RDDpred: A condition-specific RNA-editing prediction model from RNA-seq data

Min-su Kim, Benjamin Hur and Sun Kim*

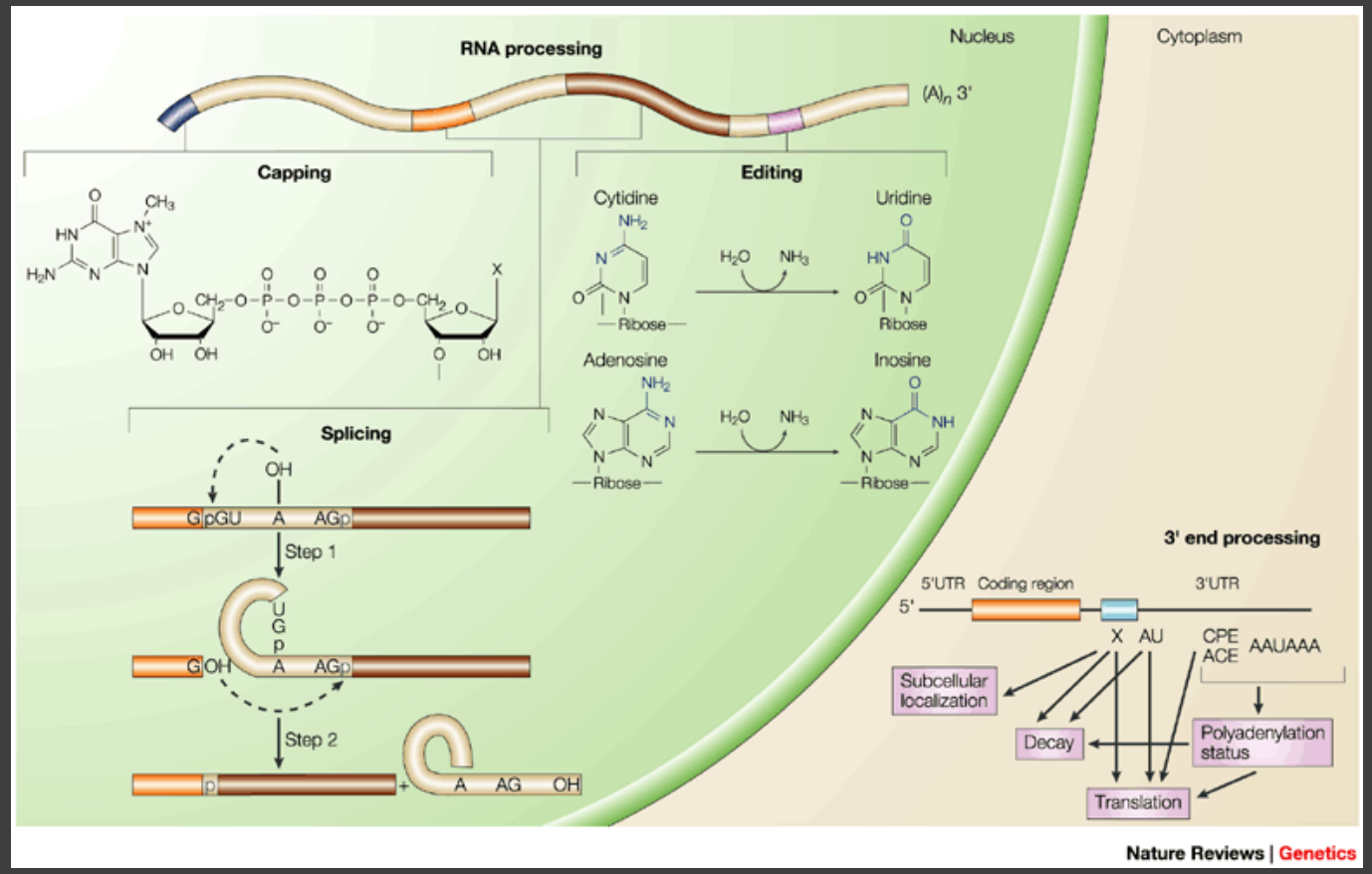
January, 2016

Bio & Health Informatics Lab., Seoul National University

Background

What is RNA-editing?

1) An innate post-transcriptional sequence modification mechanism



Keegan, Liam P., Angela Gallo, and Mary A. O'Connell. et. al., 2001

1) RNA-editing is a part of innate RNA processing pipeline, along with capping, splicing, and tailing.

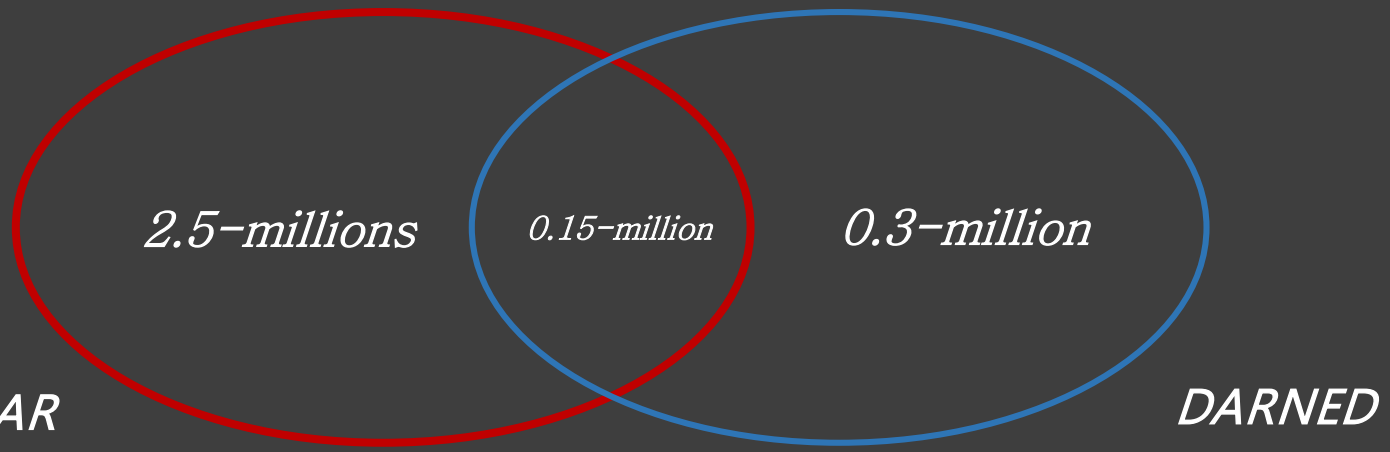
2) There are two classes of editing mechanisms, which is performed by ADAR, and APOBEC enzymes respectively.

3) The most common type of editing in metazoans is ADAR type.

What is RNA-editing?

2) Highly prevalent events in metazoans (including human)

# of currently reported editing-sites in hg19



Ramaswami, Gokul, and Jin Billy Li. et al., 2013

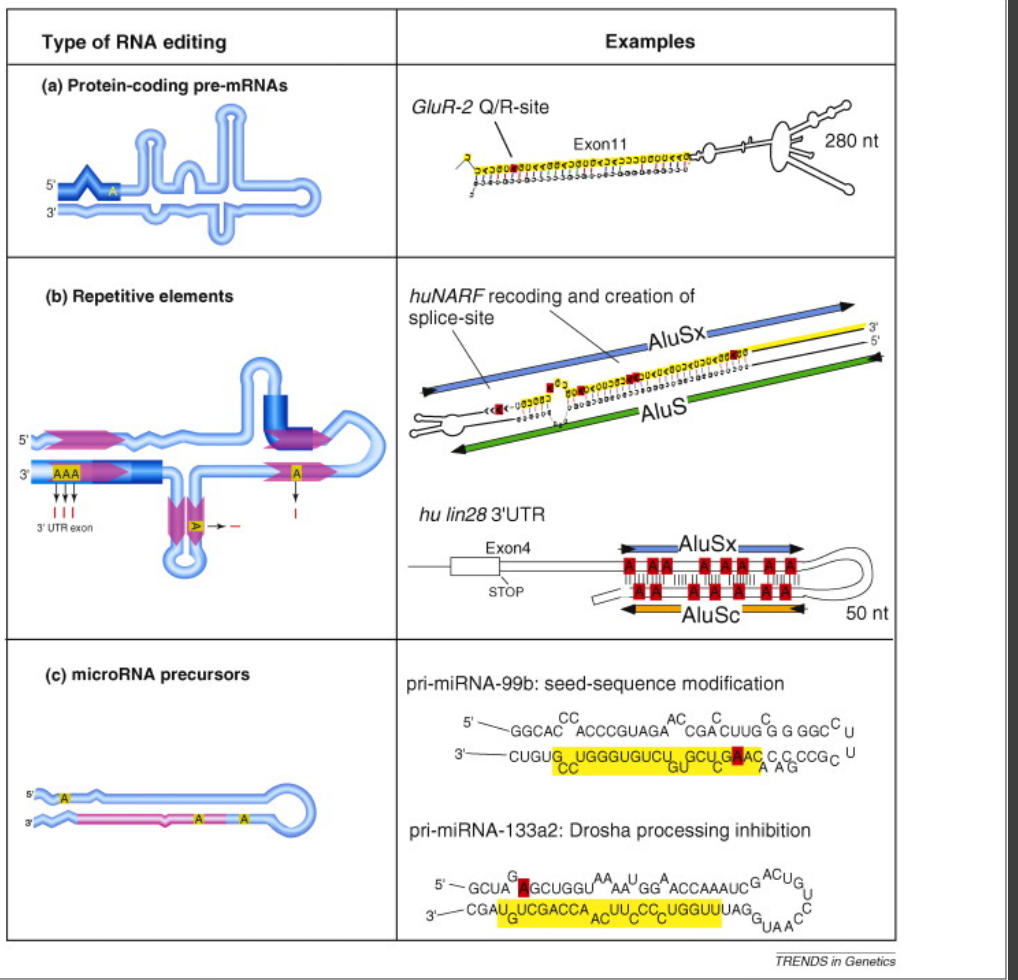
Kiran, Anmol, and Pavel V. Baranov. et al., 2010

According to two well-organized public databases of RNA-editing (RADAR, DARNED)

> The number of reported editing-sites have been accumulated up-to 2.65-millions (human genome 19)

What is RNA-editing?

3) Biologically crucial and tightly regulated events



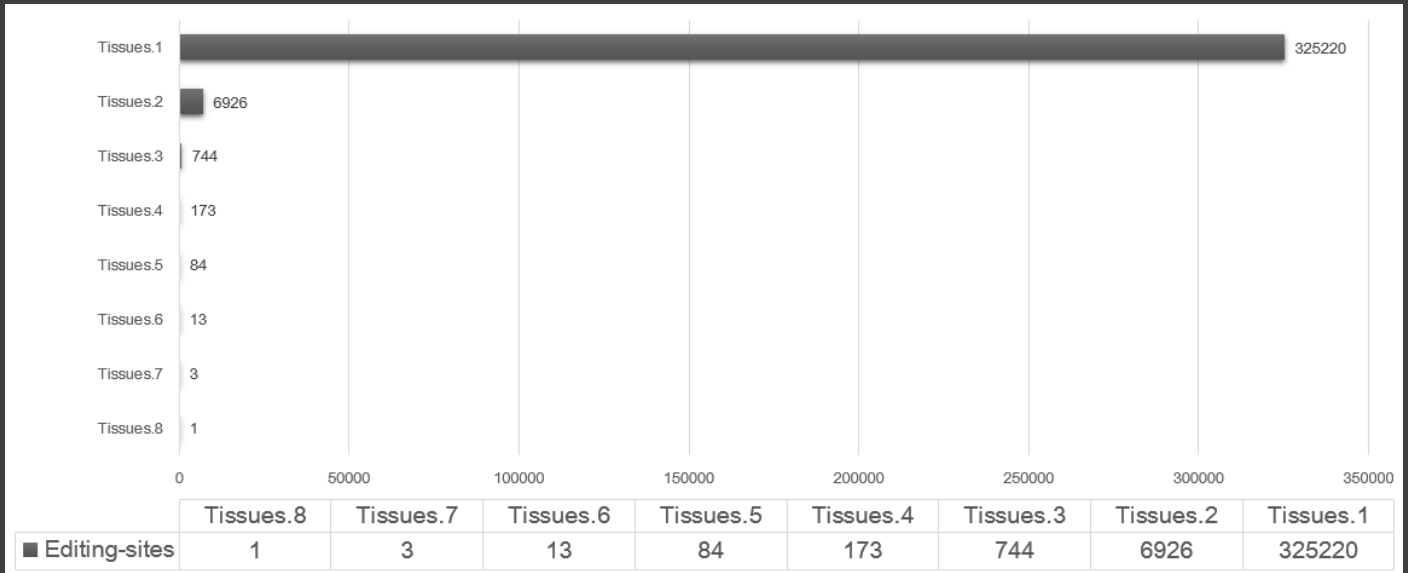
Farajollahi, Sanaz, et al., 2010

1) RNA-editing is known to **control various cellular processes**, such as protein activity, miRNA stability, miRNA target substitution, and Alu-repeats.

2) And recently, it is considered as **tightly regulated rather than random events** *(Nishikura, Kazuko. et. al., 2010).*

What is RNA-editing?

4) Highly condition-specific events

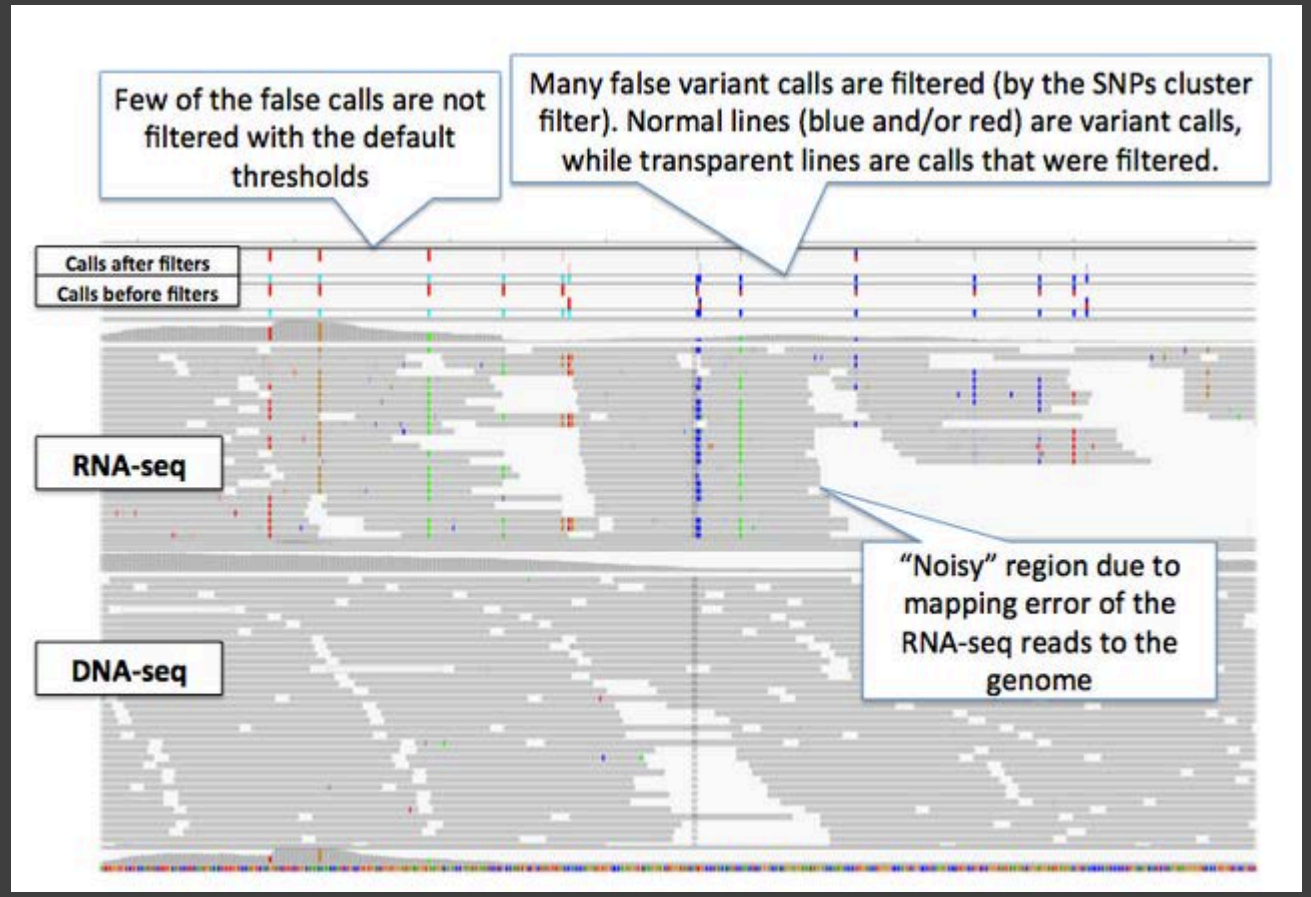


	Tissues.8	Tissues.7	Tissues.6	Tissues.5	Tissues.4	Tissues.3	Tissues.2	Tissues.1
Editing-sites	1	3	13	84	173	744	6926	325220

According to DARNED database, **97.62% of total editing-sites are detected in a single tissue**. Hence, showing significantly low conservation rates, which indicates high-level of condition-specificity (*Kiran, Anmol, and Pavel V. Baranov. et al., 2010*).

RNA-seq

1) The most effective way for investigating RNA-editome in specific condition

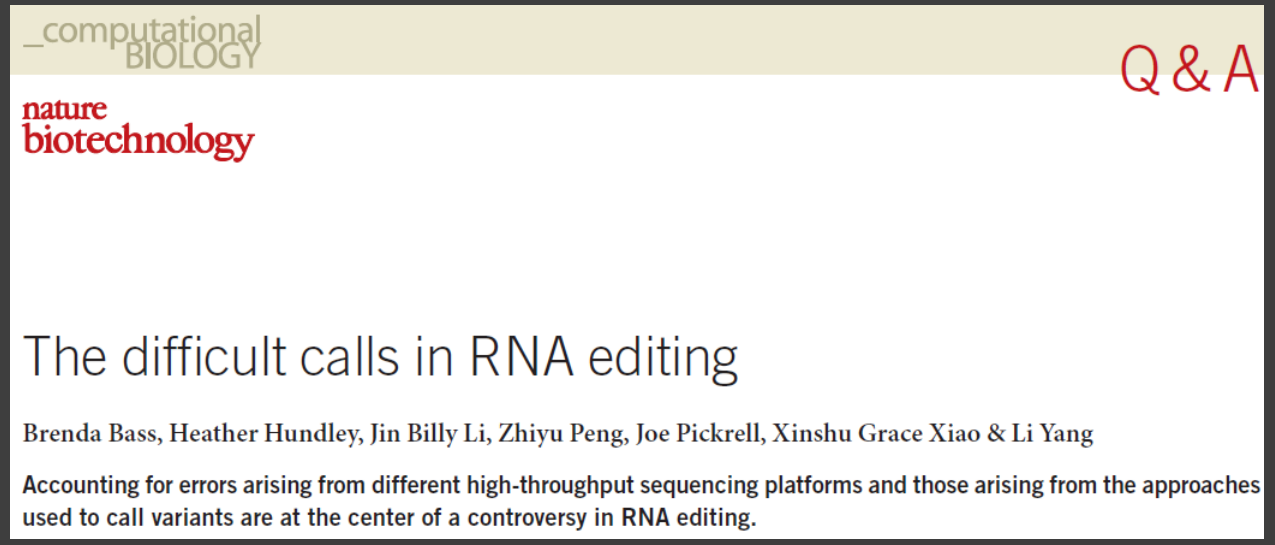


from GATK Forum

Since the nature of the technology is taking a snapshot of cells with massive sequencing reads, it is **suitable for detecting condition-specific events** in whole-transcriptome scale.

RNA-seq

2) Inherent mis-alignment risks of RNA-seq has confounded the RNA-editing researchers


_computational BIOLOGY

Q & A

nature biotechnology

The difficult calls in RNA editing

Brenda Bass, Heather Hundley, Jin Billy Li, Zhiyu Peng, Joe Pickrell, Xinshu Grace Xiao & Li Yang

Accounting for errors arising from different high-throughput sequencing platforms and those arising from the approaches used to call variants are at the center of a controversy in RNA editing.

Bass, Brenda, et al., 2012

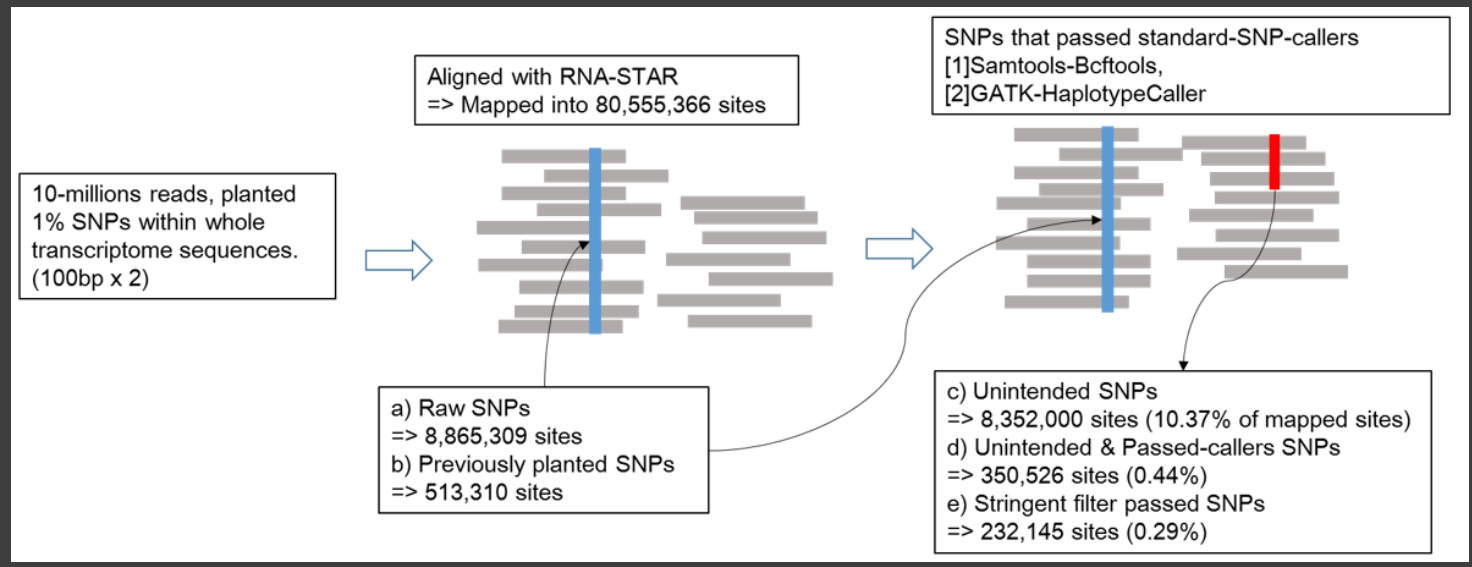
What sources of error confound variant calling in RNA-seq data?

J.B. Li: Based on our experience, **mapping error is the main source**, although sequencing errors inevitably affect accuracy. Once the reads are actually mapped, the challenge is to distinguish RNA editing events from genomic SNPs ...

RNA-seq

3) We demonstrated the inherent risks of mapping-errors by a simple simulation test.

A Simulation Test with hg19



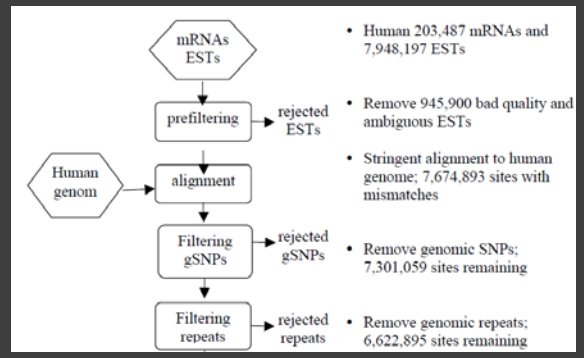
1) We measured the innate mis-alignment risks of human genome 19 with a published method, called MES (*Peng, Zhiyu, et al., 2012*)

2) As a result, we identified 232-thousands false-positives per 10-millions RNA-seq reads, which is not distinguishable by any standard methods (i.e. by standard callers or stringent filtering)

Motivation

1) Overview of three distinct approaches to deal with false-positive callings

1. A priori knowledge based filtering

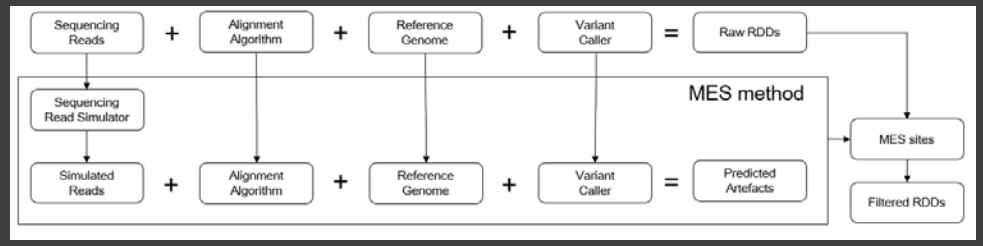


Li, Jin Billy, et al., 2009

Directly assesses RDD candidates with **public genomic features** (ex: genomic repeats)

2. Computational simulation of artefacts

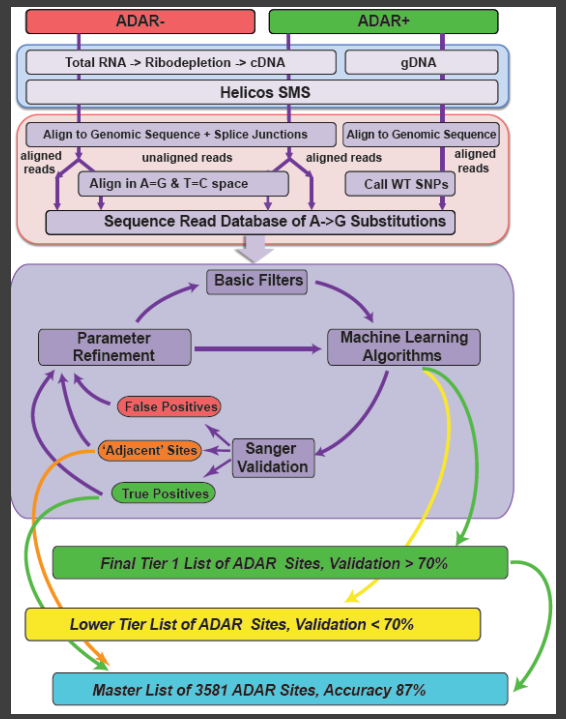
Assesses RDD candidates with **calculated features** (MES method)



Peng, Zhiyu, et al., 2012

3. Machine-learning based prediction model

Assesses RDD candidates with pre-defined **machine-learning classifier**, which is trained with massive experimental validations

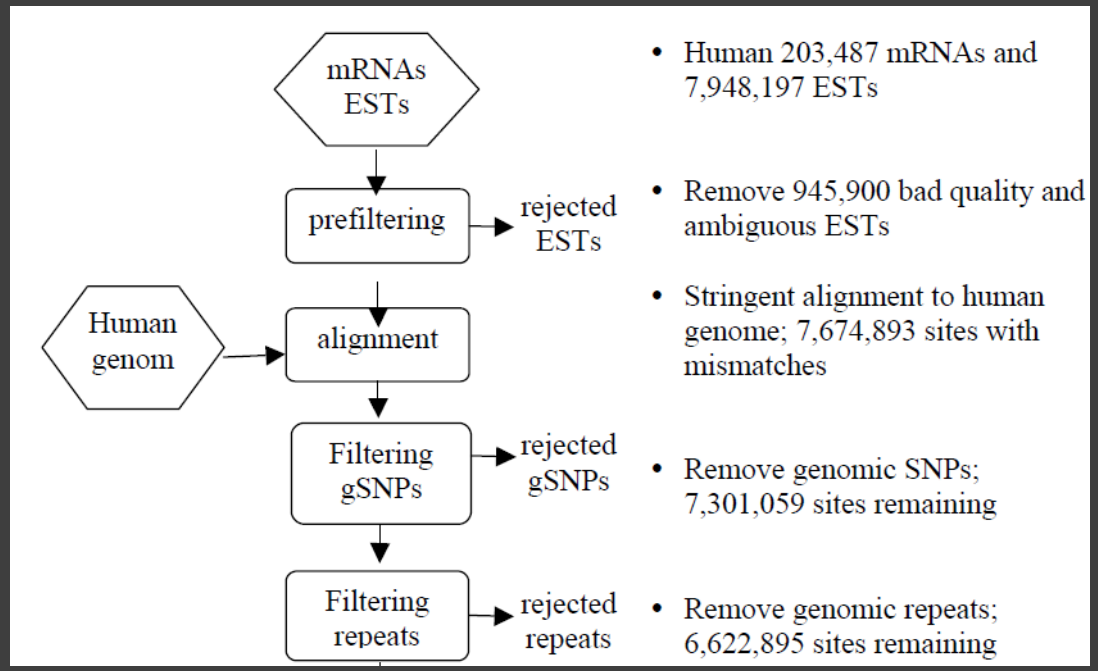


St Laurent, Georges, et al., 2013

Motivation

2) A priori knowledge based filtering

1. A priori knowledge based filtering



Li, Jin Billy, et al., 2009

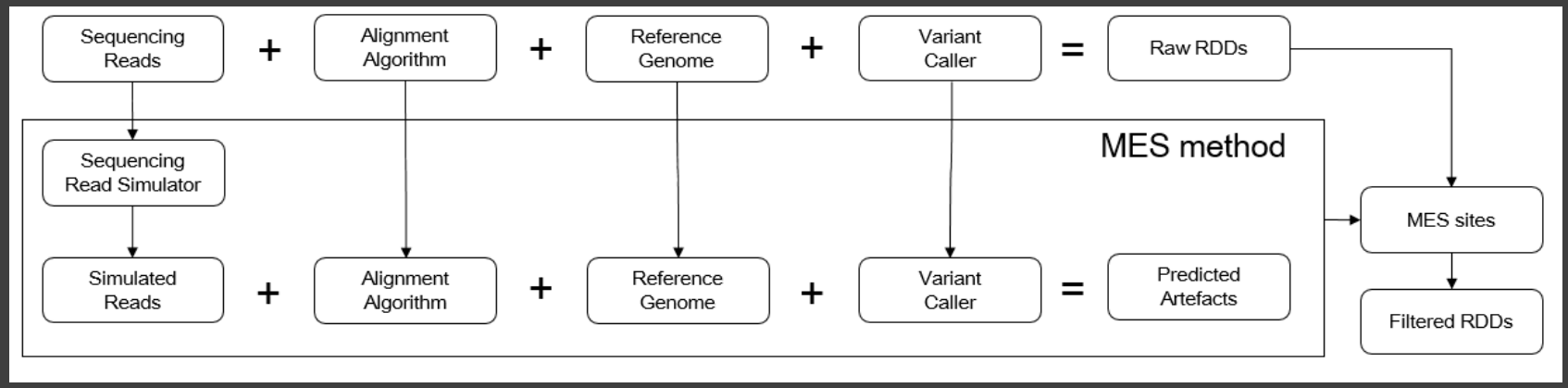
1) Using **public genomic features (such as genomic repeats, duplicates)** to assess the candidates directly (*Li, Jin Billy, et al., 2009)*

2) An effective method for excluding potential mapping-errors, which did not require any computation, but has **significant false-negative risks**.

Motivation

3) Computational simulation of artefacts

2. Computational simulation of artefacts



Peng, Zhiyu, et al., 2012

1) Using calculated features such as MES-sites (*Peng, Zhiyu, et al., 2012*).

2) MES method simulates RNA-seq from a target genome and align them to the genome sequence retrospectively to pinpoint which locus of the genome can be induce mis-callings.

3) However, calculating MES-sites in every possible occasions is almost impossible, because the mapping-errors are affected by many individual parameters, such as individual polymorphisms (SNPs, InDels).

Motivation

4) Machine-learning based prediction model

3. Machine-learning based prediction model

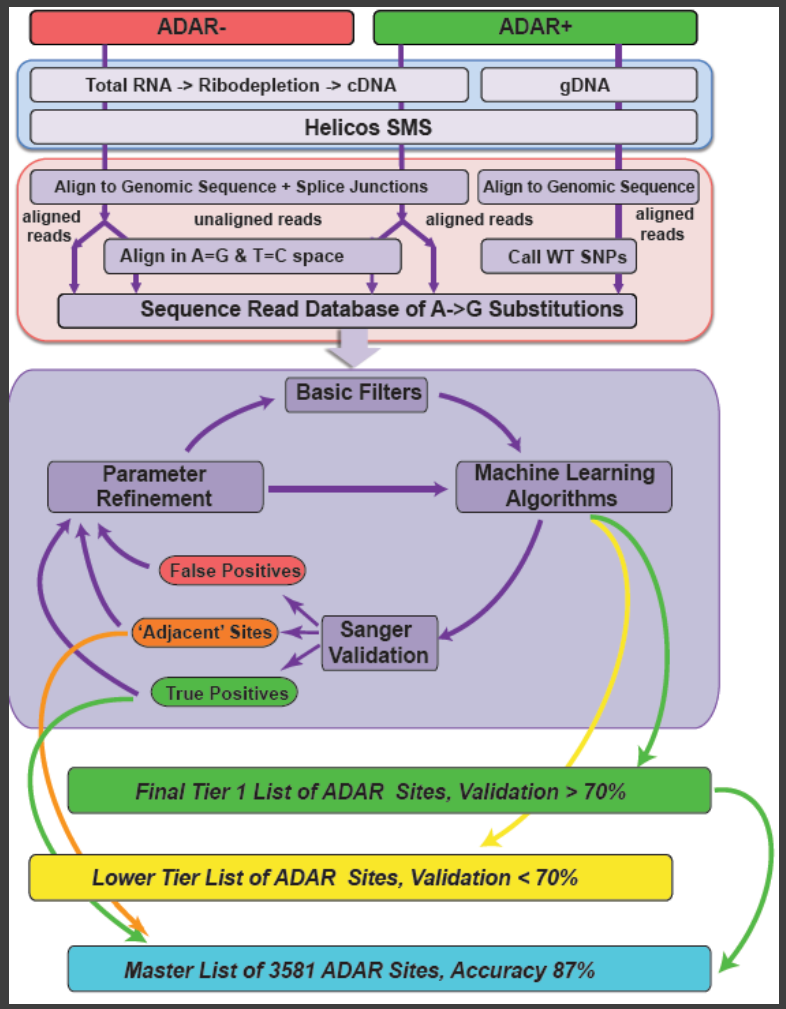
1) Unlike previous two approaches that used pre-defined filters, *St Laurent's* group rather generated a predictor in advance.

2) They reported 87% ACC showing relatively good performances (*St Laurent, Georges, et al., 2013*).

3) However, the model they build requires massive amounts of experimental validations to prepare training examples.

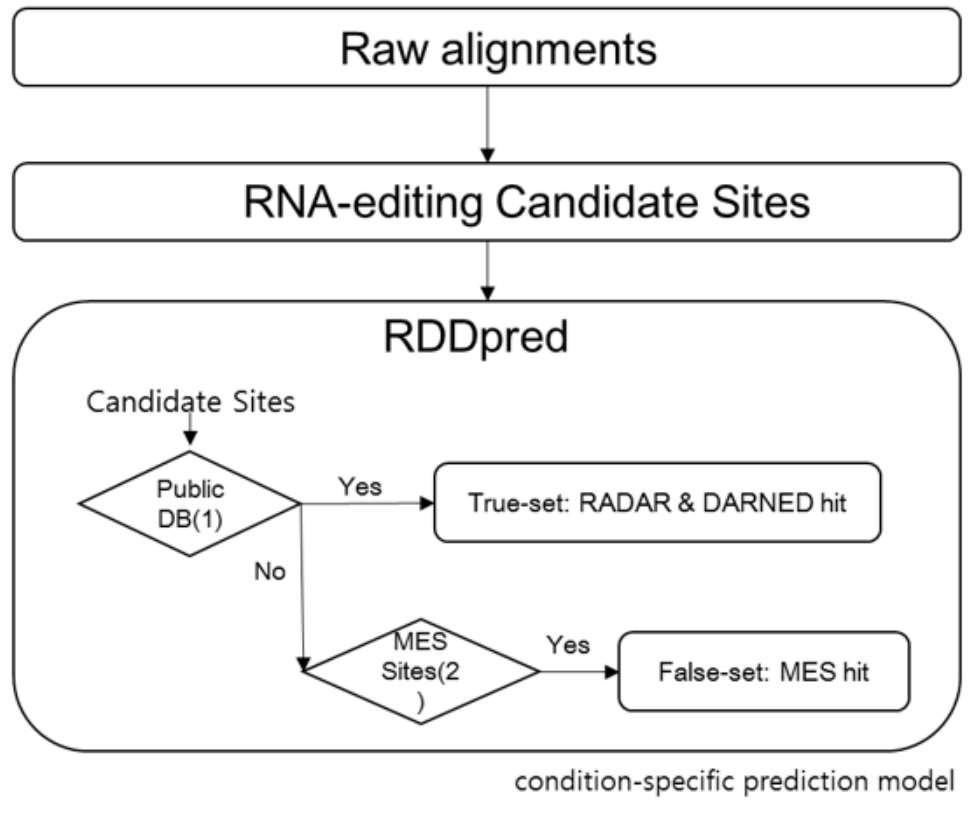
4) And also, their model is fully-customized to their proprietary dataset, which is not generally applicable.

St Laurent, Georges, et al., 2013



Motivation

5) We developed a generally usable pipeline which absorbs the benefits of previous works



1) RDDpred is a machine-learning based model, which is motivated by the *St Laurent's* work (*St Laurent, Georges, et al., 2013*).

2) The difference is that RDDpred does not require any experimental validations to collect training examples, instead, it deduces them from input instances.

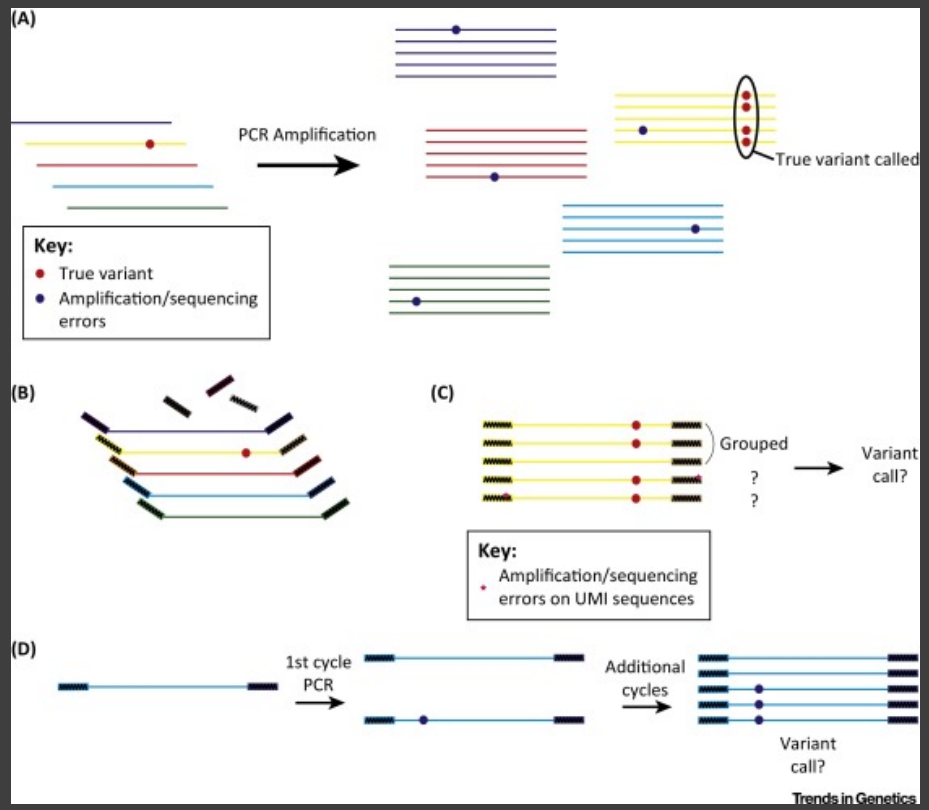
3) Also, RDDpred is provided as fully-automated pipeline, which is generally applicable.

Methods

Systematic Artefacts

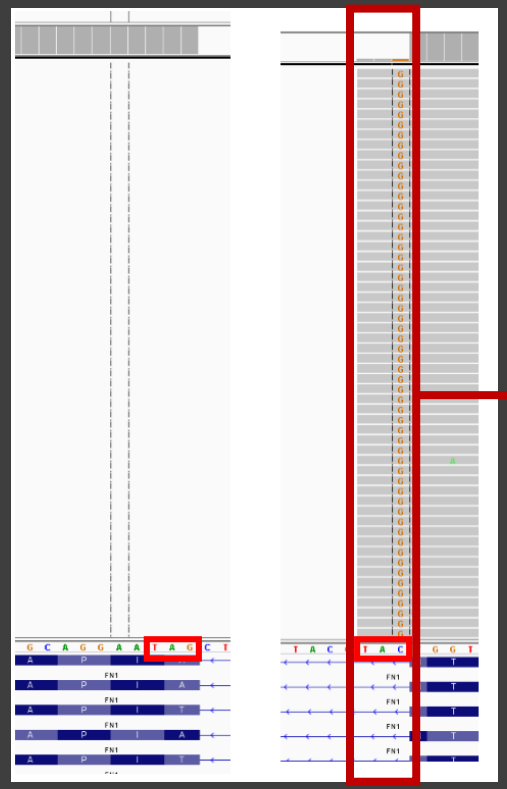
1) We named the errors caused by mis-alignments as “Systematic Artefacts”.

Transient Errors
Ex: Sequencing Errors



Cai, Xuyu, et al., 2015

Systematic Artefacts



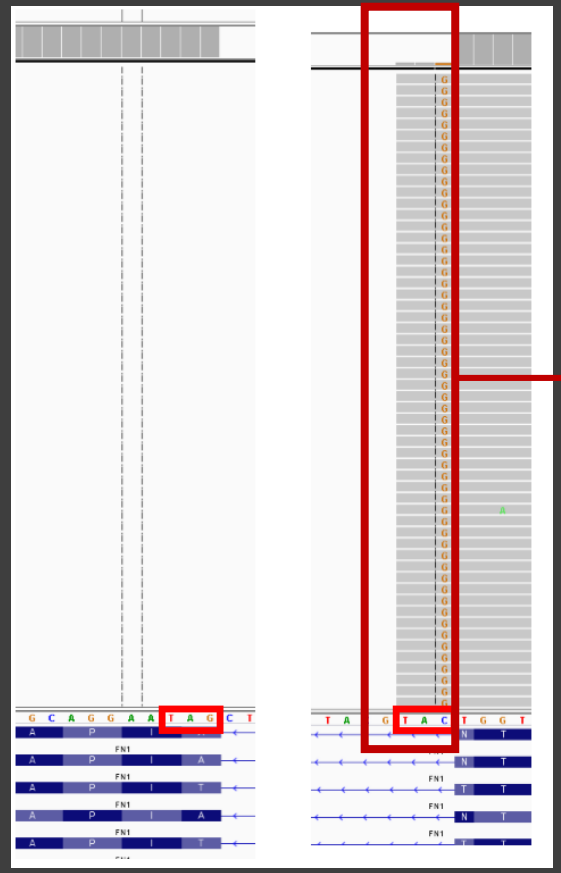
1) An example of systematic artefacts, which is induced by splicing junction

2) We named these errors as “systematic artefacts” for its **reproducible and inherent nature.**

Systematic Artefacts

2) Systematic artefacts leave distinct local alignment patterns, or signatures

Ex) Read Position Bias



1) In the case of systematic artefacts caused by splicing-junction, it **can be distinguished by read-position biases, a type of local alignment signatures** (i.e. All the variants are placed in the same position of reads).

2) We calculated **metrics that well-represents these characteristics** to build a predictor classifying artefact-sites from true-sites.

Systematic Artefacts

3) We utilize these alignment pattern signatures to recognize systematic artefacts

6 classes of 15 metrics for recognizing SAs

Class	Metric	Description
Allele Segregation	CallQual	Variant/reference QUALity
Allele Segregation	FQ	Phred probability of all samples being the same
Allele Segregation	SGB	Segregation based metric
Allele Segregation	VAF	Variant allele frequency
Base Quality	BQB	Mann-Whitney U test of Base Quality Bias
Base Quality	PV2	Base quality bias
Mapping Quality	MQ	Root-mean-square mapping quality of covering reads
Mapping Quality	MQ0F	Fraction of MQ0 reads
Mapping Quality	MQB	Mann-Whitney U test of Mapping Quality Bias
Mapping Quality	PV3	Mapping quality bias
Read Depth	ReadDepth	Read depth
Read Position	PV4	Tail distance bias
Read Position	RPB	Mann-Whitney U test of Read Position Bias
Read Position	VDB	Variant Distance Bias for filtering splice-site artefacts in RNA-seq data
Read Strand	PV1	Read strand bias

1) It is well known that **read-alignment patterns have significant classification power** in recognizing sequence variant mis-calling (*Li, Heng. et al., 2011*).

2) All these listed metrics are calculated by utilizing **samtools-bcftools pipeline** (*Li, Heng. et al., 2011*).

RDDpred

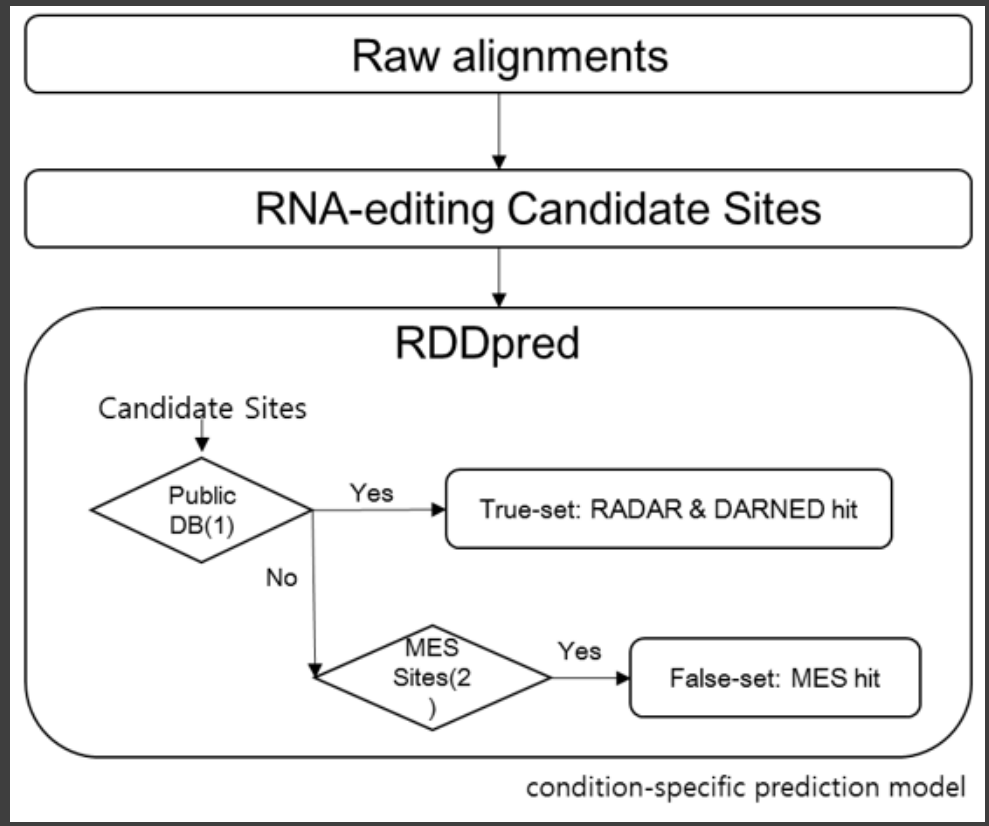
1) RDDpred deduces positive-examples by utilizing public databases

Positive Examples

1) Since RNA-editing is not occurred in random residues, we can reasonably assume that the publicly known editing-sites have editing potential.

2) Therefore, RDDpred considers the publicly known sites in input instances (or Candidates) as positive-examples.

3) Fortunately, there are two well-organized public databases (RADAR, DARNED).



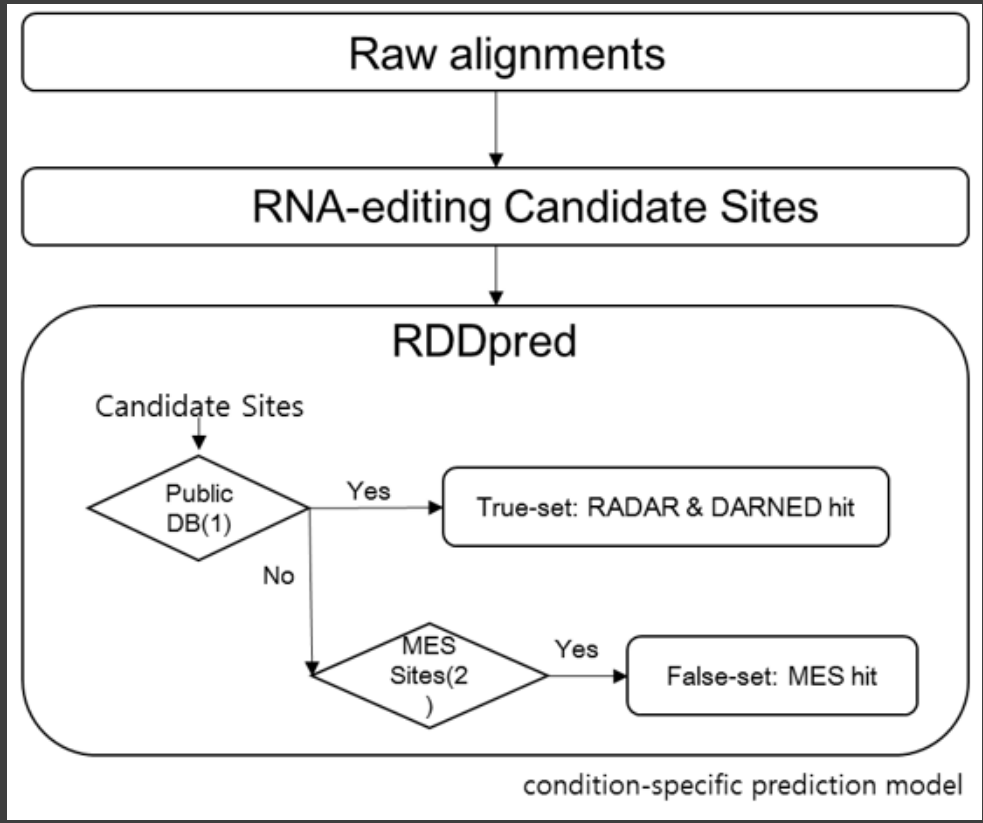
RDDpred

2) RDDpred deduces negative examples by utilizing MES method.

Negative Examples

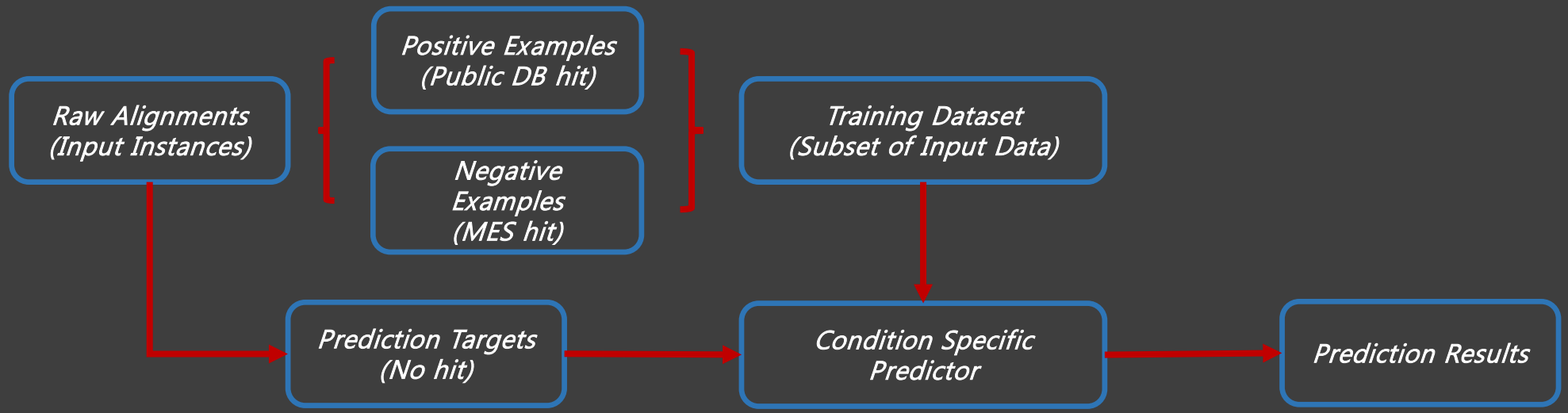
1) To deduce negative examples from input instances, we prepared adjust amounts of MES-sites with proper parameters specific to input data in advance.

2) As mentioned, since MES predicts the genomic locus having inherent mis-alignment risks, RDDpred considers MES locus in input instances (or Candidates) as negative-examples.



RDDpred

3) RDDpred is a condition-specific model.

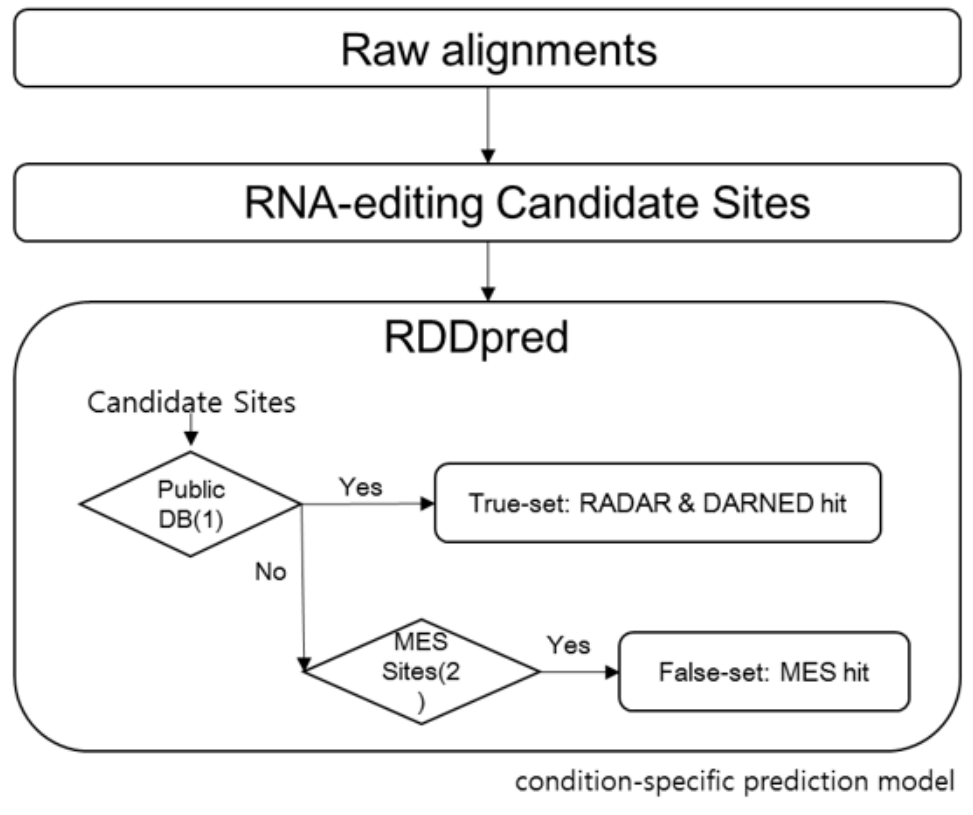


1) RDDpred dynamically generates predictors specific to each given input data.

2) Hence, RDDpred is generally applicable to any input data, as long as adjust amounts of Positive/Negative hits can be made from the given inputs.

RDDpred

4) RDDpred utilizes “Random Forest” algorithm to build a predictor



1) RDDpred trained a **predictor from the deduced training-examples** with calculated alignment-signature metrics as input features.

2) The algorithm and parameters are followed by the default settings of **WEKA data-mining package** (*Hall, Mark, et al., 2009*).

Results

Evaluation of RDDpred

1) RDDpred was tested with two previous studies including experimental validations

Bahn's Study

Condition	SRA	PMID
Human glioblastoma astrocytoma	SRP009659	21960545

Reads	Bases	Raw RDDs
115,132,348	13,815,881,760	6,856,440

Reported	False Discovery
4,141	19

1) Banh's group detected RNA-editing with 115-millions reads, which contain about 6.8-millions RDD(RNA/DNA Difference) as our estimation.

2) Of that, Bahn's group reported 4,141 RNA-editing sites as true-events.

3) They also published 19 falsely discovered sites which is confirmed by Sanger-seq.

Bahn, Jae Hoon, et al., 2012

Peng's Study

Condition	SRA	PMID
Human lymphoblastoid	SRP007605	22327324

Reads	Bases	Raw RDDs
583,640,030	1.01787E+11	58,666,976

Reported	False Discovery
22,688	29

1) Peng's group detected RNA-editing with 583-millions reads, which contain about 58-millions RDD(RNA/DNA Difference) as our estimation.

2) Of that, Peng's group reported 22,688 RNA-editing sites as true-events.

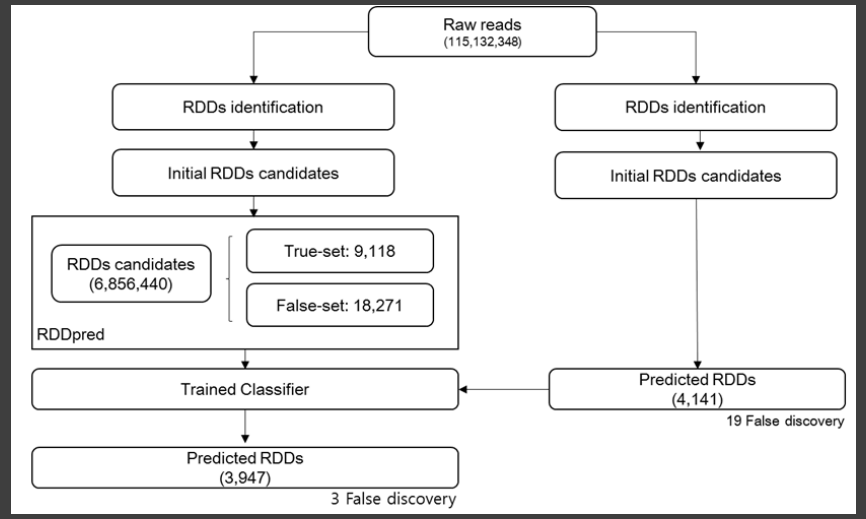
3) They also published 29 falsely discovered sites which is confirmed by Sanger-seq.

Peng, Zhiyu, et al., 2012

Evaluation of RDDpred

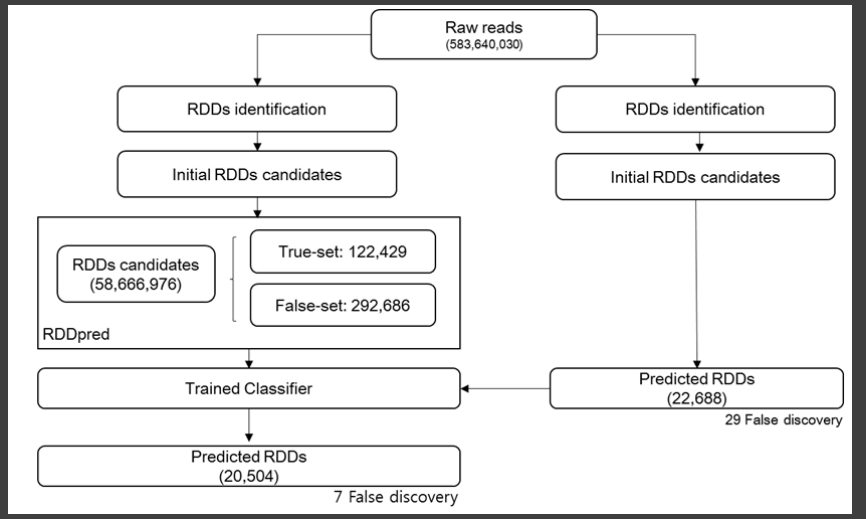
2) RDDpred successfully reduced false-positives while reproducing the results in both cases

Bahn's Study



RDDpred **reproduced 3,947 sites of 4,141 (95.32%)** of their results, while **rejecting 16 sites of 19 (84.21%)** of their false discoveries.

Peng's Study



RDDpred **reproduced 20,504 sites of 22,688 (90.37%)** of their results, while **rejecting 22 sites of 29 (75.86%)** of their false discoveries.

** Note that we **excluded the positive/negative results of each studies (i.e. test datasets) from our training datasets** for fair comparisons.

Evaluation of RDDpred

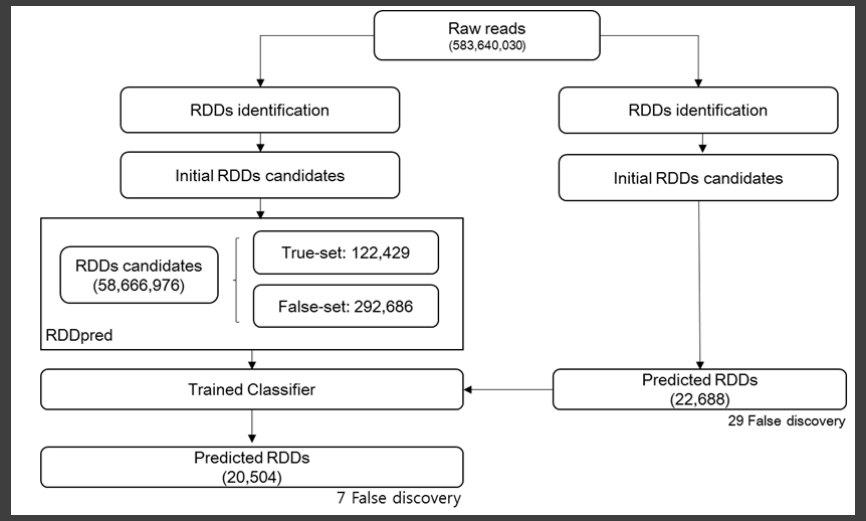
4) RDDpred is well-parallelized to handle massive RNA-seq data efficiently.

Test Dataset

Peng, Zhiyu, et al., 2012

Condition	SRA	PMID
Human lymphoblastoid	SRP007605	22327324

Reads	Bases	Raw RDDs
583,640,030	1.01787E+11	58,666,976



Evaluating Software Performances

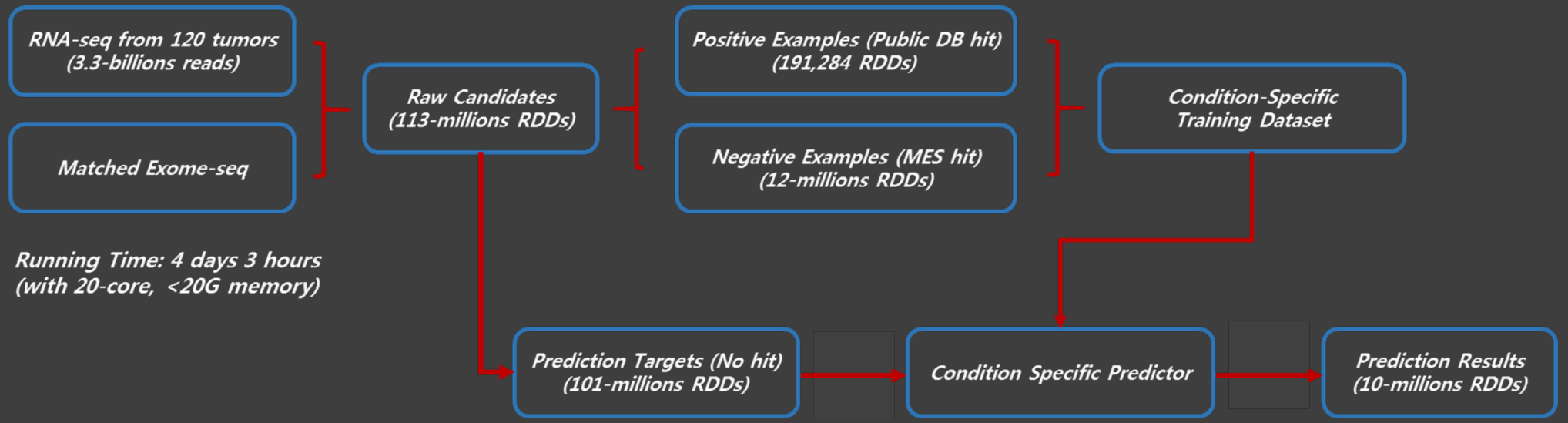
Linux version	Linux version 2.6.32-358.el6.x86_64 (CentOS release 6.4)
Memory usage	20GB (in maximum)
CPU usage	20-cores [Intel(R) Xeon(R) CPU E5645 @2.40GHz]
Running Time	18.33 hours

1) Note that the running time does not include alignment and MES procedures (but, extensively calculated MES sites will be provided by our web-site, so no need to concern).

2) As we thought, the running time is not too excessive to utilize for ordinary research groups.

**Case Study (not in manuscript)

Recently, we have been applying RDDpred to our proprietary data of breast cancer



1) We applied RDDpred to 120 tumor RNA-seq, which contains 113-millions sites of raw RDDs.

2) As a result, we accepted 10-millions RDDs (9.39% Acceptance Rate).

** Note that the above numbers are pooled results from 120 individual tumors and the number of RDDs from individual tumor is far less (which also indicates diversities of RNA-editing).

Conclusion

Conclusion

1) RDDpred deduces condition-specific training examples without any experimental validations to construct a predictor.

2) As far as we know, RDDpred is the very first machine-learning based automated pipeline for RNA-editing prediction.

3) RDDpred successfully reproduced the results of two previous studies (95%, 90%), with showing significant NPV (84%, 75%) and the prediction procedures are finished in reasonable time (18 hrs).

4) The source code and every required accessory data of RDDpred are available at http://biohealth.snu.ac.kr/software/RDDpred.

THANK YOU