De novo identification of SNPs from RNA-seq data in non-model species

H. Lopez-Maestre, L. Brinza, C. Marchet,
J. Kielbassa, S. Bastien, M. Boutigny, D. Monin,
A. El Filali, C. Carareto, C. Vieira, F. Picard,
N. Kremer, F. Vavre, M. Sagot, V. Lacroix

Lyon 1

Colib'read

informatics mathematics

LBBE
BIOMETRIE ET BIOLOGIE EVOLUTIVE

Single Nucleotide Polymorphism

…ATTGTGCATGAATTCTGAATACGGCTACGCCAT…

…ATTGTGCATGAATTCTAAATACGGCTACGCCAT…

genetic marker for association studies, population structure etc.

Single Nucleotide Polymorphism and NGS

…ATTGTGCATGAATTCTGAATACGGCTACGCCAT…

…ATTGTGCATGAATTCTAAATACGGCTACGCCAT…

genetic marker for association studies, population structure etc.

With NGS : massive access to sequencing, cheaper and easier

Single Nucleotide Polymorphism and NGS

...ATTGTGCATGAATTCTGAATACGGCTACGCCAT...

...ATTGTGCATGAATTCTAAATACGGCTACGCCAT...

genetic marker for association studies, population structure etc.

With NGS : massive access to sequencing, cheaper and easier

Methods well developped for model species :

- rely on a (good) reference genome
- mostly for genome (DNAseq) re-sequencing

Why working with RNAseq ?

- Lower cost
- SNPs from **expressed regions**
- SNPs with a more direct **functional impact**

How to detect SNPs in RNAseq data ?

How to detect SNPs in RNAseq data ? for non model species ?

SNP identification from RNAseq

reads from RNAseq

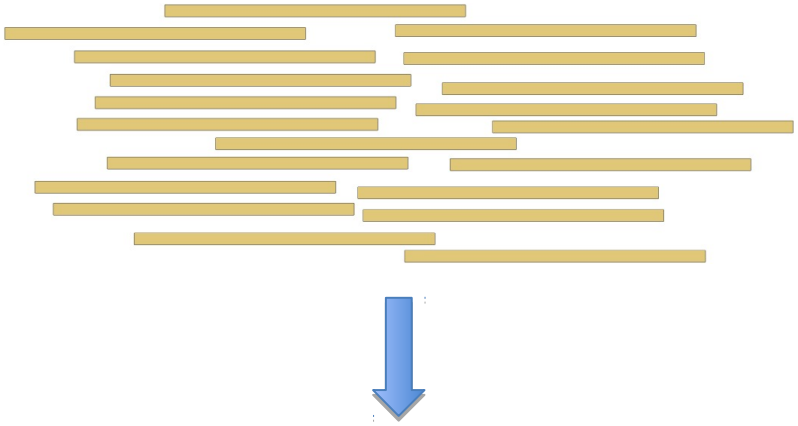
SNP identification from RNAseq
1) with a reference genome

reads from RNAseq

reference genome

SNP identification from RNAseq
1) with a reference genome

reads from RNAseq



reference genome

Several softwares :

- SNPiR
- GATK
- SAMtools mpileup

etc.

SNP identification from RNAseq
2) with an assembled transcriptome

reads from RNAseq

transcriptome assembly :

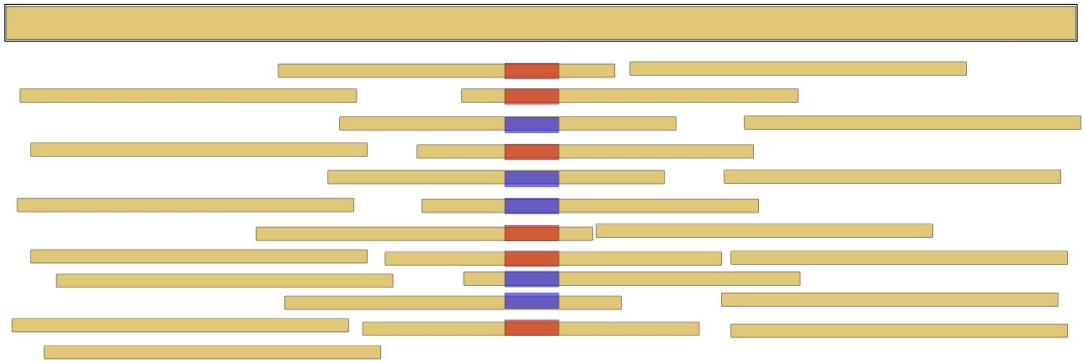
Trinity

Oases

SNP identification from RNAseq
2) with an assembled transcriptome

reads from RNAseq

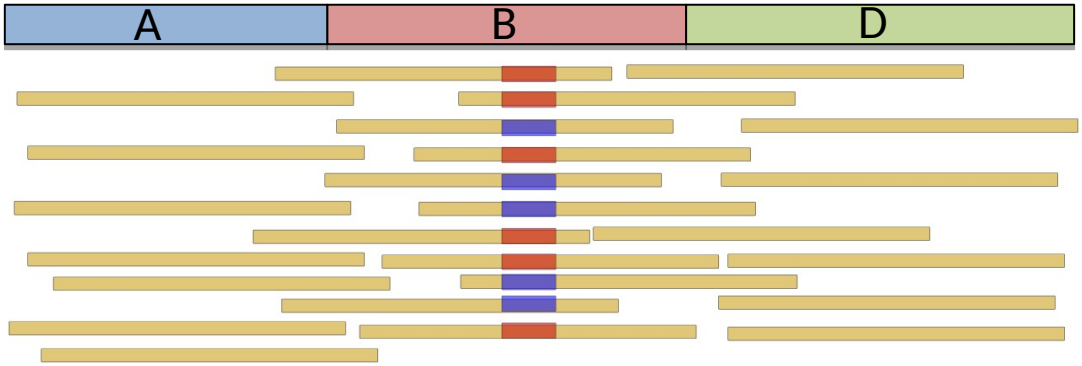
transcriptome assembly :



SNP identification from RNAseq
2) with an assembled transcriptome

reads from RNAseq

transcriptome assembly :

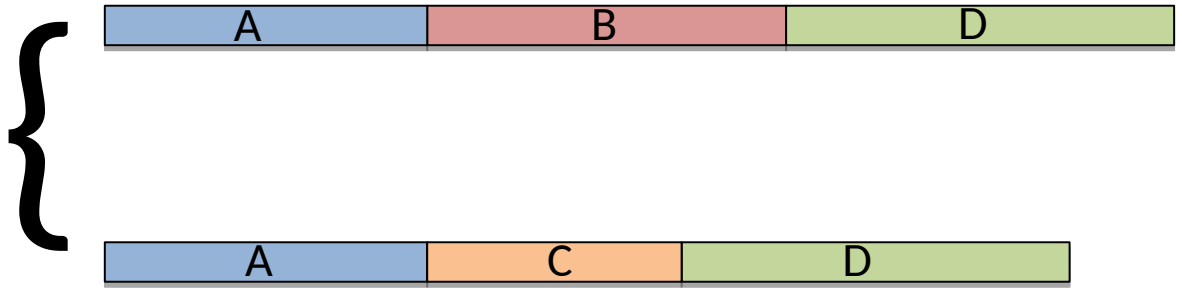


SNP identification from RNAseq
2) with an assembled transcriptome

reads from RNAseq

transcriptome assembly :

two assembled transcripts
from one gene
due to alternative splicing

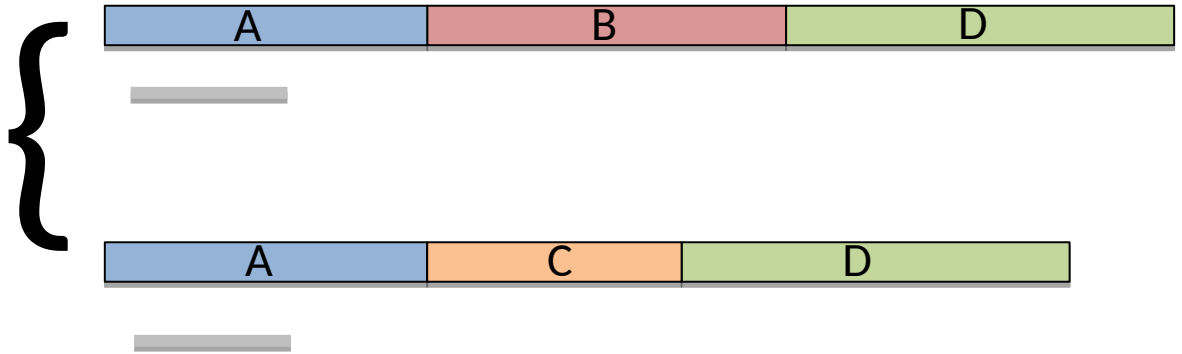


SNP identification from RNAseq
2) with an assembled transcriptome

reads from RNAseq

transcriptome assembly :

two assembled transcripts
from one gene
due to alternative splicing

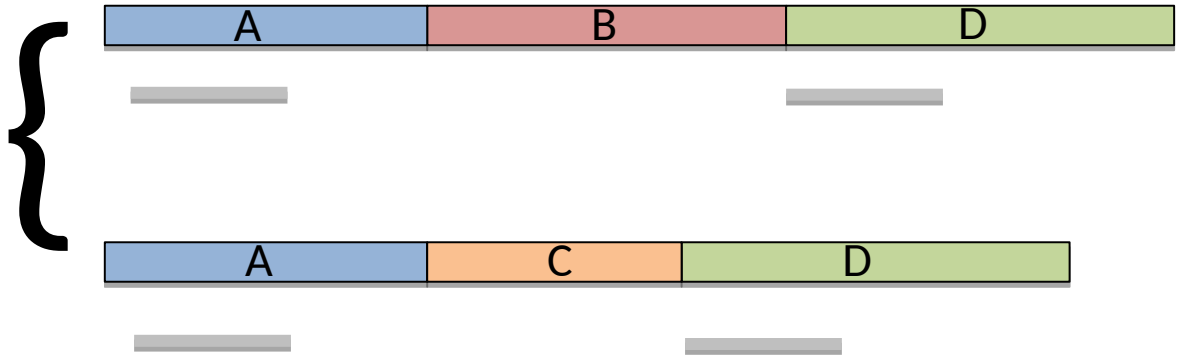


SNP identification from RNAseq
2) with an assembled transcriptome

reads from RNAseq

transcriptome assembly :

two assembled transcripts
from one gene
due to alternative splicing

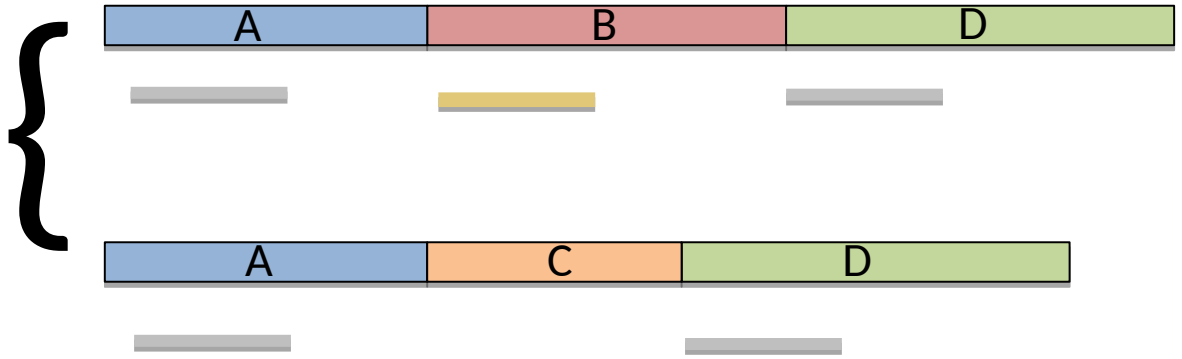


SNP identification from RNAseq
2) with an assembled transcriptome

reads from RNAseq

transcriptome assembly :

two assembled transcripts
from one gene
due to alternative splicing

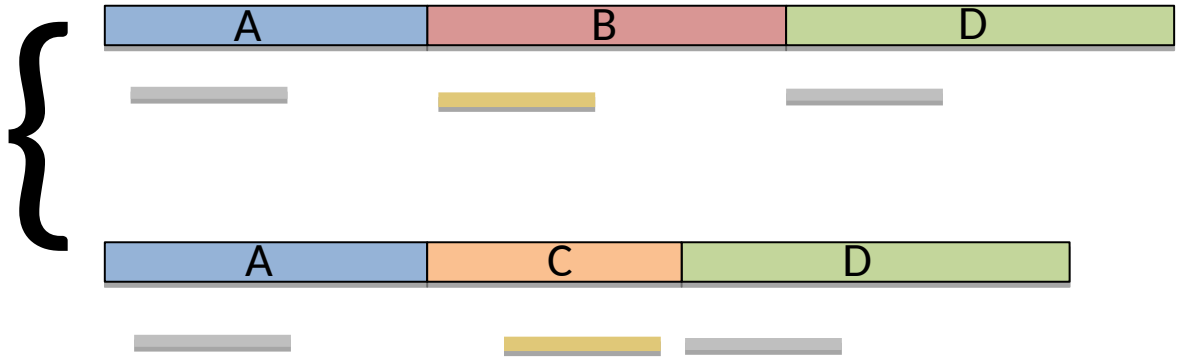


SNP identification from RNAseq
2) with an assembled transcriptome

reads from RNAseq

transcriptome assembly :

two assembled transcripts
from one gene
due to alternative splicing

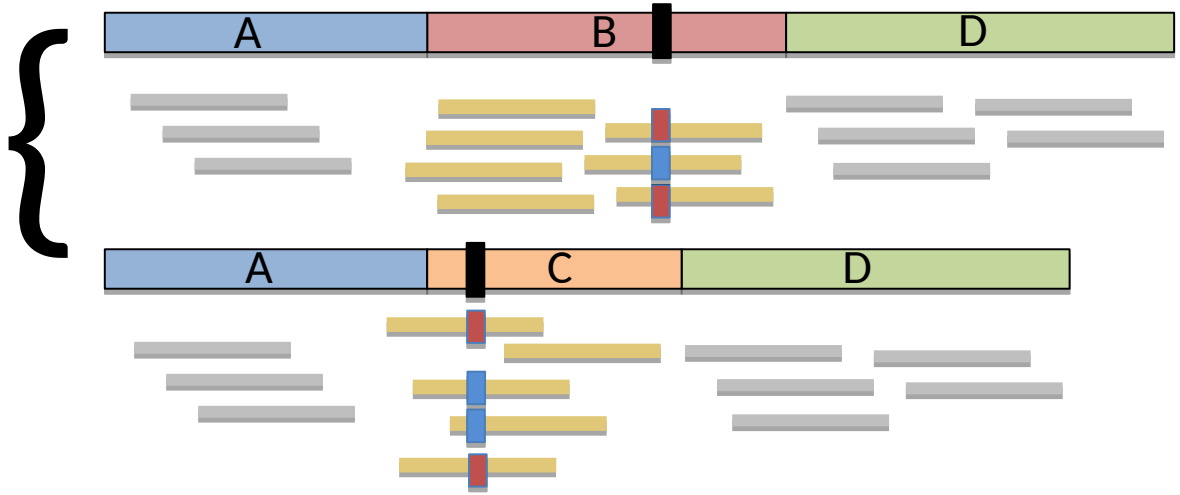


SNP identification from RNAseq
2) with an assembled transcriptome

reads from RNAseq

transcriptome assembly :

two assembled transcripts
from one gene
due to alternative splicing

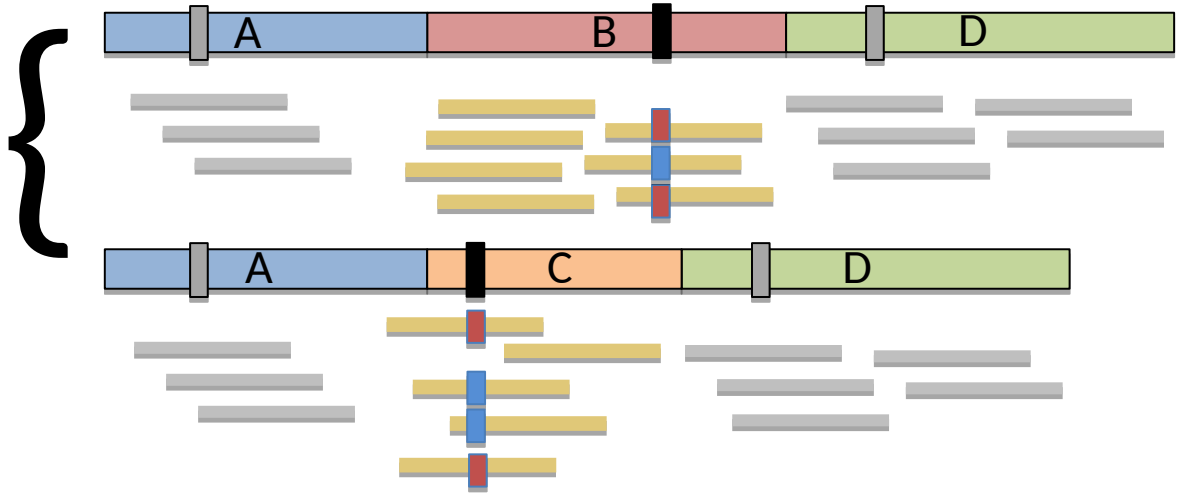


SNP identification from RNAseq
2) with an assembled transcriptome

reads from RNAseq

transcriptome assembly :

two assembled transcripts
from one gene
due to alternative splicing

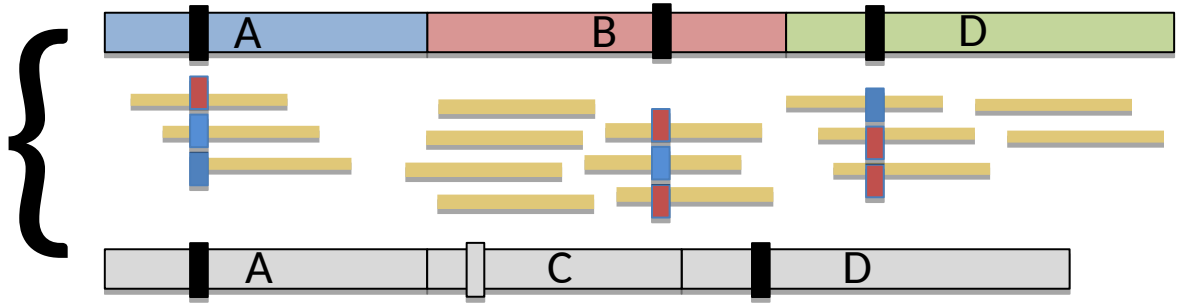


SNP identification from RNAseq
2) with an assembled transcriptome

reads from RNAseq

transcriptome assembly :

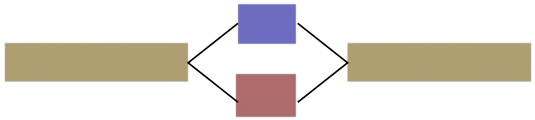
two assembled transcripts
from one gene
due to alternative splicing



SNP identification from RNAseq
3) from a De Bruijn graph

reads from RNAseq

local assembly



SNP identification from RNAseq
3) from a De Bruijn graph

ATTCCTGCAATAC

GGTAGGTACATCGTGT

GTTGTGCATGAATTCTG

TGCATGAATTCTACAATA

ATGAGGTACATGCAT

GCATGAATTCTGCAATAC

GTTGTGCATGAATTCTA

ATTCCTACAATAC

KisSplice

SNP identification from RNAseq
3) from a De Bruijn graph

ATTCCTGCAATAC

GGTAGGTACATCGTGT

GTTGTGCATGAATTCTG

TGCATGAATTCTACAATA

ATGAGGTACATGCAT

GCATGAATTCTGCAATAC

GTTGTGCATGAATTCTA

ATTCCTACAATAC

ATT
TTC
TCC
CCT
...

KisSplice

SNP identification from RNAseq
3) from a De Bruijn graph

ATTCCTGCAATAC

GGTAGGTACATCGTGT

GTTGTGCATGAATTCTG

TGCATGAATTCTACAATA

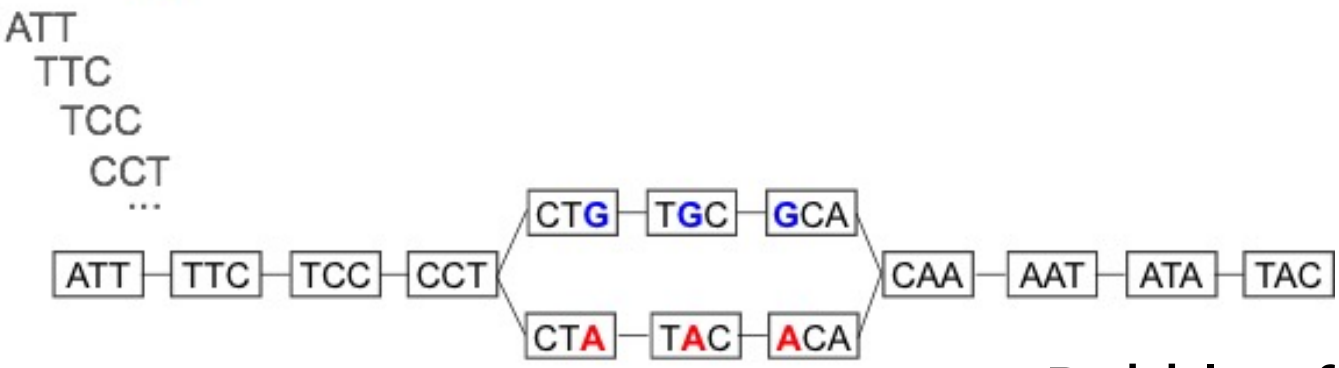
ATGAGGTACATGCAT

GCATGAATTCTGCAATAC

GTTGTGCATGAATTCTA

KisSplice

ATTCCTACAATAC



Bubble of size = 2k+1

SNP identification from RNAseq
3) from a De Bruijn graph

ATTCCTGCAATAC

GGTAGGTACATCGTGT

GTTGTGCATGAATTCTG

TGCATGAATTCTACAATA

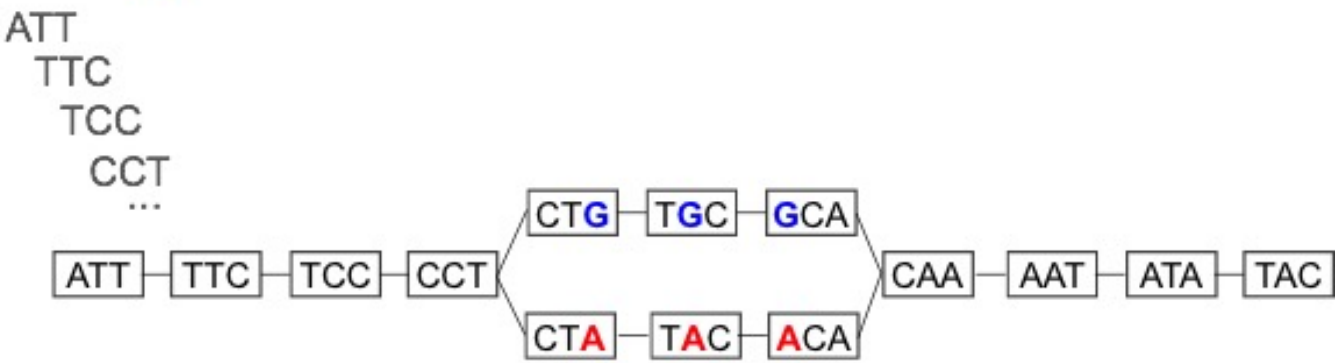
ATGAGGTACATGCAT

GCATGAATTCTGCAATAC

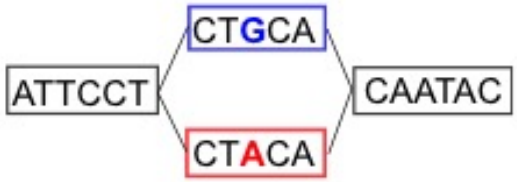
GTTGTGCATGAATTCTA

KisSplice

ATTCCTACAATAC



Bubble of size = 2k+1



SNP identification from RNAseq
3) from a De Bruijn graph

ATTCCTGCAATAC

GGTAGGTACATCGTGT

GTTGTGCATGAATTCTG

TGCATGAATTCTACAATA

ATGAGGTACATGCAT

GCATGAATTCTGCAATAC

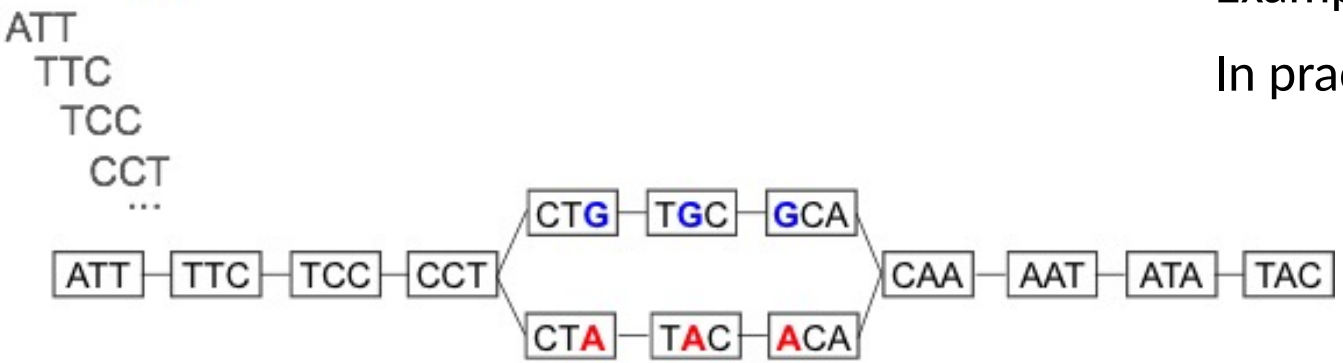
GTTGTGCATGAATTCTA

KisSplice

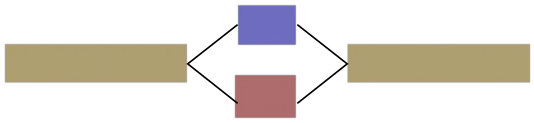
ATTCCTACAATAC

Example with k=3

In practice k=41 (default)



Bubble of size = 2k+1



SNP identification from RNAseq

1) with a reference genome

2) with an assembled transcriptome

3) directly from the De Bruijn graph

Validation of the method

Recall : Out of all the « True » SNPs, the proportion predicted by the method

Precision : Out of all the predicted SNPs, the proportion that is correctly predicted

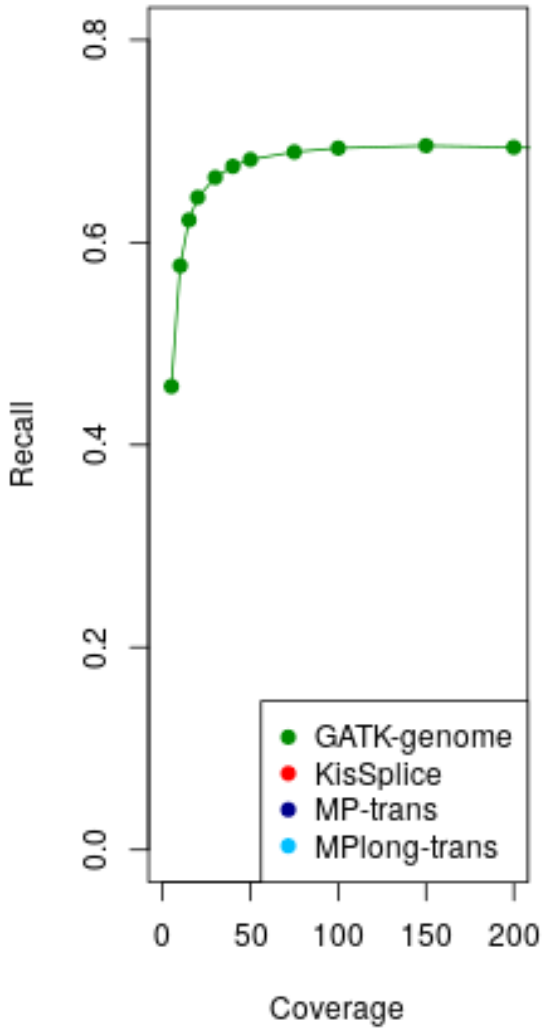
Validation of the method

Human dataset

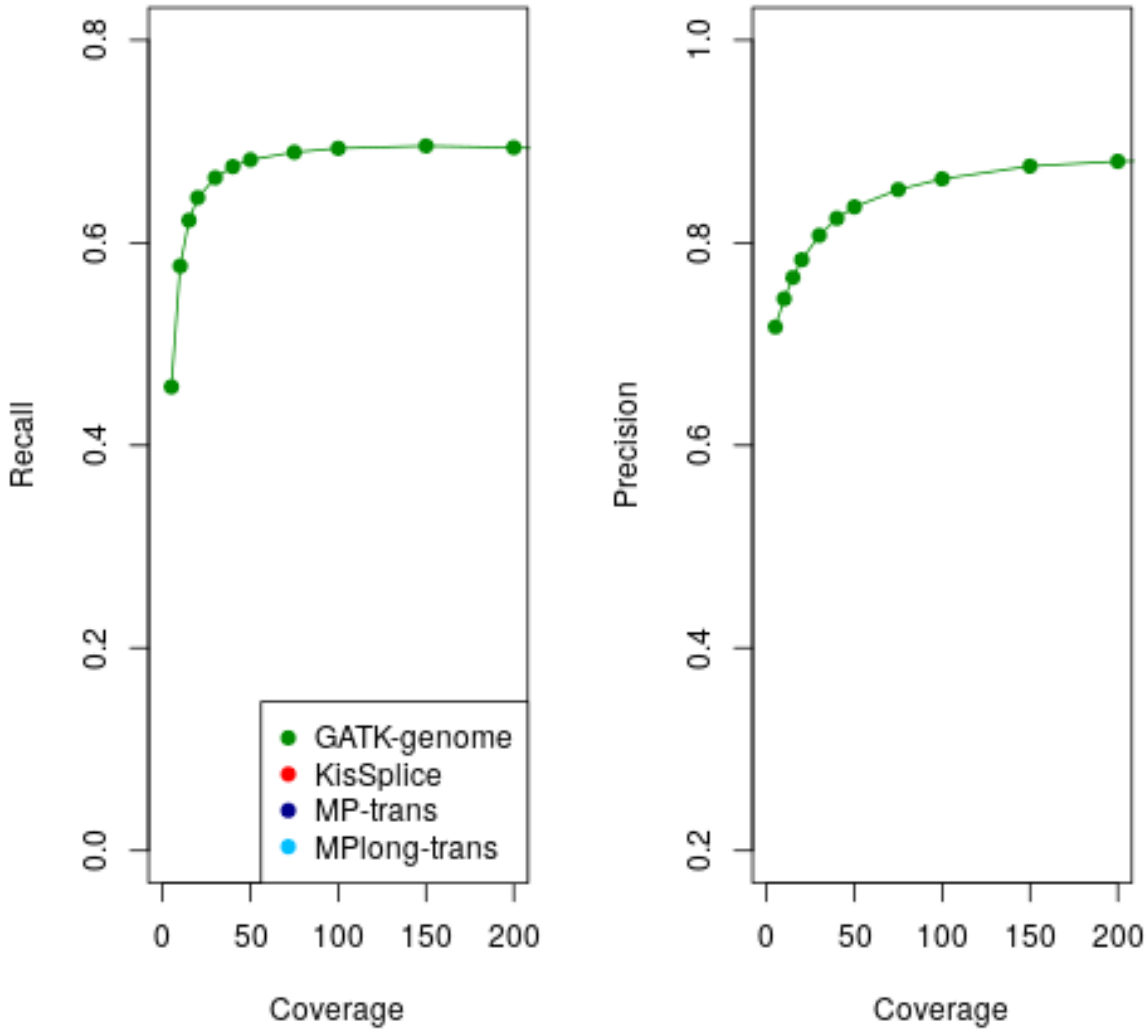
- genome and annotation available
- **1000G** project : genotype of 1000 individuals
- **Geuvadis** : RNAseq of these individuals

RNAseq from 2 x 5 central europeans and 2 x 5 toscans

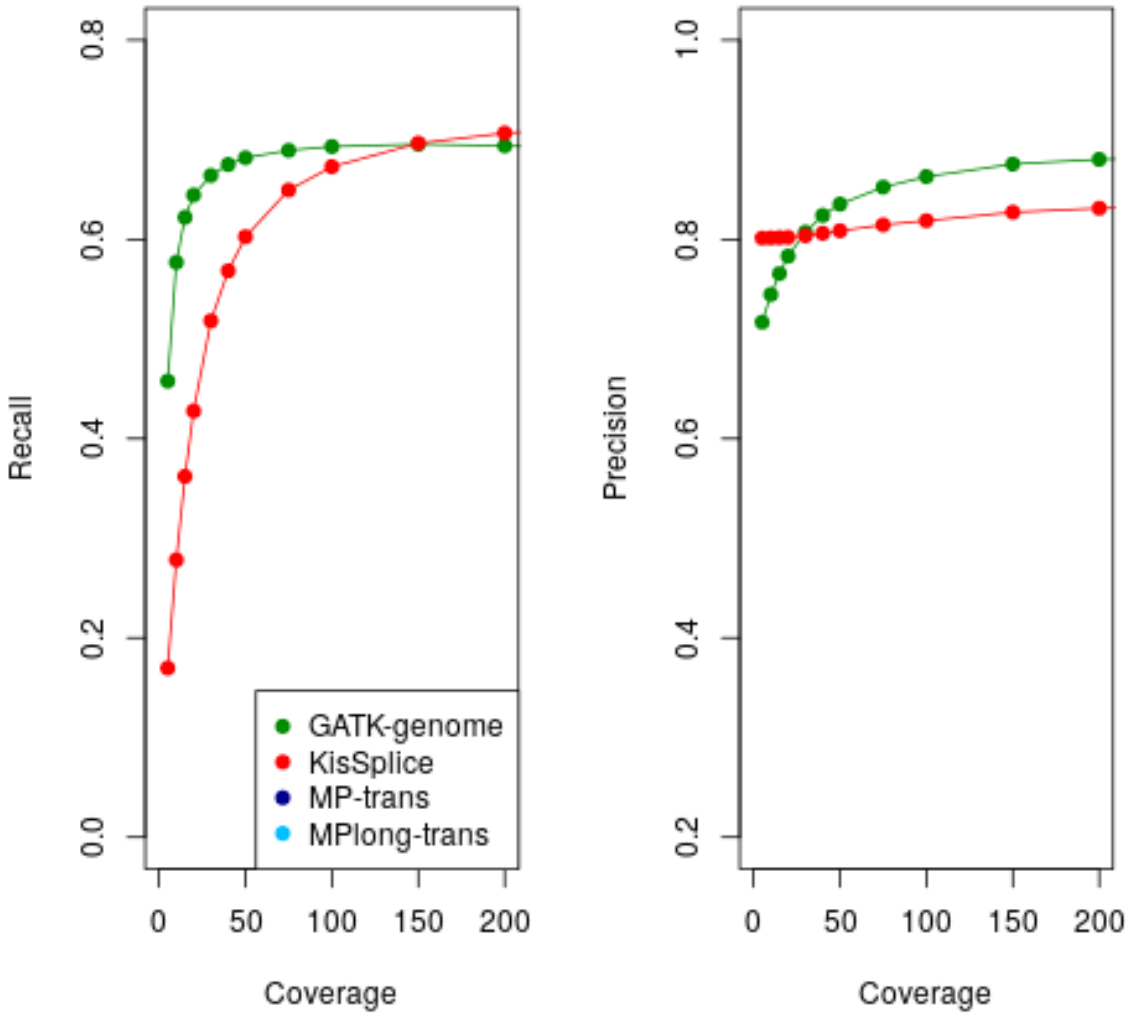
Recall


Recall
0.0
0.2
0.4
0.6
0.8
0
50
100
150
200
Coverage
GATK-genome
KisSplice
MP-trans
MPlong-trans

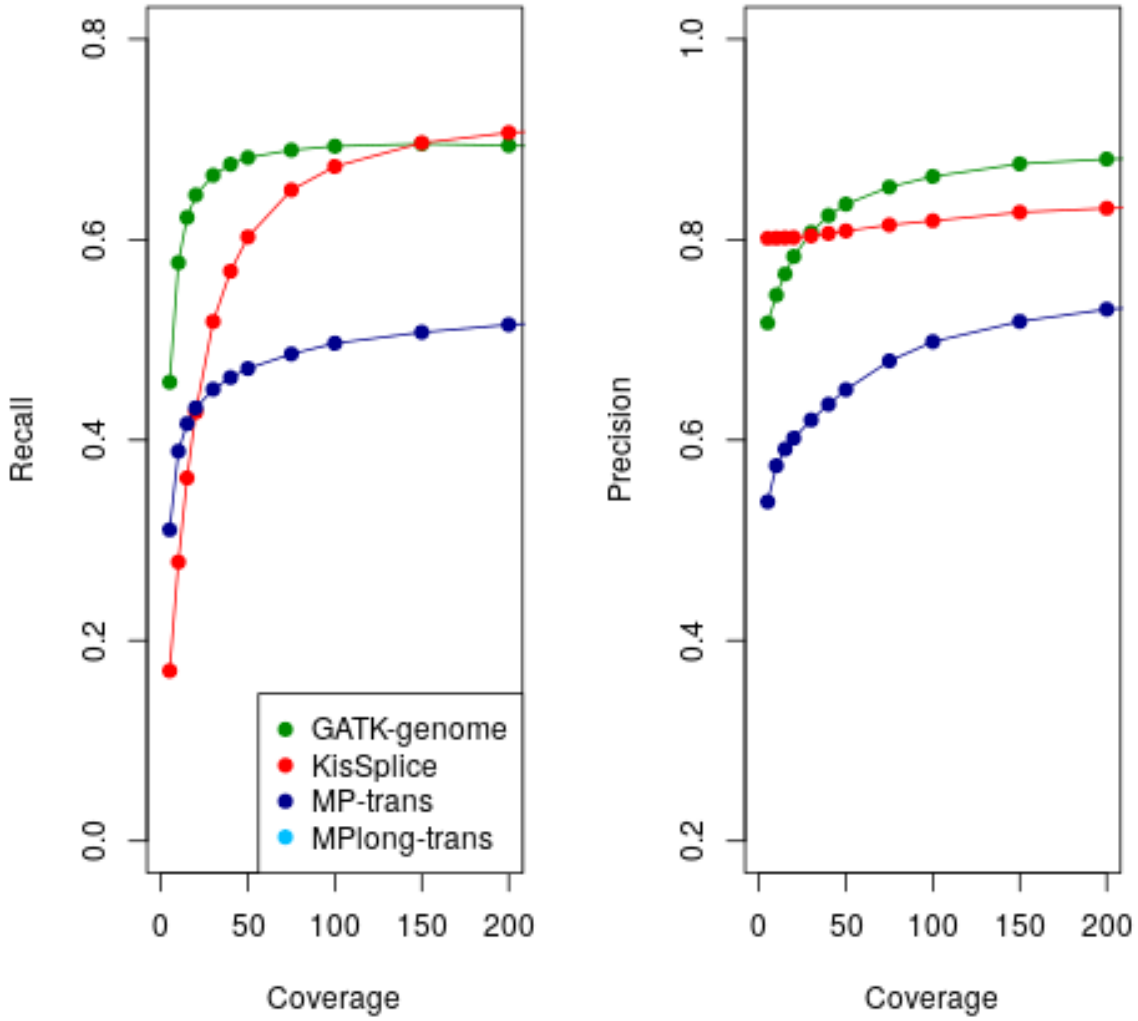
Recall and precision



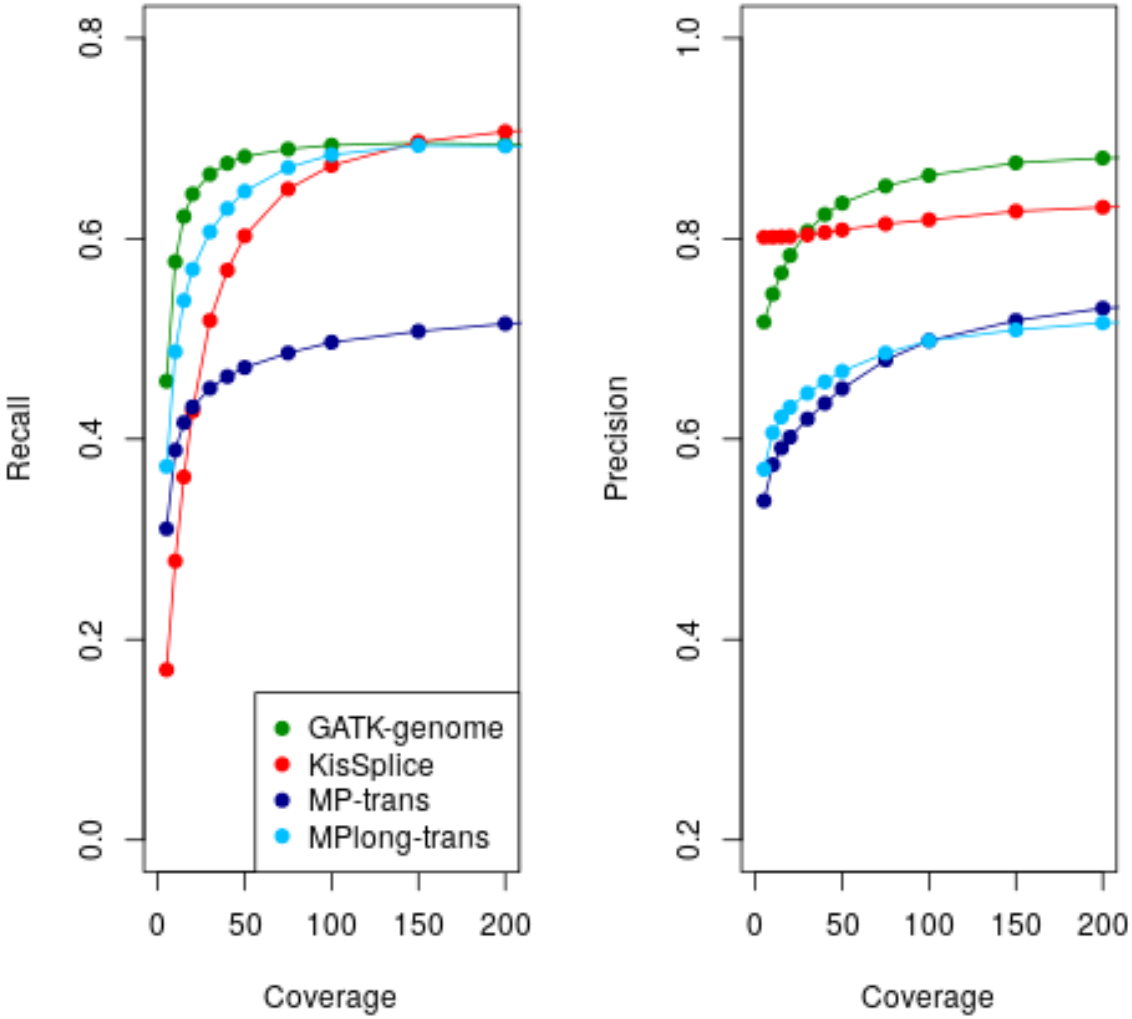
Recall and precision



Recall and precision



Recall and precision



Recall and precision

- GATK > KisSplice > Mplong > MP
- False negatives include rare variants, and SNP clusters
- False positives include sequencing errors and inexact repeats

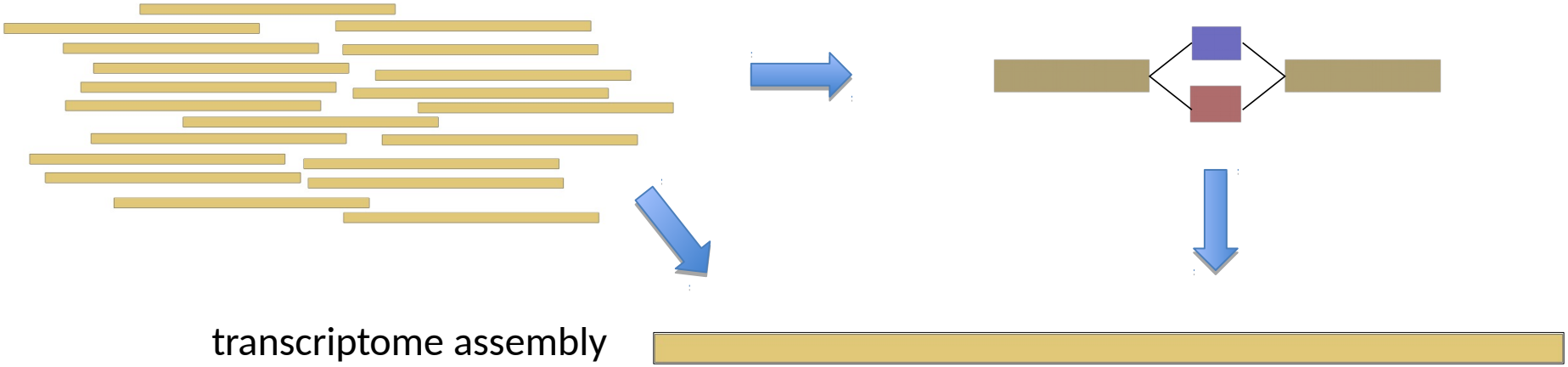
Downstream analyses

i) What is the impact on the protein ?

ii) Is my variant associated to a given biological condition ?

Prediction of the amino acid changes

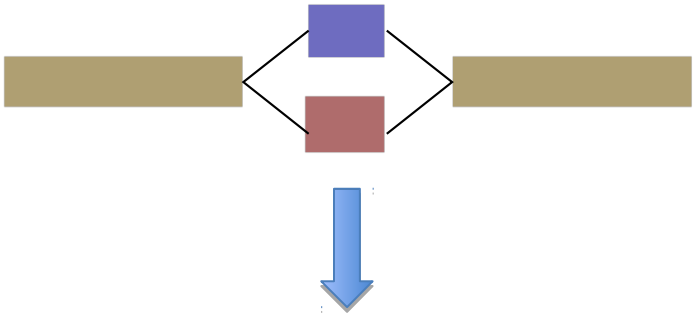
reads from RNAseq



transcriptome assembly

Prediction of the amino acid changes

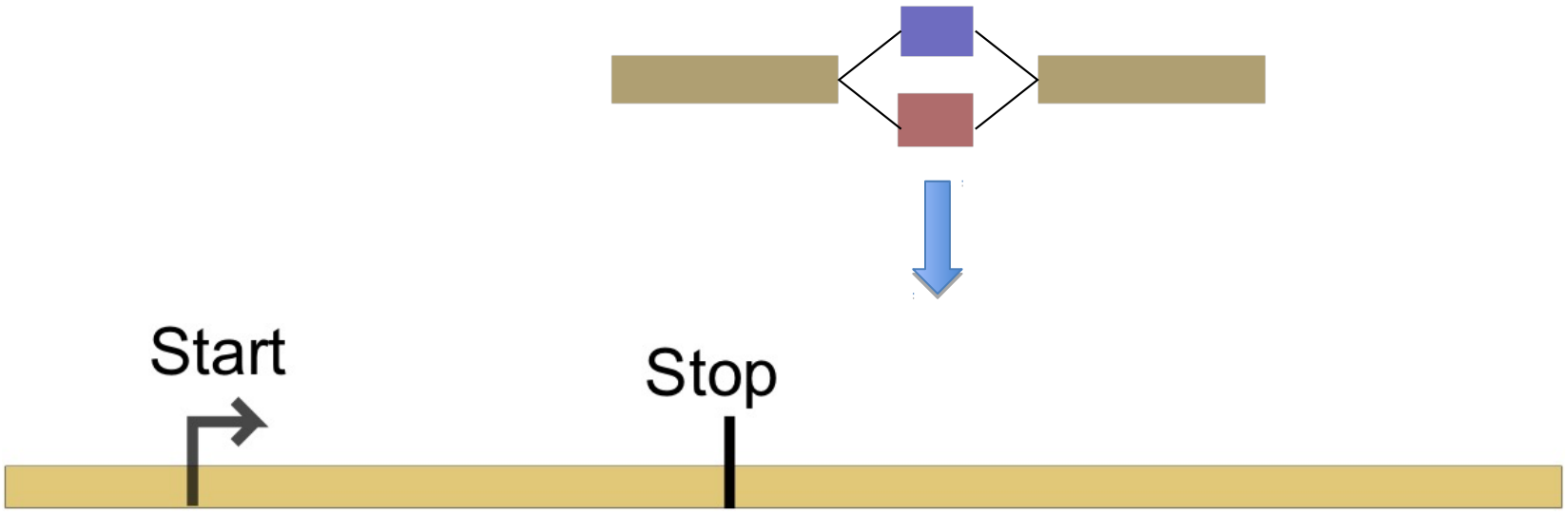
KisSplice2refTranscriptome



transcriptome assembly

Prediction of the amino acid changes

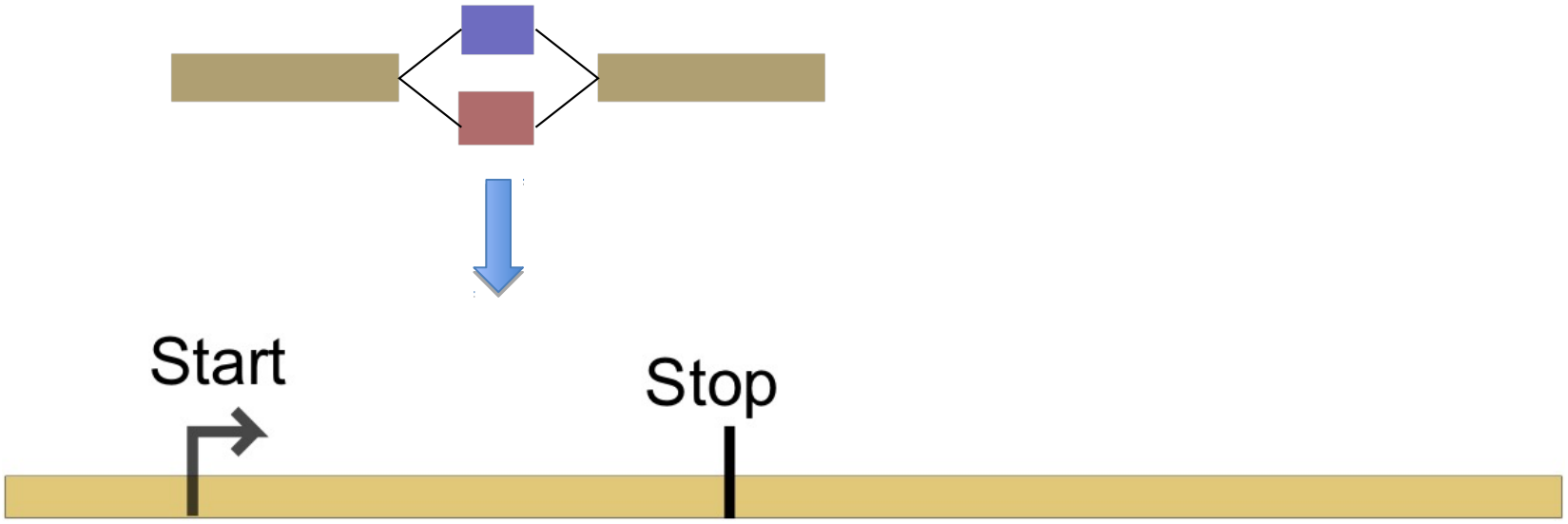
KisSplice2refTranscriptome



transcriptome assembly

Prediction of the amino acid changes

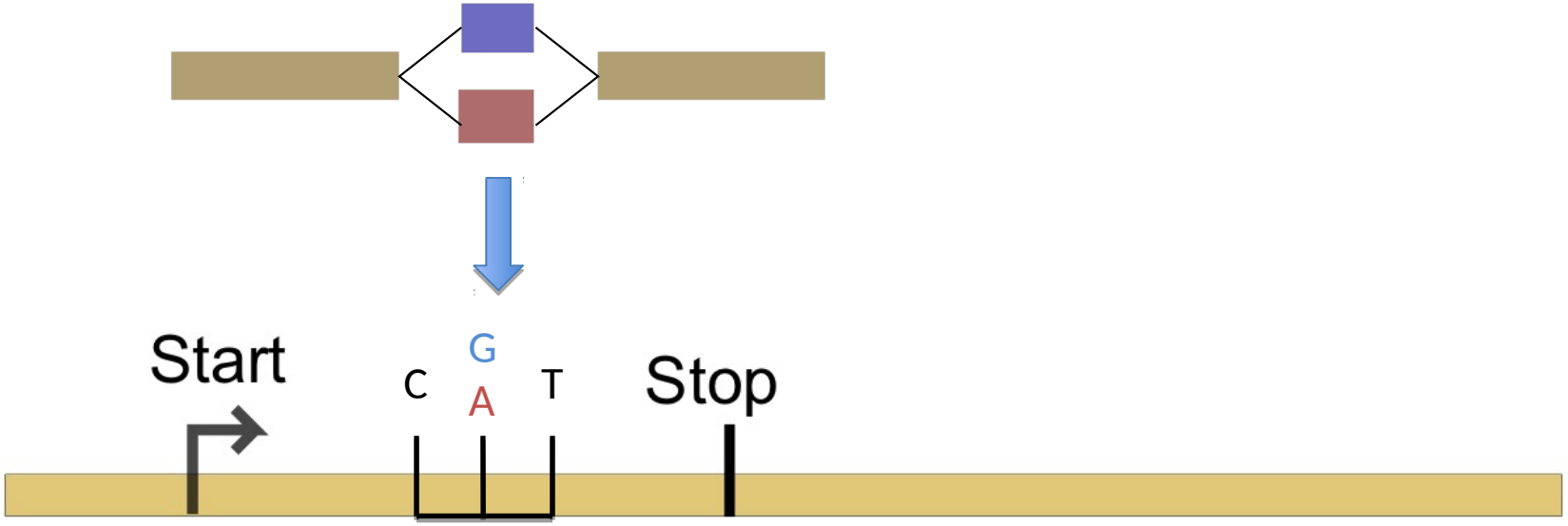
KisSplice2refTranscriptome



transcriptome assembly

Prediction of the amino acid changes

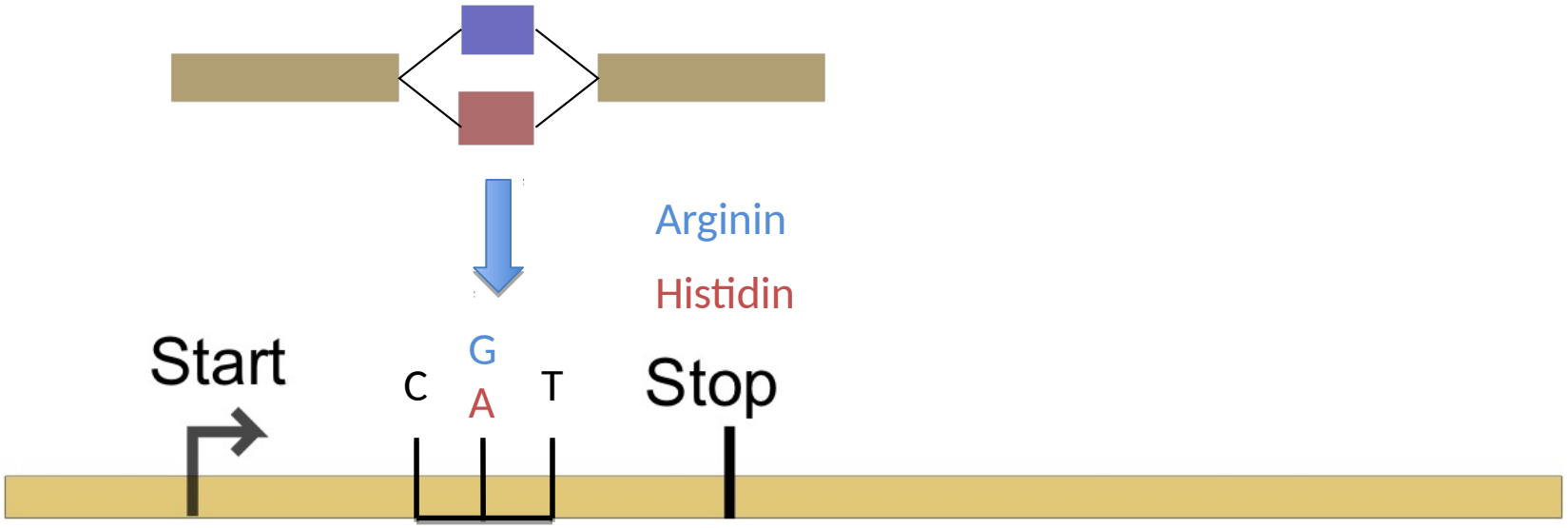
KisSplice2refTranscriptome



transcriptome assembly

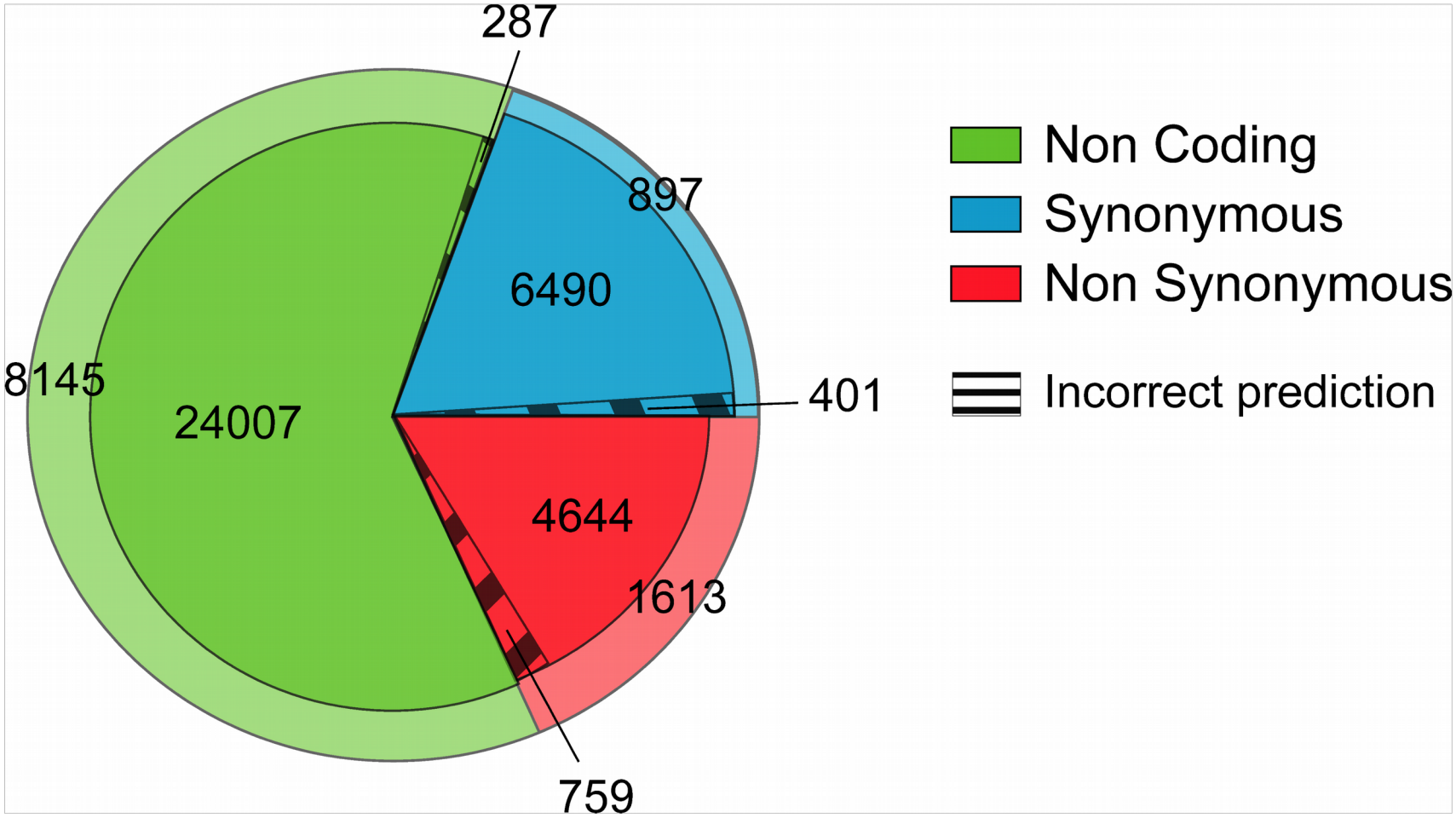
Prediction of the amino acid changes

KisSplice2refTranscriptome

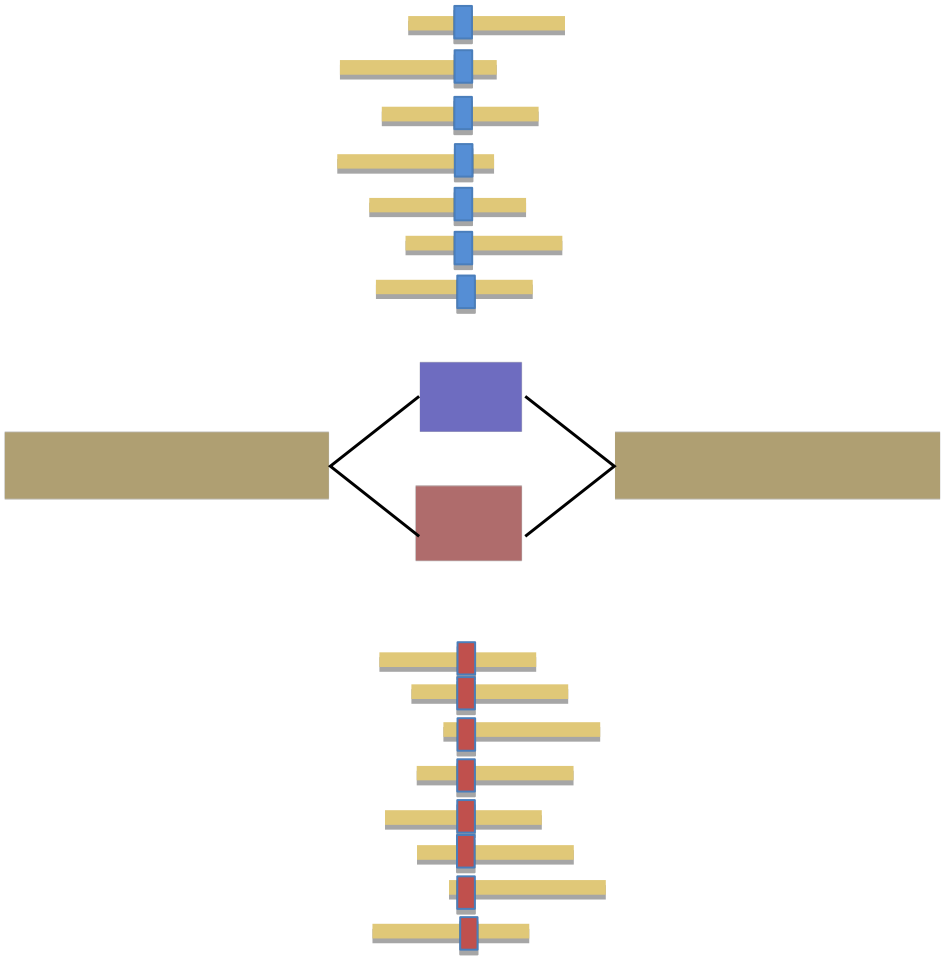


transcriptome assembly

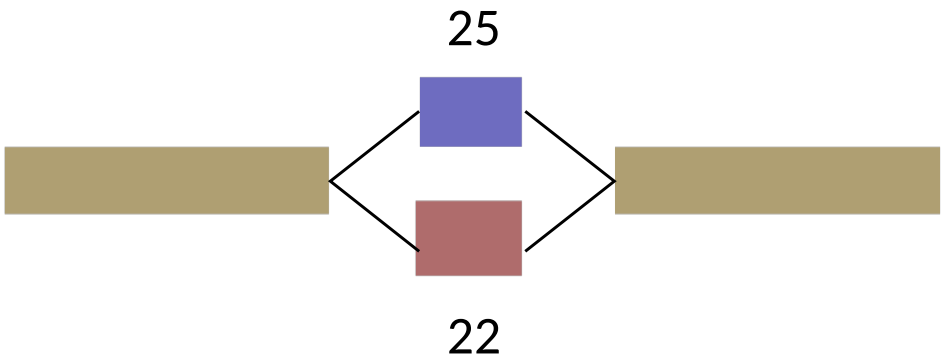
Prediction of the amino acid changes



Quantification of variants



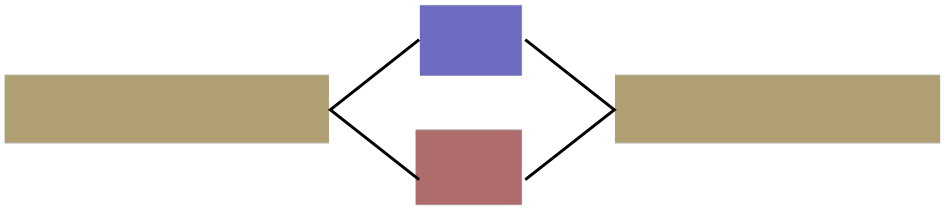
Quantification of variants



Is my variant specific to a condition ?

Condition 1

Condition 2



KissDE

Statistical Analysis

- Count regression with negative binomial distribution
- Generalised linear model, 2 way design, with interaction

$$\log \lambda_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij}$$

Mean gene expression

Contribution of variant i

Contribution of condition j

Interaction term

$$H_0 : \{(\alpha\beta)_{ij} = 0\}$$

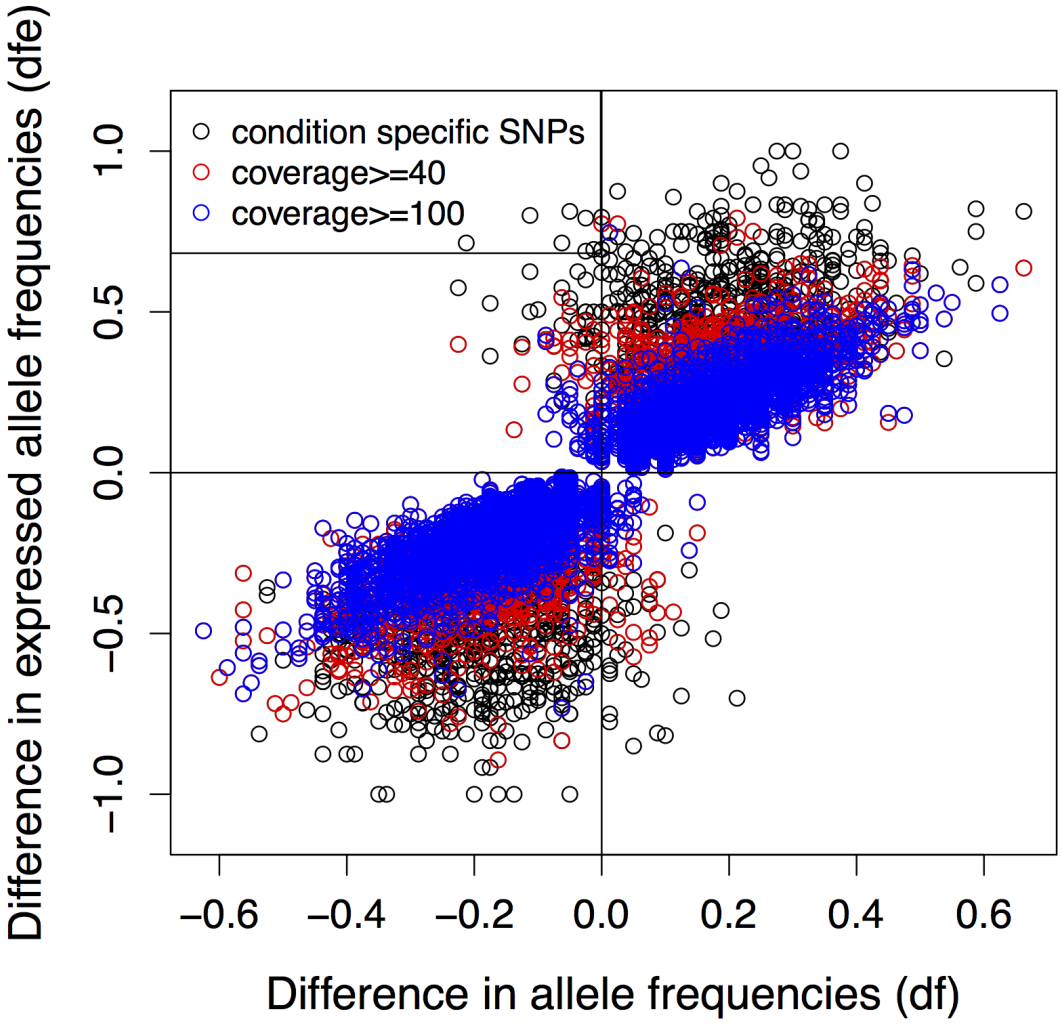
- Target hypothesis:
- Likelihood ratio test
- P-values adjusted with benjamini-hochberg procedure

Magnitude of the effect : Dfe

	C1	C2
Var. 1	23 20	2 1
Var. 2	4 2	18 16

Frequency of variant 1 in C1 : 90%
Frequency of variant 1 in C2 : 10%
Difference of allele frequency: 90 – 10 = 80 %

Dfe is a good proxy for Df when genes are highly expressed



Pipeline : Input

fastq files :

@HWI-D00393:39:C3PHDACXX:2:1101:8716:2263 1:N:0:CTATAC
TCAGTCTTTTTCTCTCTCCTAGTCCCATAATGCCTTGACGAACATCAGTCTGTCCCATAATGCTCTCCTATTA
+
CCCCCCCCCCCCCCCCCCCBBBBBBBBBBBBBBFFFFFFFFFFFFFFFFFJJJJJJJJJJJJJJJJJJJJJJJJJJJHHI92:?EGIJ
@HWI-D00393:39:C3PHDACXX:2:1101:11867:2372 1:N:0:CTATAC
TACGTCGGAAATATATAAAAAAATGGCACTGGAAGAATTCACTGGCACTGGAAGAATTCATGGCACTGGA
+
CCCCCCCCCCCCCCCCCCCCCB?BCBBBBBBBBCBBCBBFFFFFHHHHHJJJJJJIJJJHFHGIJJIIIHBGHIHDFH
@HWI-D00393:39:C3PHDACXX:2:1101:3047:2629 1:N:0:CTATAA
CCGTGGTCCCATAATGCCTTGACACGGTCTTGGCACTGGAAGAATTCAAGAATTCATGGCACTGGAAGAA
+
CCCCCCCCCCCCCCCBBBBBBBB?BBBBBBBBBBBBBBBBBBBCBBDDDDHHHHHJJJJJJJJJJJJIJFHIJJIJDH;
@HWI-D00393:39:C3PHDACXX:2:1101:8009:2719 1:N:0:CTATAC
CCCGACTGAAGGTGTACTCATGGCCAGGTCCCATAATGCCTTGACCTCTCCTATTACTGAAGGGGTGTACT
+
CCCCCCCCCCCCCCCCCCBCCCCCCCCCCBC?CBBBBBBBBBBBBBCFFFFFGGHHFHHHHHHHHJJIIJIIJJJI

Pipeline : Output

KisSplice + KissDE + Kissplice2refTranscriptome output :

c54770_g1_i3	1128	bcc_487|Cycle_259CAA	CCA	Q	P	NonSyn.	No

Pipeline : Output

KisSplice + KissDE + Kissplice2refTranscriptome output :

c54770_g1_i3	1128	bcc_487|Cycle_259	CAA	CCA	Q	P	NonSyn.	No	
c54495_g1_i1	1245	bcc_6275|Cycle_1	GCA	GCC		A	A	Syn.	No

Pipeline : Output

KisSplice + KissDE + Kissplice2refTranscriptome output :

c54770_g1_i3	1128	bcc_487|Cycle_259	CAA	CCA	Q	P	NonSyn.	No	
c54495_g1_i1	1245	bcc_6275|Cycle_1	GCA	GCC	A	A	Syn.	No	
c59969_g1_i1	1073	bcc_9888|Cycle_0	AGC	TGC	S	C	NonSyn.	Yes	

Pipeline : Output

KisSplice + KissDE + Kissplice2refTranscriptome output :

c54770_g1_i3	1128	bcc_487|Cycle_259	CAA	CCA	Q	P	NonSyn.	No	
c54495_g1_i1	1245	bcc_6275|Cycle_1	GCA	GCC		A	A	Syn.	No
c59969_g1_i1	1073	bcc_9888|Cycle_0	AGC	TGC		S	C	NonSyn.	Yes
c65293_g5_i2	847	bcc_6446|Cycle_0	N/A	N/A		N/A	N/A	NonCod.	Yes

Experimental Validations

Asobara tabida

ovaries infected with Wolbachia

wildtype

Asobara tabida

ovaries infected with Wolbachia

2 lines with variable dependency

no eggs

some sterile eggs

Asobara tabida

ovaries infected with Wolbachia
2 lines with variable dependency

RNAseq (ovaries) pool of 30 individuals

no eggs

some sterile eggs

Asobara tabida

ovaries infected with Wolbachia
2 lines with variable dependency

RNAseq (ovaries) pool of 30 individuals

no eggs

some sterile eggs

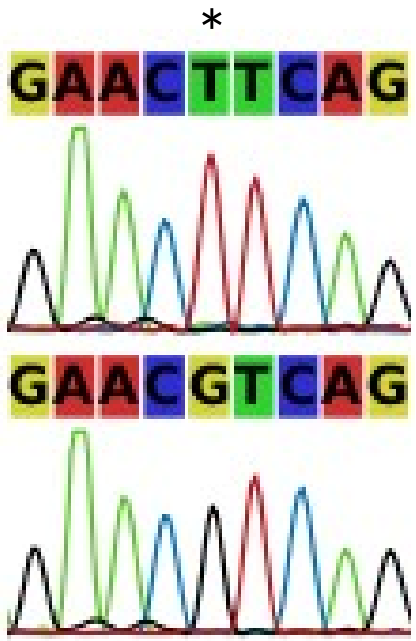
54532 SNPs of which 28392 condition specific

36/36 SNPs validated (rt-PCR + sequencing)
of which 7 predicted by Kissplice only

Experimental validation

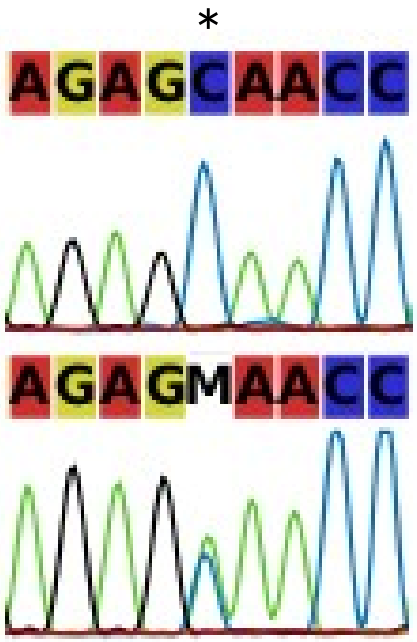
SFR2

SFR3



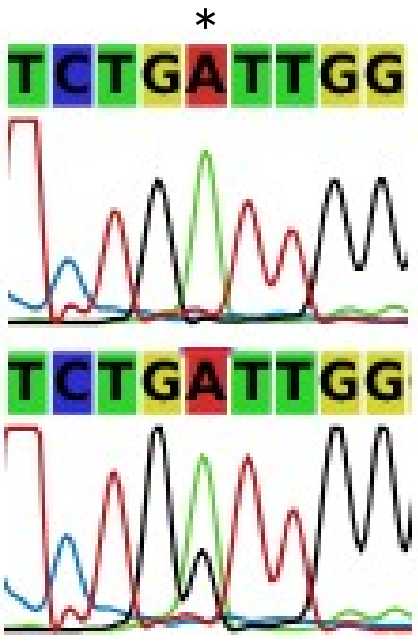
Nitric oxide synthase

SFR2: 0% / SFR3: 100%



Protein ovarian tumor

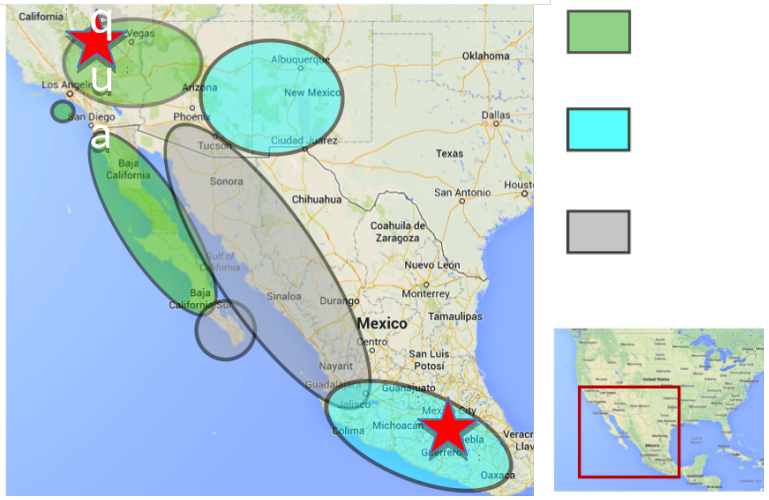
SFR2: 0% / SFR3: 33%



Pyrimidodiazepine synthase

SFR2: 100% / SFR3: 67%

Drosophila mojavensis & *D. arizonae*



D. mojavensis

D. arizonae

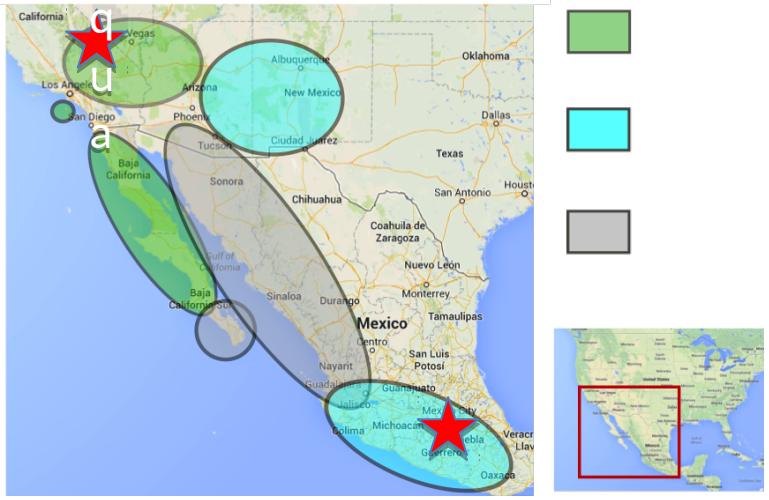
Sympatry

Divergence < 0.5 MYA

D. mojavensis

D. arizonae

Drosophila mojavensis & *D. arizonae*



D. mojavensis

D. arizonae

Sympatry

Divergence < 0.5 MYA

D. mojavensis *D. arizonae*

25 SNPs choosen for validation :

- 3/4 inexact repeats validated
- 19/21 SNPs and divergent sites validated
- 3 non validated cases

Conclusion

Detecting SNPs in non model species from (pooled) RNAseq data is possible

- Similar performances (precision and recall) compared to methods using a ref. genome

- Average precision is 77 %
- Filtering out SNPs which are not condition specific increases the precision to 96 %

Lopez-Maestre et al. NAR 2016 - http://kissplice.prabi.fr/TWAS/

Open questions

Two main challenges for SNP detection

1- Rare variants Vs sequencing errors

2- Inexact repeats Vs Close SNPs

What is the recall and precision of our method in species with large effective population size ?

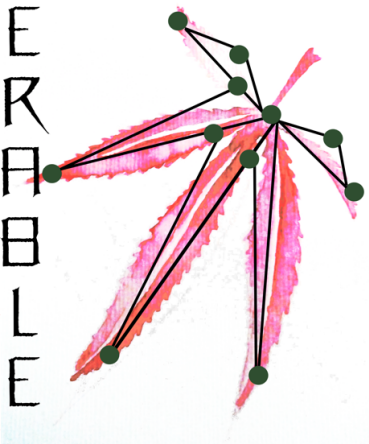
People

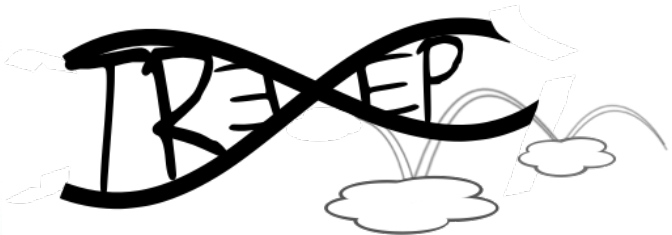
- KisSplice models and algorithms
  - Leandro Ishi, Gustavo Sacomoto, Rayan Chikhi, Pierre Peterlongo, Marie-France Sagot
- KisSplice2RefTranscriptome
  - Hélène Lopez-Maestre, Mathilde Boutigny
- KissDE
  - Camille Marchet, Clara Benoit, Audric Cologne, Janice Kielbassa, Lilia Brinza, Franck Picard
- Data analysis & Experimental Validation
  - Cristina Vieira, Hélène Lopez-Maestre (Drosophila)
  - Fabrice Vavre, David Monin, Natacha Kremer (Asobara)

Thanks to

H. Lopez-Maestre, L. Brinza,
C. Marchet, J. Kielbassa,
S. Bastien, M. Boutigny,
D. Monin, A. El Filali,
C. Carareto, C. Vieira, F. Picard,
N. Kremer, F. Vavre, M. Sagot,
V. Lacroix

ANR Colib'read

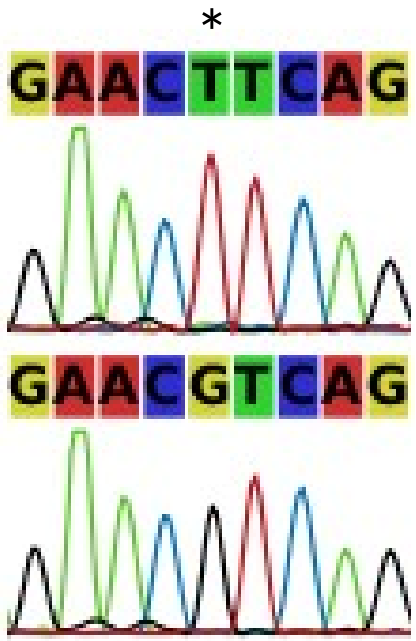




Experimental validation

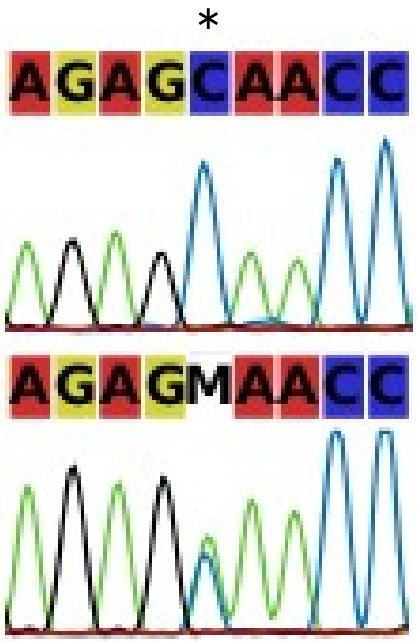
SFR2

SFR3



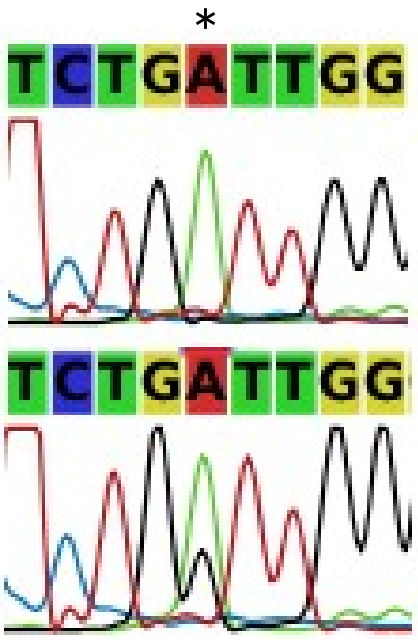
Nitric oxide synthase

SFR2: 0% / SFR3: 100%



Protein ovarian tumor

SFR2: 0% / SFR3: 33%



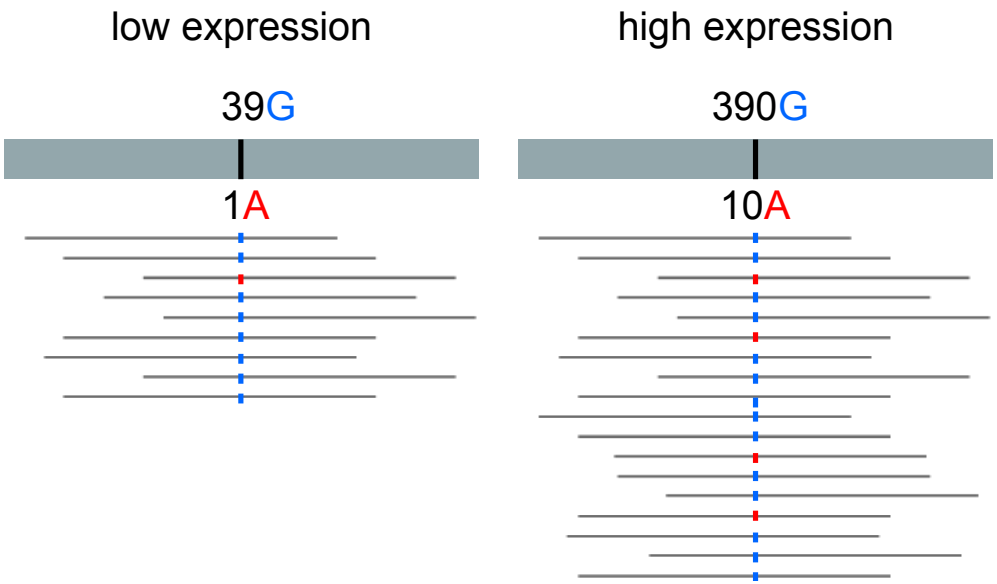
Pyrimidodiazepine synthase

SFR2: 100% / SFR3: 67%

Kissplice Filtres

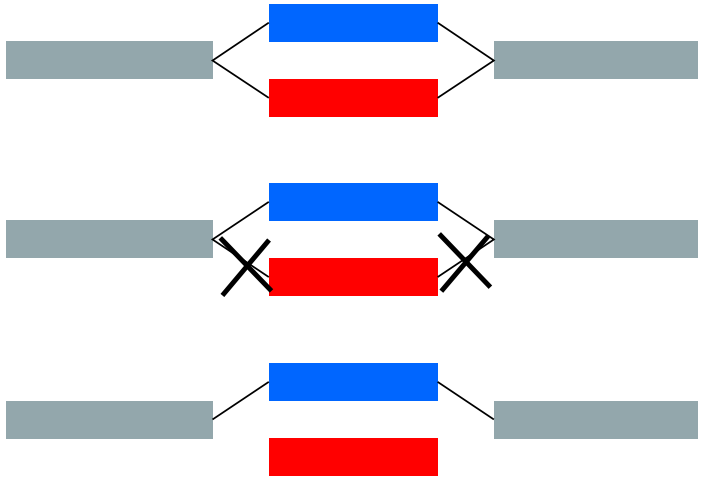
- Sequencing errors
- Inexacts repeats

Sequencing errors



absolute cutoff

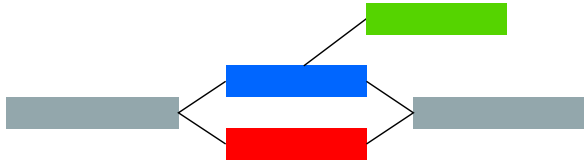
relative cutoff



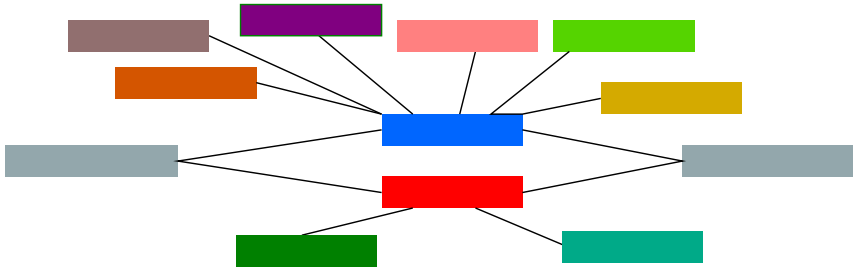
Inexacts repeats

...ATTGCTGCAATTC.....ATTGCTACAATTC...
...ATTGCTGCAATTC.....ATTGCTACAATTC...

...ATTGCTGCACTTC.....ATTGCTGCAATTC.....ATTGCTACAATTC...
...ATTGCTGCACTTC.....ATTGCTGCAATTC.....ATTGCTACAATTC...



X10 copies



Multiple SNPs

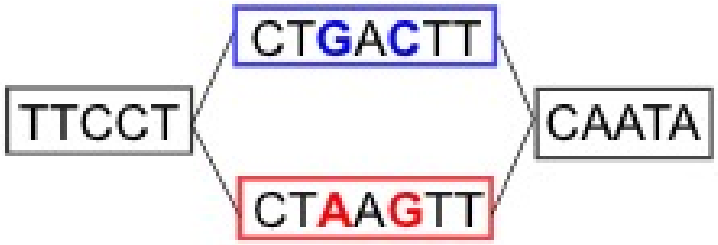
Exclus de l‘étude :

- Fort taux de faux positifs
- Probablement des répétitions inexactes

Close SNPs : phased

...TTCCTGACTTTGA...
...TTCCTAAGTTTGA...

distance <k



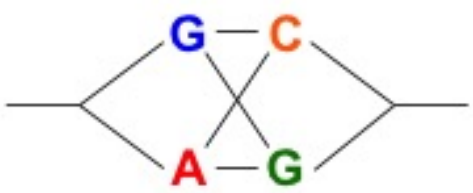
Bubble of size >2k+1

Close SNPs : non-phased

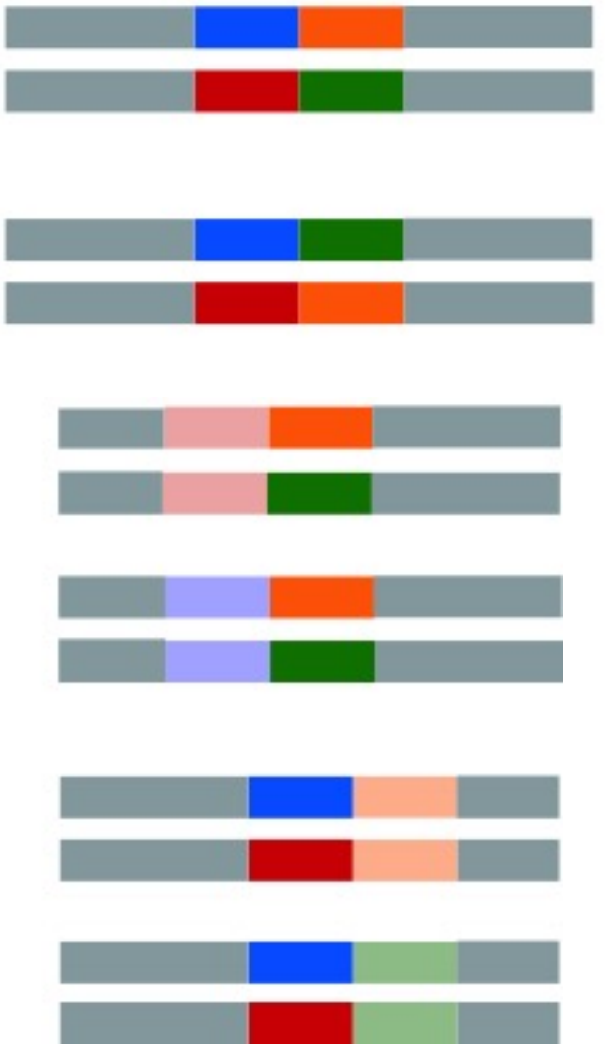
A)

...TCCTGACTTTG...
...TCCTAAGTTTG...
...TCCTGAGTTTG...
...TCCTAACTTTG...

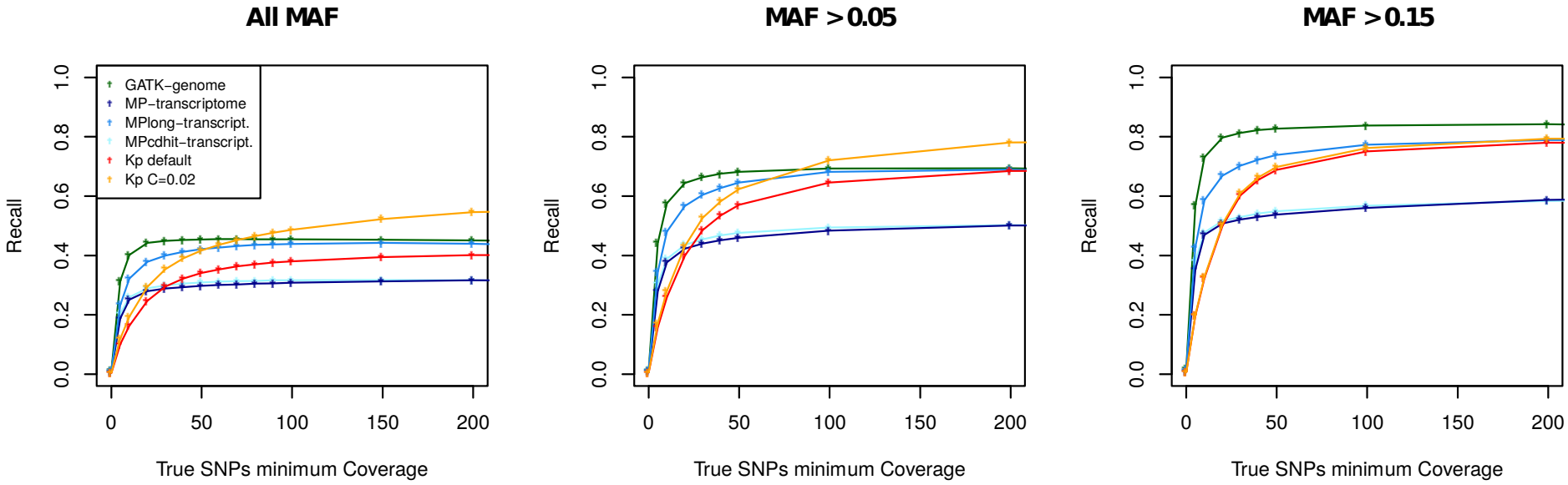
B)



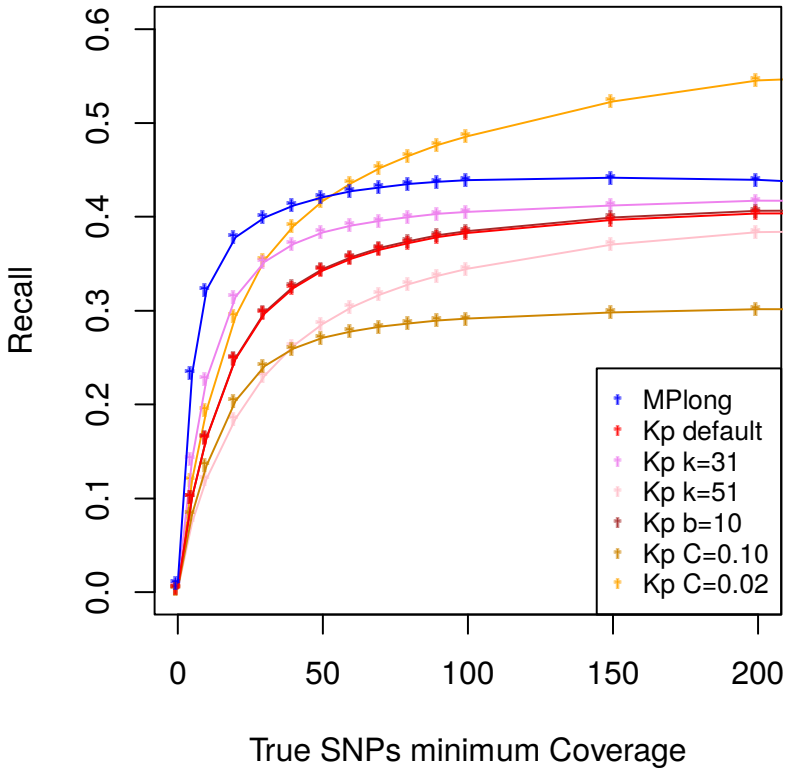
C)

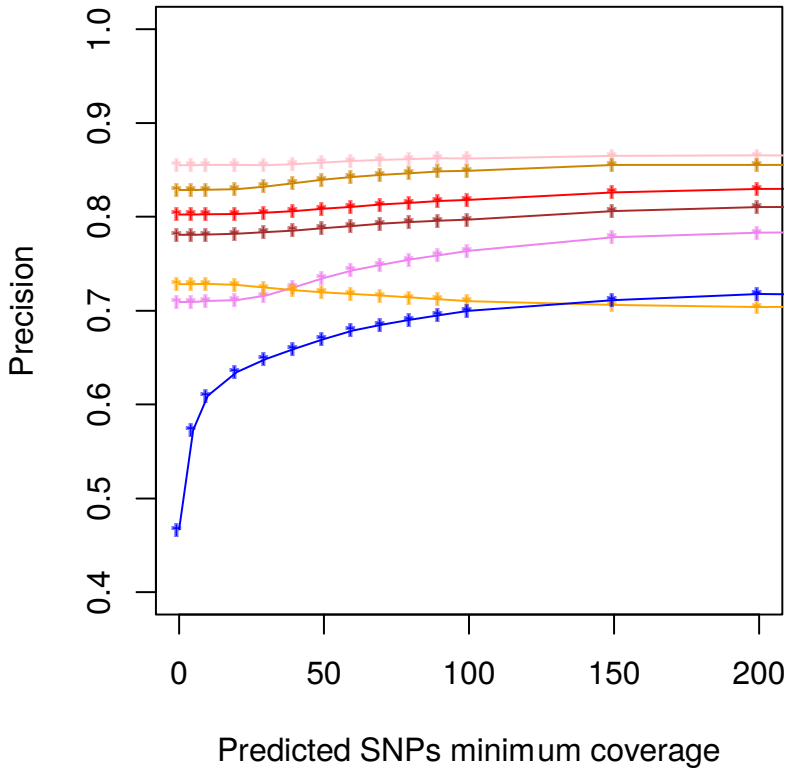


Recall vs MAF



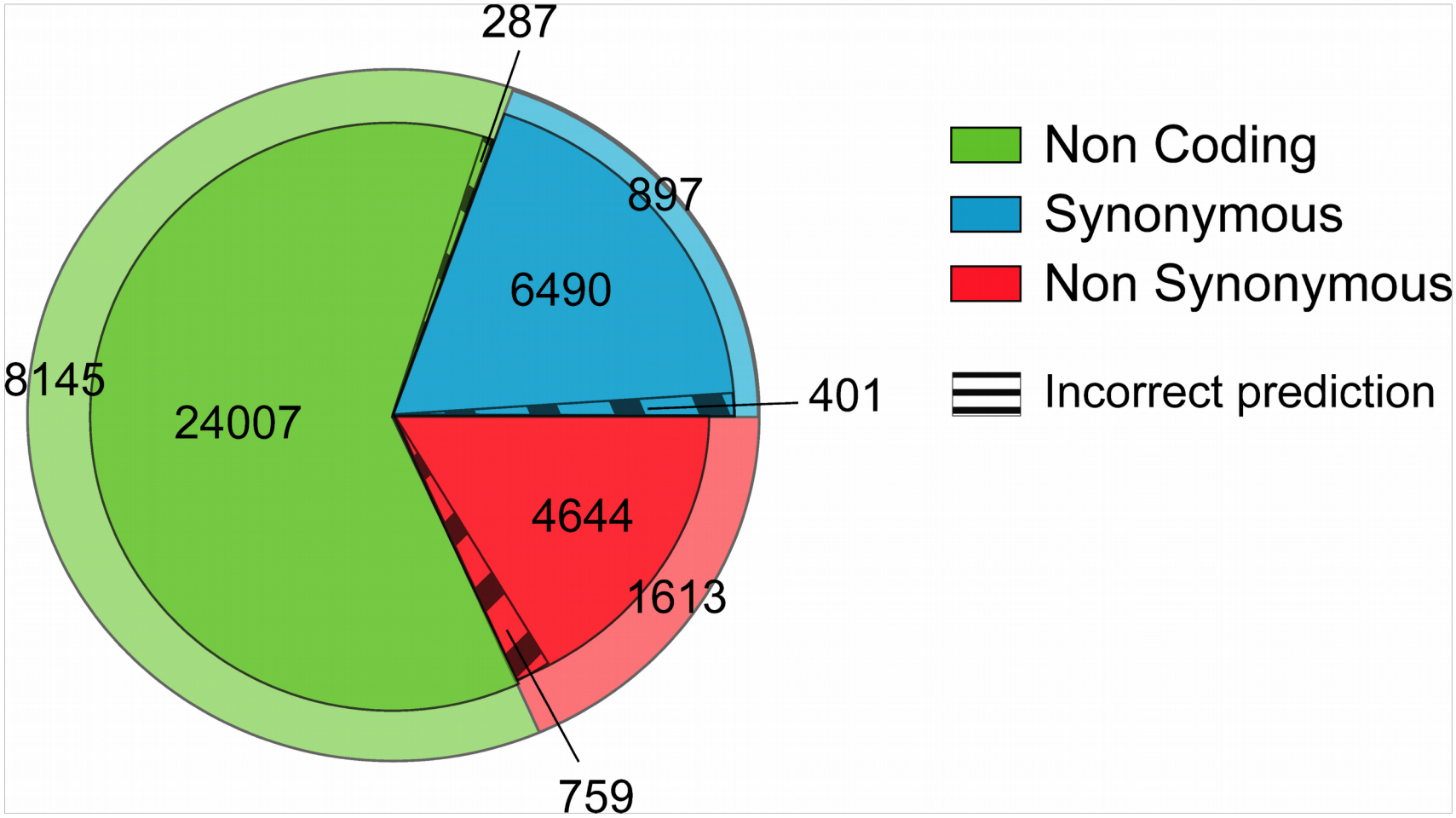
Kissplice parameters



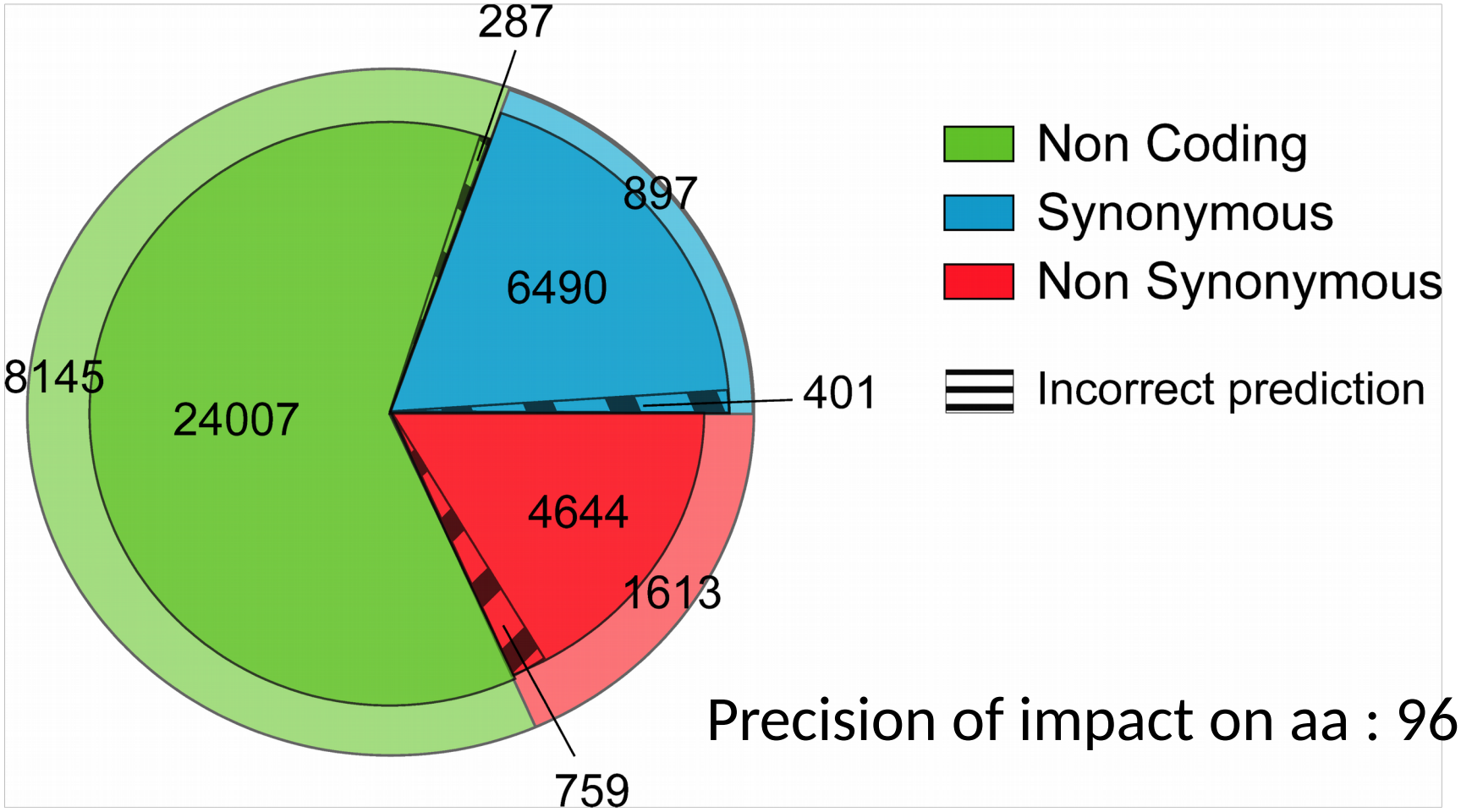


Impact on amino acids

Impact on amino acids

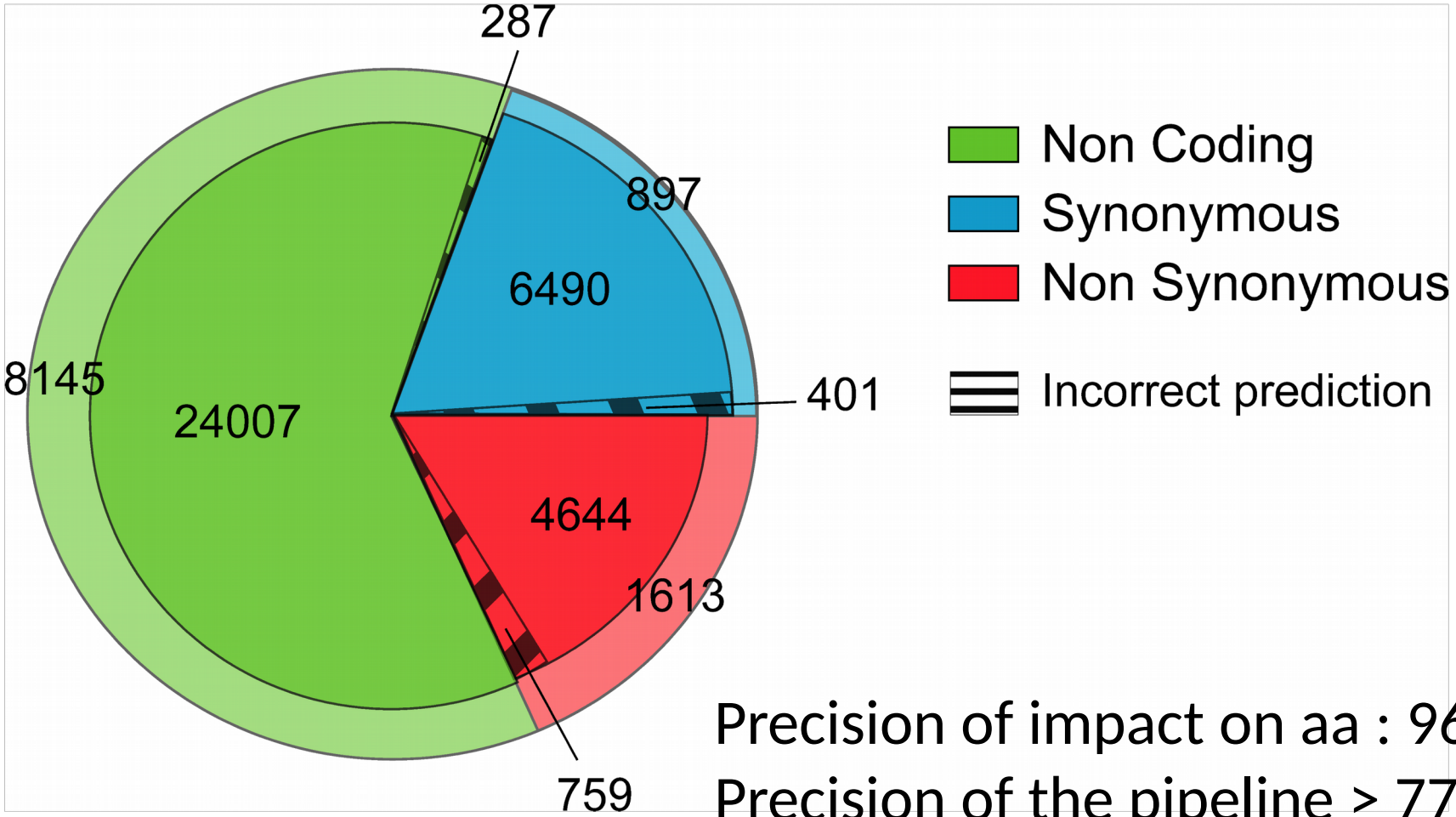


Impact on amino acids



Precision of impact on aa : 96%

Impact on amino acids



Precision of impact on aa : 96%

Precision of the pipeline > 77%

KissDE

	C1	C2
Var. 1	21	0
Var. 2	0	23

100% – 0%

KissDE

	C1	C2
Var. 1	21	0
Var. 2	0	23

100% – 0%

	C1	C2
Var. 1	21	7
Var. 2	0	16

100% – 30%

KissDE

	C1	C2
Var. 1	21	0
Var. 2	0	23

100% – 0%

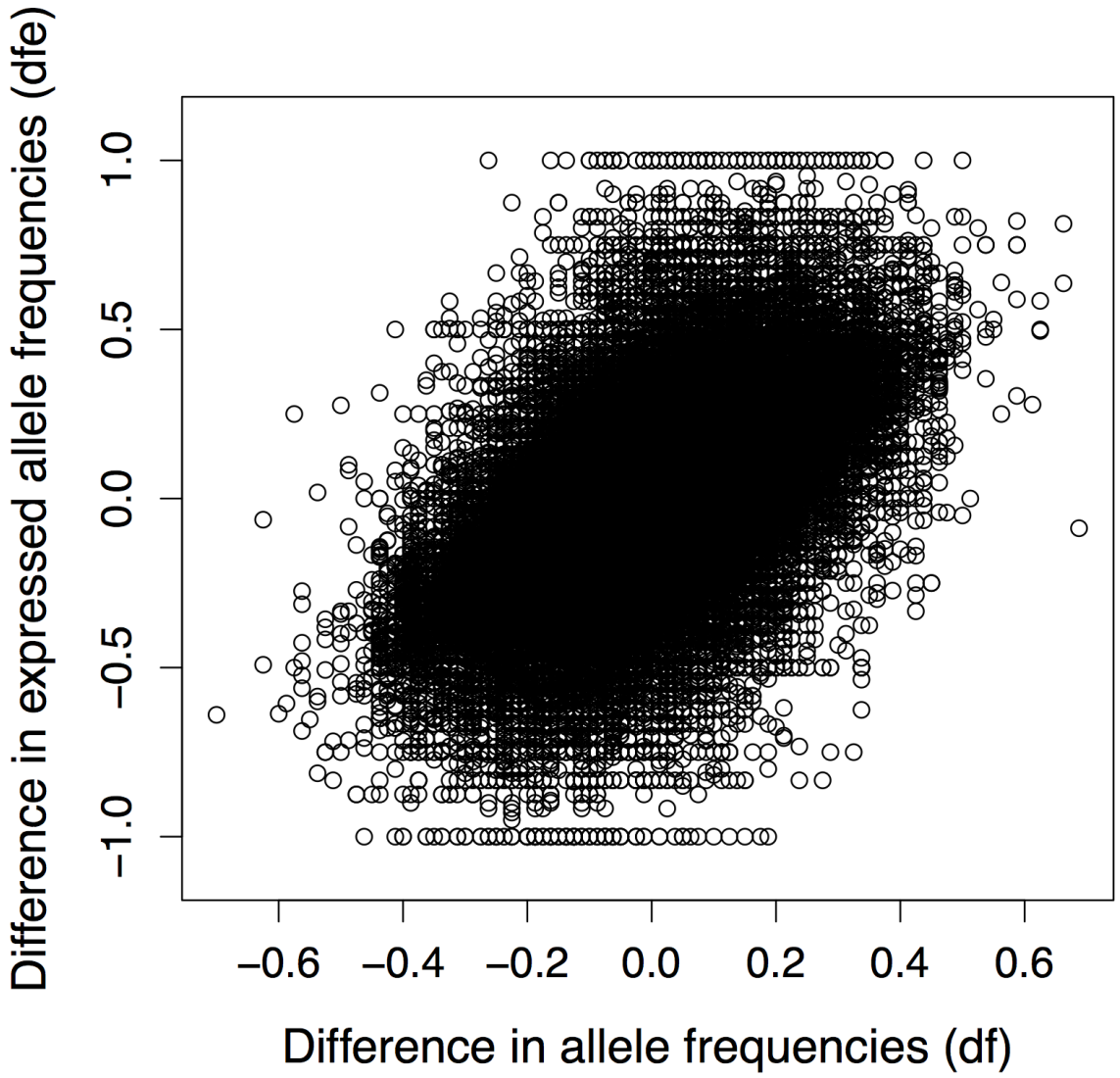
	C1	C2
Var. 1	21	7
Var. 2	0	16

100% – 30%

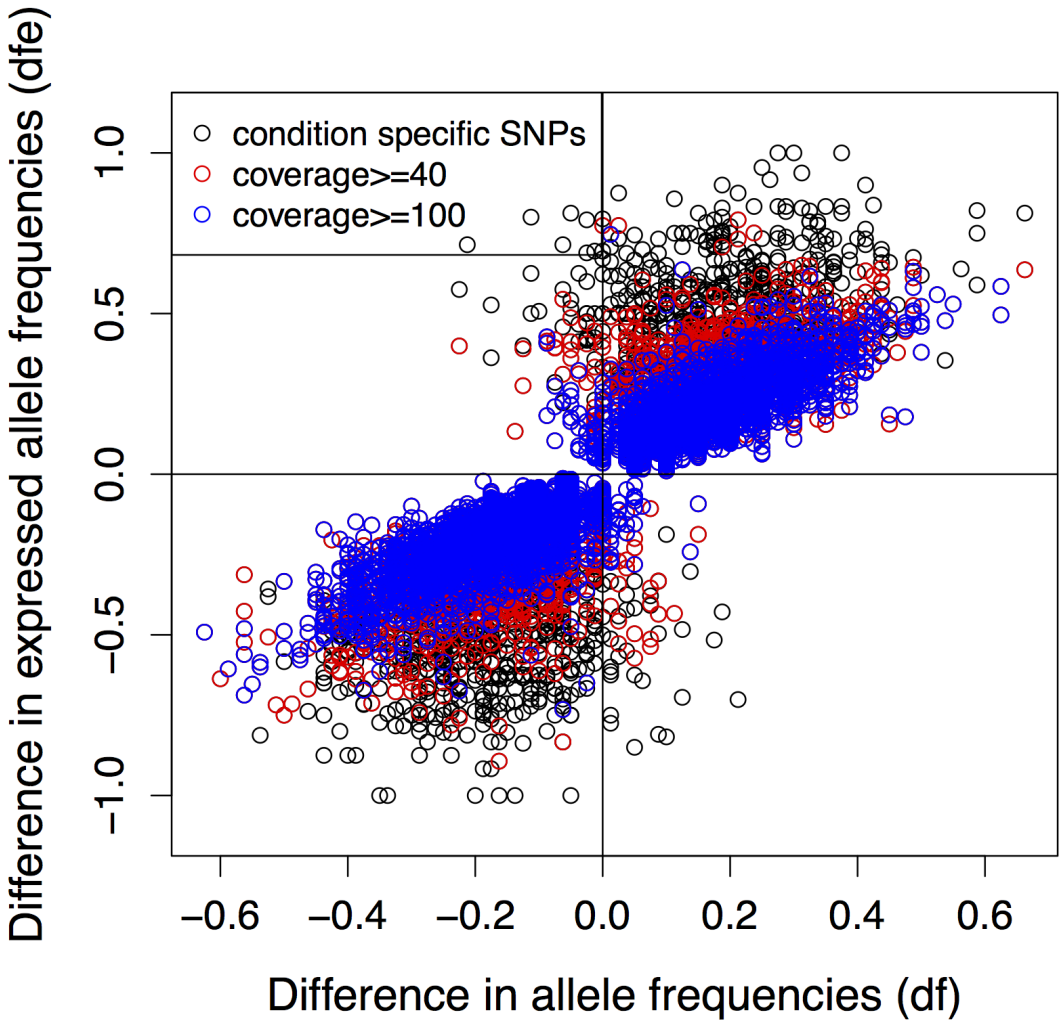
	C1	C2
Var. 1	15	7
Var. 2	6	16

71% – 30%

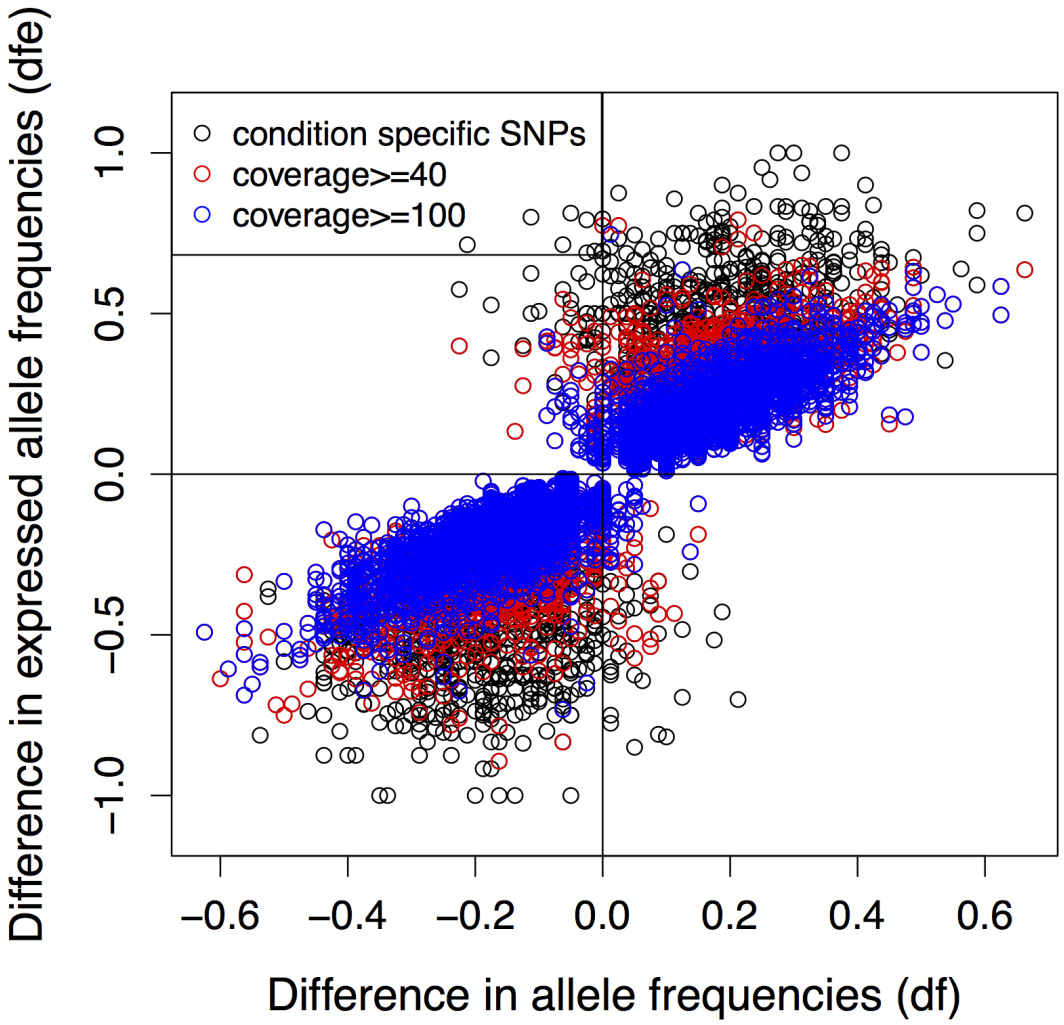
Difference in allele frequency



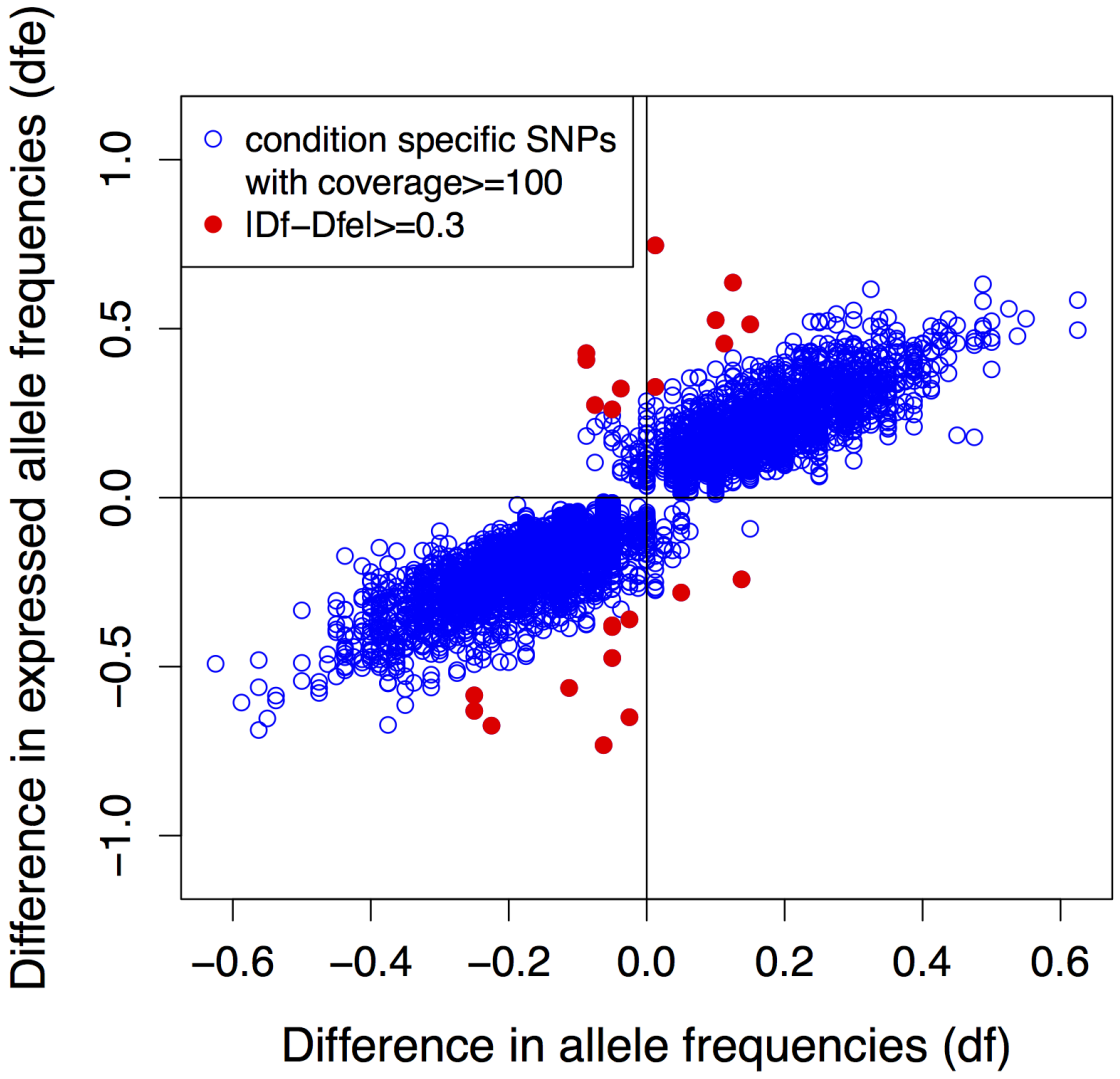
Difference in allele frequency



Difference in allele frequency



Difference in allele frequency



Pvalue ≤ 0.05

#reads ≥ 100