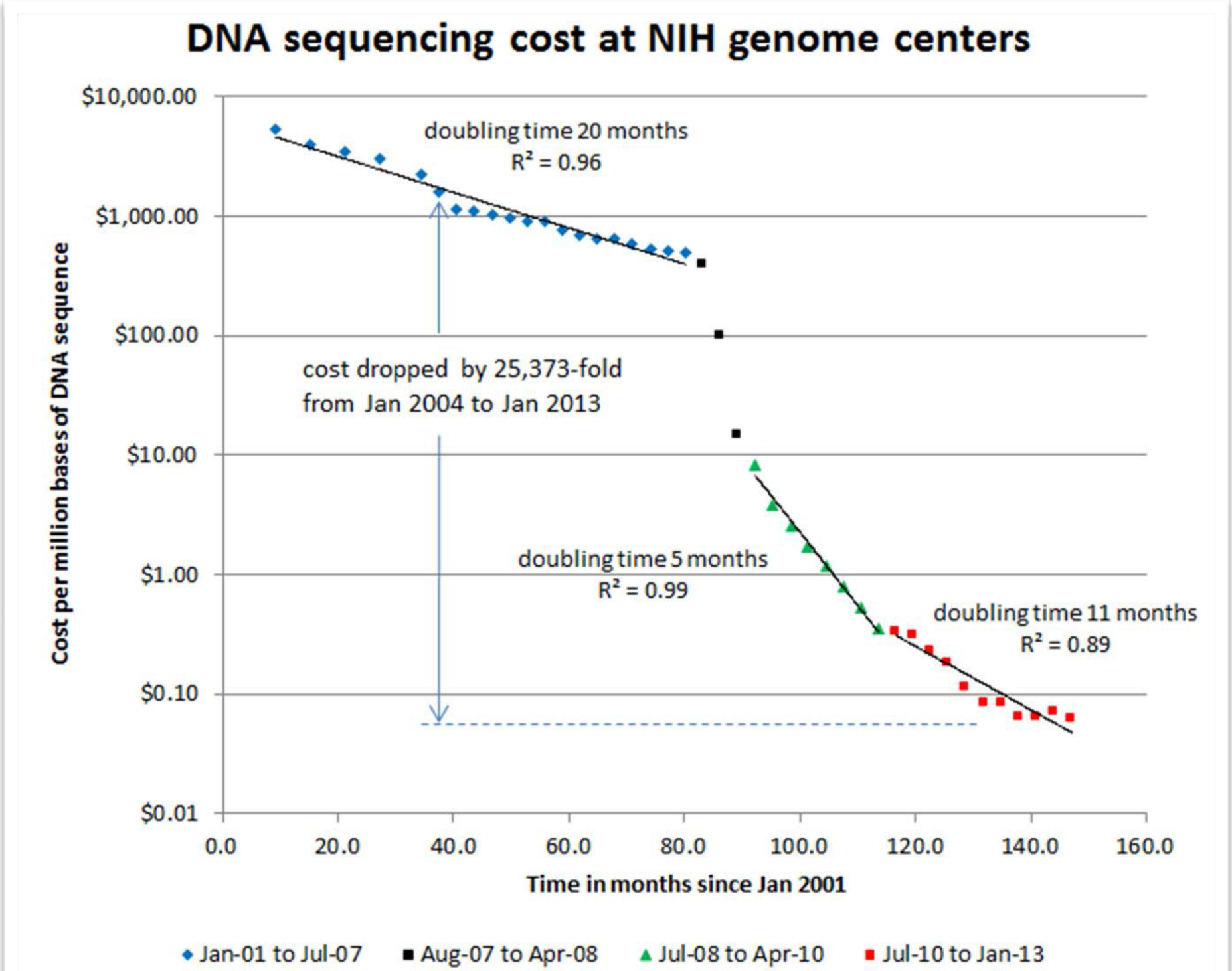
Why NGS??

Why NGS??


DNA sequencing cost at NIH genome centers
$10,000.00
$1,000.00
$100.00
$10.00
$1.00
$0.10
$0.01
Cost per million bases of DNA sequence
doubling time 20 months
R² = 0.96
cost dropped by 25,373-fold
from Jan 2004 to Jan 2013
doubling time 5 months
R² = 0.99
doubling time 11 months
R² = 0.89
0.0
20.0
40.0
60.0
80.0
100.0
120.0
140.0
160.0
Time in months since Jan 2001
Jan-01 to Jul-07
Aug-07 to Apr-08
Jul-08 to Apr-10
Jul-10 to Jan-13

Resequencing and Mapping

Andreas Gisel
Inernational Institute of Tropical Agriculture (IITA)
Ibadan, Nigeria

The Principle of Mapping

reads

```
good, ood_, d_mo, morn, orni, ning, ing_,
g_be, beau, auti, utif, iful, ul_w, _wor orld
```

reference

```
good_morning_beautiful_world
```

mapping

```
          ing_    utif
   d_mo ning    auti    _wor
 ood_ orni    beau    ul_w
good morn   g_be    iful   orld

good_morning_beautiful_world
```

consensus

```
good_morning_beautiful_world
```

What can we find with Mapping

reads

```
good, ood_, d_ev, even, veni, ning, ing_,
g_be, beau, auti, utif, iful, ul_w, _wor orld
```

reference

```
good_morning_beautiful_world
```

mapping

```
          ing_    utif
   d_ev ning    auti    _wor
 ood_ veni    beau    ul_w
good even  g_be    iful  orld

good_morning_beautiful_world
```

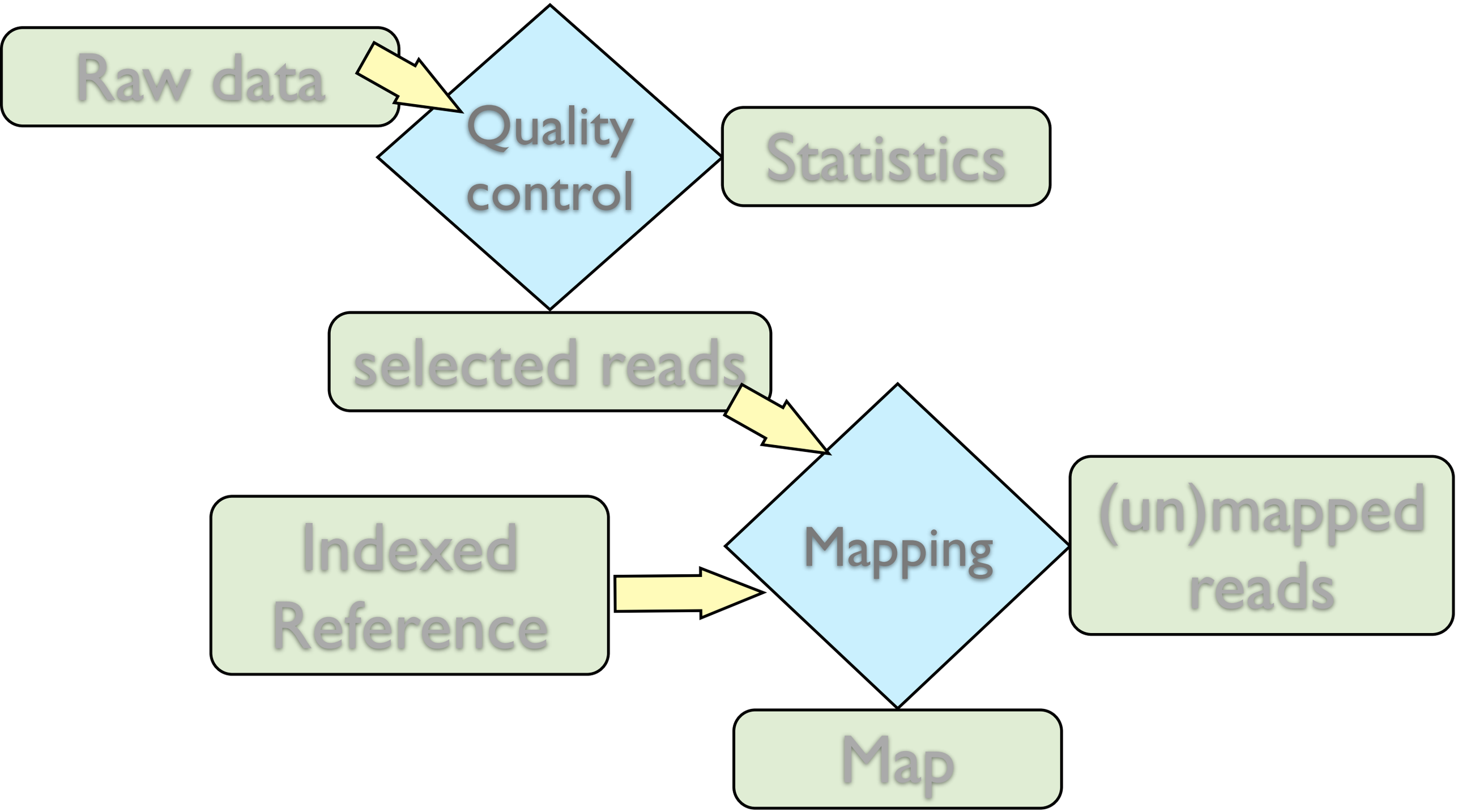
consensus

```
good_evening_beautiful_world
```

What can we find with Mapping

reference	`good_morning_beautiful_world`	
consensus	`good_evening_beautiful_world`	**change**
consensus	`mood_morning_beautiful_world`	**SNP**
consensus	`good_morning _world`	**deletion**
consensus	`good_morning beautiful_world` + not mappable reads	**insert**
consensus	`good_evening dlrow_lufituaeb` + not mappable reads	**inversion**

Workflow for Mapping



Compare Mapping Tools

Table 1. Evaluation on simulated data

	Single-end			Paired-end		
Program	Time (s)	Conf (%)	Err (%)	Time (s)	Conf (%)	Err (%)
Bowtie-32	1271	79.0	0.76	1391	85.7	0.57
BWA-32	823	80.6	0.30	1224	89.6	0.32
MAQ-32	19797	81.0	0.14	21589	87.2	0.07
SOAP2-32	256	78.6	1.16	1909	86.8	0.78
Bowtie-70	1726	86.3	0.20	1580	90.7	0.43
BWA-70	1599	90.7	0.12	1619	96.2	0.11
MAQ-70	17928	91.0	0.13	19046	94.6	0.05
SOAP2-70	317	90.3	0.39	708	94.5	0.34
bowtie-125	1966	88.0	0.07	1701	91.0	0.37
BWA-125	3021	93.0	0.05	3059	97.6	0.04
MAQ-125	17506	92.7	0.08	19388	96.3	0.02
SOAP2-125	555	91.5	0.17	1187	90.8	0.14

One million pairs of 32, 70 and 125 bp reads, respectively, were simulated from the human genome with 0.09% SNP mutation rate, 0.01% indel mutation rate and 2% uniform sequencing base error rate. The insert size of 32 bp reads is drawn from a normal distribution $N(170,25)$, and of 70 and 125 bp reads from $N(500,50)$. CPU time in seconds on a single core of a 2.5 GHz Xeon E5420 processor (Time), percent confidently mapped reads (Conf) and percent erroneous alignments out of confident mappings (Err) are shown in the table.

Compare Mapping Tools

Table 1: Popular short-read alignment software

Program	Algorithm	SOLiD	Long[a]	Gapped	PE[b]	Q[c]
Bfast	hashing ref.	Yes	No	Yes	Yes	No
Bowtie	FM-index	Yes	No	No	Yes	Yes
BWA	FM-index	Yes[d]	Yes[e]	Yes	Yes	No
MAQ	hashing reads	Yes	No	Yes[f]	Yes	Yes
Mosaik	hashing ref.	Yes	Yes	Yes	Yes	No
Novoalign[g]	hashing ref.	No	No	Yes	Yes	Yes

[a]Work well for Sanger and 454 reads, allowing gaps and clipping. [b]Paired end mapping. [c]Make use of base quality in alignment. [d]BWA trims the primer base and the first color for a color read. [e]Long-read alignment implemented in the BWA-SW module. [f]MAQ only does gapped alignment for Illumina paired-end reads. [g]Free executable for non-profit projects only.

Tools for Mapping

bowtie - an ultrafast, memory-efficient short read aligner
http://bowtie-bio.sourceforge.net/index.shtml

BWA - Burrows-Wheeler Aligner
http://bio-bwa.sourceforge.net/

SOAPaligner - Short Oligonucleotide Analysis Package
http://soap.genomics.org.cn/soapaligner.html

De novo sequencing and Assembly

Andreas Gisel
International Institute of Tropical Agriculture (IITA)
Ibadan, Nigeria

The Principle of Assembly

reads

```
good, ood_, d_mo, morn, orni, ning, ing_,
g_be, beau, auti, utif, iful, ul_w, _wor orld
```

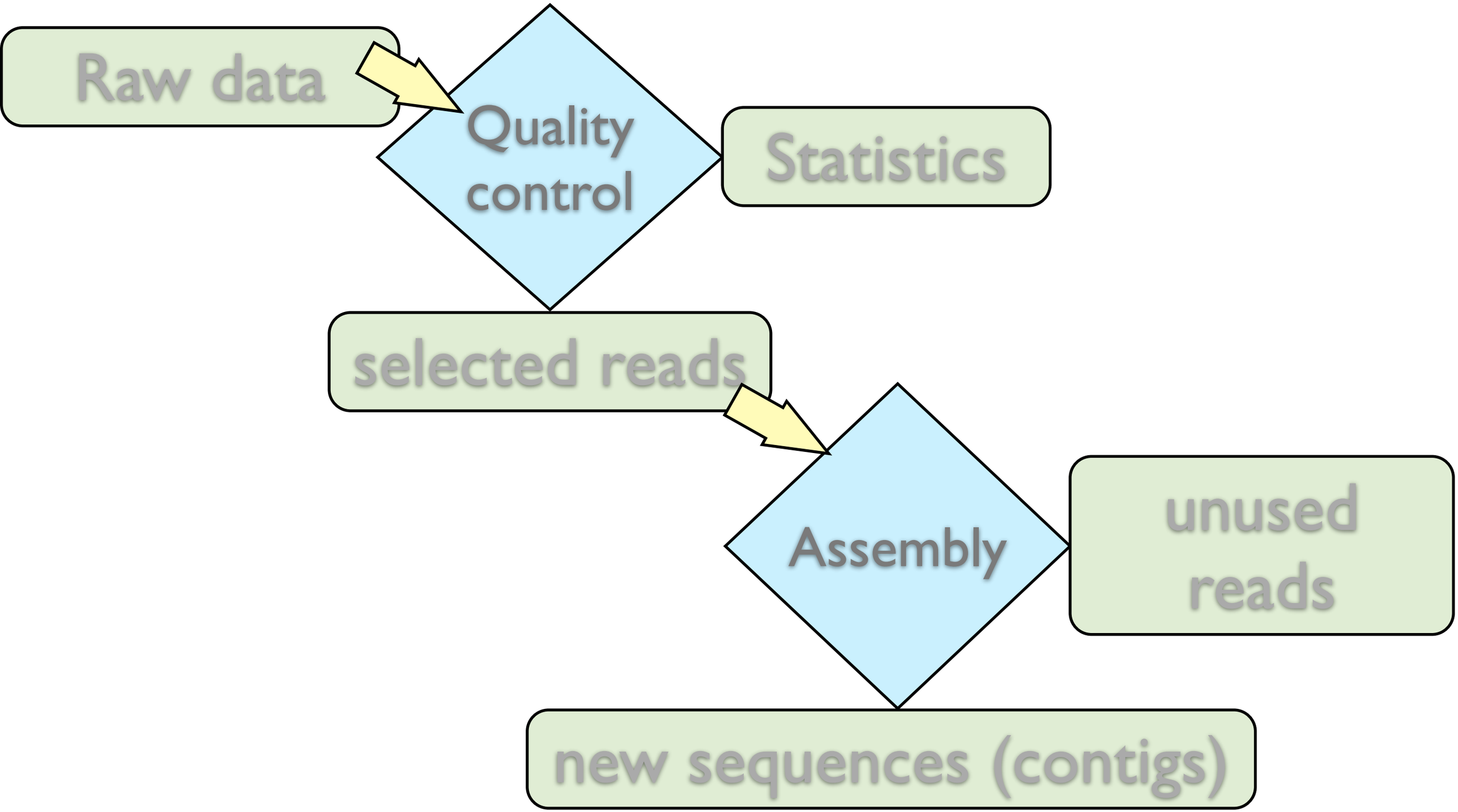
assembly

```
good
 ood_
   d_mo
    morn
     orni
       ning
        ing_
          g_be
            beau
              auti
               utif
                 iful
                   ul_w
                     _wor
                       orld
```

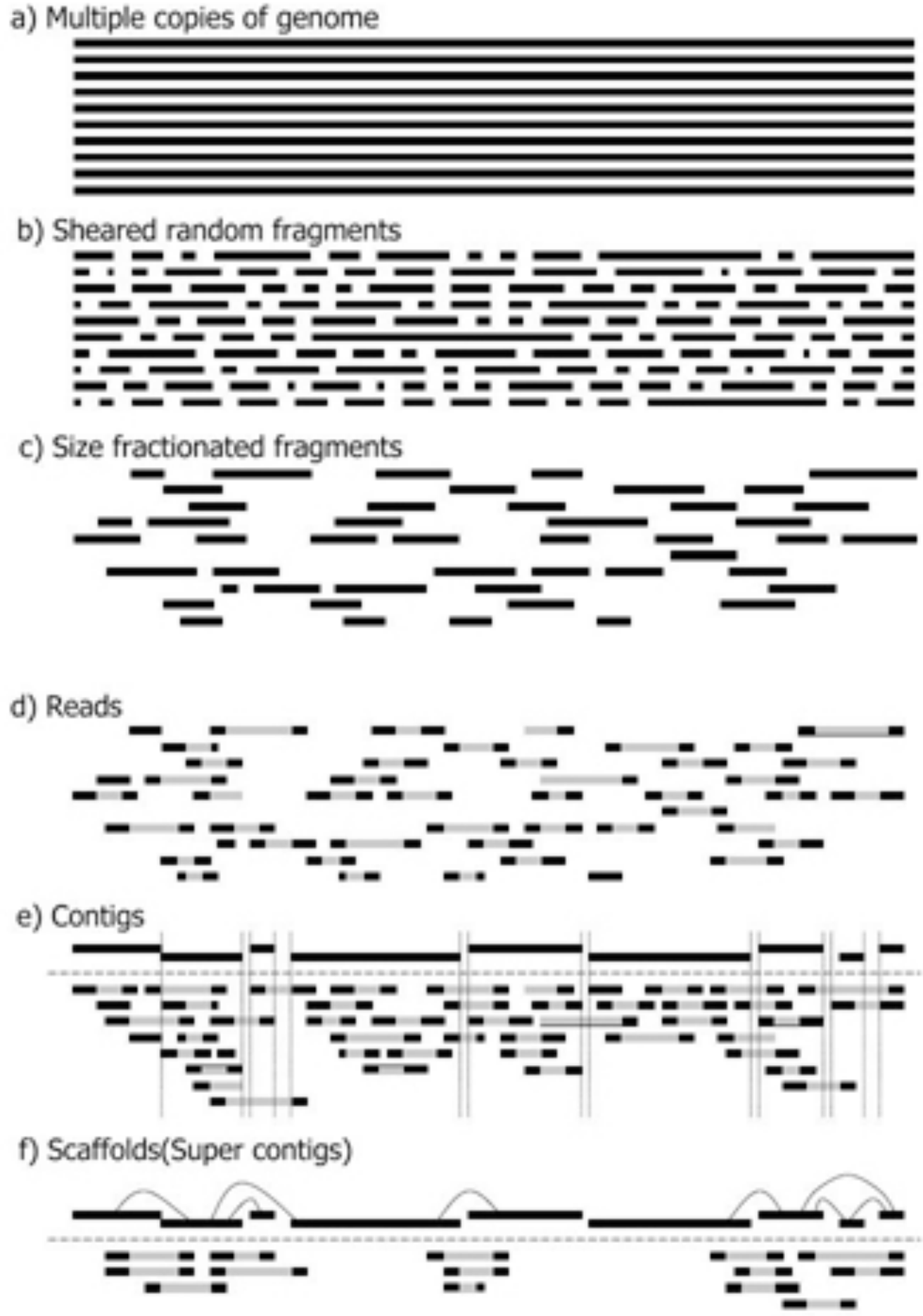
consensus

```
good_morning_beautiful_world
```

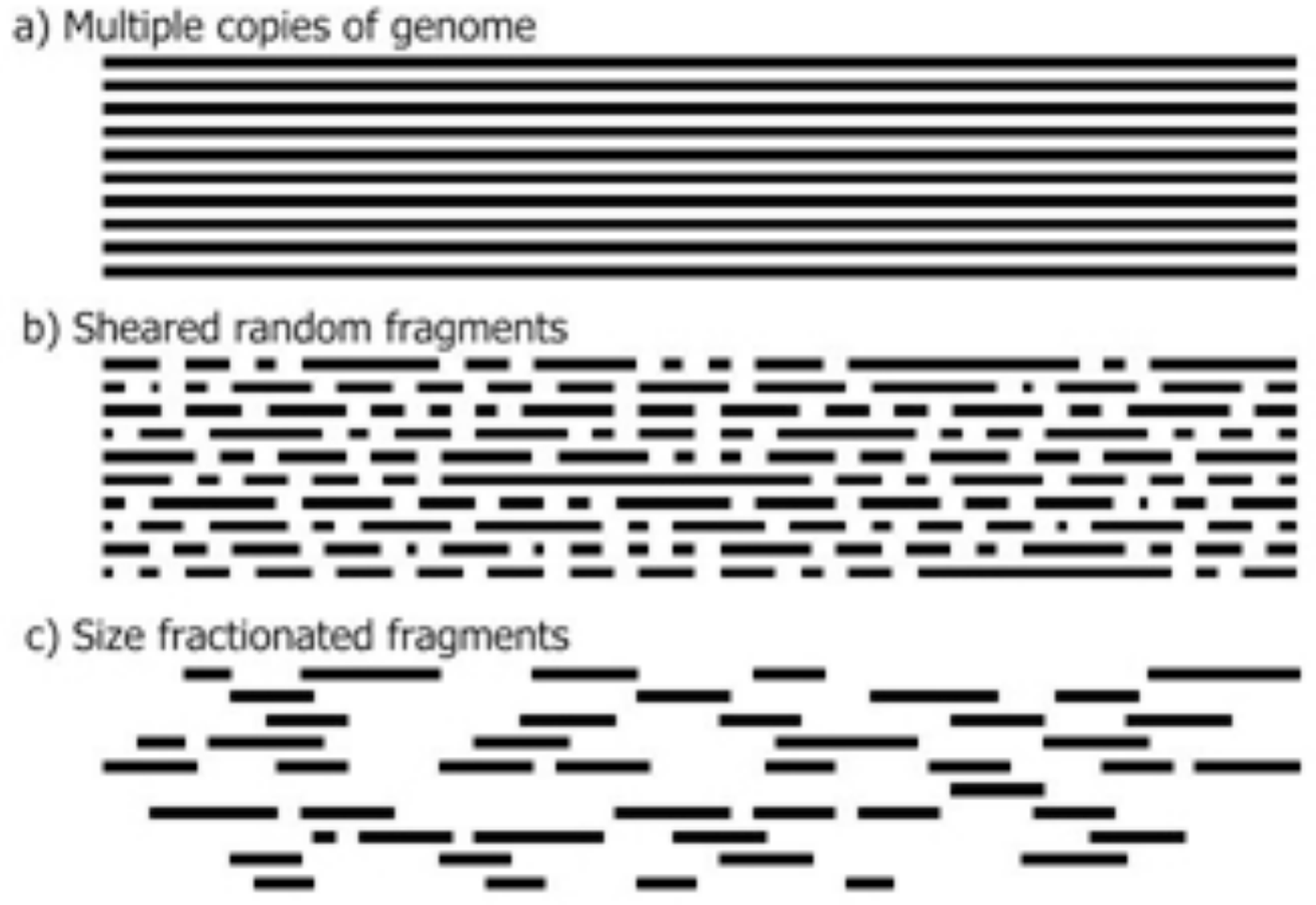
Workflow for Assembly

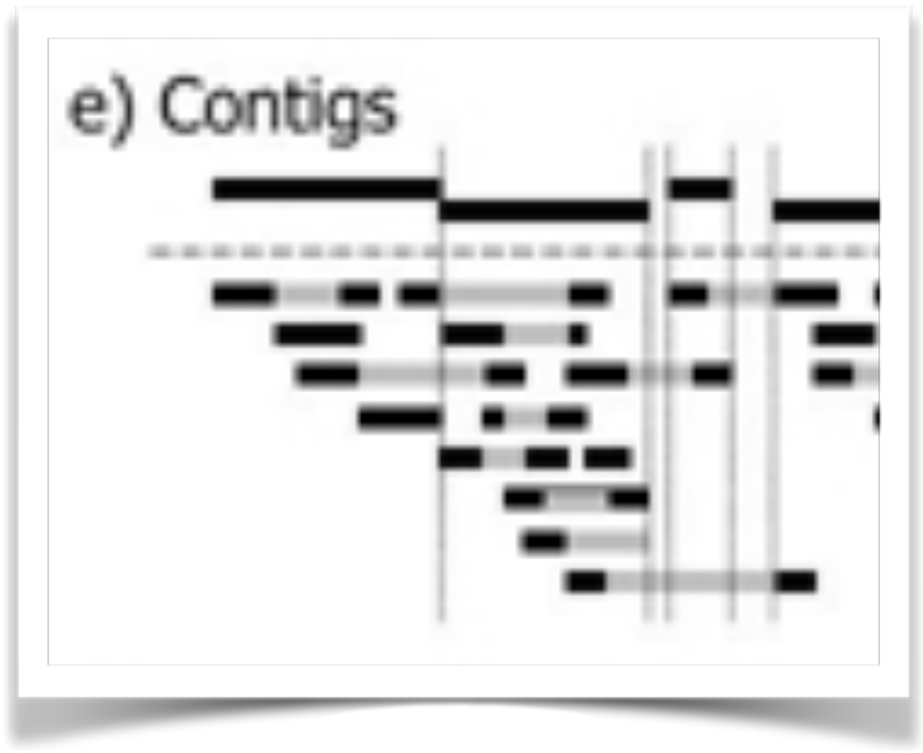


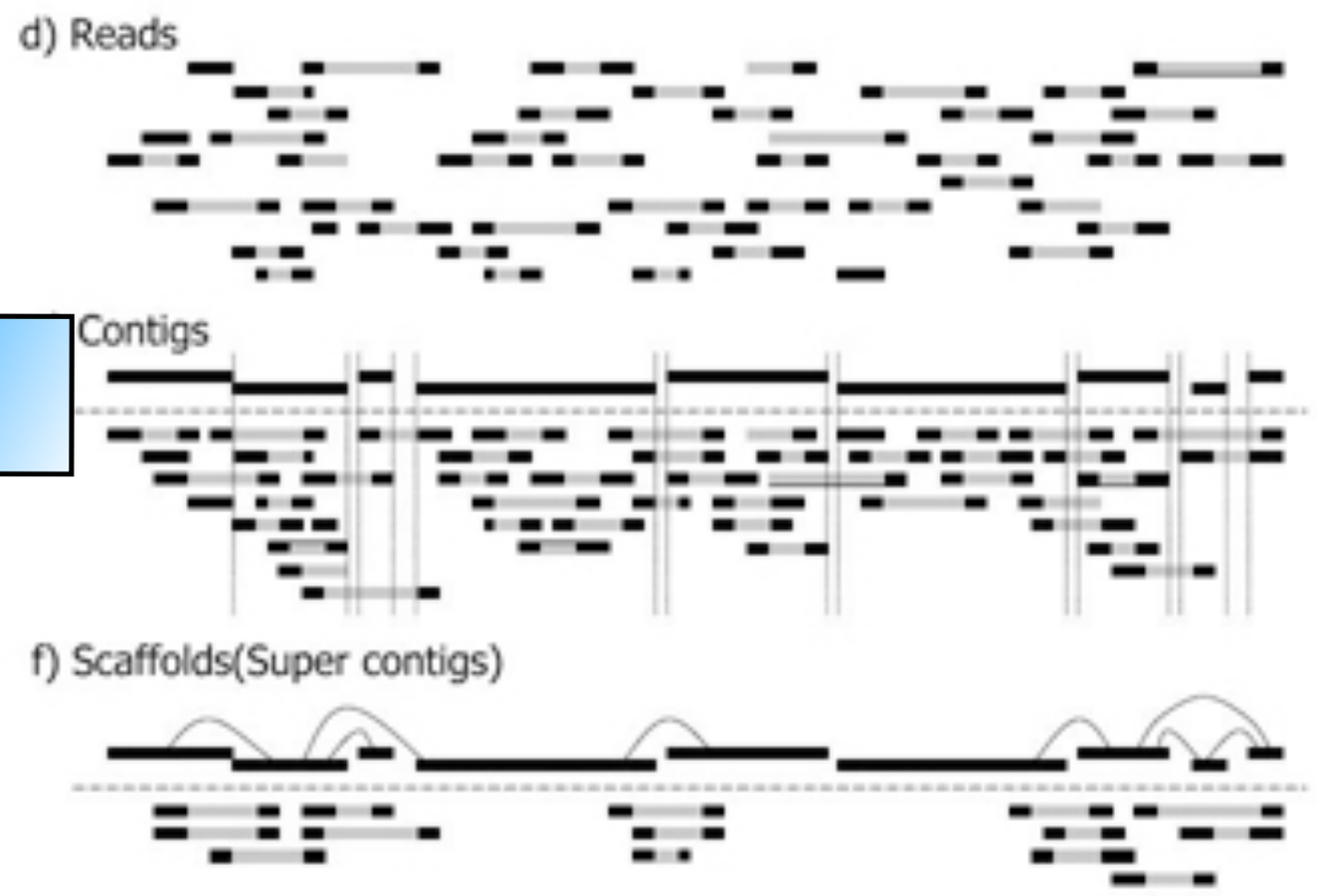
Workflow


a) Multiple copies of genome
b) Sheared random fragments
c) Size fractionated fragments
d) Reads
e) Contigs
f) Scaffolds(Super contigs)

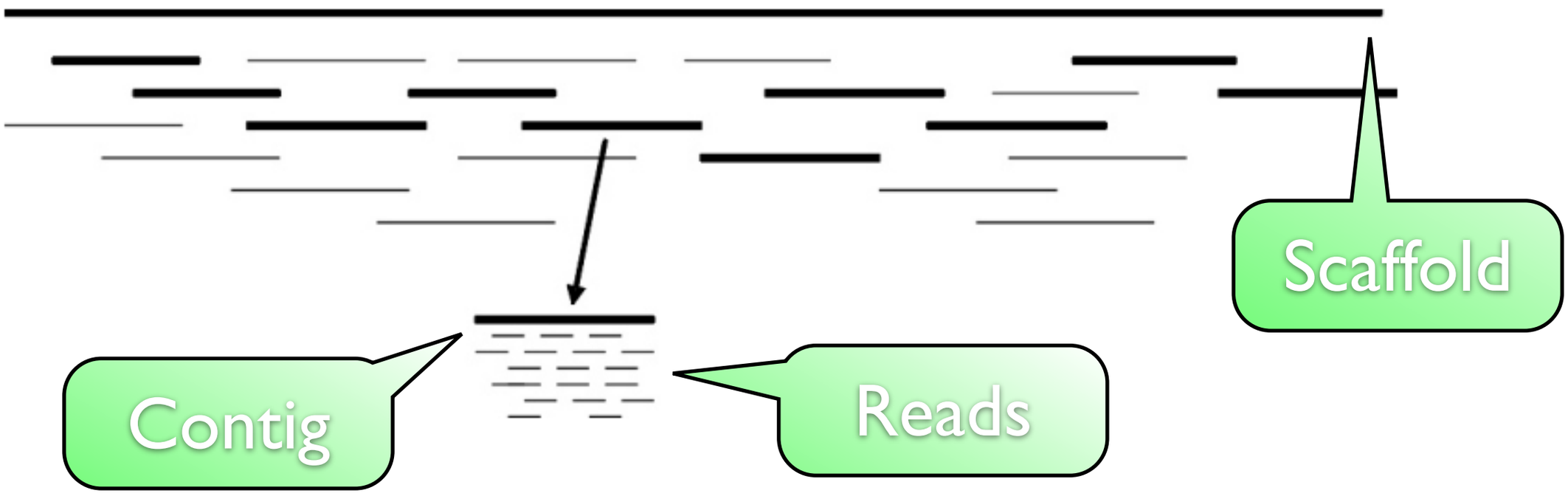
Workflow


a) Multiple copies of genome
b) Sheared random fragments
c) Size fractionated fragments


e) Contigs


d) Reads
Contigs
f) Scaffolds(Super contigs)

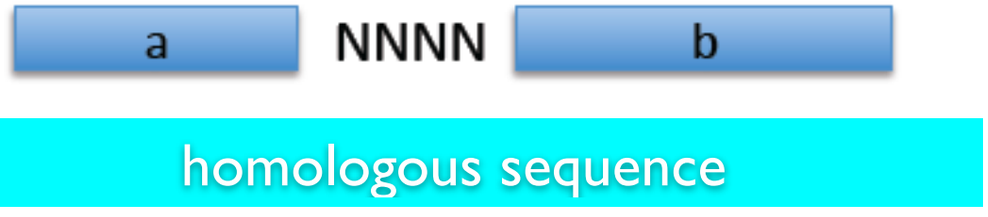
Contigs - Scaffolds



Contigs - Scaffolds

Connect Contigs with:

- ☑ mate-pair information
- ☑ homology data
- ☑ physical maps
- ☑ gene synteny

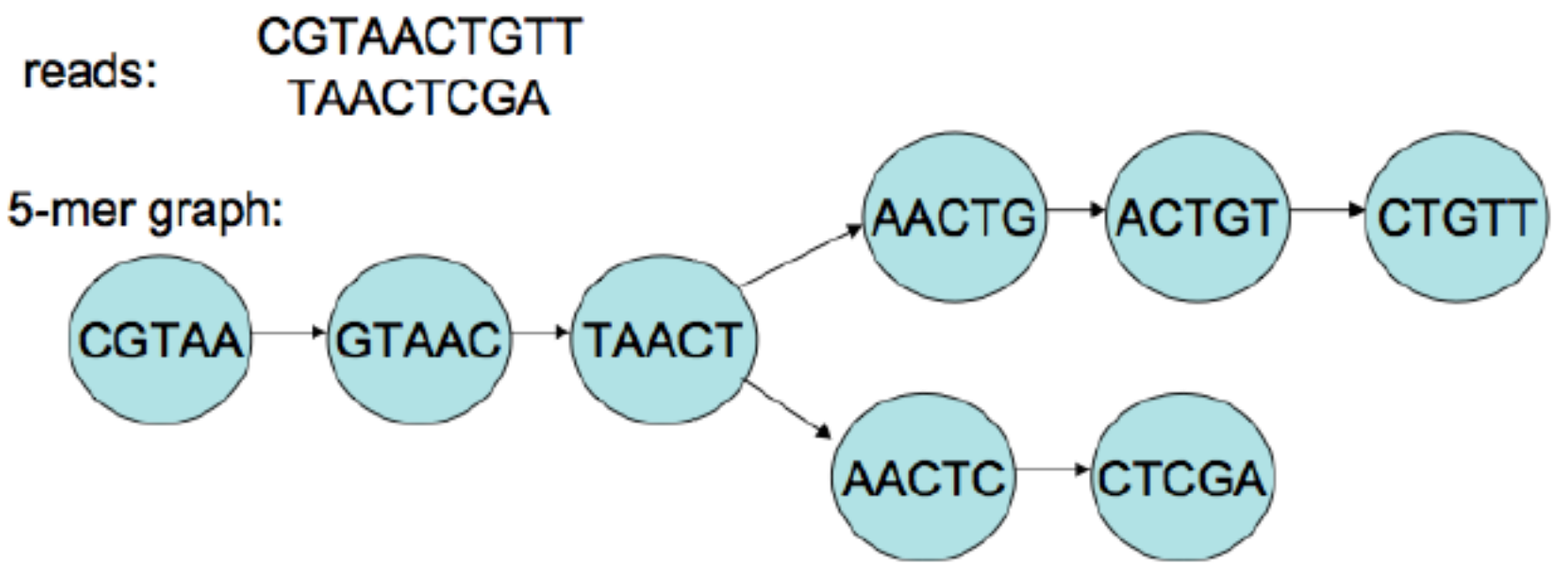


deBruijn graph

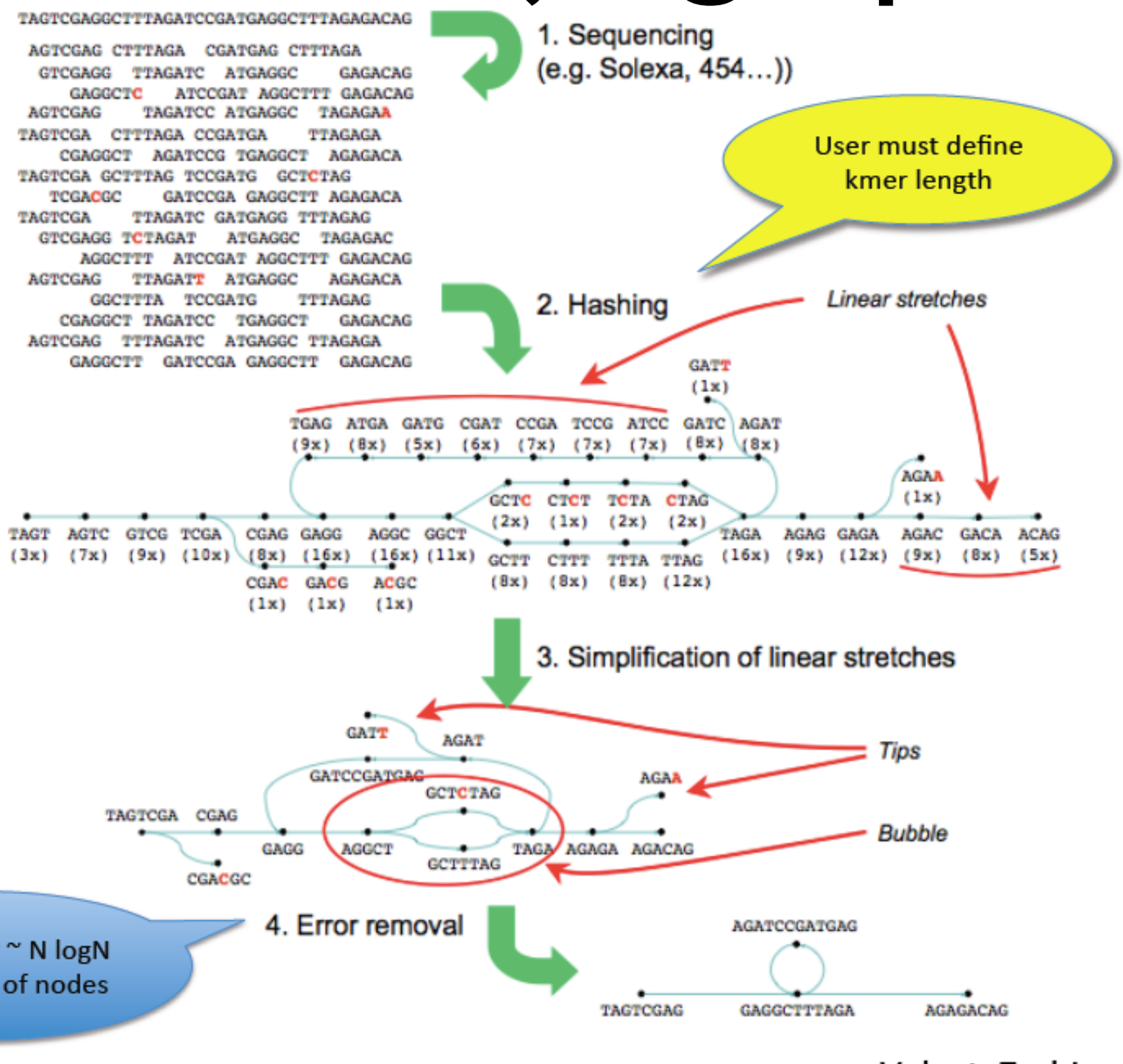
Nodes are k-mers and not reads

☑ small k-mers dense graph (not good)

☑ large k-mers sparse graph (good, results in larger contigs, but need more reads)



deBruijn graph



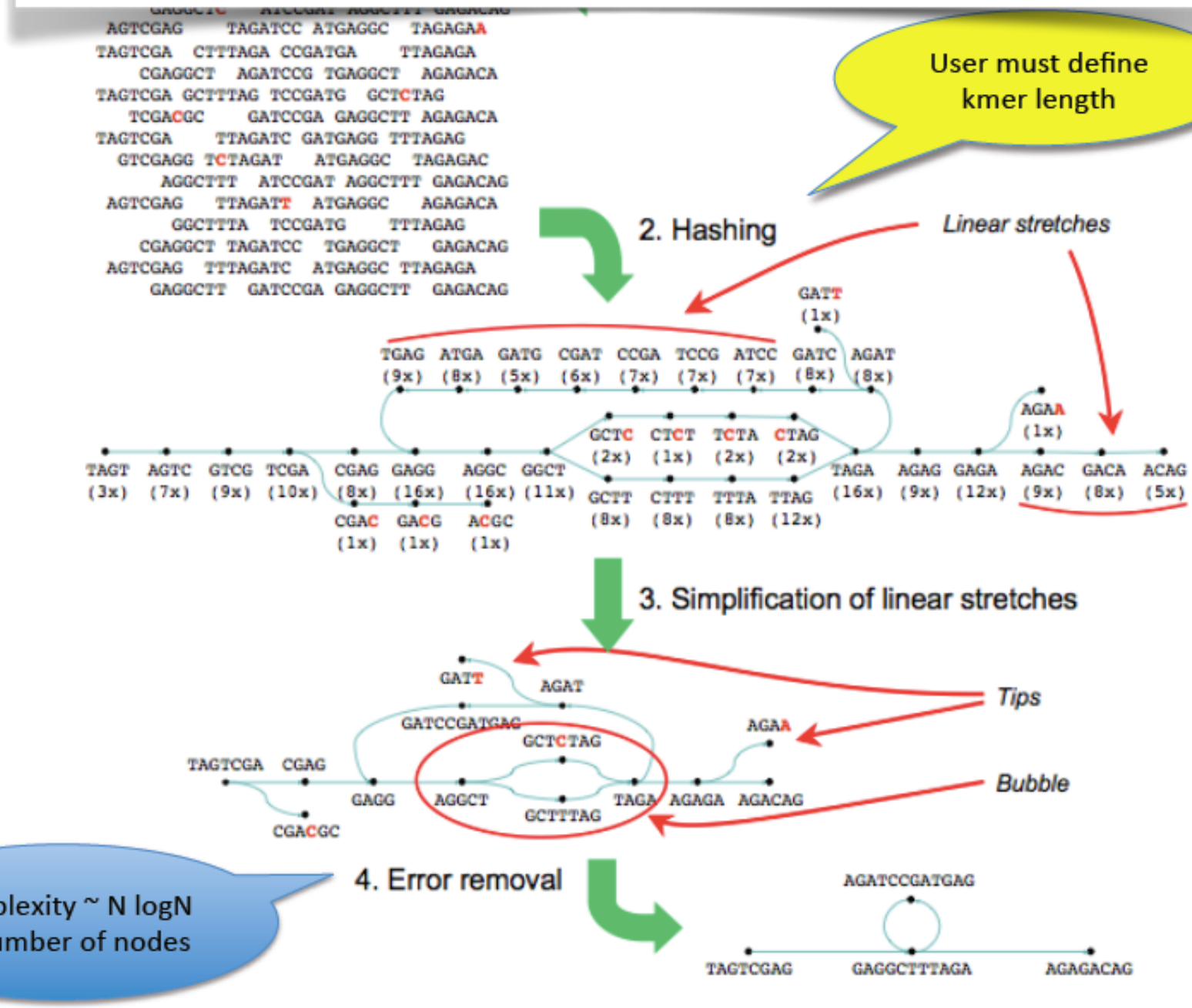
Velvet: Zerbino & Birney

TAGTCGAGGCTTTAGATCCGATGAGGCTTTAGAGACAG

TAGTCGAG

GAGGCTTTAGA AGAGACAG

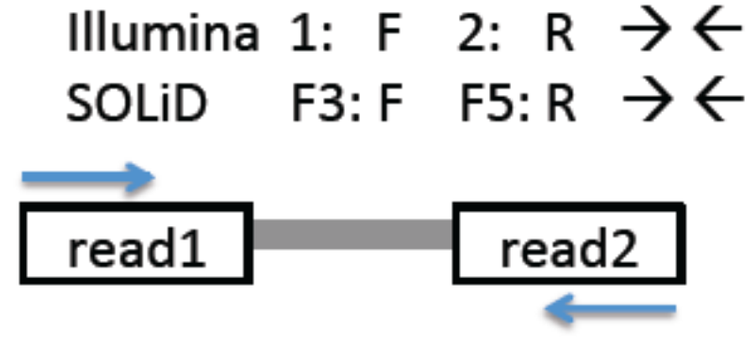
AGATCCGATGAG



Velvet: Zerbino & Birney

Data

- Pair-end
  - 200bp
  - 600bp
  - 800bp
- Mate pair
  - 3Kb
  - 8Kb
  - 20Kb



Assembly measures

- ☑ Sum of Contig length
  - Theoretical genome size
- ☑ Number of contigs
- ☑ N50
  - Contig or scaffold N50 is a weighted median statistic such that 50% of the entire assembly is contained in contigs or scaffolds equal to or larger than this value
- ☑ Accuracy

Assembly measures

Grapevine clone: 6 lanes ($100bp$), insert size 200 ± 50
Coverage: $89\times$

	AbySS	SOAPdenovo	CLC
# Scaf num	289,854 (244k)	127,648 (368k)	151,288 (423k)
Tot Scaf. length (bp)	562M (158M)	257M (285M)	339M (382M)
Max Scaf length (bp)	89,700 (12k)	59,054 (36k)	69,474 (70k)
Mean Scaf lgth (bp)	1942 (649)	2014 (776)	2241 (904)
N50 length	2634 (872)	3186 (2038)	3328 (1823)
time	18h 49m (12h)	8h 57m (1d)	6h 45m (7h)
RAM available (GB)	130 (240)	240 (120)	120 (120)
RAM used (GB)	$\sim$ 90 (102)	143 (70)	$\sim$ 80 (60)
CPUs	80 (80)	8 (8)	8 (8)

Assemblers

- ☑ Phrap
- ☑ CAP3
- ☑ Celera assembler
- ☑ CABOG (modified Celera assembler for 454)
- ☑ Newbler
- ☑ Arachne
- ☑ AMOS (A Modular Open-.−Source whole genome assembler)
- ☑ ABBA (Assembly Boosted by Amino Acid Sequences)
- ☑ MIRA
- ☑ ABySS
- ☑ Euler
- ☑ Velvet
- ☑ SOAPdenovo
- ☑ ALLPATHS, ALLPATHS-.−LG

Assembler

☑ Velvet

- http://www.ebi.ac.uk/~zerbino/velvet/

☑ ABySS

- http://www.bcgsc.ca/platform/bioinfo/software/abyss/

☑ SOAPdenovo

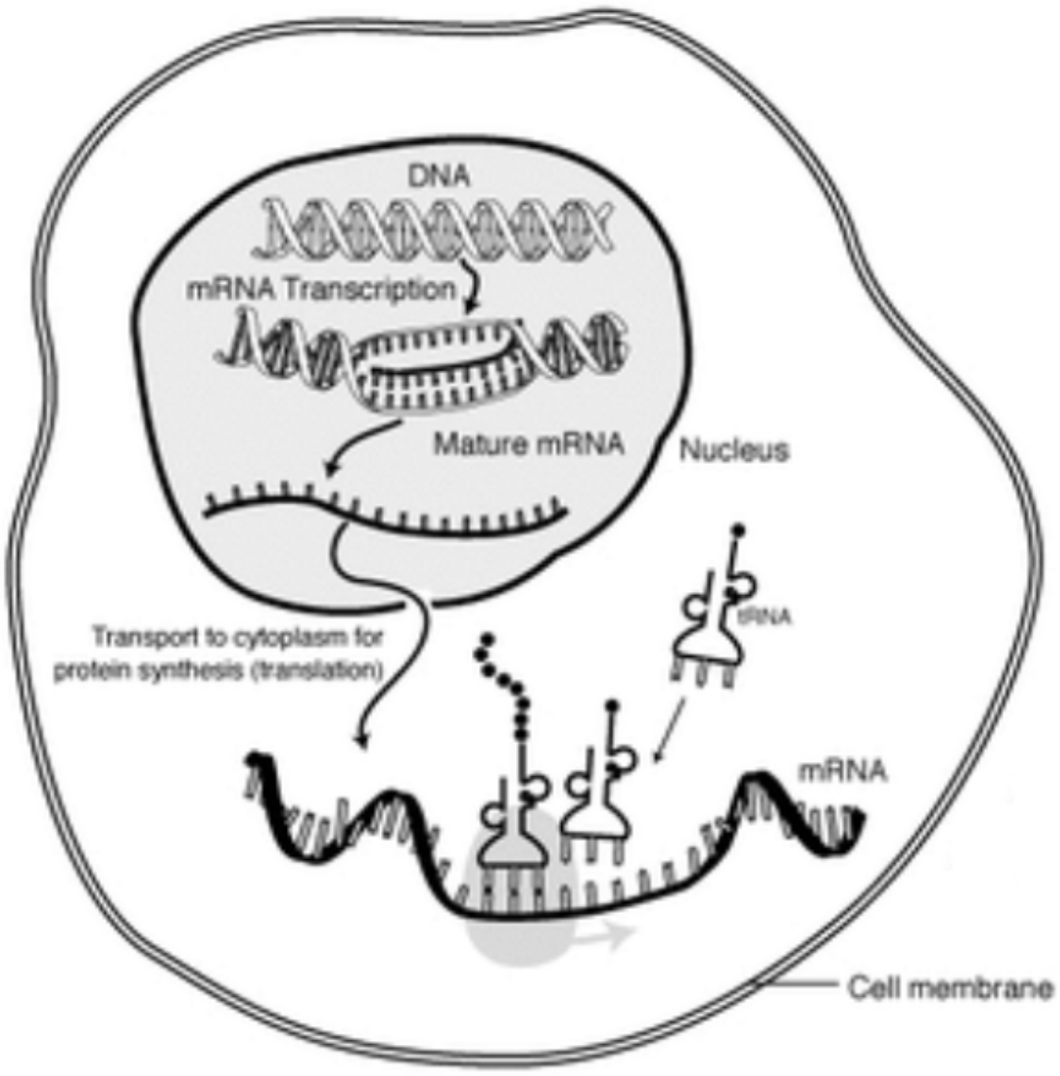
- http://soap.genomics.org.cn/soapdenovo.html

RNA-seq

Andreas Gisel
International Institute of Tropical Agriculture (IITA)
Ibadan, Nigeria

IITA
Research to Nourish Africa
CGIAR

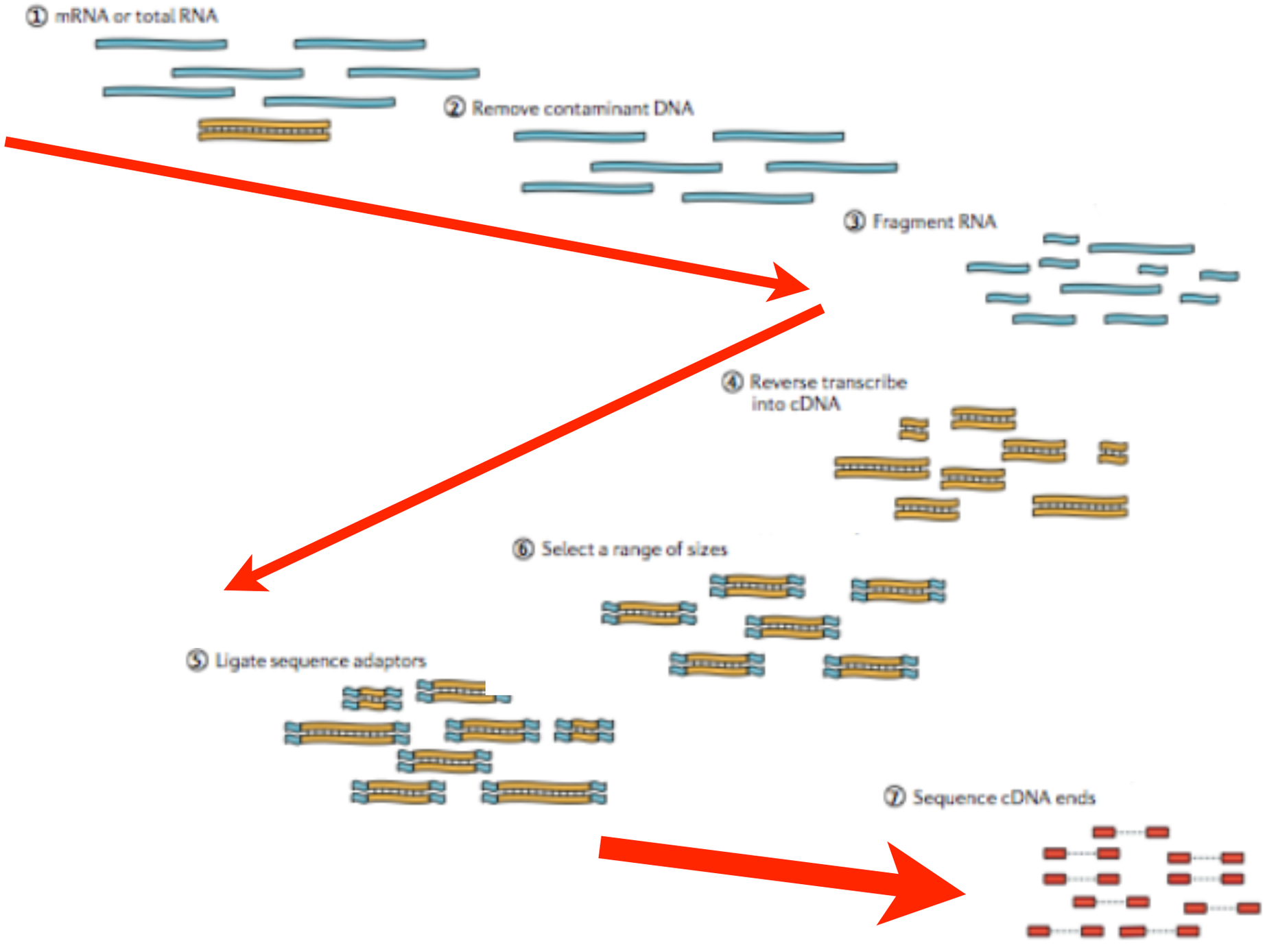
Transcriptome



Methods for Transcriptome Analysis

- EST (Expressed Sequence Tag
  - cDNA library from cloned mRNA sanger sequencing
- Microarray
- qPCR
- SAGE (Serial Analysis of Gene Expression)
- RNA-seq
  - Applying NGS technology

(m)RNA Processing



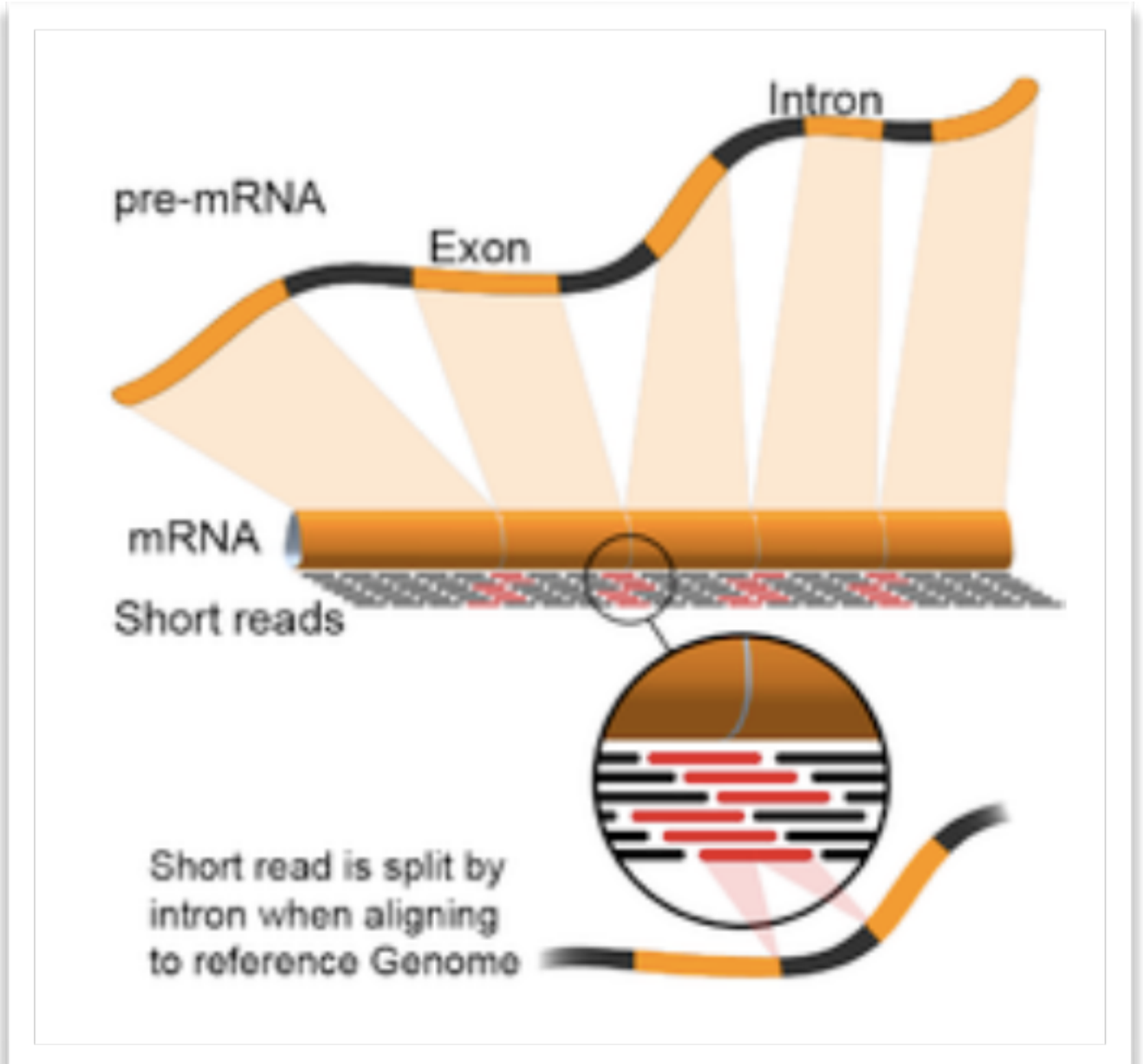
RNA-seq Data Processing

- Transcriptome sequence
  - Gene models
  - Alternativne splicing, novel splice junctions
- Gene expression
  - Transcript abundances
- non-coding RNA
  - miRNA, siRNA, lncRNA, piRNA etc.

RNA-seq Data Mapping

- Reference
  - Genomic
  - Transcriptome (usually incomplete)
- Split-read mapping
- Software:
  - Casava (Illumina), LifeScope (SOLiD), GS Reference Mapper (454)
  - TopHat
  - GSNAP
  - etc

RNA-seq Data Mapping



RNA-seq Software

TopHat (mapping)

- Based on bowtie
- Single reads & PE reads
- Can find novel splice junctions
- Different versions

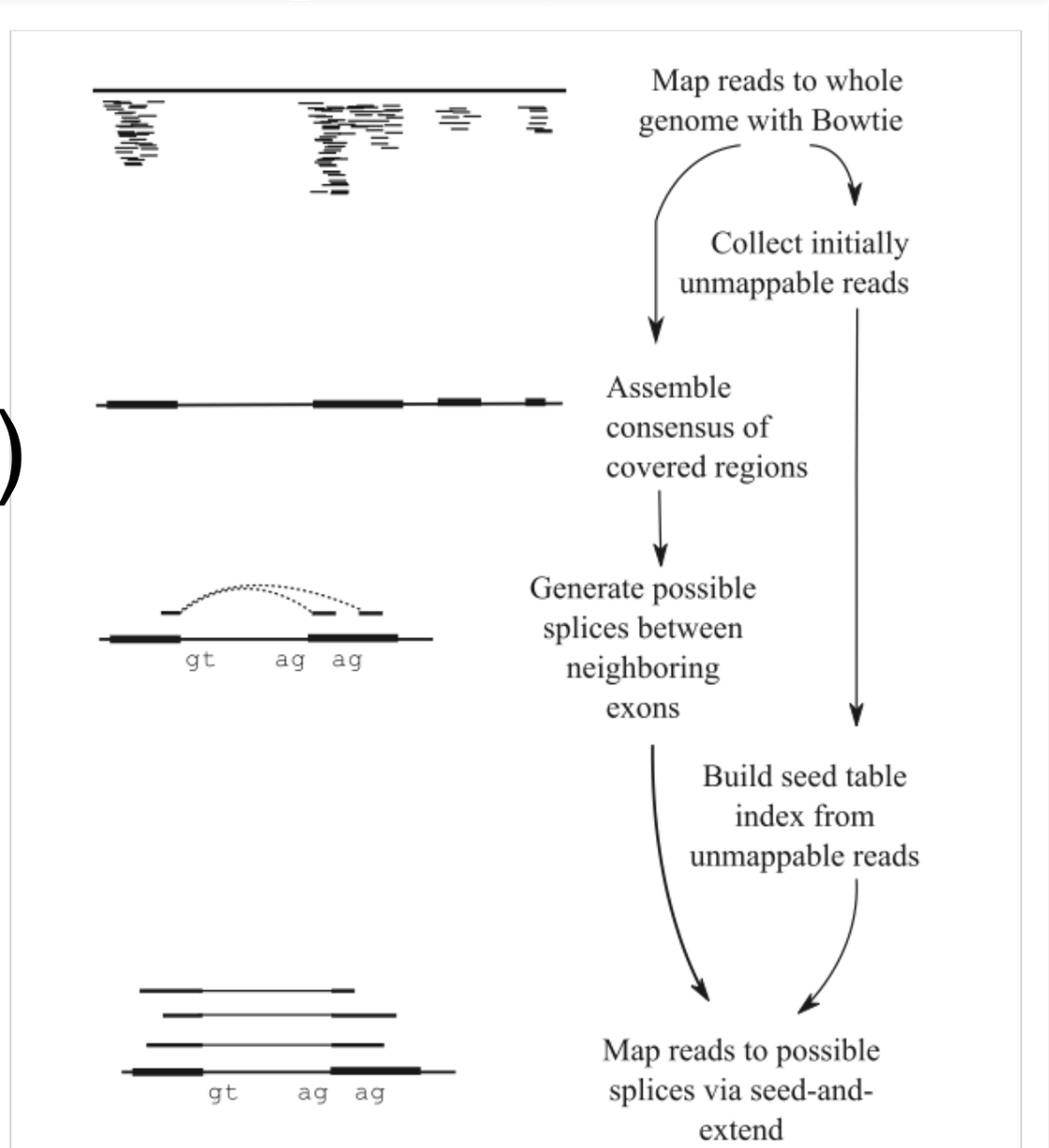


Fig. 1. The TopHat pipeline. RNA-Seq reads are mapped against the whole reference genome, and those reads that do not map are set aside. An initial consensus of mapped regions is computed by Maq. Sequences flanking potential donor/acceptor splice sites within neighboring regions are joined to form potential splice junctions. The IUM reads are indexed and aligned to these splice junction sequences.

Trapnell et al., 2009

RNA-seq Software

TopHat (splice call)

- Based on bowtie
- Single reads & PE reads
- Can find novel splice junctions
- Different versions

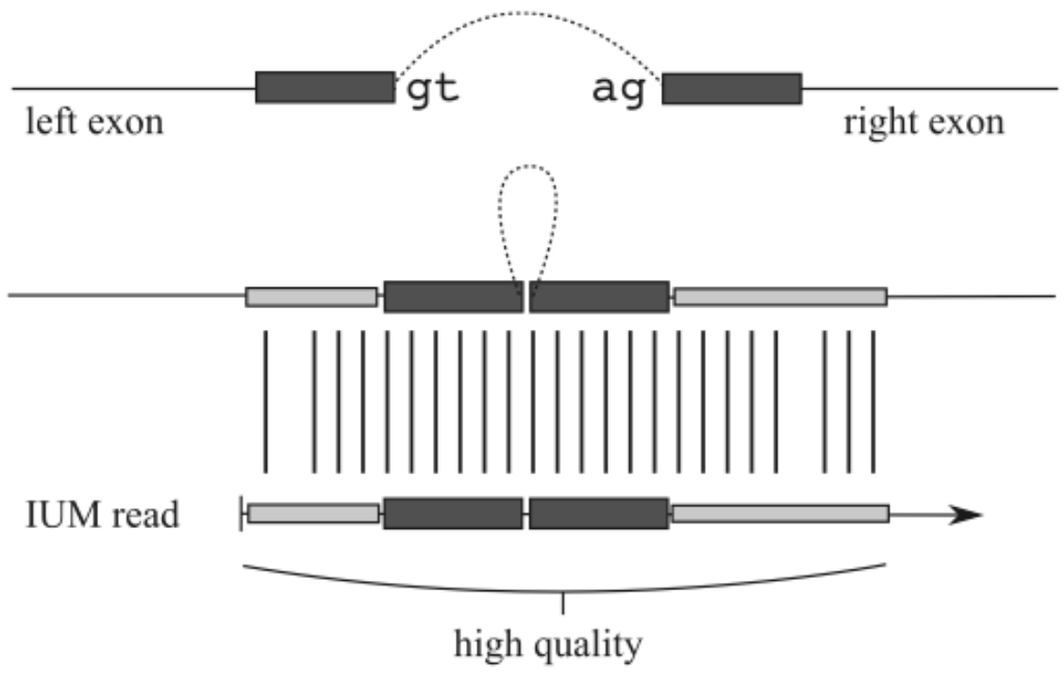


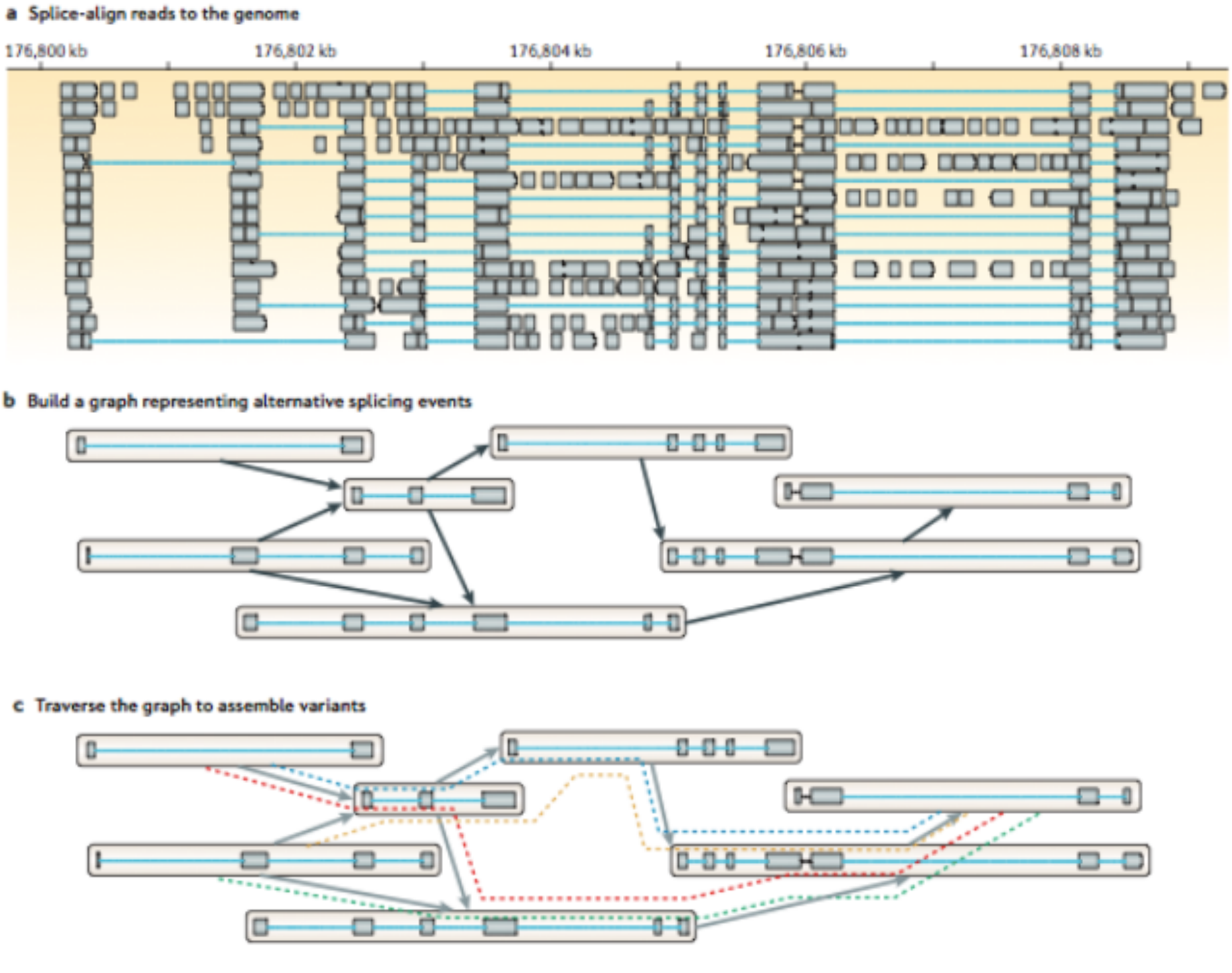
Fig. 3. The seed and extend alignment used to match reads to possible splice sites. For each possible splice site, a seed is formed by combining a small amount of sequence upstream of the donor and downstream of the acceptor. This seed, shown in dark gray, is used to query the index of reads that were not initially mapped by Bowtie. Any read containing the seed is checked for a complete alignment to the exons on either side of the possible splice. In the light gray portion of the alignment, TopHat allows a user-specified number of mismatches. Because reads typically contain low-quality base calls on their 3′ ends, TopHat only examines the first 28 bp on the 5′ end of each read by default.

Trapnell et al., 2009

RNA-seq Software

CuffLinks

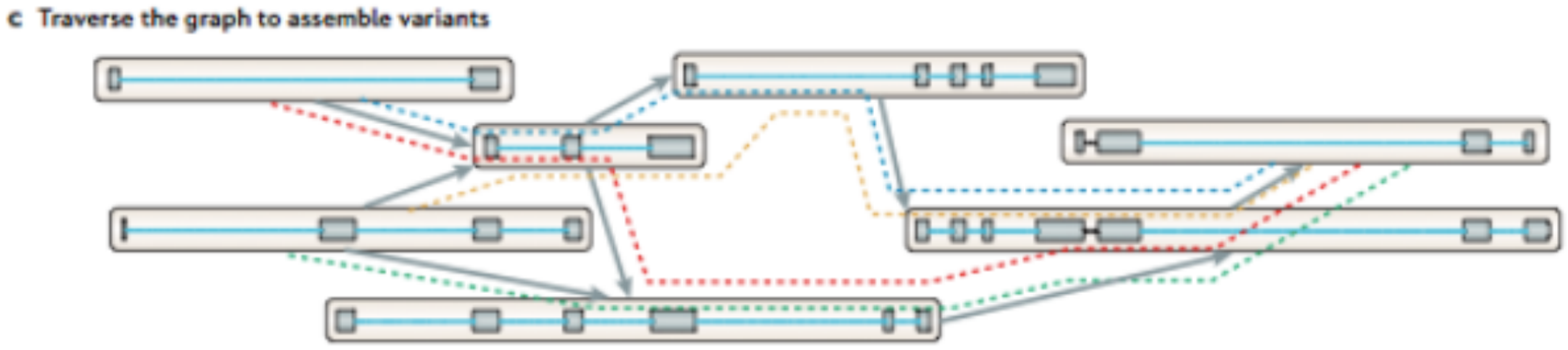
- Transcript call

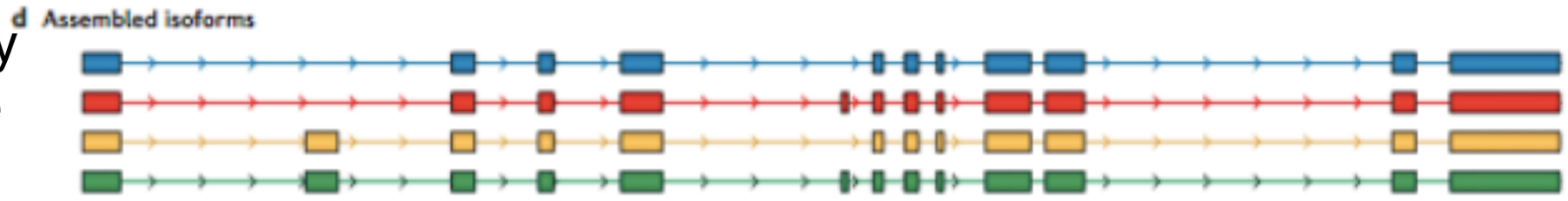


RNA-seq Software

CuffLinks

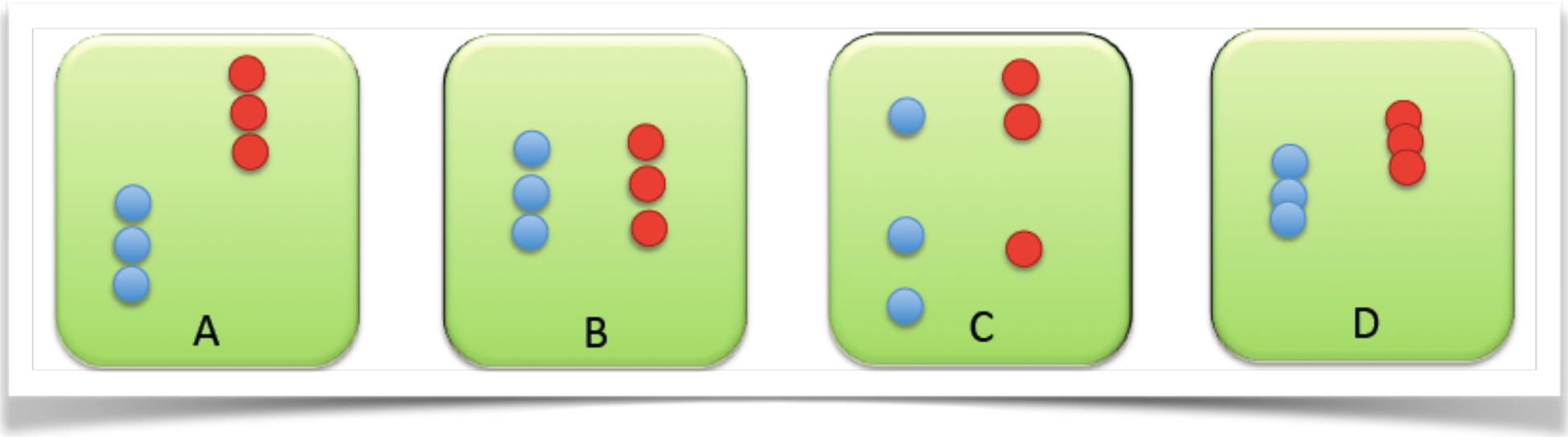
- Transcript assembly based on reference mapping
- Trapnell et al., 2010





RNA-seq Gene Expression

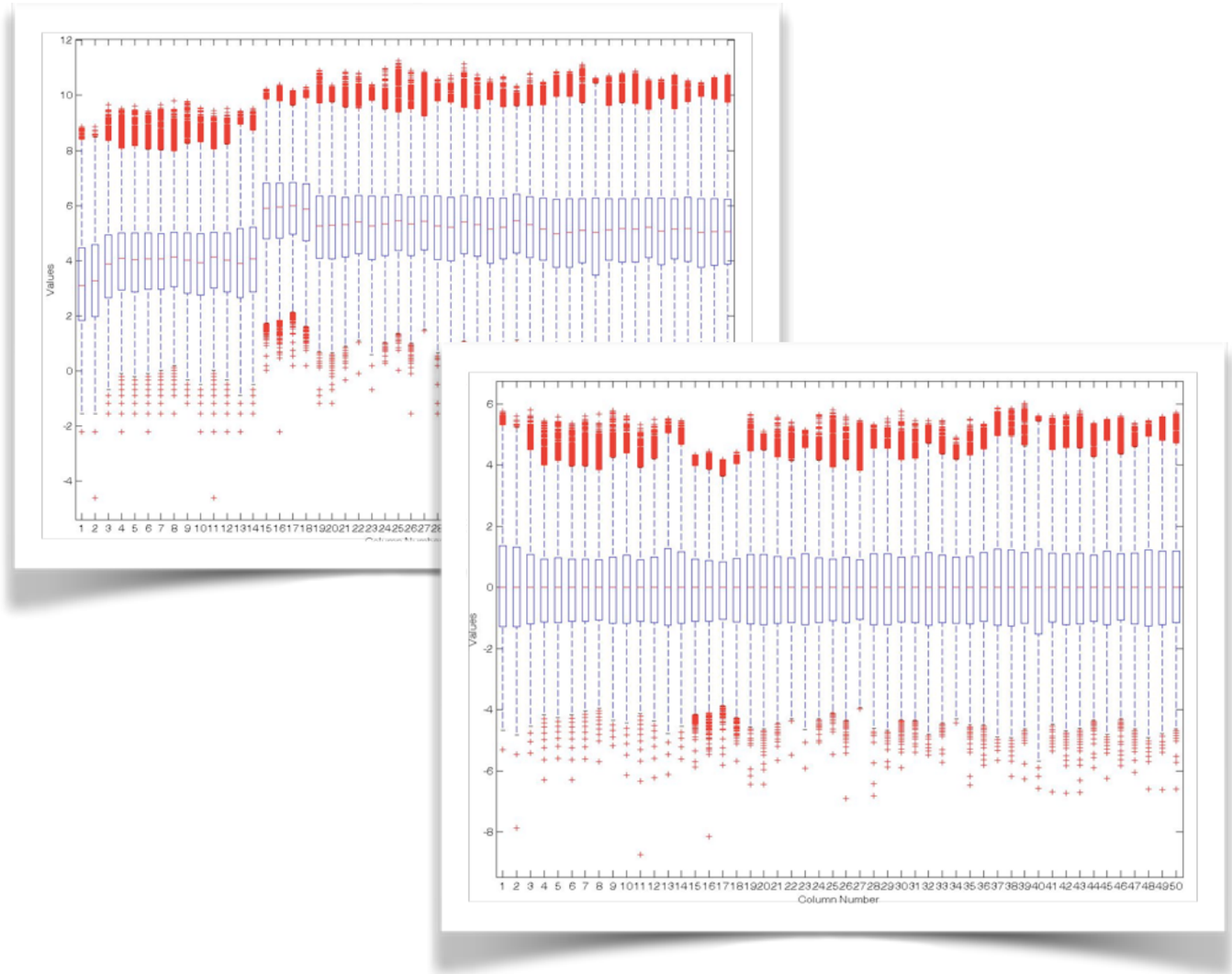
- Sequence count data
  - Mapping
  - Read count - abundance of target transcript
- Differential expression
  - Is difference statistically significant?
  - What is good distribution for count data?



RNA-seq Gene Expression

Data normalization

- Goal
  - Remove technical artefacts
  - Keep biological variation
- Methods
  - Total number of counts
  - House keeping gene(s)
  - Distribution of counts



All normalization methods are based on some assumptions!

RNA-seq Gene Expression

Software

R packages

- edgeR
  - Robinson, McCarthy, Smyth; Bioinformatics 2010: edgeR: a Bioconductor package for differential analysis of digital gene expression data
- DESeq
  - Anders and Huber; Genome Biology 2010: Differential expressio analysis for sequence count data sequence
- baySeq
  - Thomas J Hardcastle* and Krystyna A Kelly; BMC Bioinformatics 2010: baySeq: Empirical Bayesian methods for identifying differential expression in sequence count data

Analyzing known and new small RNAs from Deep Sequencing Data

Andreas Gisel
Inernational Institute of Tropical Agriculture (IITA)
Ibadan, Nigeria

Discovery

1986

Proc. Natl. Acad. Sci. USA
Vol. 83, pp. 5372–5376, August 1986
Biochemistry

Inhibition of gene expression in plant cells by expression of antisense RNA

(chimeric genes/electroporation/plant transformation/transient chloramphenicol

JOSEPH R. ECKER AND RONALD W. DAVIS
Department of Biochemistry, Stanford University School of Medicine, Stanford, CA 94305

1990

The Plant Cell, Vol. 2, 279–289, April 1990 © 1990 American Society of Plant Physiologists

Introduction of a Chimeric Chalcone Synthase Gene into Petunia Results in Reversible Co-Suppression of Homologous Genes *in trans*

Carolyn Napoli,[1] Christine Lemieux, and Richard Jorgensen[2]
DNA Plant Technology Corporation, 6701 San Pablo Avenue, Oakland, California 94608

1992

Molecular Microbiology (1992) 6(22), 3343–3353

Quelling: transient inactivation of gene expression in *Neurospora crassa* by transformation with homologous sequences

Nicoletta Romano and Giuseppe Macino*
Dipartimento di Biopatologia Umana, Sezione di Biologia Cellulare, Policlinico Umberto 1, Università di Roma 'La Sapienza', 00161 Rome, Italy.

whether these are tandemly arranged or located on different chromosomes (Faugeron *et al.*, 1990; Selker, 1990). Pre-meiotic inactivation appears to involve at least two different steps: an initial interaction between homologous sequences followed by sequence modifications, either cytosine methylation as in *A. immersus*, or both methyla-

Discovery

1998

NATURE | VOL 391 | 19 FEBRUARY 1998

Potent and specific genetic interference by double-stranded RNA in *Caenorhabditis elegans*

Andrew Fire*, SiQun Xu*, Mary K. Montgomery*, Steven A. Kostas*†, Samuel E. Driver‡ & Craig C. Mello‡

* *Carnegie Institution of Washington, Department of Embryology, 115 West University Parkway, Baltimore, Maryland 21210, USA*
† *Biology Graduate Program, Johns Hopkins University, 3400 North Charles Street, Baltimore, Maryland 21218, USA*
‡ *Program in Molecular Medicine, Department of Cell Biology, University of Massachusetts Cancer Center, Two Biotech Suite 213, 373 Plantation Street, Worcester, Massachusetts 01605, USA*

Experimental introduction of RNA into cells can be used in certain biological systems to interfere with the function of an endogenous gene[1,2]. Such effects have been proposed to result from a simple antisense mechanism that depends on hybridization between the injected RNA and endogenous messenger RNA transcripts. RNA interference has been used in the nematode *Caenorhabditis elegans* to manipulate gene expression[3,4]. Here we investigate the requirements for structure and delivery of the interfering RNA. To our surprise, we found that double-stranded RNA was substantially more effective at producing interference than was either strand individually. After injection into adult animals, purified single strands had at most a modest effect, whereas double-stranded mixtures caused potent and specific interference. The effects of this interference were evident in both the injected animals and their progeny. Only a few molecules of injected double-stranded RNA were required per affected cell, arguing against stochiometric interference with endogenous mRNA and suggesting that there could be a catalytic or amplification component in the interference process.

The Nobel Prize in Physiology or Medicine 2006
Andrew Z. Fire, Craig C. Mello

Nobelförsamlingen
The Nobel Assembly at Karolinska Institutet

Karolinska Institutet

English
Swedish

Press Release

2 October 2006

The Nobel Assembly at Karolinska Institutet has today decided to award The Nobel Prize in Physiology or Medicine for 2006 jointly to

Andrew Z. Fire and **Craig C. Mello**

for their discovery of **"RNA interference – gene silencing by double-stranded RNA"**

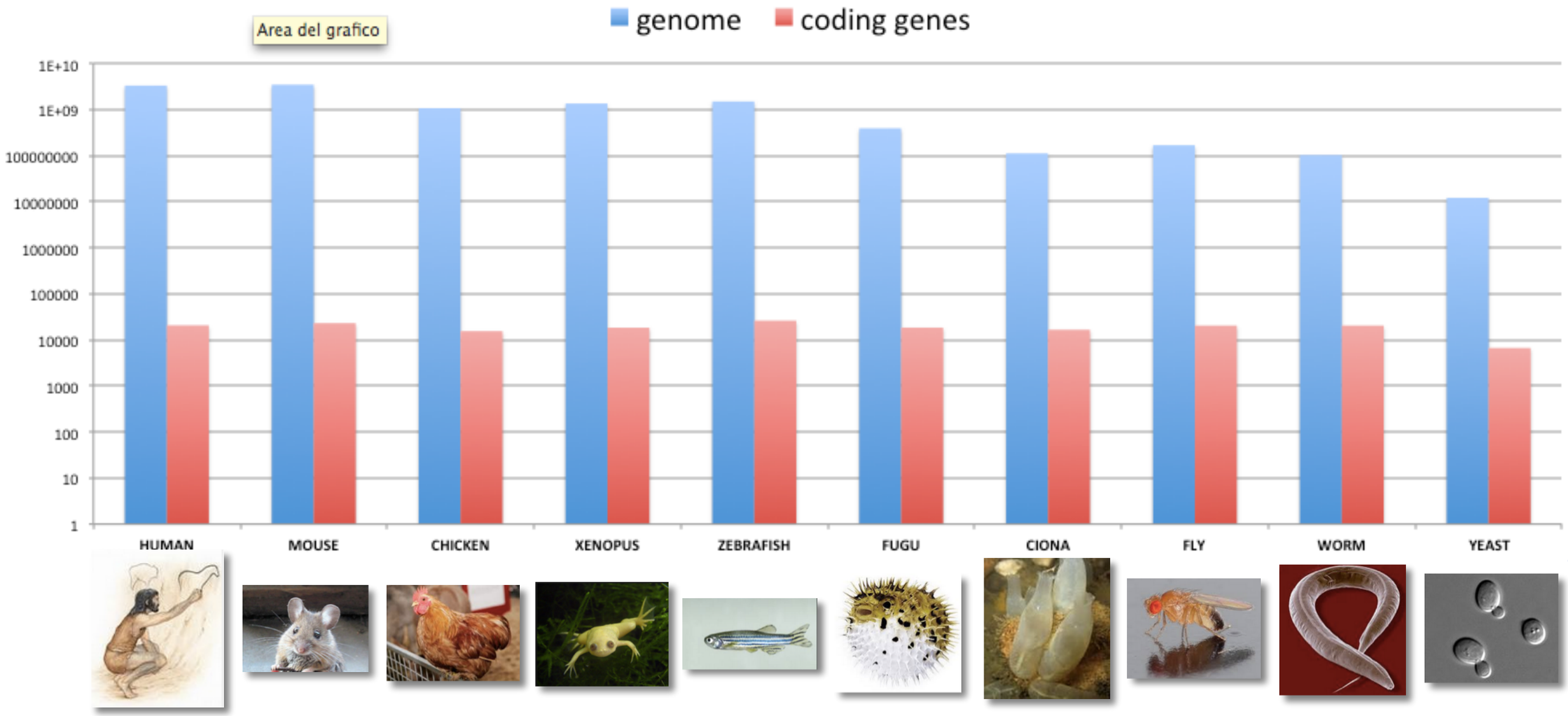
Summary

PTGS - Post Transcriptional Gene Silencing (plants)
RNAi - RNA interference (animals)

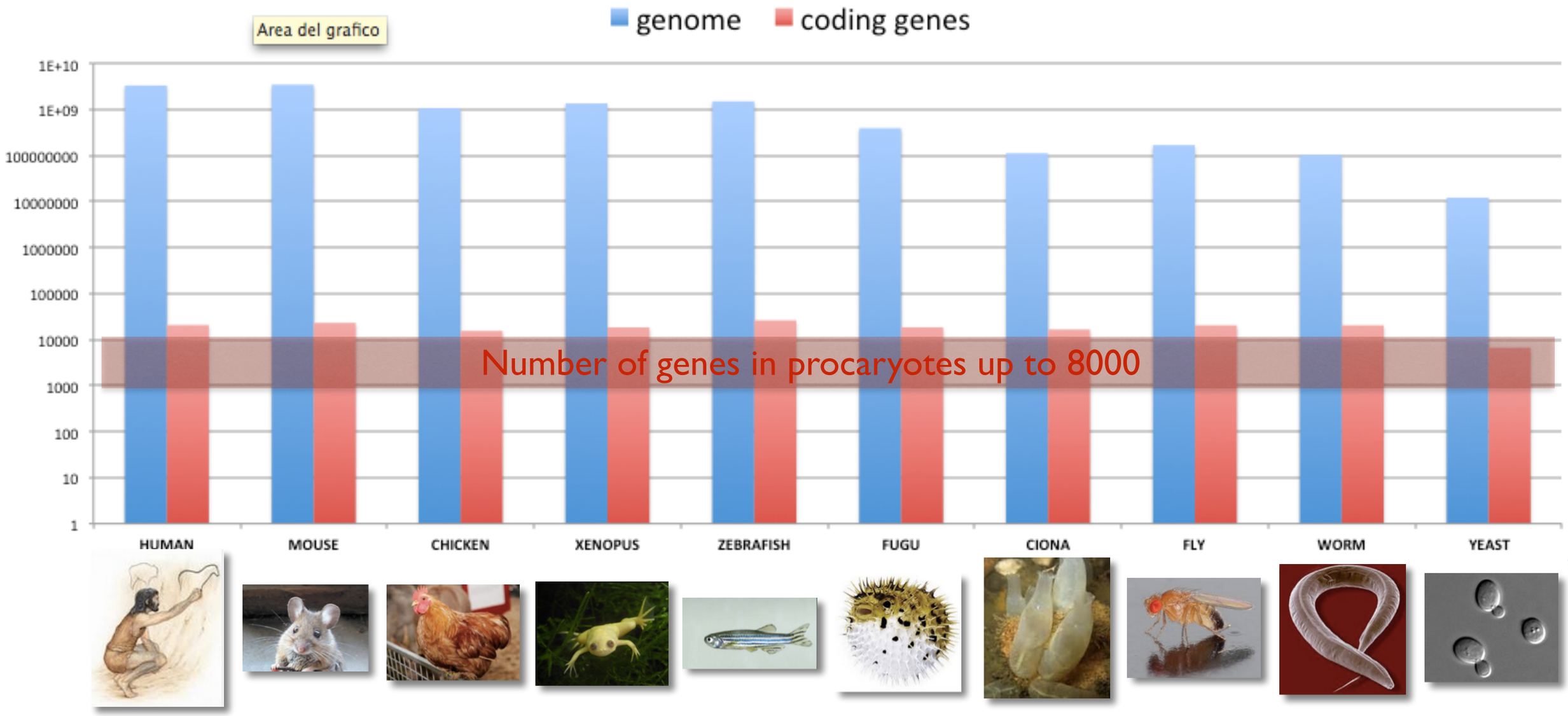
both

result from the same mechanism for the down-regulation of a gene at the RNA level by cleavage or blocking of the mRNA with help of short (19 - 25) single stranded RNA fragments.

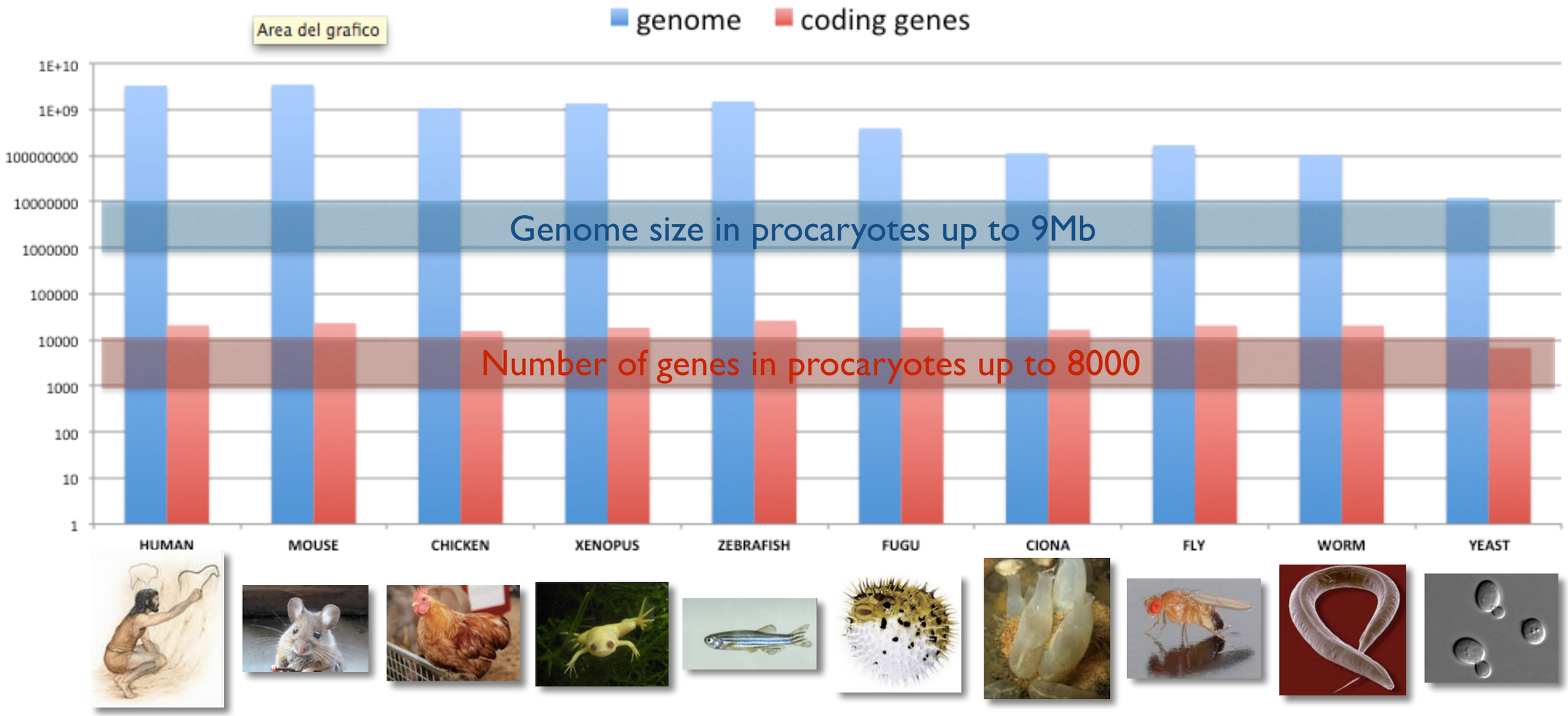
Coding vs. non-coding RNA



Coding vs. non-coding RNA



Coding vs. non-coding RNA



Coding vs. non-coding RNA

An explanation for this apparent paradox comes from two unexpected findings:

☑ that biological complexity generally correlates with the proportion of the genome that is non-protein-coding; and

☑ that, while only 2% of the mammalian genome encodes mRNAs, the vast majority is transcribed, largely as long and short non-protein-coding RNAs (ncRNAs)

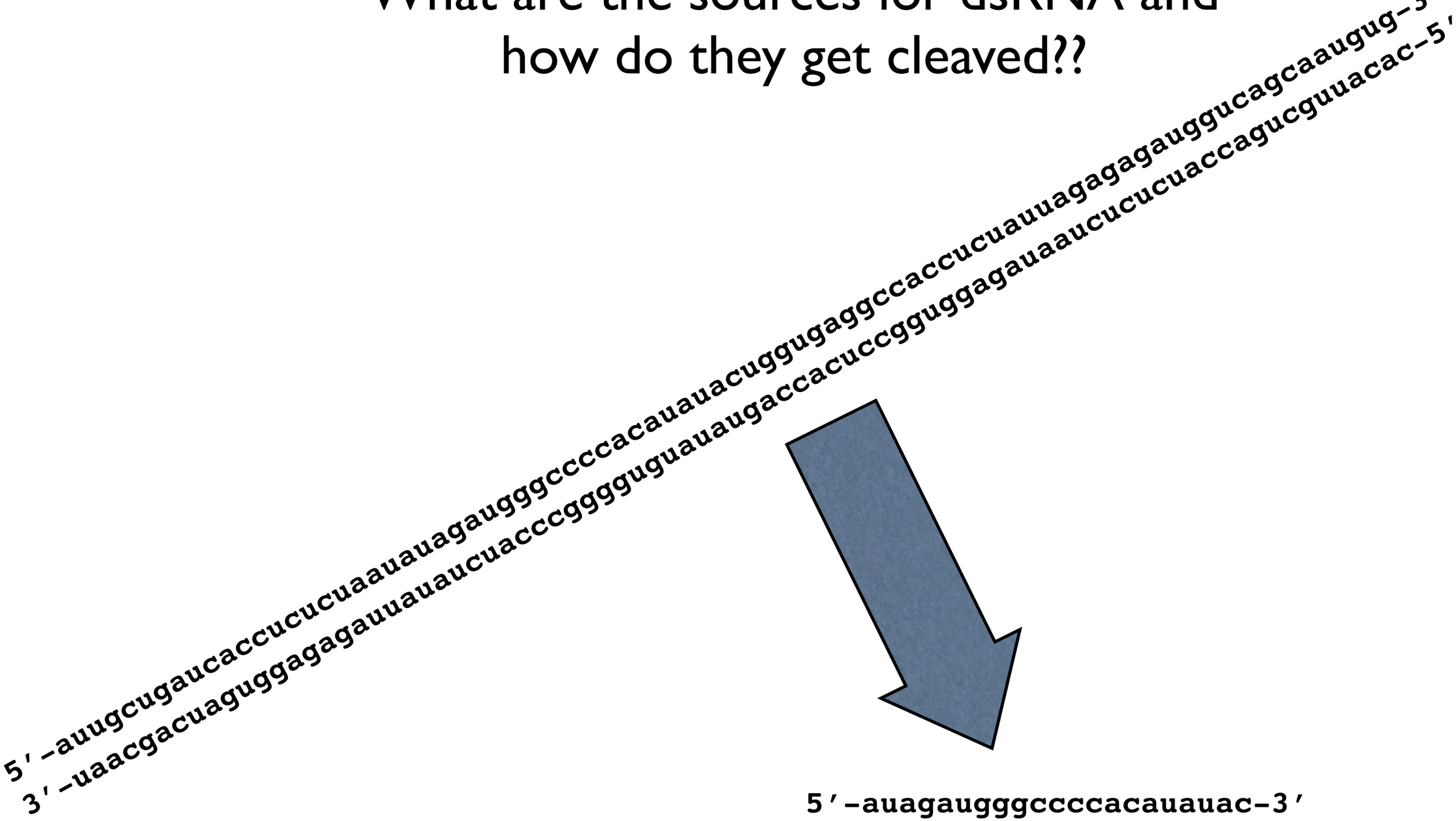
Coding vs. non-coding RNA

An explanation for this apparent paradox comes from two unexp

- ☑ tha he pro and
- ☑ tha es mR ng and

small RNAs

What are the sources for dsRNA and
how do they get cleaved??



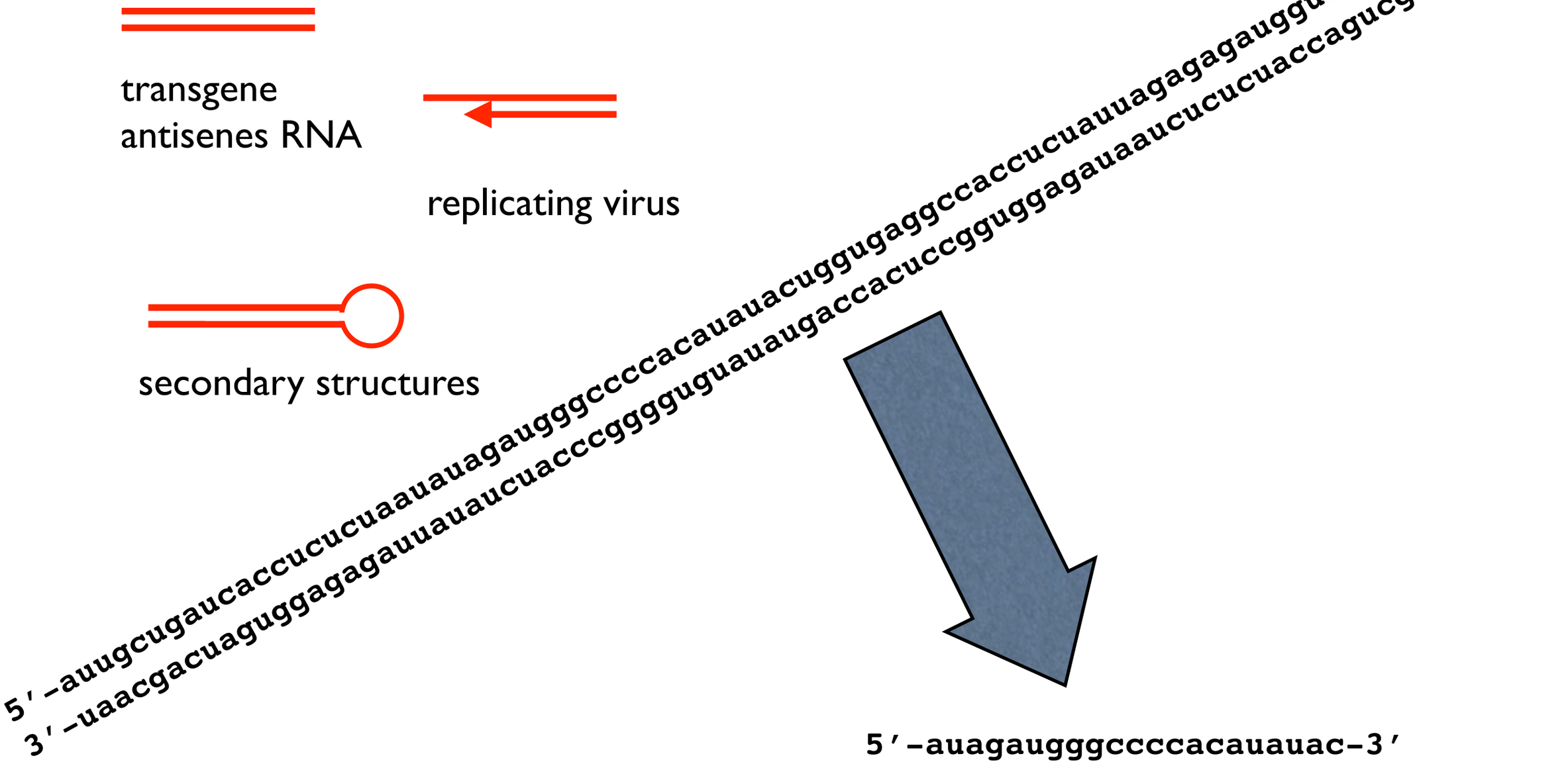
small RNAs

What are the sources for dsRNA and how do they get cleaved??

transgene antisenes RNA

replicating virus

secondary structures



small RNAs

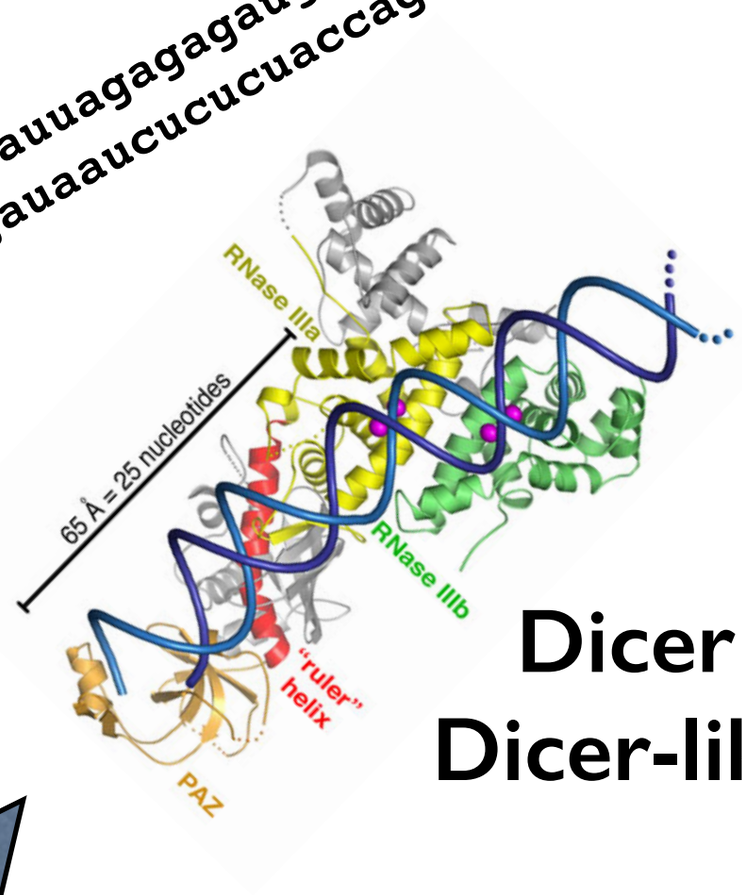
What are the sources for dsRNA and how do they get cleaved??

transgene
antisenes RNA

replicating virus

secondary structures

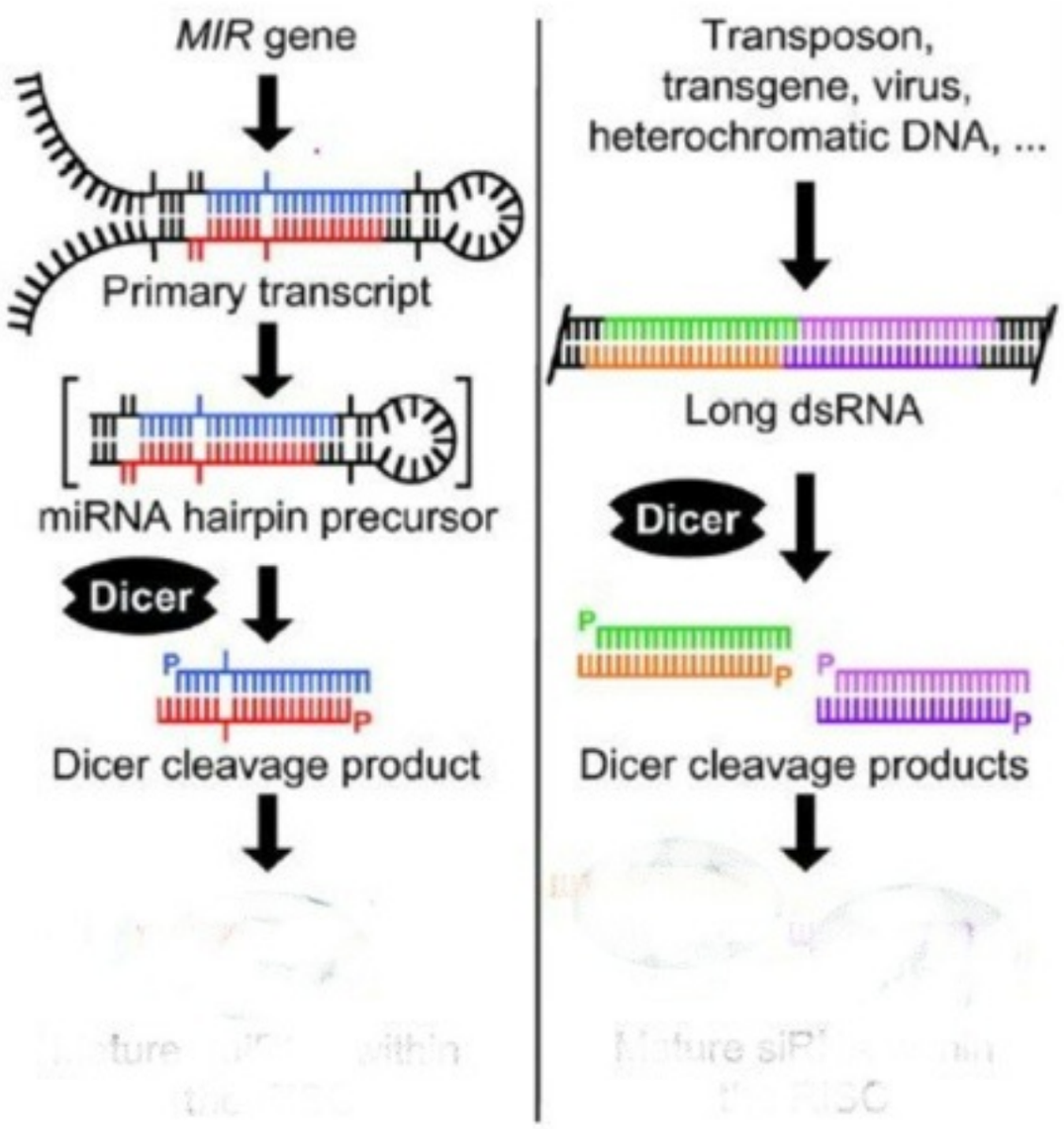
5′–auugcugaucaccucucuaauauagaugggccccacauauacuggugaggccaccucuauuagagagauggucagcaaugug–3′
3′–uaacgacuaguggagagauuauaucuacccgggguguauaugaccacuccgguggagauaaucucucuaccagucguuacac–5′



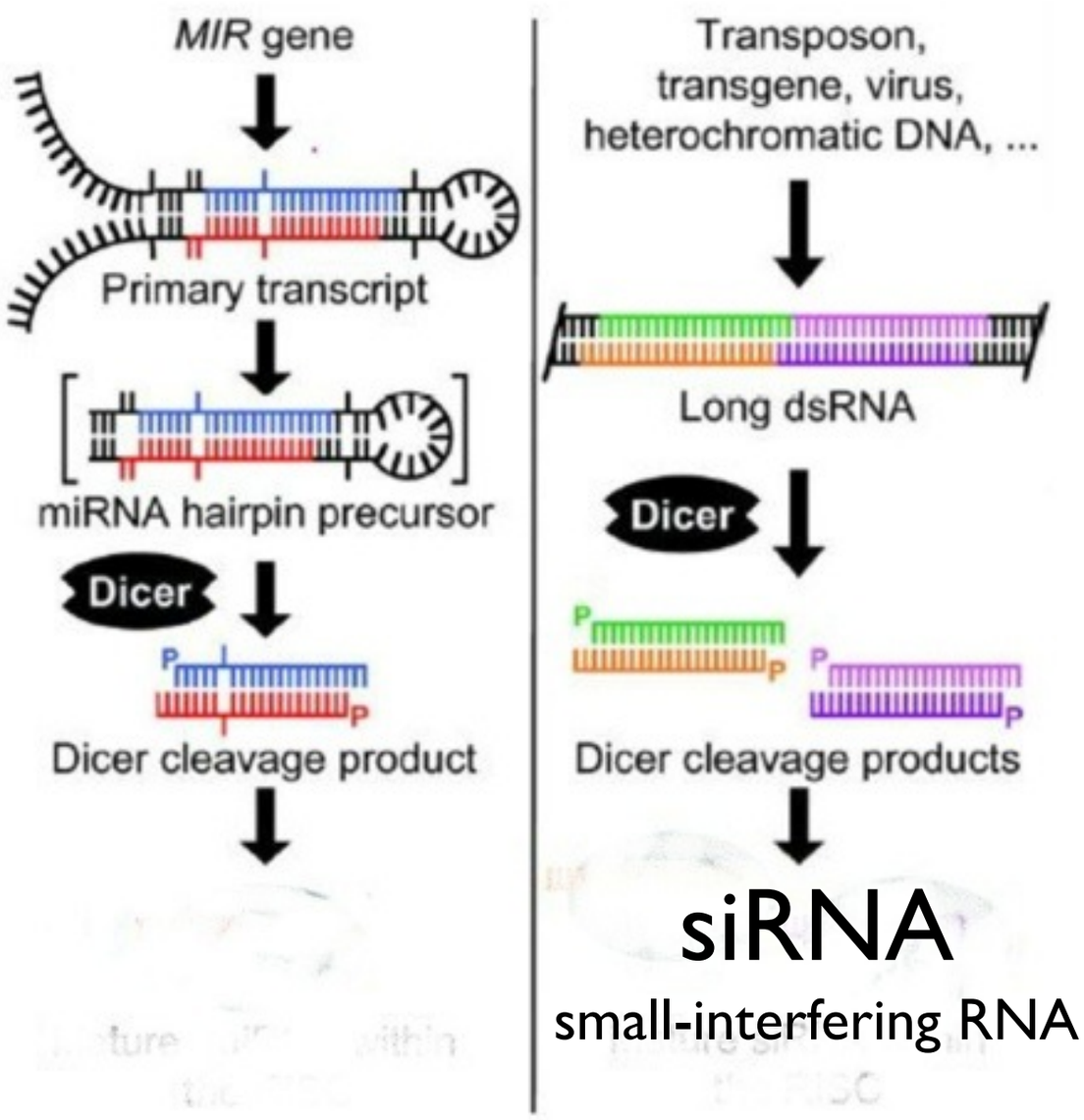
Dicer
Dicer-like

5′–auagaugggccccacauauac–3′

small RNAs

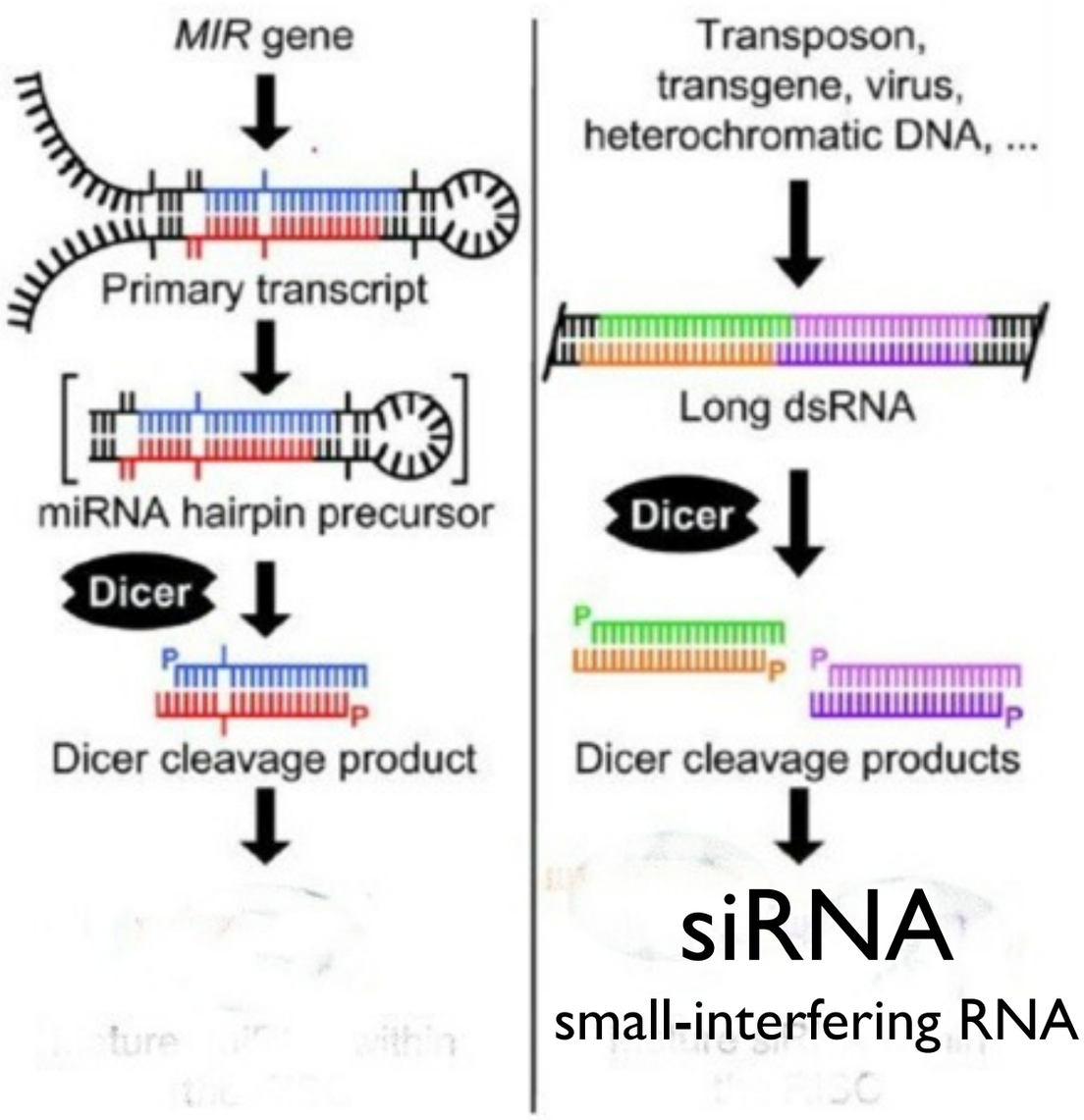

MIR gene
Primary transcript
miRNA hairpin precursor
Dicer
P
P
Dicer cleavage product
Transposon,
transgene, virus,
heterochromatic DNA, ...
Long dsRNA
Dicer
P
P
P
P
Dicer cleavage products

small RNAs


MIR gene
Primary transcript
miRNA hairpin precursor
Dicer
P
P
Dicer cleavage product
Transposon,
transgene, virus,
heterochromatic DNA, ...
Long dsRNA
Dicer
P
P
P
P
Dicer cleavage products
siRNA
small-interfering RNA

small RNAs

Drosha



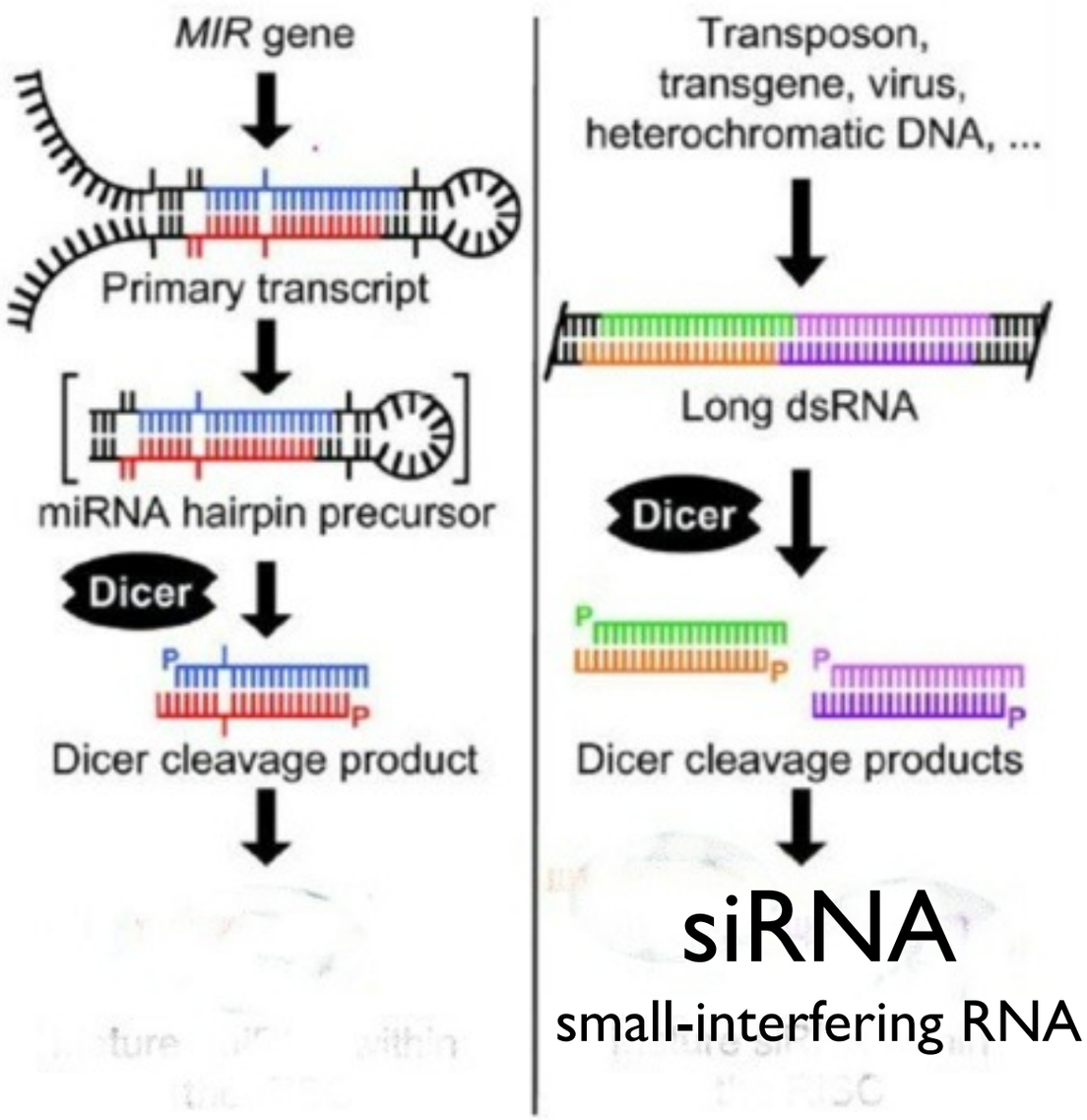
siRNA

small-interfering RNA

small RNAs

Drosha

two-base overhang on the 3' end



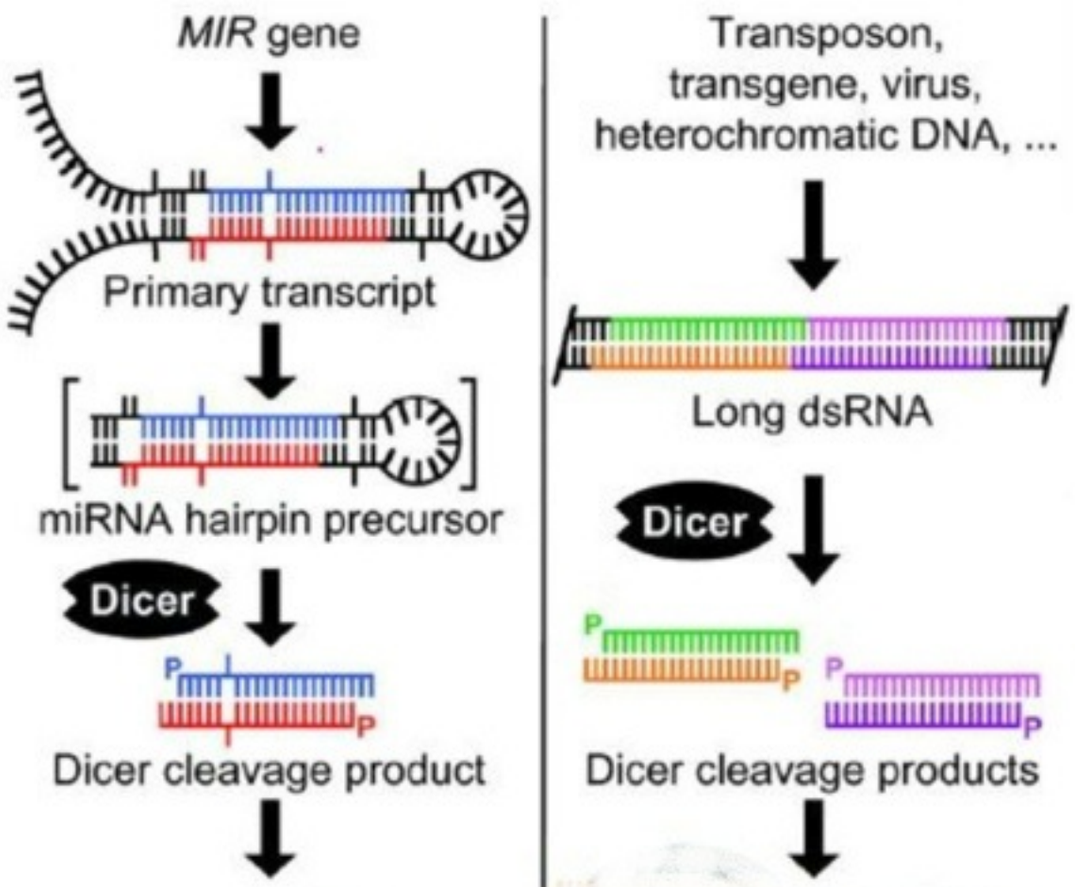
siRNA

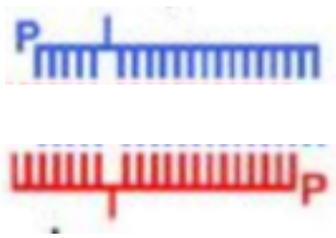
small-interfering RNA

small RNAs

Drosha

two-base overhang on the 3' end





mature miRNA

miRNA* (star)

micro RNA

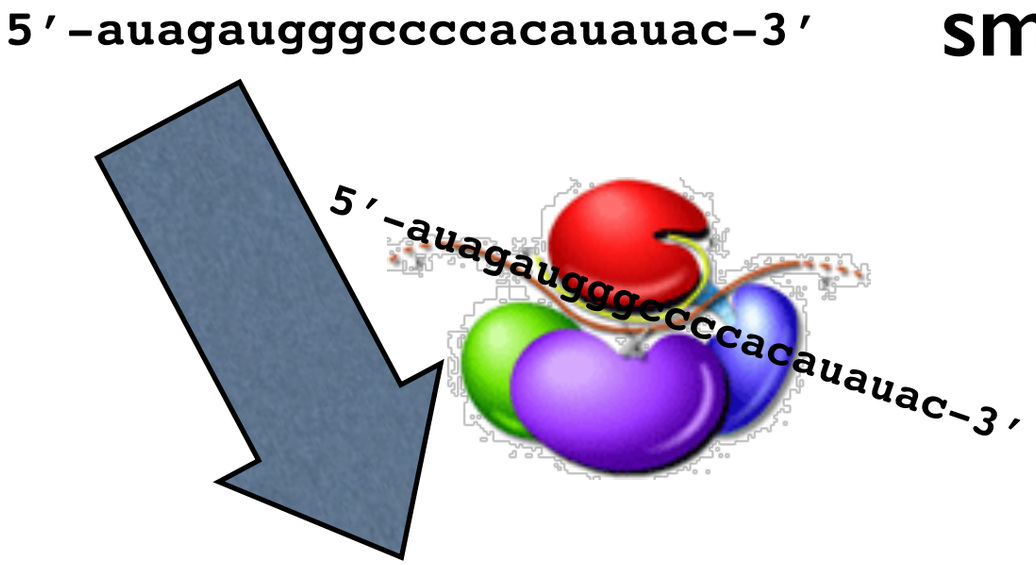
siRNA

small-interfering RNA

small RNAs

5′–auagaugggcccccacauauac–3′ **small RNA**

small RNAs



small RNA

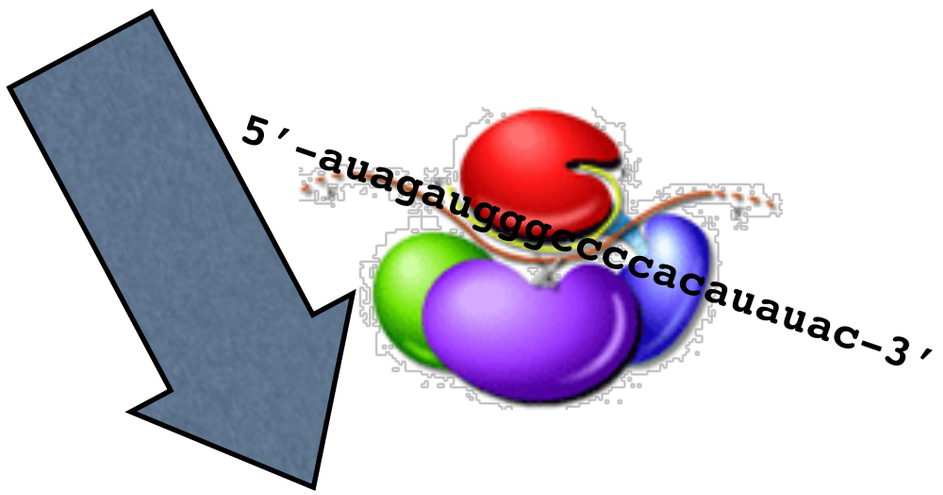
RISC

RNA induced silencing complex

small RNAs

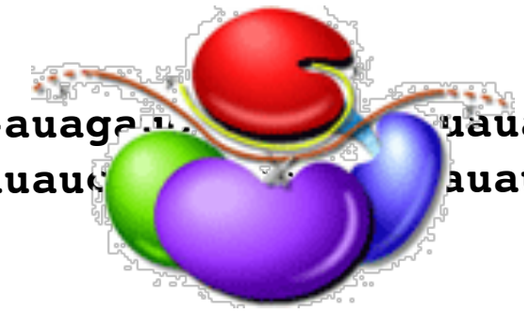
5′–auagaugggcccacauauac–3′

small RNA



RISC

RNA induced silencing complex



5′–auaga…uauac–3′

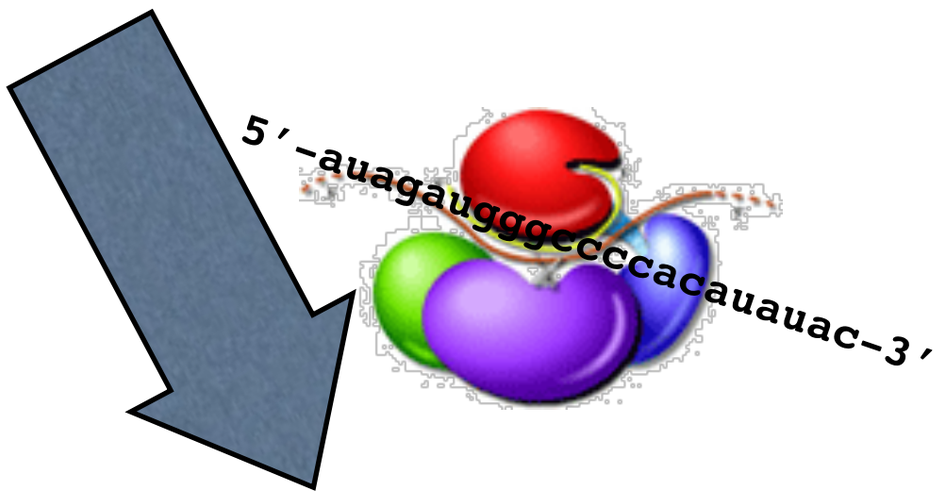
target

3′–uaacgacuaguggagagauuauauc…auaugaccacuccgguggagauaaucucucuaccagucguuacac–5′

small RNAs

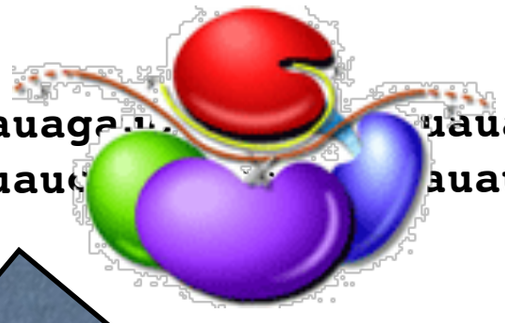
5′–auagaugggcccacauauac–3′

small RNA



RISC

RNA induced silencing complex



5′–auaga…uauac–3′

3′–uaacgacuaguggagagauuauauc…auaugaccacuccgguggagauaaucucucuaccagucguuacac–5′

target

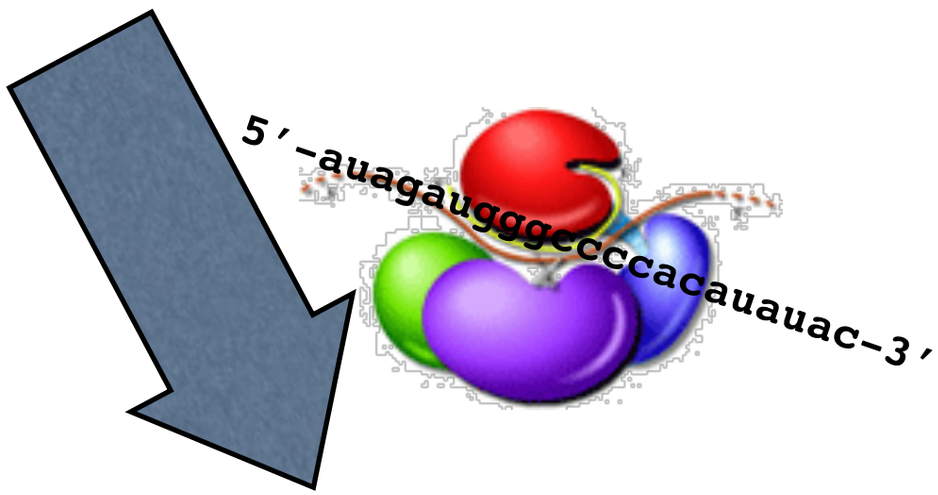
cleave

3′–uaacgacuaguggagagauuauaucuacccgg gguguauaugaccacuccgguggagauaaucucucuaccagucguuacac–5′

small RNAs

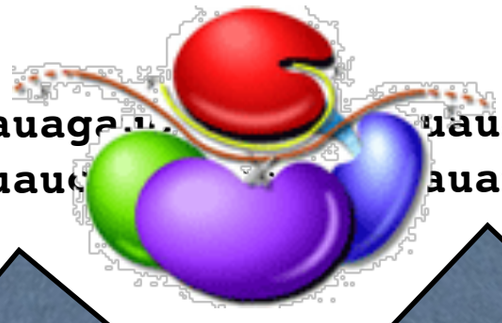
5′–auagaugggcccacauauac–3′

small RNA



RISC

RNA induced silencing complex



5′–auaga…uauac–3′

3′–uaacgacuaguggagagauuauauc…auaugaccacuccgguggagauaaucucucuaccagucguuacac–5′

target

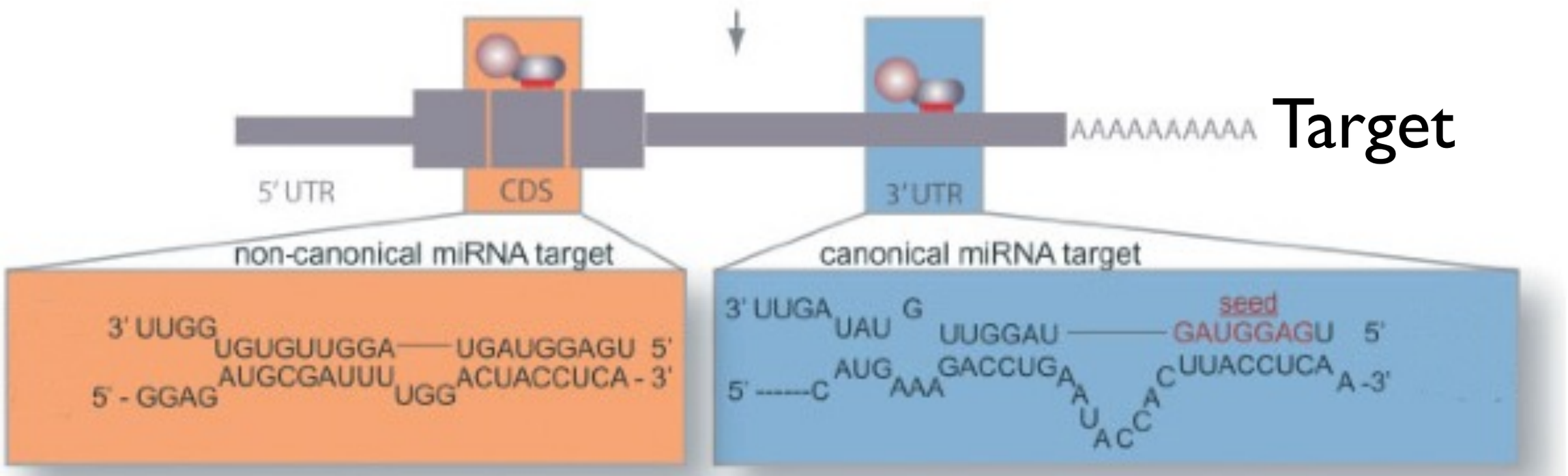
cleave

block

3′–uaacgacuaguggagagauuauaucuac…uaugaccacuccgguggagauaaucucucuaccagucguuacac–5′

3′–uaacgacuaguggagagauuauaucuacccgg gguguauaugaccacuccgguggagauaaucucucuaccagucguuacac–5′

small RNAs



no seed but more extensive base pairing

perfect seed match

non-canonical and canonical targeting

other small RNAs

- promoter-associated small RNAs
- sno-derived small RNAs
- ta-siRNAs (Trans-acting siRNA)
- piRNAs (Piwi-interacting RNA)

Bioinformatics of miRNAs

Parameters for the miRNA analysis

- specific sequence length
- two-base overhang on the 3' end between mature and star sequences
- precursor sequence folds in stem loop manner
- high abundance of mature miRNA

Bioinformatics of miRNAs

Steps for the miRNA data analysis

- Read processing
- miRNA identification
- Target search
- Digital Gene Expression

Bioinformatics of miRNAs

- Read processing

Bioinformatics of miRNAs

- Read processing

Bioinformatics of miRNAs

Read processing

- Reads are typically around 50 bases and therefore read into the 3’-adaptor
- If multiplexed detect barcode sequence and separate experiments
- Trimming barcode and 3’-adaptor fragment (adaptor sequence not always present since “small RNA” too long ==> discharge!!!!
- Since sequencing errors tends to be more frequent in 3’ part, 3‘-adaptor trimming need to accept mismatches

Bioinformatics of miRNAs

Read processing

- Small RNA are regulated and therefore can appear as redundant sequences in the deep sequencing data set.

- Therefore the cleaned reads are clustered so that the new data set consists on unique sequences.

Bioinformatics of miRNAs

Read processing

- Small RNA are regulated and therefore can appear as redundant sequences in the deep sequencing data set.

- Therefore the cleaned reads are clustered so that the new data set consists on unique sequences.

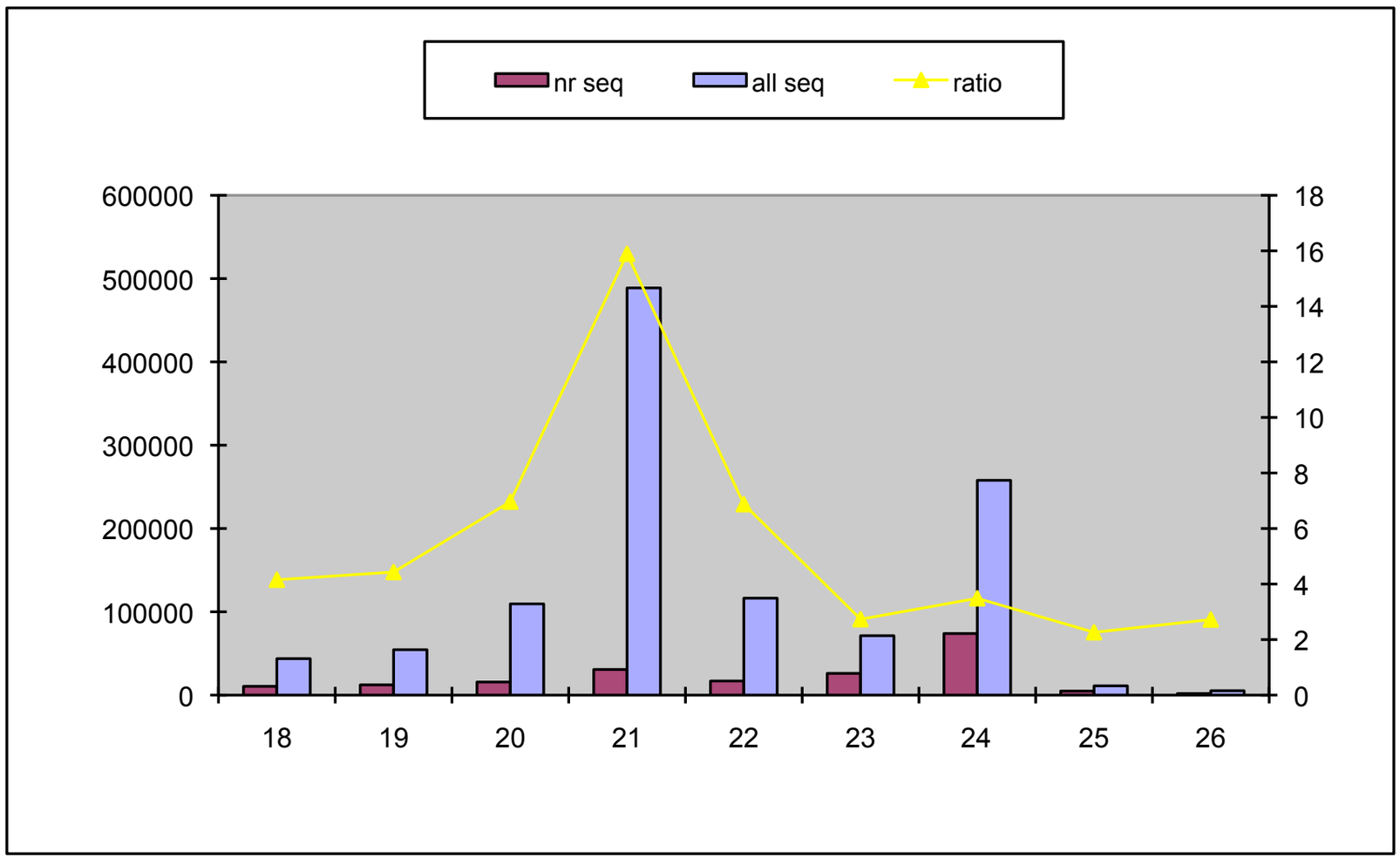
1	153224
2	19430
3	6872
4	3718
5	2443
.	
,	
,	
47048	1
68067	1
76800	1
220414	1

Bioinformatics of miRNAs

- Read processing
  - Small RNAs have specific sizes

Bioinformatics of miRNAs

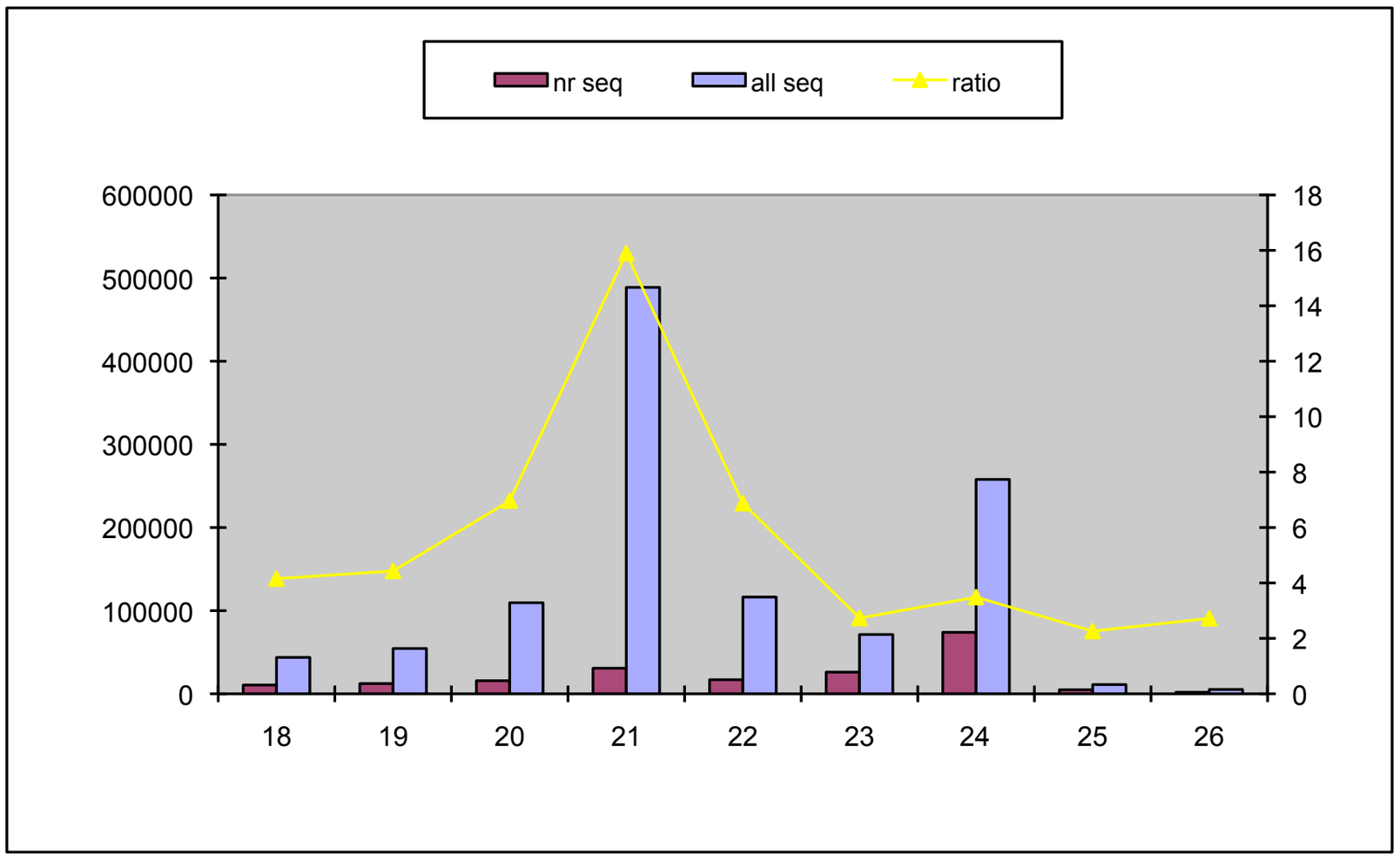
- Read processing
- Small RNAs have specific sizes



Bioinformatics of miRNAs

Read processing

- Small RNAs have specific sizes

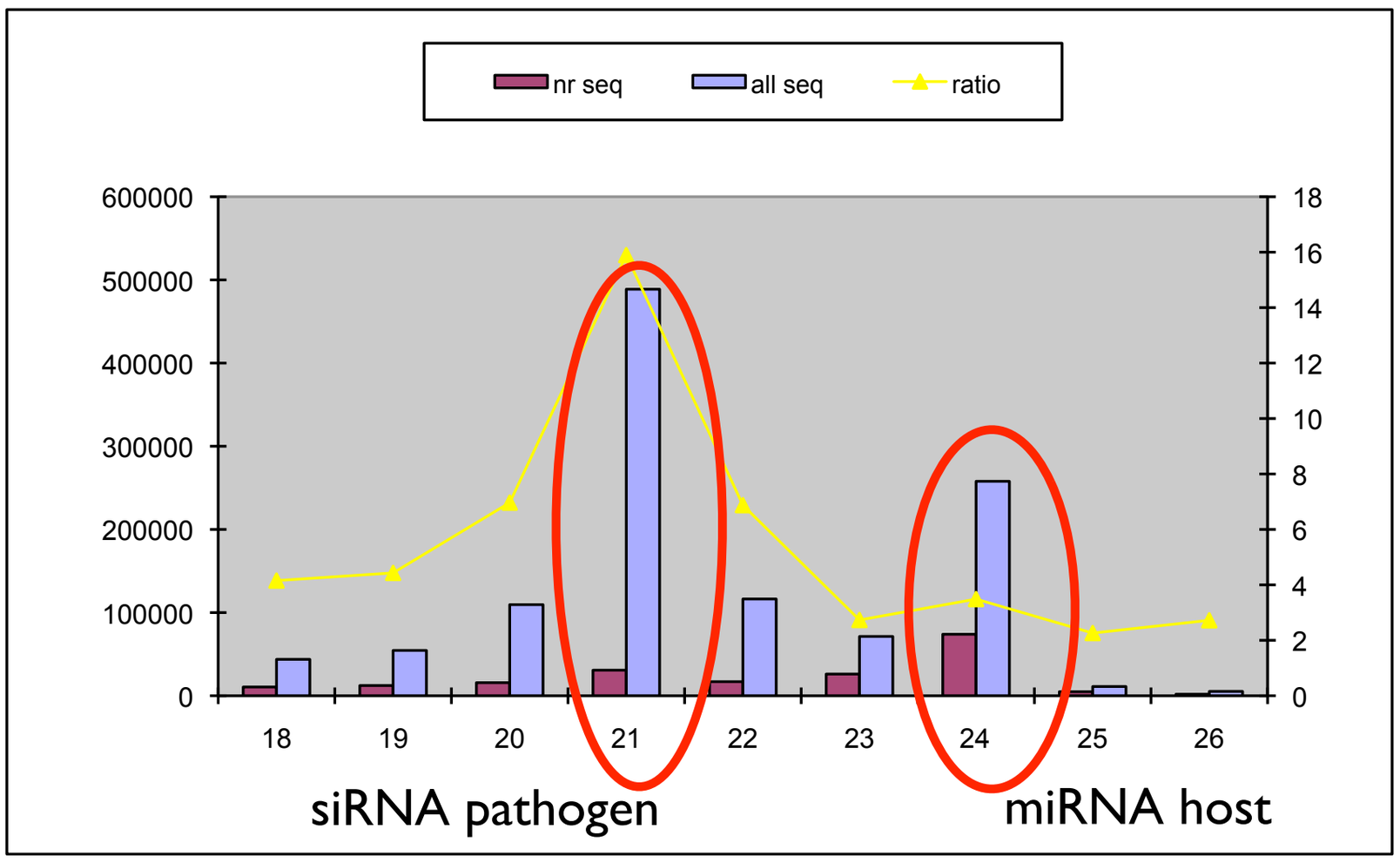


size	nr seq	all seq
18	10543	43795
19	12301	54504
20	15727	109552
21	30757	488808
22	16927	116372
23	26094	71308
24	73943	257841
25	4912	11109
26	1936	5269
total	193140	1158558

Bioinformatics of miRNAs

Read processing

- Small RNAs have specific sizes



size	nr seq	all seq
18	10543	43795
19	12301	54504
20	15727	109552
21	30757	488808
22	16927	116372
23	26094	71308
24	73943	257841
25	4912	11109
26	1936	5269
total	193140	1158558

Bioinformatics of miRNAs

Target search

- Cleavage of mRNA requires a high degree of miRNA:target base-pairing however allowing bulges and gaps.

- Diffucult problem, particularly in animals, where the degree of miRNA:target complementarity is limited

- Based on stability of miRNA:target hybrid including empirical rules of hybridisation

Tools and Databases

miRNA databases

miRBase	miRBase database is a searchable database of published miRNA sequences and annotation.	database	http://www.mirbase.org/
deepBase	deepBase is a database for annotating and discovering small and long ncRNAs (microRNAs, siRNAs, piRNAs...) from high-throughput deep sequencing data.	database	http://deepbase.sysu.edu.cn/
microRNA.org	microRNA.org is a ddatabase for Experimentally observed microRNA expression patterns and predicted microRNA targets & target downregulation scores.	database	http://www.microrna.org/microrna/getExprForm.do
miRGen 2.0	miRGen 2.0: a database of microRNA genomic information and regulation	database	http://www.microrna.gr/mirgen/
miRNAMap	miRNAMap: genomic maps of microRNA genes and their target genes in mammalian genomes	database	http://mirnamap.mbc.nctu.edu.tw/
PMRD	PMRD: plant microRNA database	database	http://bioinformatics.cau.edu.cn/
Rfam	The Rfam database is a collection of RNA families, each represented by multiple sequence alignments, consensus secondary structures and covariance models (CMs).	database	http://rfam.sanger.ac.uk/

Tools and Databases

miRNA detecting tools

mirDeep2	miRDeep2 is a completely overhauled tool which discovers microRNA genes by analyzing sequenced RNAs. The tool reports known and hundreds of novel microRNAs with high accuracy in seven species representing the major animal clades.	tool	http://www.mdc-berlin.de/en/research/research_teams/systems_biology_of_gene_regulatory_elements/projects/miRDeep/index.html
miRcheck	MIRcheck is a perl script designed to identify RNA sequences with secondary structures similar to plant miRNAs.	tool	http://web.wi.mit.edu/bartel/pub/software.html
miRanalyzer	A microRNA detection and analysis tool for next-generation sequencing experiments	tool, webserver	http://web.bioinformatics.cicbiogune.es/microRNA/miRanalyser.php
UEA slRNA toolkit	This site provides access to our software tools for the analysis of high-throughput small RNA data. This is the plant-specific version of the site. We provide versions for animal and plant datasets.	tool, webserver	http://srna-tools.cmp.uea.ac.uk/

Tools and Databases

miRNA target databases

targetScan	targetScan is Search for predicted microRNA targets in animals	database, webserver	http://www.targetscan.org/
StarBase	starBase is a database for exploring microRNA-target interaction maps from Argonaute (Ago) CLIP-Seq (HITS-CLIP) and degradome sequencing (Degradome-Seq, PARE) data.	database	http://starbase.sysu.edu.cn
TarBase	A comprehensive database of experimentally supported animal microRNA targets	database	http://diana.cslab.ece.ntua.gr/tarbase/
miRecords	an integrated resource for microRNA-target interactions.	database	http://mirecords.biolead.org
PicTar	PicTar is Combinatorial microRNA target predictions.	database, webserver, predictions	http://pictar.bio.nyu.edu/

Tools and Databases

miRNA target databases

RepTar	A database of inverse miRNA target predictions, based on the RepTar algorithm that is independent of evolutionary conservation considerations and is not limited to seed pairing sites.	database	http://reptar.ekmd.huji.ac.il/
miRDB	miRDB is an online database for miRNA target prediction and functional annotations in animals.	database	http://mirdb.org/miRDB/
miRGen	miRGen is an integrated database of:a) positional relationships between animal miRNAs and genomic annotation sets, b) animal miRNA targets according to combinations of widely used target prediction programs	database	http://www.diana.pcbi.upenn.edu/miRGen.html
miRNA – Target Gene Prediction at EMBL	This website provides access to our 2003 and 2005 miRNA-Target predictions for Drosophila miRNAs.	database	http://www.russelllab.org/miRNAs/

Tools and Databases

miRNA target prediction tools

miRanda	miRanda is an algorithm for finding genomic targets for microRNAs. This software will be further developed under the open source model, coordinated by Anton Enright and Chris Sander	tool	http://www.microrna.org/microrna/home.do
RNAhybrid	RNAhybrid is a tool for finding the minimum free energy hybridization of a long and a short RNA. The hybridization is performed in a kind of domain mode, ie. the short sequence is hybridized to the best fitting part of the long one.	tool	http://bibiserv.techfak.uni-bielefeld.de/rnahybrid/
PITA	PITA, incorporates the role of target-site accessibility, as determined by base-pairing interactions within the mRNA, in microRNA target recognition.	webserver, predictions	http://genie.weizmann.ac.il/pubs/mir07/mir07_data.html
RNA22	First finds putative microRNA binding sites in the sequence of interest, then identifies the targeted microRNA.	webserver, predictions	http://cbcsrv.watson.ibm.com/rna22.html
Diana-microT	DIANA-microT 3.0 is an algorithm based on several parameters calculated individually for each microRNA and it combines conserved and non-conserved microRNA recognition elements into a final prediction score.	webserver	http://diana.cslab.ece.ntua.gr/microT/

Workflow

Raw Sequencing Data

Adaptor & Barcode Search (step1)

Size filtering & Statistics (step2)

Cleaned Sequence Data (fastq, fasta)

miRNA detection with mirDEEP2

Known & Novel miRNA

Cleaned Sequence Data (fastq, fasta)

Mapping on Genome with BWA

Mapping on ncRNA Reference with BWA

fold-change Calculation

Genome Position & other ncRNA

DATABASE

Genotyping by Sequencing

Andreas Gisel
Inernational Institute of Tropical Agriculture (IITA)
Ibadan, Nigeria

Development

2007 – Complexity Reduction of Polymorphic Sequences van Orsouw et al., PLoS ONE 2(11): e1172. SNP discovery using 454 sequencing, genotyping via Keygene SNPWave –> patent application

2008 – Rapid SNP Discovery and Genetic Mapping Using Sequenced RAD Markers. Baird et al. PLoS ONE 3(10): e3376 Direct SNP genotyping by Illumina sequencing

2009 - High-throughput genotyping by whole-genome resequencing. Huang et al., Genome Res 19:1068–1076. Low-coverage whole genome resequencing of rice RILs

2011 – Multiplex shotgun genotyping for rapid and efficient genetic mapping. Andolfatto et al., Genome Res. 21(4): 610–617 Single restriction enzyme digest, HMM model for data analysis

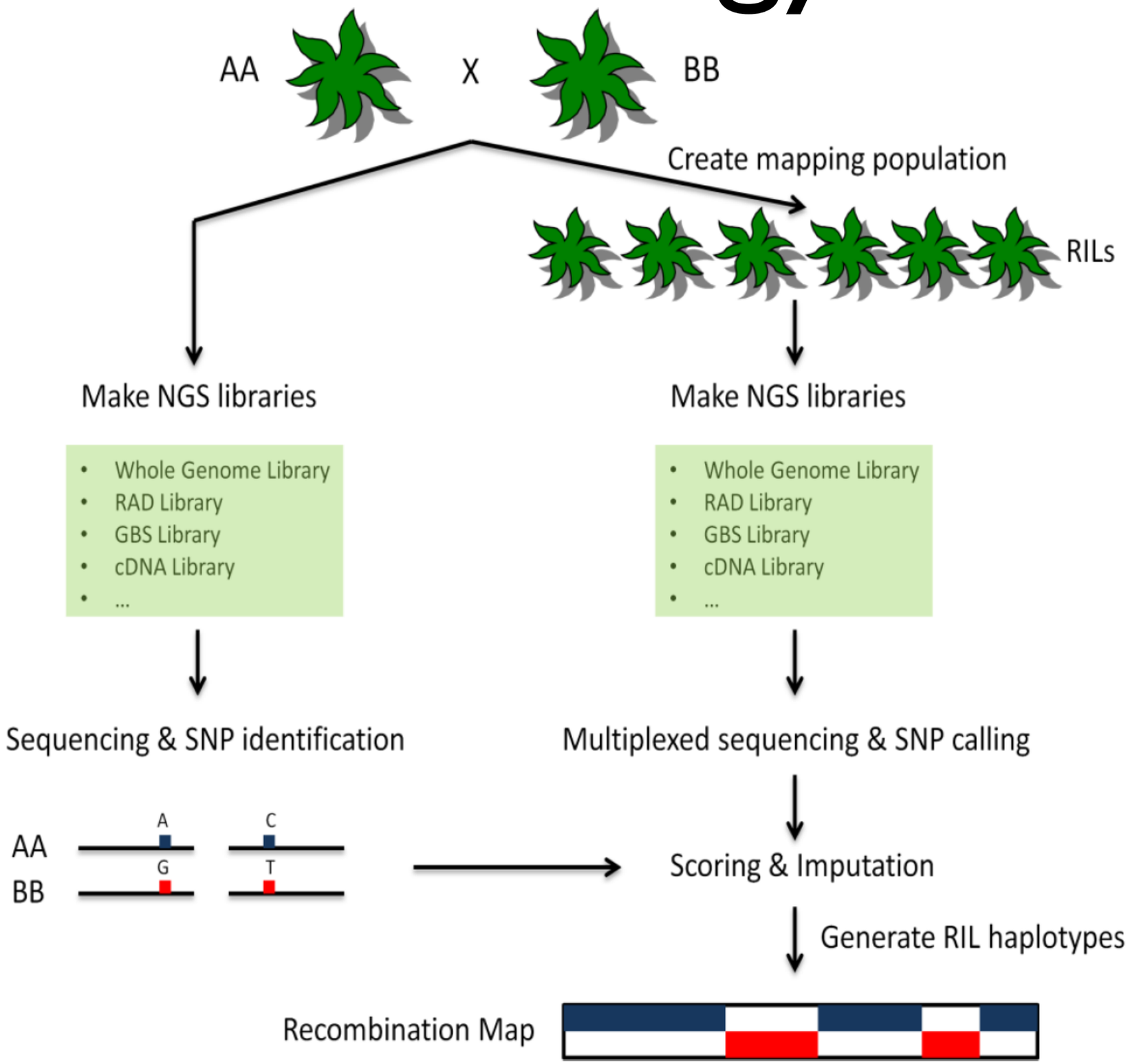
2011 – A Robust, Simple Genotyping-by-Sequencing (GBS) Approach for High Diversity Species. Elshire et al., PLoS ONE 6(5): e19379. Simplified protocol for high throughput

2012 –Development of High-Density Genetic Maps for Barley and Wheat Using a Novel Two-Enzyme Genotyping-by-Sequencing Approach. Poland et al., PLoS ONE 7(2): e32253 Two-enzyme version of Elshire et al protocol

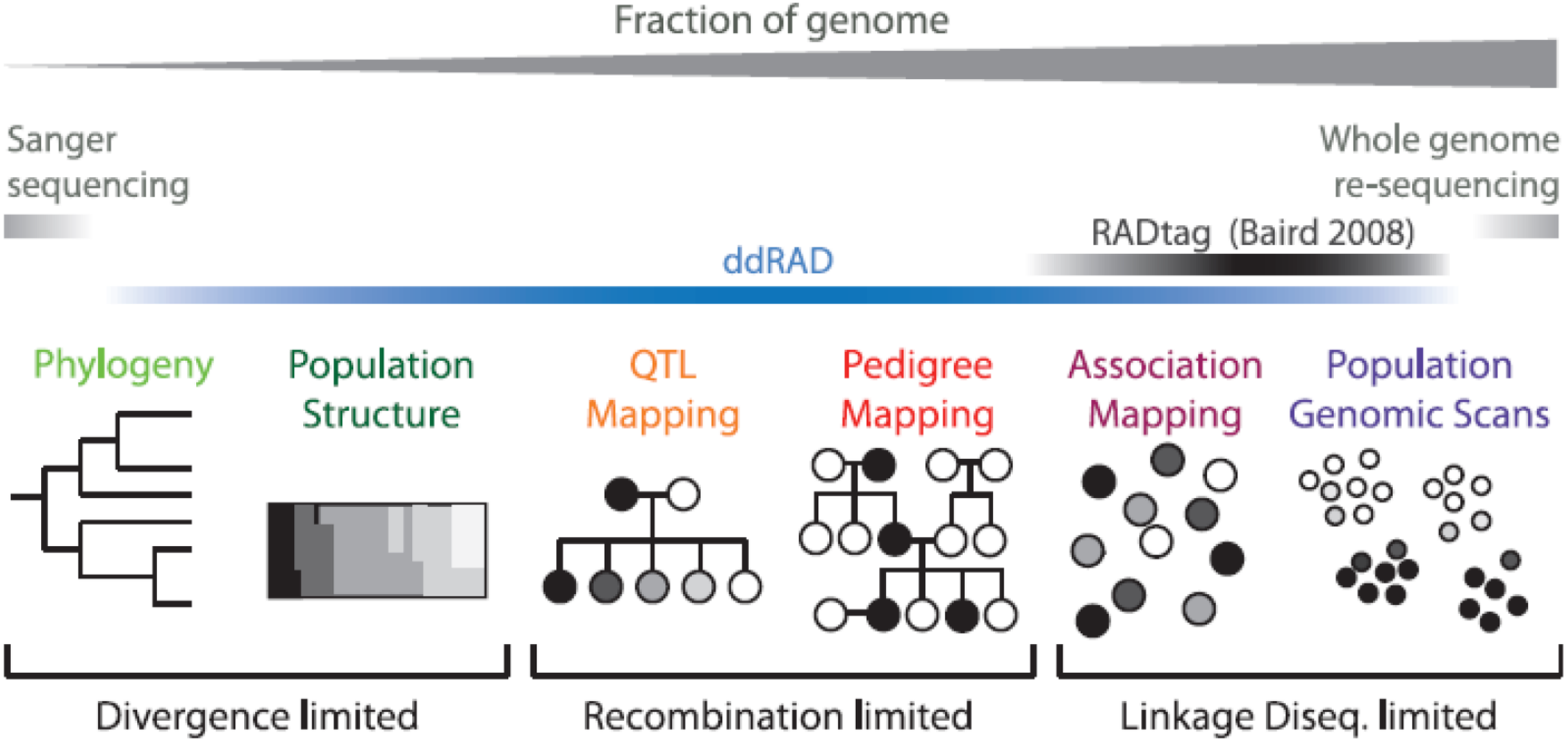
2012 – Double-digest RAD-seq. Peterson et al., PLoS ONE 7(5): e37135 Two-enzyme method similar to Poland et al, with size-selection to increase reproducibility of genotyping

2013 – RESTseq – Efficient Benchtop Population Genomics with RESTriction Fragment SEQuencing. Stolle & Moritz, PLoS ONE 8(5): e63960 Complexity reduction for fewer markers, higher multiplexing, reduced costs

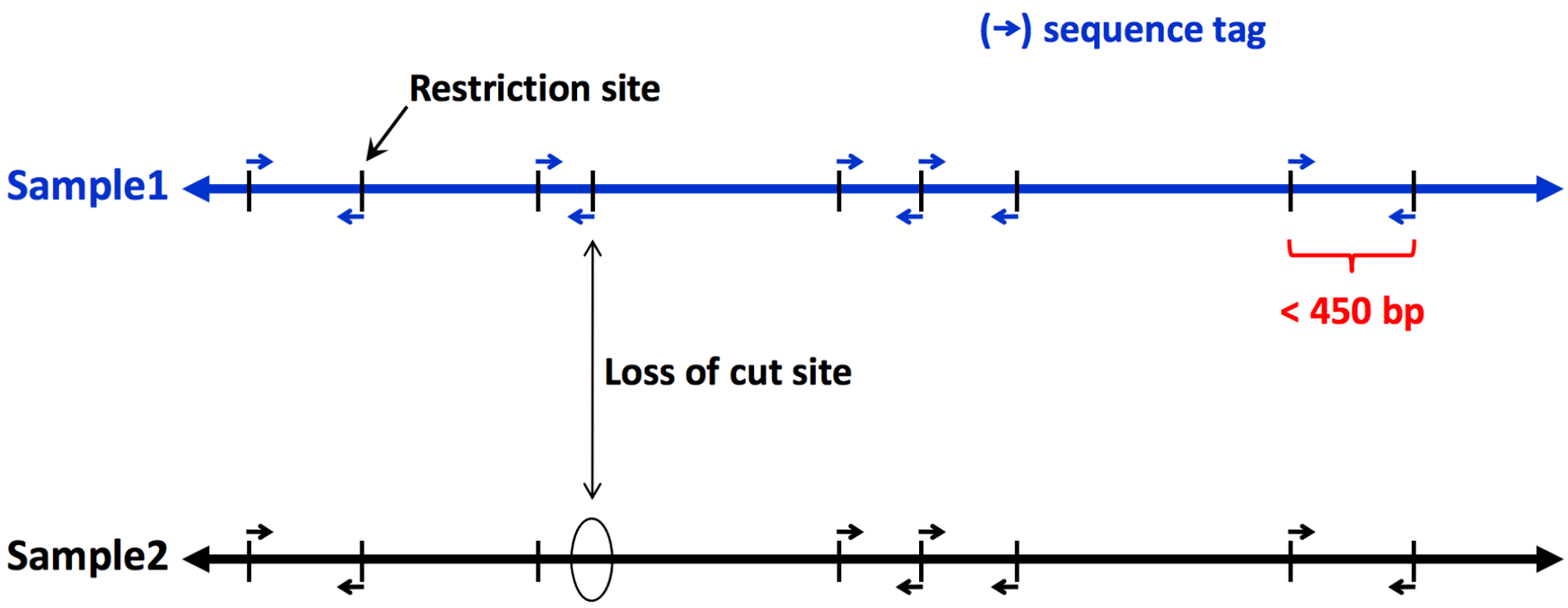
Strategy


AA
X
BB
Create mapping population
RILs
Make NGS libraries
Make NGS libraries
• Whole Genome Library
• RAD Library
• GBS Library
• cDNA Library
• …
• Whole Genome Library
• RAD Library
• GBS Library
• cDNA Library
• …
Sequencing & SNP identification
Multiplexed sequencing & SNP calling
AA
BB
A
C
G
T
Scoring & Imputation
Generate RIL haplotypes
Recombination Map

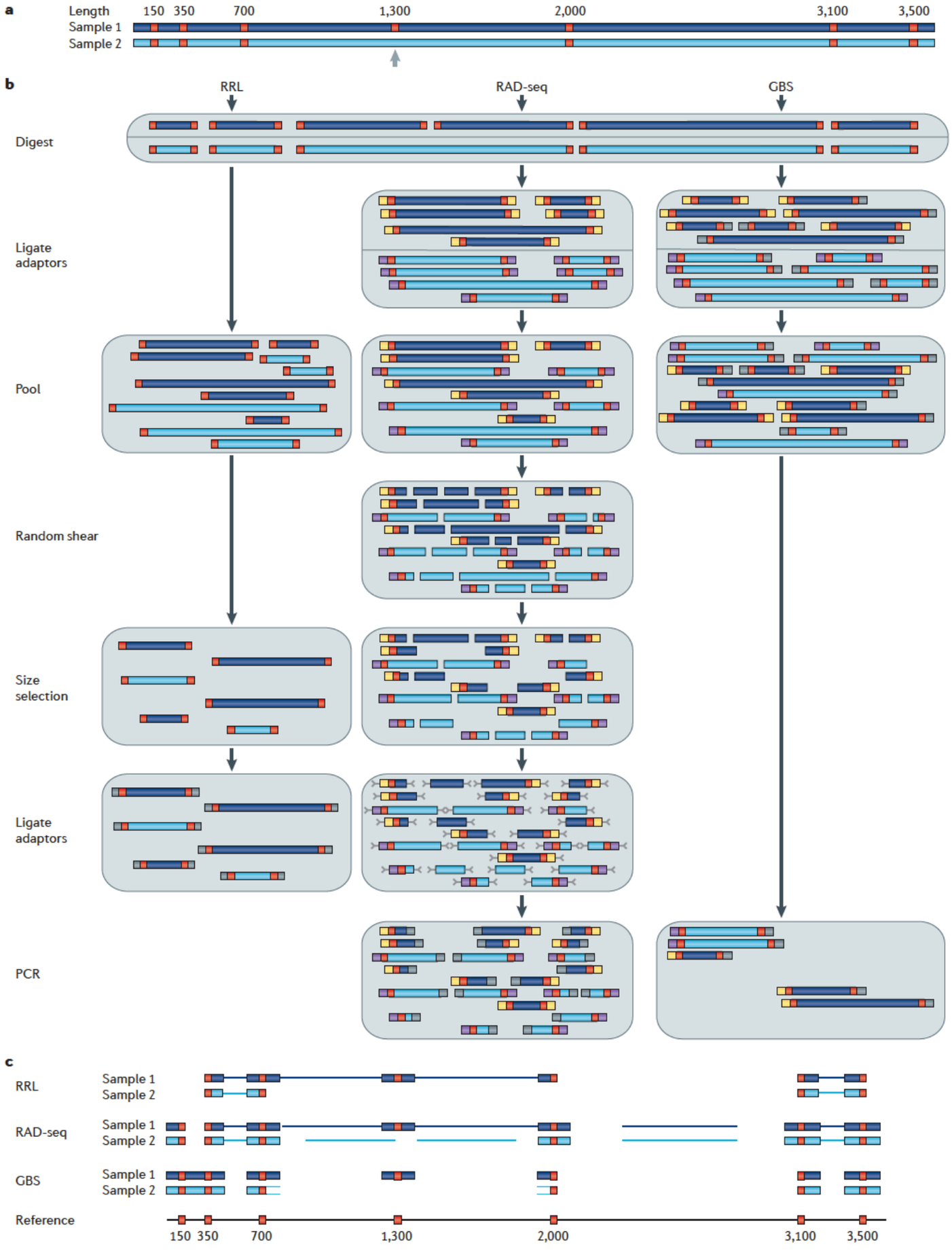
Genome Coverage vs. Sample Size



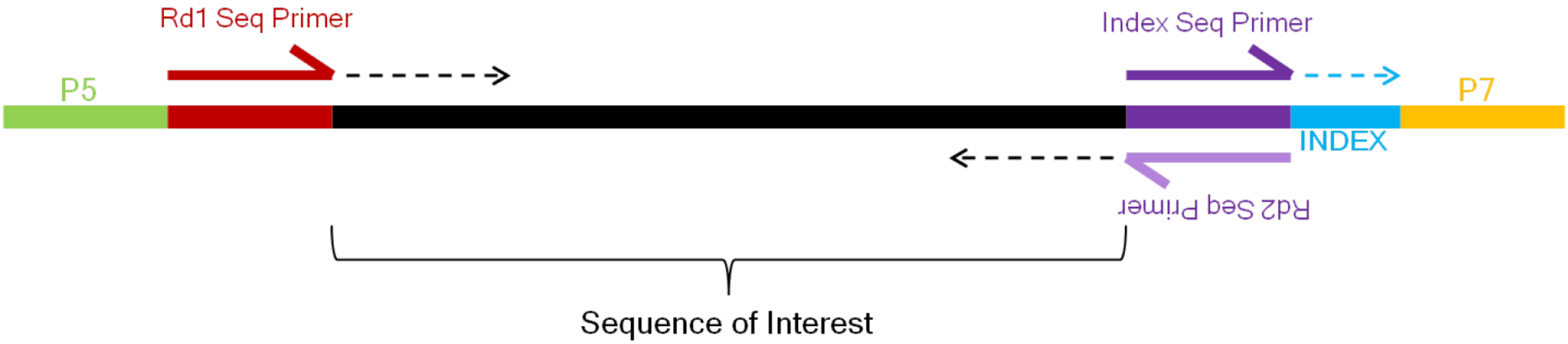
GBS



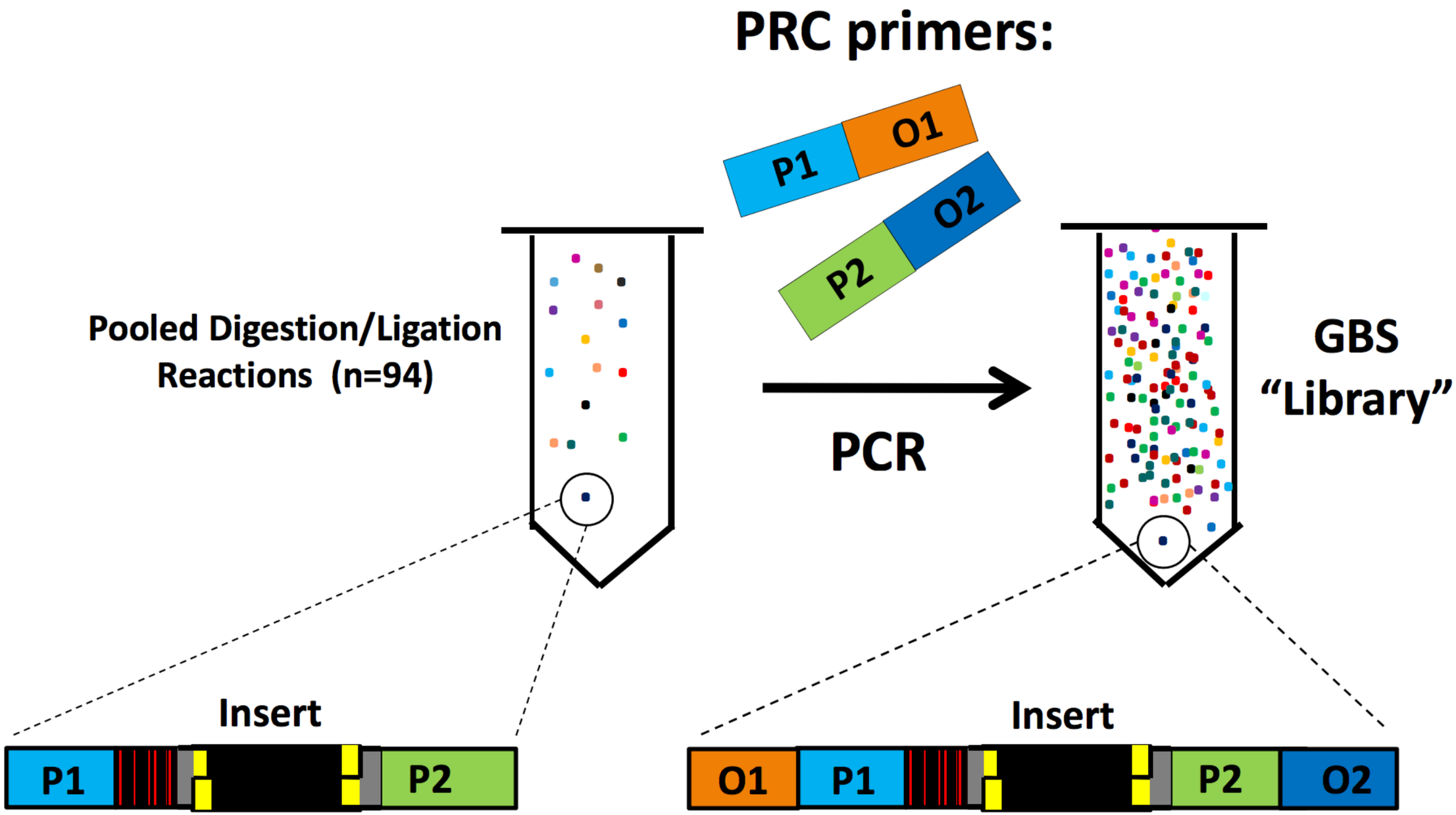
- Focuses NextGen sequencing power to ends of restriction fragments
- Scores both SNPs and presence/absence markers


a
Length
150 350 700 1,300 2,000 3,100 3,500
Sample 1
Sample 2
b
RRL
RAD-seq
GBS
Digest
Ligate adaptors
Pool
Random shear
Size selection
Ligate adaptors
PCR
c
RRL
Sample 1
Sample 2
RAD-seq
Sample 1
Sample 2
GBS
Sample 1
Sample 2
Reference
150 350 700 1,300 2,000 3,100 3,500

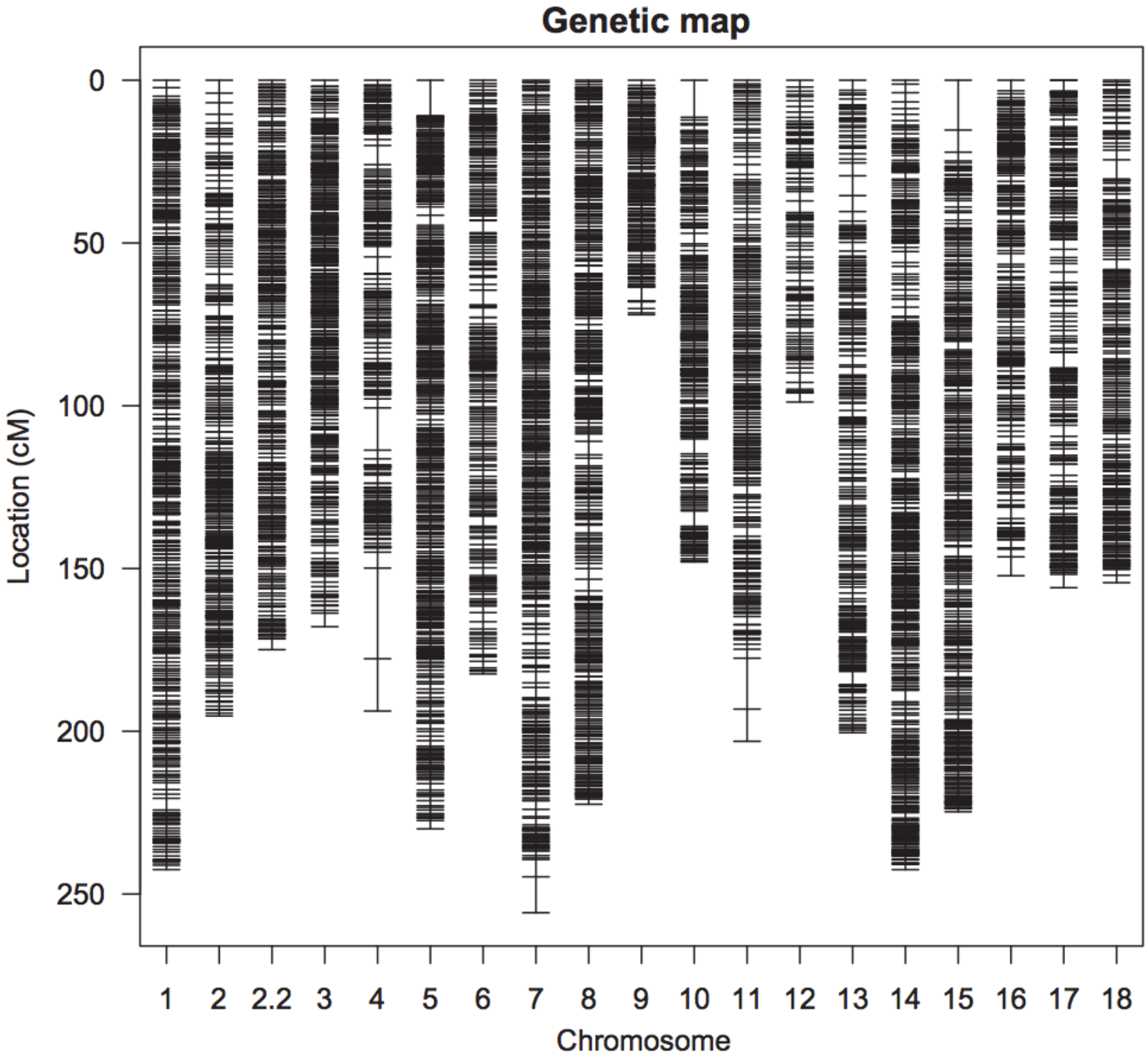
Multiplexing



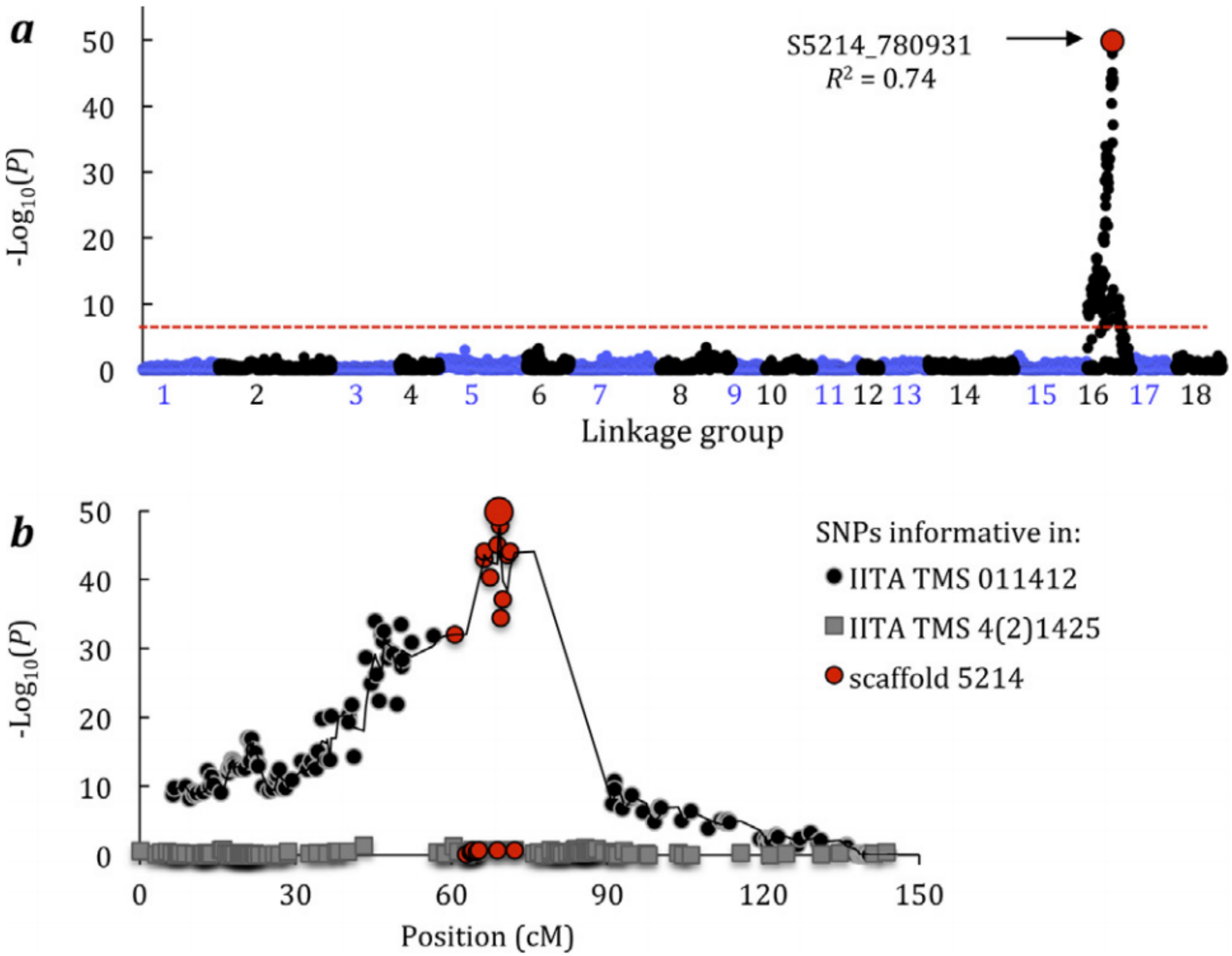
Multiplexing



Genetic Map


Genetic map
Location (cM)
0
50
100
150
200
250
1
2
2.2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
Chromosome

Linkage Map

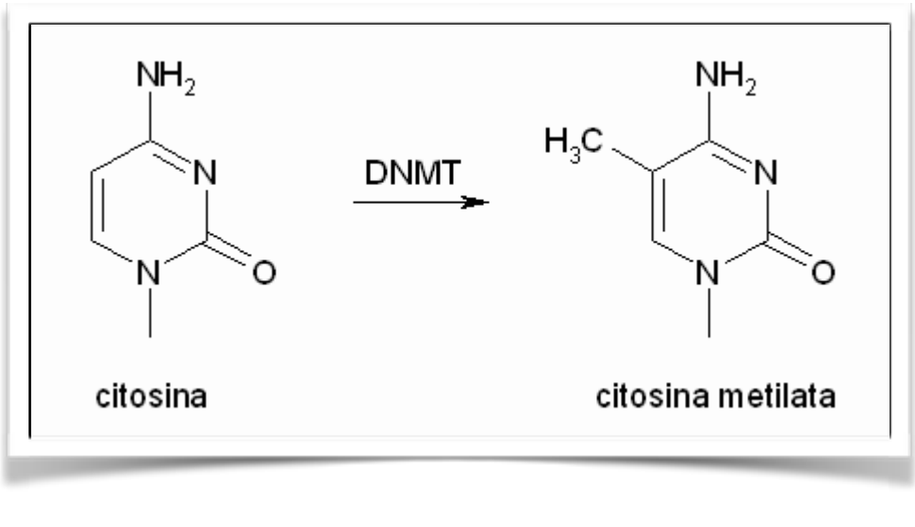

a
S5214_780931
R^2 = 0.74
-Log10(P)
50
40
30
20
10
0
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18
Linkage group
b
SNPs informative in:
IITA TMS 011412
IITA TMS 4(2)1425
scaffold 5214
-Log10(P)
50
40
30
20
10
0
0 30 60 90 120 150
Position (cM)

Methylation profiling

Andreas Gisel
Inernational Institute of Tropical Agriculture (IITA)
Ibadan, Nigeria

IITA Research to Nourish Africa CGIAR

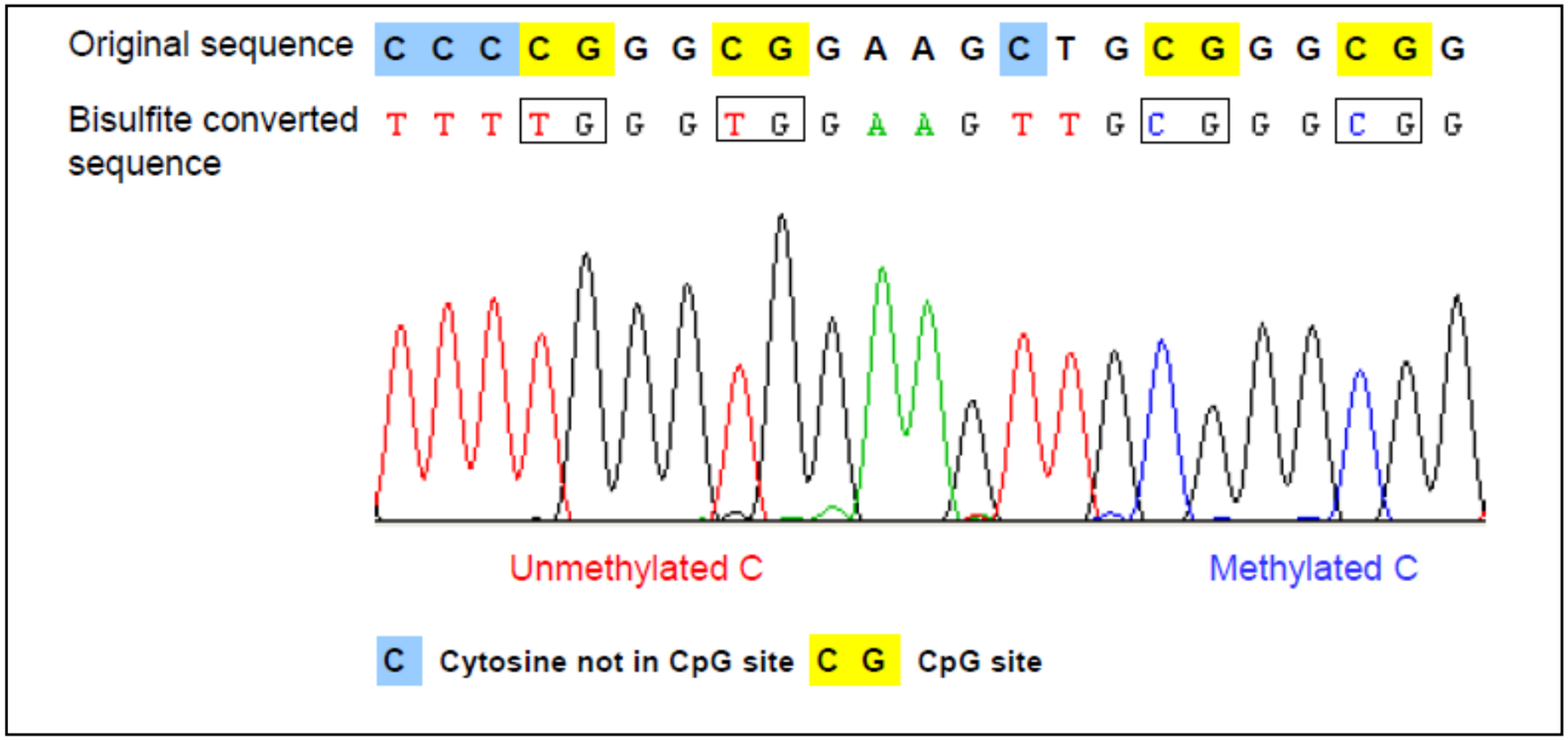
Methylation



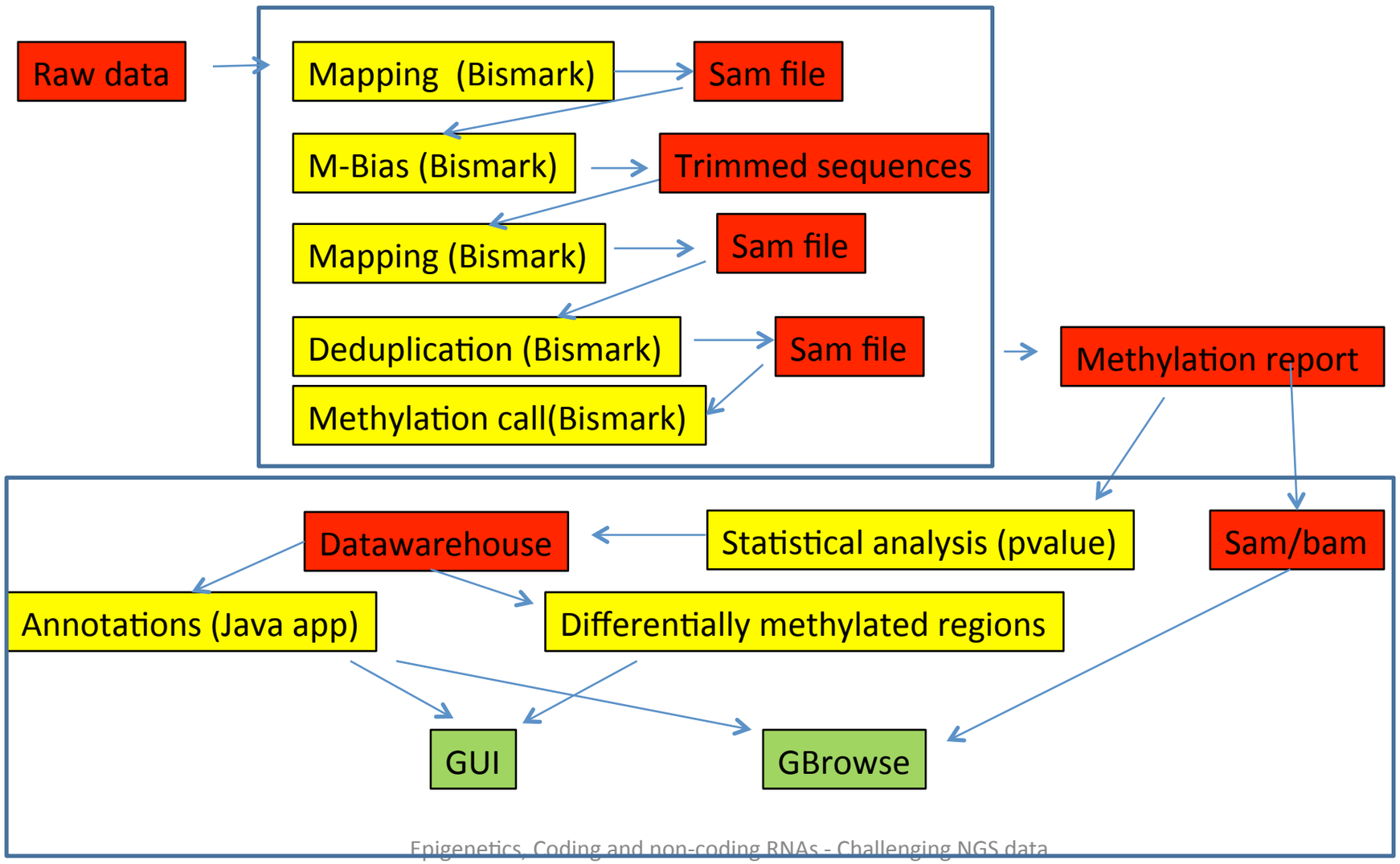
CG, CHG and CHH (H = A, C or T)

MeDIP-Seq - Immuno precipitation
MBD-Seq - methylation specific DNA binding
BS-Seq - Bisulfit conversion

BS-seq

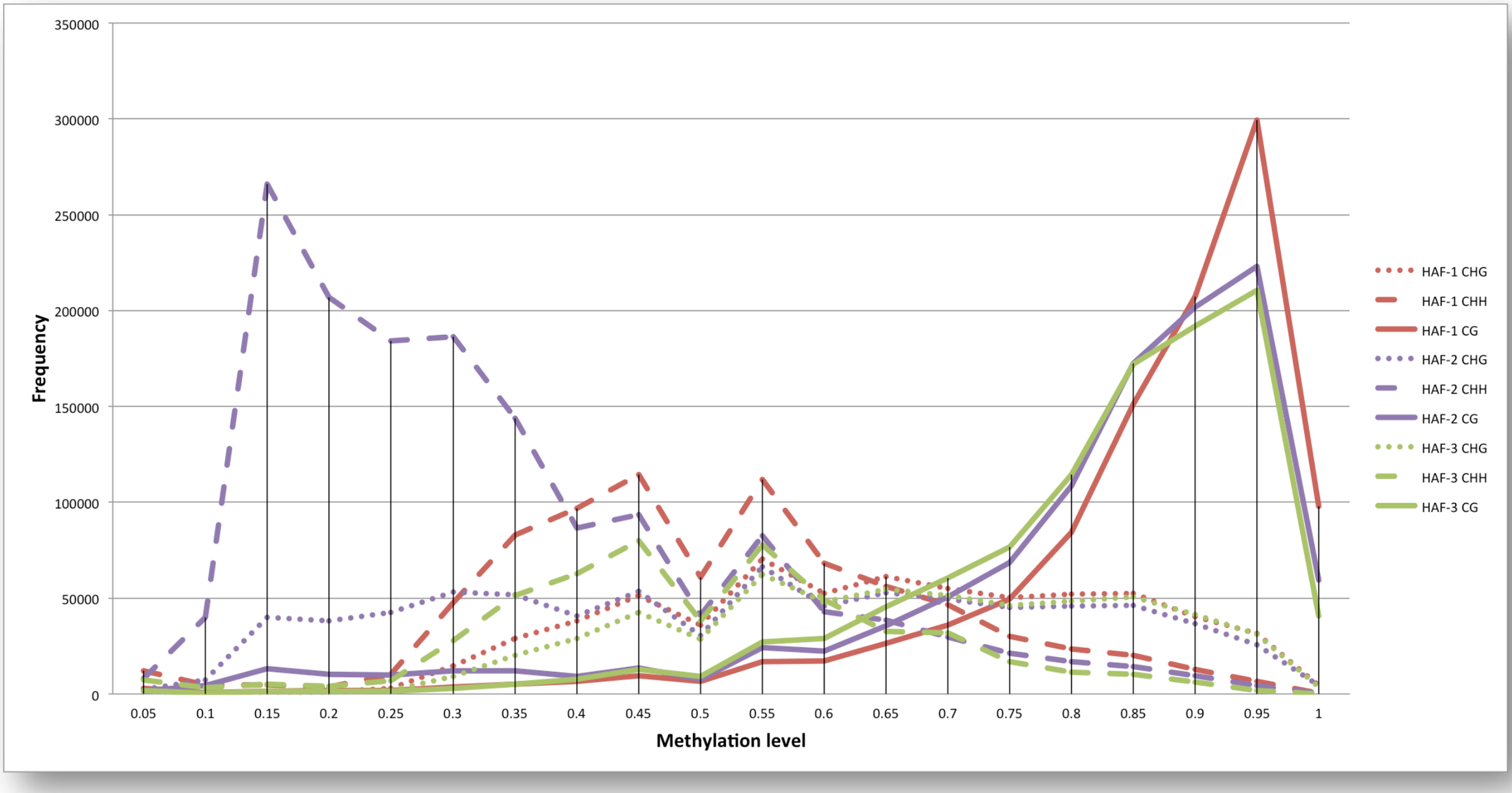

Original sequence C C C C G G G C G G A A G C T G C G G G C G G
Bisulfite converted sequence T T T T G G G T G G A A G T T G C G G G C G G
Unmethylated C
Methylated C
C Cytosine not in CpG site
C G CpG site

Data Analysis Workflow



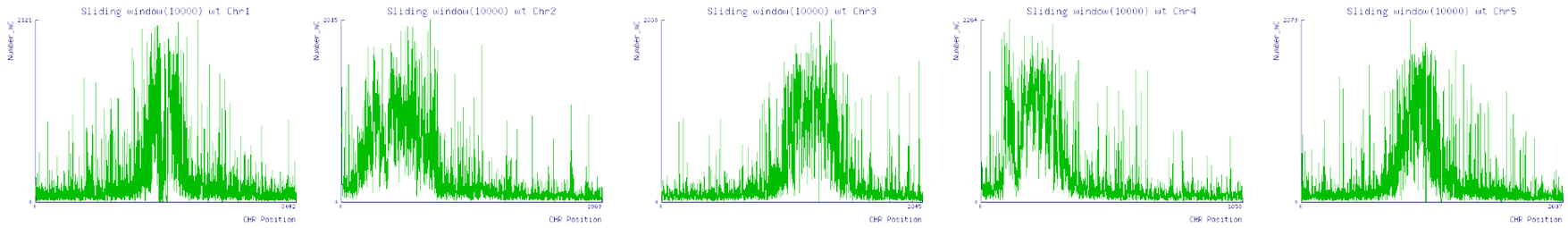
Results

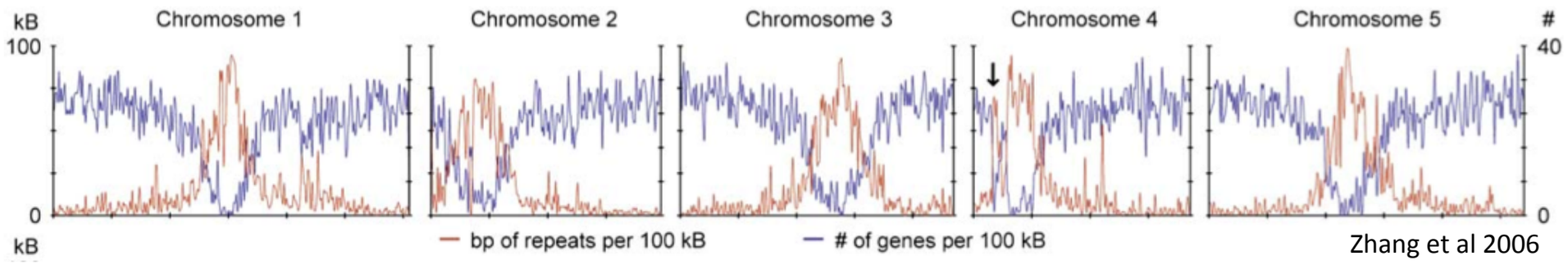
Methylation Level



Results

Methylated regions





Zhang et al 2006