nsq
  - *nin
  - *nhr
- Tabular results generated from the 3 aligners (blat, megablast, blastn)
  - blat: blatoutput_blast.psl
  - megablast: megablast_tabular.txt
  - blastn: blastn_result.txt
- Summary tables generated from the 3 tabular results
  - blat: blat.info.csv
  - megablast: megablast.info.csv
  - blastn: blastn_stage2.ps.key.info.csv
- ID list files (to store selected amplicon ID(s))
  - blat.all (a list of all amplicon IDs identified by blat)
  - blat.single (a list of single hit amplicons identified by blat)
  - megablast.all (a list of all amplicon IDs identified by megablast)
  - megablast.single (a list of single hit amplicons identified by megablast)
  - fileAll (an ID list of blat.all ∪ megablast.all)
  - fileSingle (an ID list of blat.single ∩ megablast.single)
  - pseudocandidates.pseudo.subset (a list of pseudogene amplicon candidates)
  - pseudolist.report (a list of identified pseudogene amplicons)
- Fasta file for stage 2
  - selected_fasta

Workflow

- Pseudogene detection pipeline
- Aligners
- Python codes



Repo: https://bitbucket.org/sandyteng/psf_2/src/master/

Pipeline execution

- PA039 (GRCh38) – 4,107 amplicons
- Onco27 (hg19)
- PA031 (hg19)

PA039 (GRCh38)

```
sandyteng@RD183:/mnt/BI3/Team_workdir/sandyteng_workdir/PseudoGene_pipeline_results/PA039$ /mnt/BI3/Team_workdir/sandyteng_workdir/execprogs/bin/nextflow run
 /mnt/BI3/Team_workdir/sandyteng_workdir/PseudoGene_pipeline/repo_dev/psf_2/main_grch38.nf -params-file /mnt/BI3/Team_workdir/sandyteng_workdir/PseudoGene_pi
peline/repo_dev/psf_2/params/PA039_GRCh38.json -c /mnt/BI3/Team_workdir/sandyteng_workdir/PseudoGene_pipeline/repo_dev/psf_2/pseudogene_localdocker.config
N E X T F L O W  ~  version 21.10.6
Launching `/mnt/BI3/Team_workdir/sandyteng_workdir/PseudoGene_pipeline/repo_dev/psf_2/main_grch38.nf` [elegant_meucci] - revision: 60bb722d2a
executor >  local (14)
[96/f6f7ae] process > Pseudo_finder:blastdb_generation (1)        [100%] 1 of 1 ✔
[15/d2906c] process > Pseudo_finder:blat (1)                      [100%] 1 of 1 ✔
[11/dad837] process > Pseudo_finder:megablast (1)                 [100%] 1 of 1 ✔
[5b/d67599] process > Pseudo_finder:write_blat (1)                [100%] 1 of 1 ✔
[00/2c19e5] process > Pseudo_finder:write_megablast (1)           [100%] 1 of 1 ✔
[04/0aad85] process > Pseudo_finder:extract_blat_IDs (1)          [100%] 1 of 1 ✔
[57/516279] process > Pseudo_finder:extract_megablast_IDs (1)     [100%] 1 of 1 ✔
[2f/db92be] process > Pseudo_finder:merge_sort_intersection (1)   [100%] 1 of 1 ✔
[bf/f98341] process > Pseudo_finder:merge_sort_union (1)          [100%] 1 of 1 ✔
[5f/ab4262] process > Pseudo_finder:extract_pseudo_candidates (1) [100%] 1 of 1 ✔
[45/984f7e] process > Pseudo_finder:generate_candidate_fasta (1)  [100%] 1 of 1 ✔
[e8/53c14f] process > Pseudo_finder:blastn (1)                    [100%] 1 of 1 ✔
[95/85cdfa] process > Pseudo_finder:summarize_results (1)         [100%] 1 of 1 ✔
[0f/ee0010] process > Pseudo_finder:create_pseudolist (1)         [100%] 1 of 1 ✔
Completed at: 25-Nov-2022 14:42:23
Duration    : 39m 54s
CPU hours   : 1.4
Succeeded   : 14
```

/mnt/BI3/Team_workdir/sandyteng_workdir/execprogs/bin/nextflow run
/mnt/BI3/Team_workdir/sandyteng_workdir/PseudoGene_pipeline/repo_dev/psf_2/main_grch38.nf
-params-file /mnt/BI3/Team_workdir/sandyteng_workdir/PseudoGene_pipeline/repo_dev/psf_2/params/PA039_GRCh38.json
-c /mnt/BI3/Team_workdir/sandyteng_workdir/PseudoGene_pipeline/repo_dev/psf_2/pseudogene_localdocker.config

Onco2M7 (hg19)

```
sandyteng@RD183:/mnt/BI3/Team_workdir/sandyteng_workdir/PseudoGene_pipeline_results/Onco2M7$ /mnt/BI3/Team_workdir/sandyteng_workdir/execprogs/bin/nextflow r
un /mnt/BI3/Team_workdir/sandyteng_workdir/PseudoGene_pipeline/repo_dev/psf_2/main.nf -params-file /mnt/BI3/Team_workdir/sandyteng_workdir/PseudoGene_pipelin
e/repo_dev/psf_2/params/Onco2M7_hg19.json -c /mnt/BI3/Team_workdir/sandyteng_workdir/PseudoGene_pipeline/repo_dev/psf_2/pseudogene_localdocker.config
N E X T F L O W  ~  version 21.10.6
Launching `/mnt/BI3/Team_workdir/sandyteng_workdir/PseudoGene_pipeline/repo_dev/psf_2/main.nf` [irreverent_galileo] - revision: a03d09ccb8
executor >  local (14)
[f6/8cf69b] process > Pseudo_finder:blastdb_generation (1)        [100%] 1 of 1 ✔
[a1/943cb7] process > Pseudo_finder:blat (1)                      [100%] 1 of 1 ✔
[7b/0530e9] process > Pseudo_finder:megablast (1)                 [100%] 1 of 1 ✔
[97/02d1e6] process > Pseudo_finder:write_blat (1)                [100%] 1 of 1 ✔
[6e/747115] process > Pseudo_finder:write_megablast (1)           [100%] 1 of 1 ✔
[26/117ddc] process > Pseudo_finder:extract_blat_IDs (1)          [100%] 1 of 1 ✔
[0b/e3e82c] process > Pseudo_finder:extract_megablast_IDs (1)     [100%] 1 of 1 ✔
[bf/273a38] process > Pseudo_finder:merge_sort_intersection (1)   [100%] 1 of 1 ✔
[d1/95aa46] process > Pseudo_finder:merge_sort_union (1)          [100%] 1 of 1 ✔
[c9/b9d717] process > Pseudo_finder:extract_pseudo_candidates (1) [100%] 1 of 1 ✔
[3f/f22ce5] process > Pseudo_finder:generate_candidate_fasta (1)  [100%] 1 of 1 ✔
[ac/0cd840] process > Pseudo_finder:blastn (1)                    [100%] 1 of 1 ✔
[96/7d47b4] process > Pseudo_finder:summarize_results (1)         [100%] 1 of 1 ✔
[1b/1de395] process > Pseudo_finder:create_pseudolist (1)         [100%] 1 of 1 ✔
Completed at: 26-Nov-2022 01:49:17
Duration    : 9h 32m 35s
CPU hours   : 19.2
Succeeded   : 14
```

/mnt/BI3/Team_workdir/sandyteng_workdir/execprogs/bin/nextflow run
/mnt/BI3/Team_workdir/sandyteng_workdir/PseudoGene_pipeline/repo_dev/psf_2/main.nf
-params-file /mnt/BI3/Team_workdir/sandyteng_workdir/PseudoGene_pipeline/repo_dev/psf_2/params/Onco2M7_hg19.json
-c /mnt/BI3/Team_workdir/sandyteng_workdir/PseudoGene_pipeline/repo_dev/psf_2/pseudogene_localdocker.config

PA031 (hg19)

```
sandyteng@RD183:/mnt/BI3/Team_workdir/sandyteng_workdir/PseudoGene_pipeline_results/PA031$ /mnt/BI3/Team_workdir/sandyteng_workdir/execprogs/bin/nextflow run
 /mnt/BI3/Team_workdir/sandyteng_workdir/PseudoGene_pipeline/repo_dev/psf_2/main.nf -params-file /mnt/BI3/Team_workdir/sandyteng_workdir/PseudoGene_pipeline/
repo_dev/psf_2/params/PA031_hg19.json -c /mnt/BI3/Team_workdir/sandyteng_workdir/PseudoGene_pipeline/pseudogene_localdocker.giant.2.config -resume
N E X T F L O W  ~  version 21.10.6
Launching `/mnt/BI3/Team_workdir/sandyteng_workdir/PseudoGene_pipeline/repo_dev/psf_2/main.nf` [ridiculous_sax] - revision: a03d09ccb8
executor >  local (2)
[c2/3ef128] process > Pseudo_finder:blastdb_generation (1)        [100%] 1 of 1, cached: 1 ✔
[88/1c8472] process > Pseudo_finder:blat (1)                      [100%] 1 of 1, cached: 1 ✔
[4d/db0ba9] process > Pseudo_finder:megablast (1)                 [100%] 1 of 1, cached: 1 ✔
[15/9acb5a] process > Pseudo_finder:write_blat (1)                [100%] 1 of 1, cached: 1 ✔
[cf/a0f674] process > Pseudo_finder:write_megablast (1)           [100%] 1 of 1, cached: 1 ✔
[15/f08e12] process > Pseudo_finder:extract_blat_IDs (1)          [100%] 1 of 1, cached: 1 ✔
[89/5e0e3c] process > Pseudo_finder:extract_megablast_IDs (1)     [100%] 1 of 1, cached: 1 ✔
[7d/17cfcf] process > Pseudo_finder:merge_sort_intersection (1)   [100%] 1 of 1, cached: 1 ✔
[9d/bc0233] process > Pseudo_finder:merge_sort_union (1)          [100%] 1 of 1, cached: 1 ✔
[b4/359a68] process > Pseudo_finder:extract_pseudo_candidates (1) [100%] 1 of 1, cached: 1 ✔
[42/a84686] process > Pseudo_finder:generate_candidate_fasta (1)  [100%] 1 of 1, cached: 1 ✔
[24/f76bf6] process > Pseudo_finder:blastn (1)                    [100%] 1 of 1, cached: 1 ✔
[55/253c0a] process > Pseudo_finder:summarize_results (1)         [100%] 1 of 1 ✔
[58/ec6e8f] process > Pseudo_finder:create_pseudolist (1)         [100%] 1 of 1 ✔
Completed at: 28-Nov-2022 09:09:40
Duration    : 27m 26s
CPU hours   : 59.6 (98.5% cached)
Succeeded   : 2
Cached      : 12
```

Pseudo_finder:blastn: 26.6 GB
Pseudo_finder:summarize_results: 64 GB

/mnt/BI3/Team_workdir/sandyteng_workdir/execprogs/bin/nextflow run
/mnt/BI3/Team_workdir/sandyteng_workdir/PseudoGene_pipeline/repo_dev/psf_2/main.nf
-params-file /mnt/BI3/Team_workdir/sandyteng_workdir/PseudoGene_pipeline/repo_dev/psf_2/params/PA031_hg19.json
-c /mnt/BI3/Team_workdir/sandyteng_workdir/PseudoGene_pipeline/repo_dev/psf_2/pseudogene_localdocker.huge.config
(-c /mnt/BI3/Team_workdir/sandyteng_workdir/PseudoGene_pipeline/pseudogene_localdocker.giant.config -resume)
(-c /mnt/BI3/Team_workdir/sandyteng_workdir/PseudoGene_pipeline/pseudogene_localdocker.giant.2.config -resume)

huge: 32 G
giant: 48 G
giant.2: 64 G

Discussion & Future Work

- Since the memory usages of the two nextflow processes “blastn” & “summarize_results” are proportional to the amount of pseudogene candidates, one may need to adjust the memory setting in the configuration file accordingly

- The memory usages of the 6 python modules may need to be optimized to process larger data
  - Possible approaches for module optimization:
    - Avoid creating any huge data frame
    - Implement some of the needed pandas methods such as merge and join using dictionary instead

- The stage 2 results can be summarized using “blastnsummarytable.py” (see repo for code usage)