Utilizing the existence of NA12878

Gamze Gürsoy

So far in the Gerstein Lab…

- ENCODEC
- Privaseq3
- EN-TEx

From Personal Genomics to Genome Privacy

Genome of an Individual

- Sequencing, analysis, interpretation
- Soon will become part of medical practice
- NCI: prevent, diagnose, and treat disease through personalized medicine

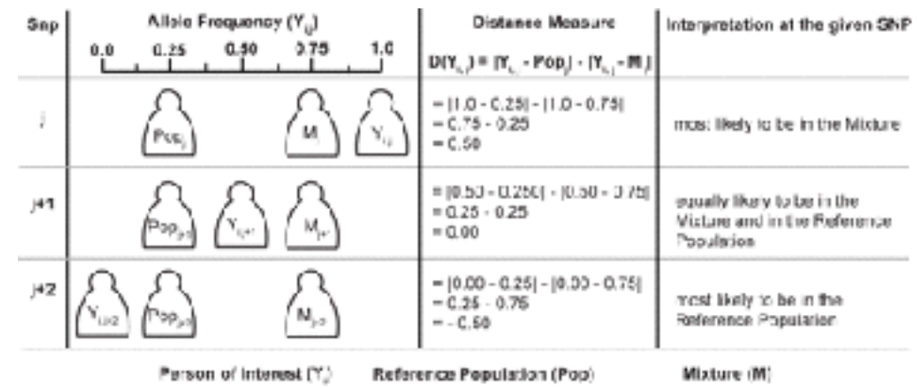
Privacy risk

- Identity tracing
  - Link between unknown genome to a panel of individual through quasi-identifiers
- Attribute Disclosure Attacks
  - Known DNA sample to private data such as HIV status or drug abuse
- Completion Techniques
  - Impute sensitive information from partial genomic data (e.g. bipolar disorder risk)

Image L: http://calorierestrictiondietplan.com/customizing-a-cr-way-lifestyle-for-your-genome/
Image R: https://www.genome.gov/27561246/privacy-in-genomics/

Genome Privacy traditionally focuses on DNA variants

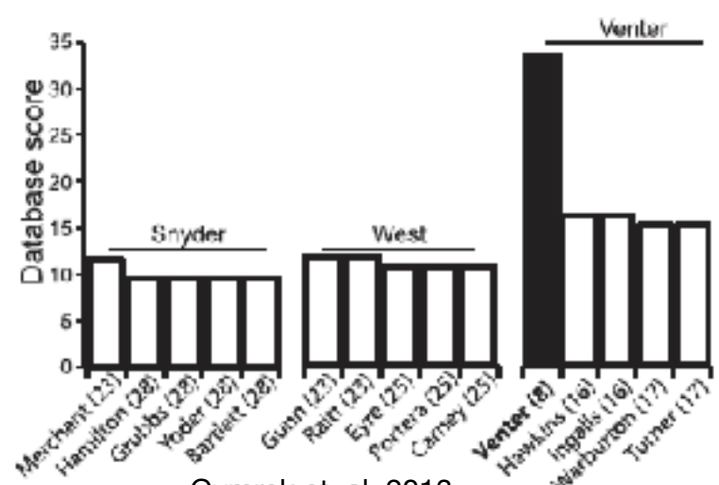
- **Detecting whether an individual with known genotypes in a complex DNA mixture**
  - Homer et. al, 2008
    Distance between genotype and dataset
  - Im et. al, 2012
    Regression coefficients of GWAS summary statistics can reveal person's participation



Homer et. al, 2008

- **Identification attacks by cross-referencing independent datasets**
  - Sweeney at al, 2013
    Cross-reference PGP profile with public voter list data
  - Gymrek et al, 2013
    Cross-reference Y-STRs with recreational genetic genealogy database

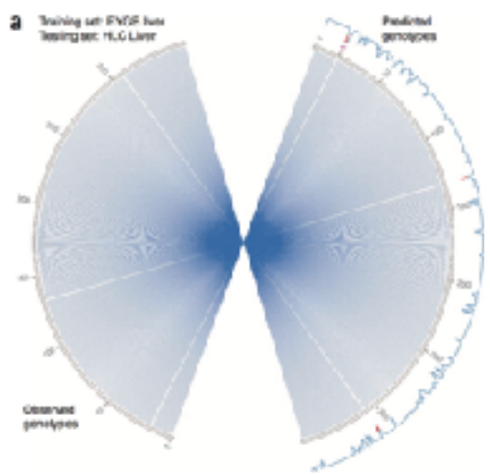
Sweeney et. al, 2013



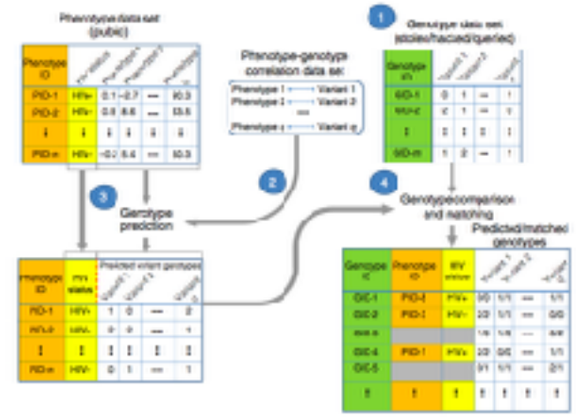
Gymrek et. al, 2013

Functional genomics era increases the number of quasi-identifiers

- **RNA-Seq is of particular interest**
  - Big consortia like ENCODE, TCGA, GTEx provide a wealth of functional genomics data, which particularly belong to individuals
  - Schadt et. al, 2012
    - SNP genotypes can be predicted from RNA-Seq expression data using known eQTLs
  - Harmanci and Gerstein, 2016
    - eQTLs and extreme expression levels can be used to do linking attacks



Schadt et. al, 2012

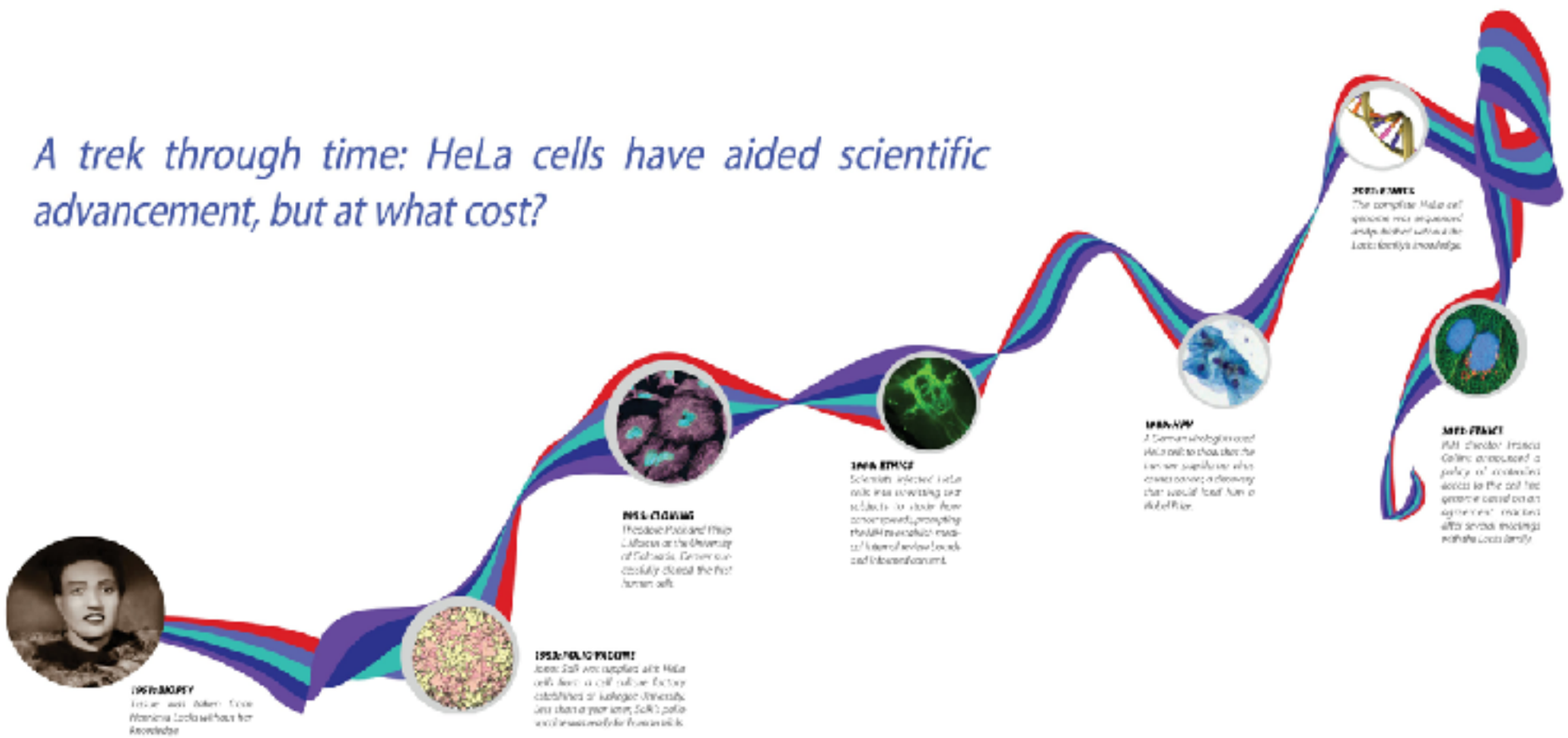


Harmanci and Gerstein, 2016

- **ChIP-Seq and Hi-C signal tracks can also leak genotype information**
  - Harmanci and Gerstein, 2017

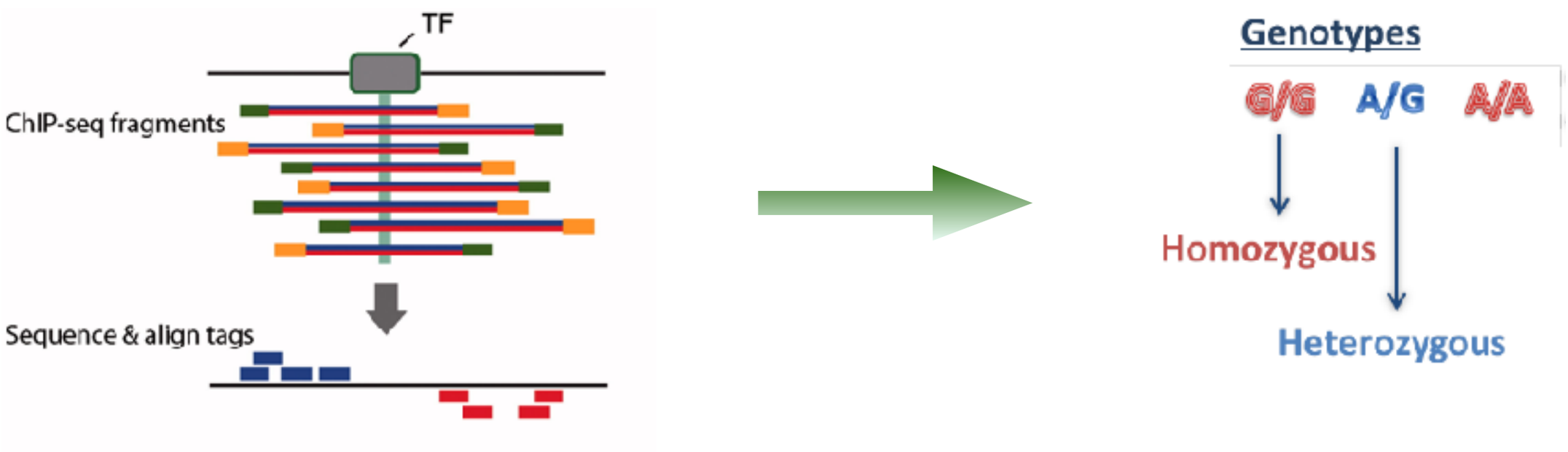
Functional genomics era attacks focus on phenotype-genotype relationship

- BUT, nobody is talking about the *"elephant in the room"*
- All the functional genomics data comes with a great deal of sequencing data
- How much information, for example, RNA-Seq reads or ChIP-Seq reads contain?
- Is that information enough to identify individuals?
- Is it safe to share the fastq/bam files from these experiments?



HeLa genome is locked, but we have access to its ChIP-Seq reads!

Image: http://berkeleysciencereview.com/article/good-bad-hela/



Private information leakage in functional genomics data

Quantification and Linking

Datasets

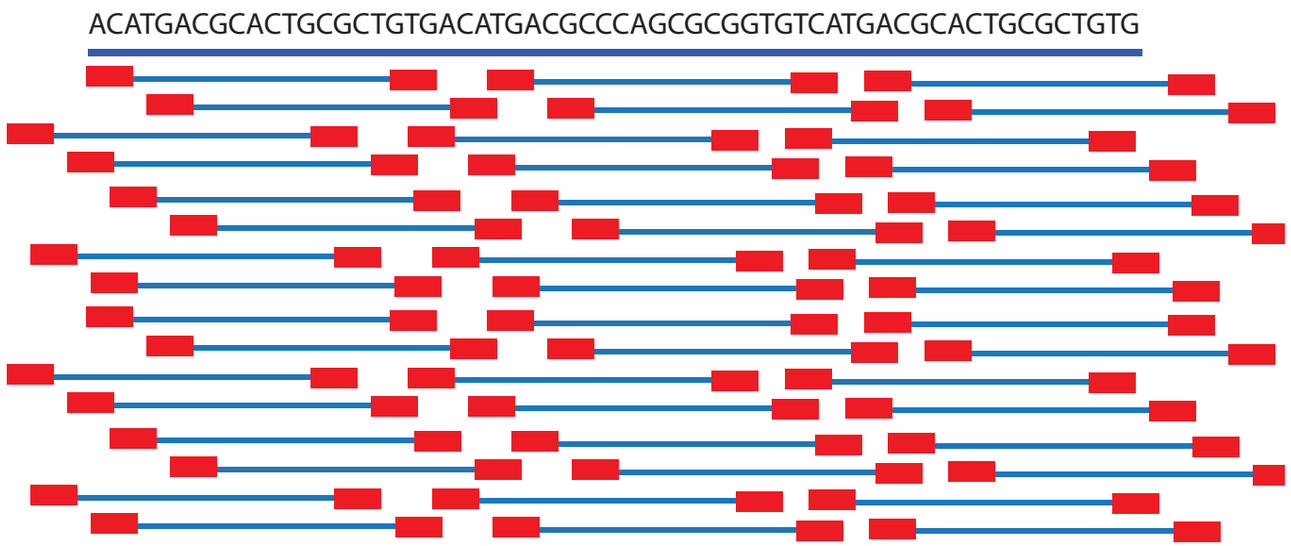
Individual: NA12878
Gold Standard: 1000 Genome genotypes
Control: WGS, # of reads= 757,704,193, read length = 250 bp

Experiment	# of Reads	Read Length
Hi-C exp 1 PE1	219,616,072	101
Hi-C exp 1 PE2	220,087,882	101
Hi-C exp 2 PE1	448,843,710	101
Hi-C exp 2 PE2	451,088,484	101
Hi-C exp 3 PE1	536,684,803	101
Hi-C exp 3 PE2	536,101,709	101
RNA-Seq	227,501,266	202

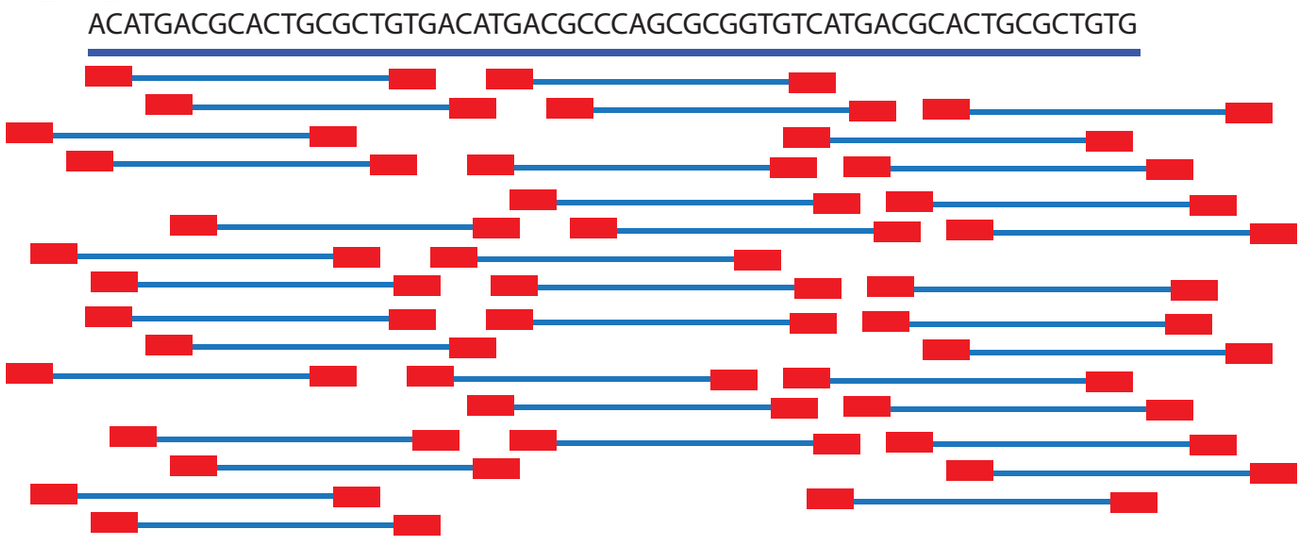
ChIP-Seq

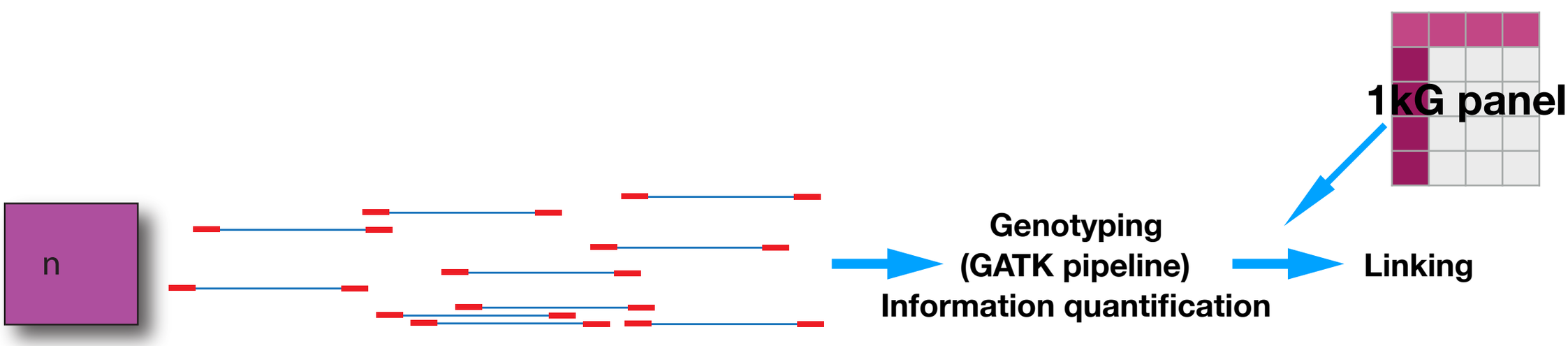
Experiment	# of Reads	Read Length
H3K4me1	42,763,056	36
HDGF	41,626,373	101
RELB	25,652,682	101
CTCF-Snyder	25,463,397	36
H3K4me3	20,221,959	36
JUND	18,701,295	36
H3K79me2	16,073,184	36
H3K36me3	15,239,685	51
H2AFZ	14,724,790	36
H3K9me3	14,049,420	36
CTCF-Broad	11,026,086	51
rnap2	10,428,778	36
H3K27ac	10,410,928	51
H3K4me2	9,815,194	51
H4K20me1	9,757,368	51
H3K27me3	8,454,639	51
H3K9ac	7,981,456	51
CTCF-Iyer	7,614,943	35
rnap2	7,516,461	36
PBX3	6,119,046	36

Approach

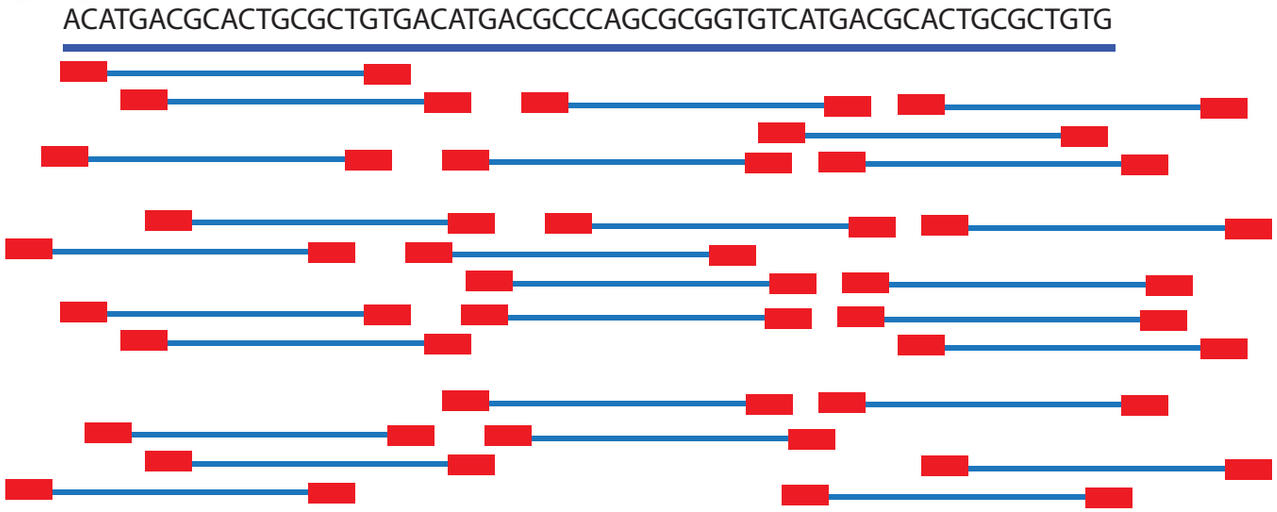

ACATGACGCACTGCGCTGTGACATGACGCCCAGCGCGGTGTCATGACGCACTGCGCTGTG

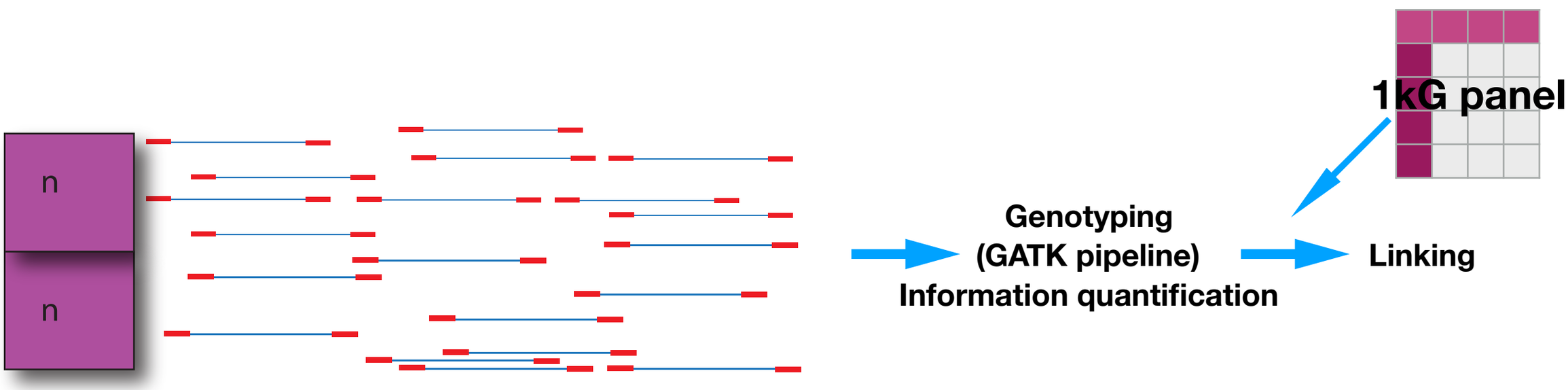
Approach


ACATGACGCACTGCGCTGTGACATGACGCCCAGCGCGGTGTCATGACGCACTGCGCTGTG

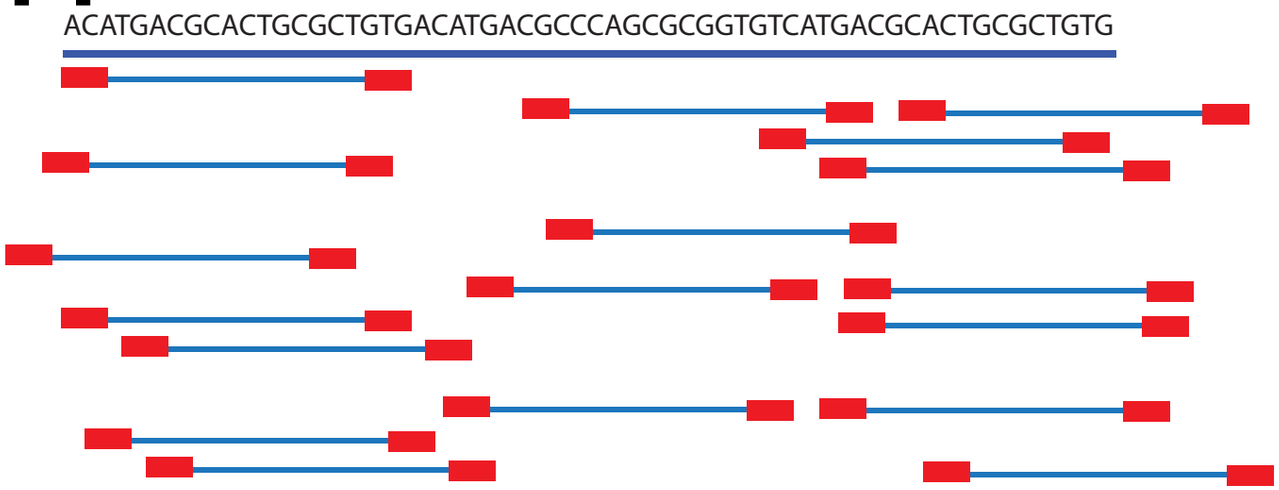

n
Genotyping
(GATK pipeline)
Information quantification
1kG panel
Linking

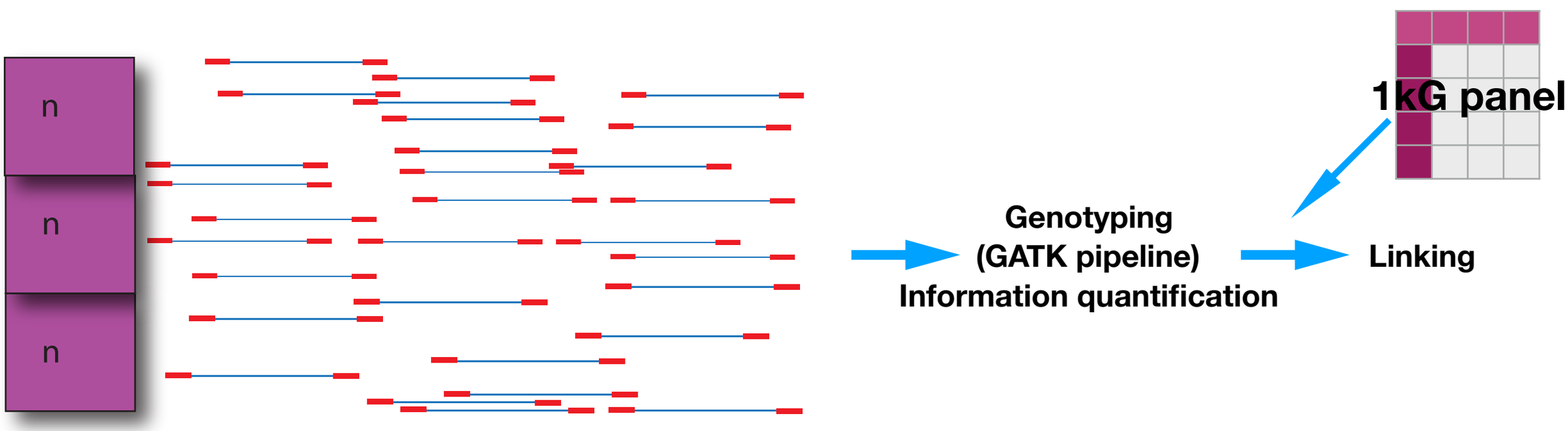
Approach


ACATGACGCACTGCGCTGTGACATGACGCCCAGCGCGGTGTCATGACGCACTGCGCTGTG

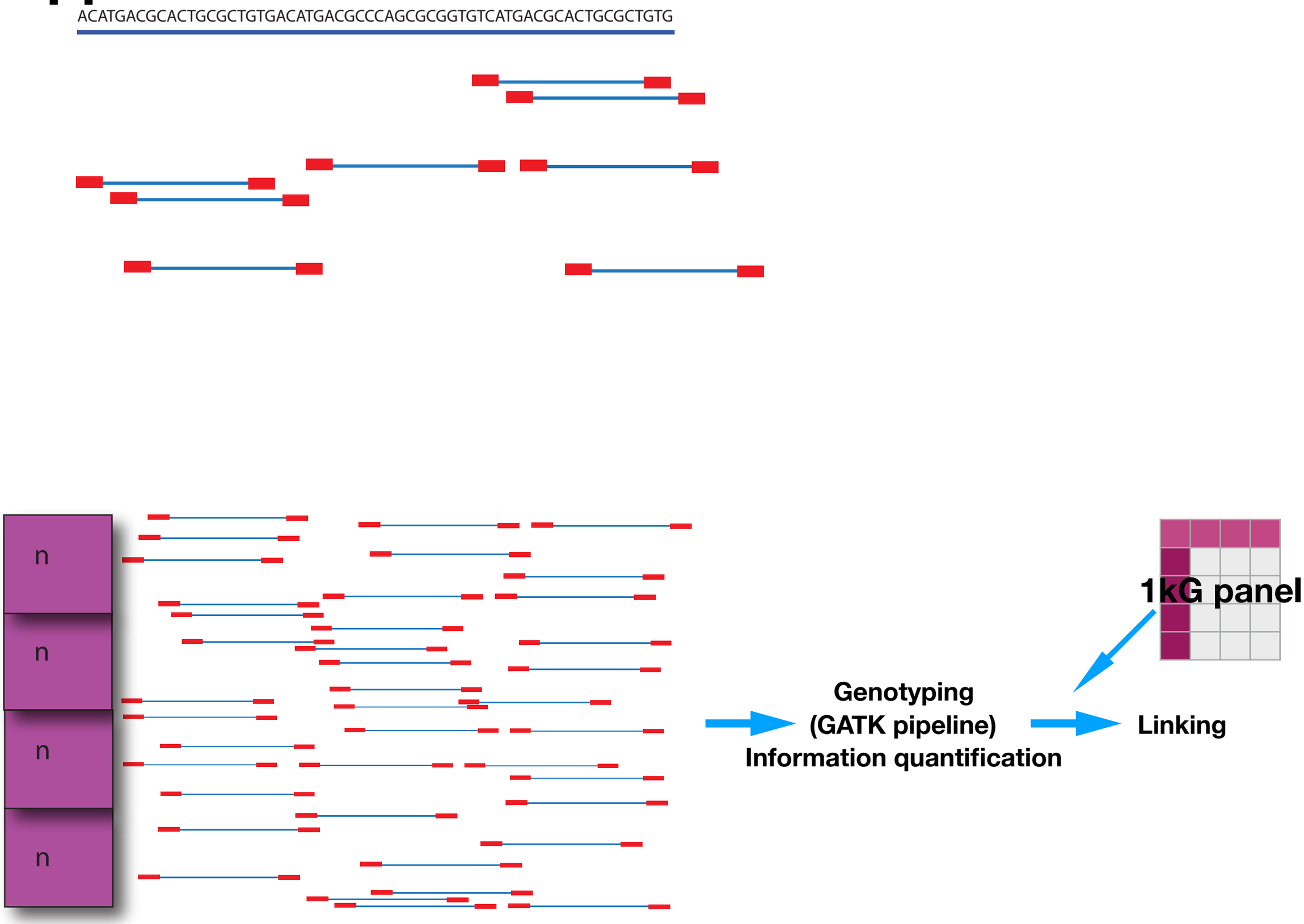

n
n
Genotyping
(GATK pipeline)
Information quantification
1kG panel
Linking

Approach


ACATGACGCACTGCGCTGTGACATGACGCCCAGCGCGGTGTCATGACGCACTGCGCTGTG


n
n
n
Genotyping
(GATK pipeline)
Information quantification
Linking
1kG panel

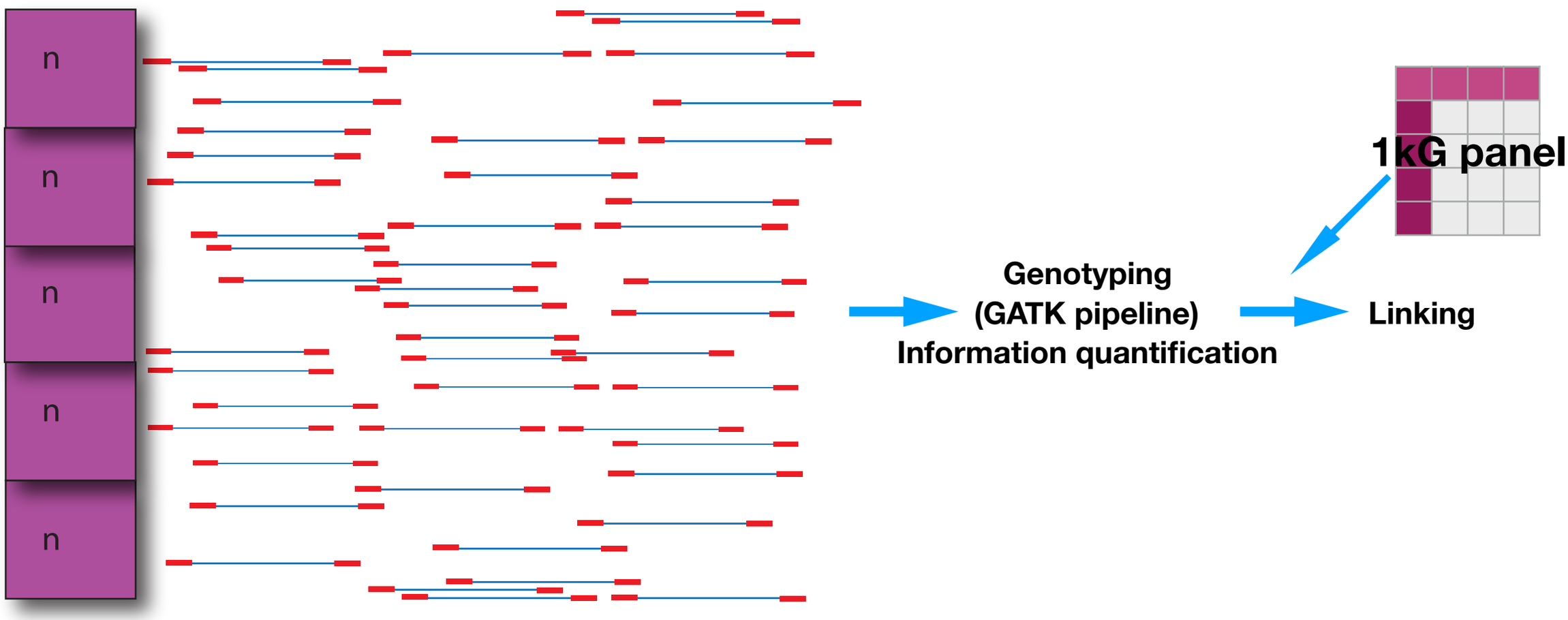
Approach


ACATGACGCACTGCGCTGTGACATGACGCCCAGCGCGGTGTCATGACGCACTGCGCTGTG
n
n
n
n
Genotyping
(GATK pipeline)
Information quantification
Linking
1kG panel

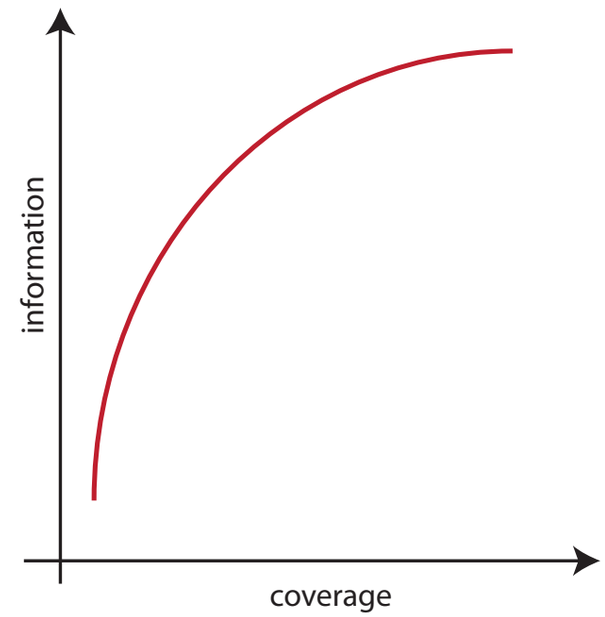
Approach

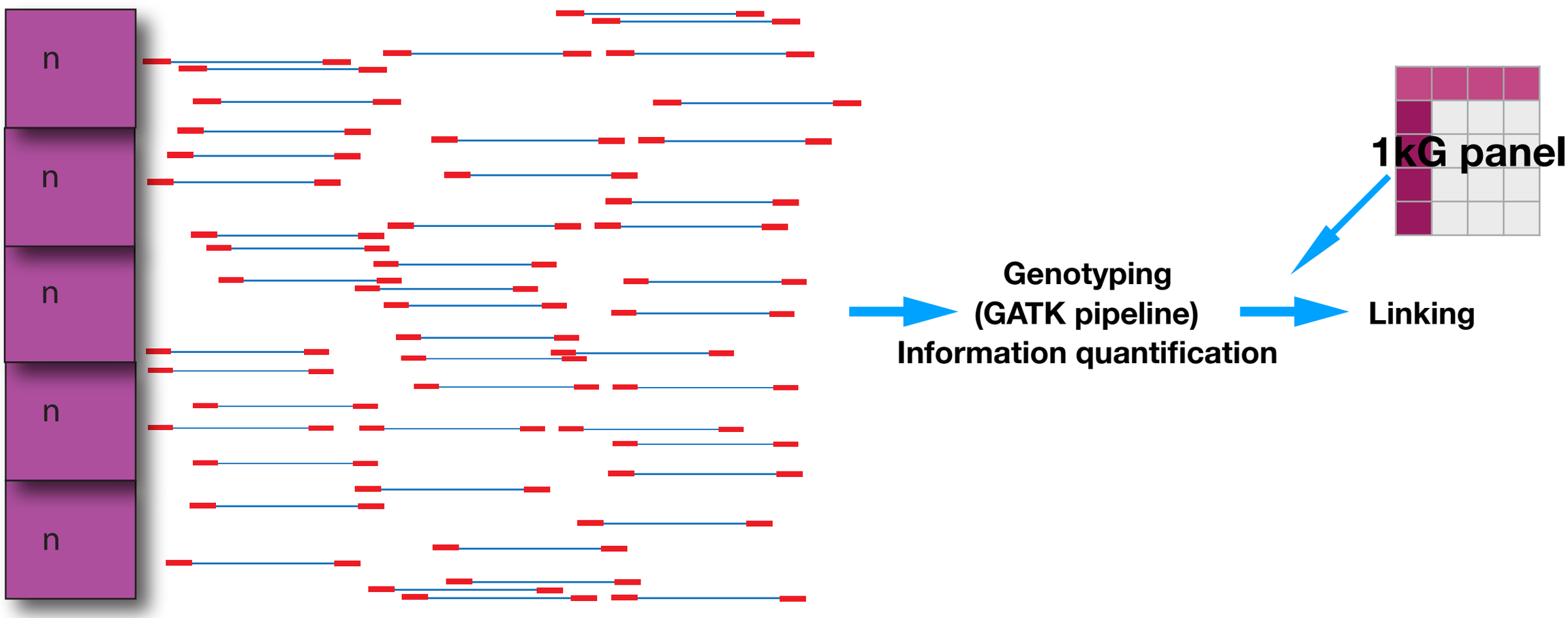
ACATGACGCACTGCGCTGTGACATGACGCCCAGCGCGGTGTCATGACGCACTGCGCTGTG

5n = total number of reads in the experiment

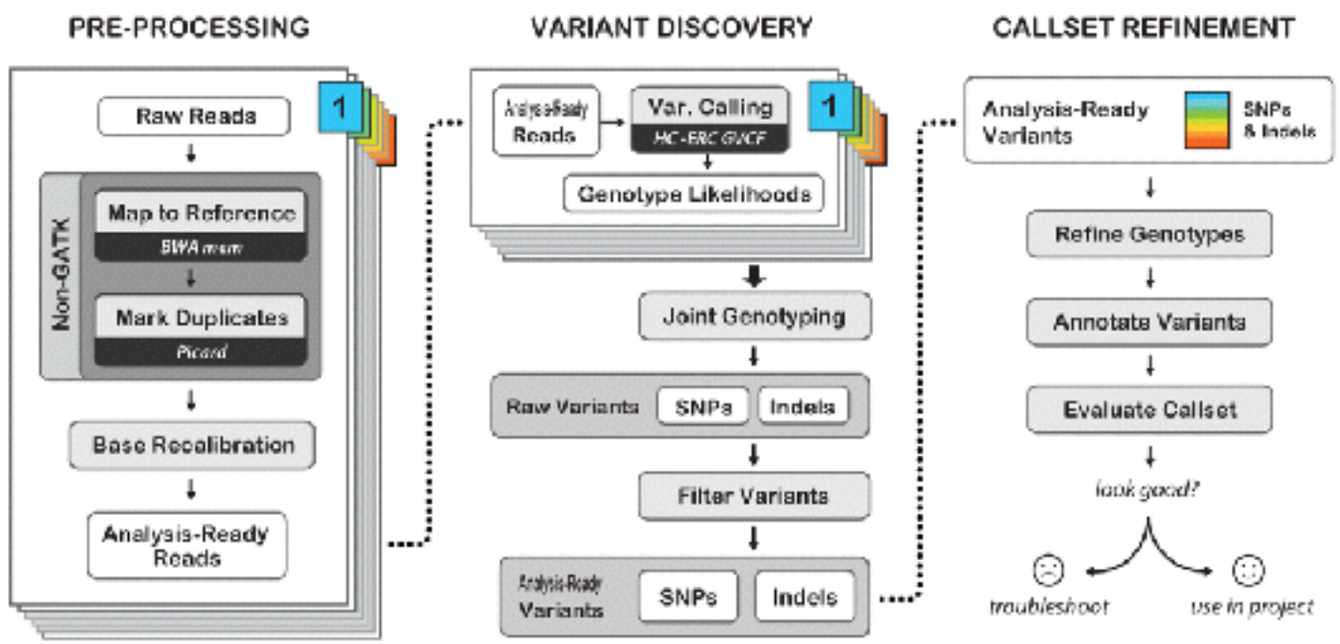


Approach


information
coverage

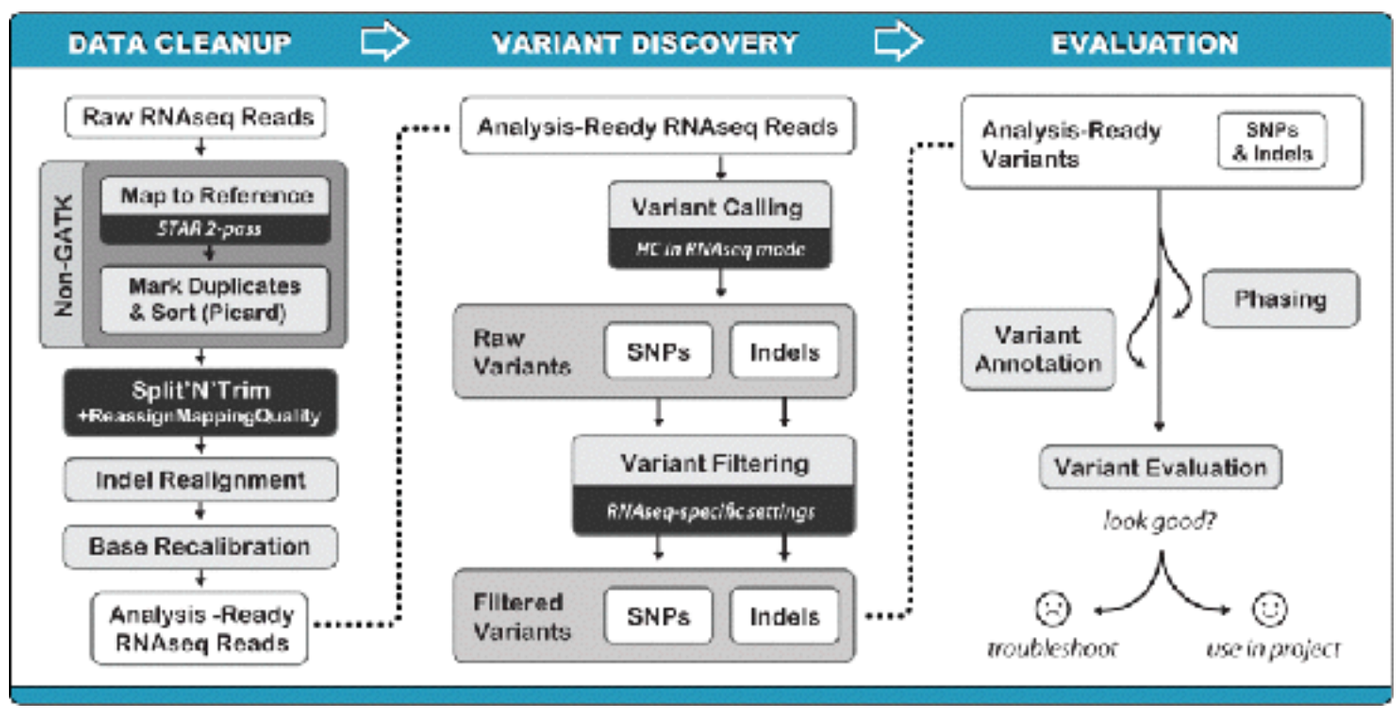

n
n
n
n
n
Genotyping
(GATK pipeline)
Information quantification
1kG panel
Linking

For WGS/Hi-C/ChIP-Seq Analysis



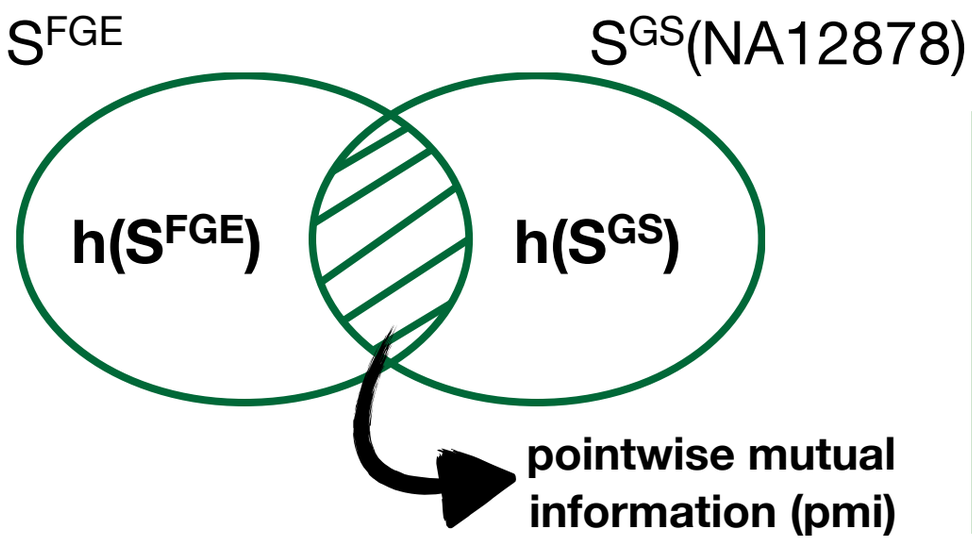
Best Practices for Germline SNPs and Indels in Whole Genomes and Exomes - June 2016

For RNA-Seq



What is “information”?

- Let S^{GS}(NA12878) be the set of SNVs determined by 1k genome (gold standard)
- Let S^{FGE} be the noisy set of SNVs called using the reads from any functional genomic experiment



$$S = \{s_1, ..., s_i, ..., s_N\}$$

$$h(S) = \sum_{i=1}^{i=N} -log(p(s_i))$$

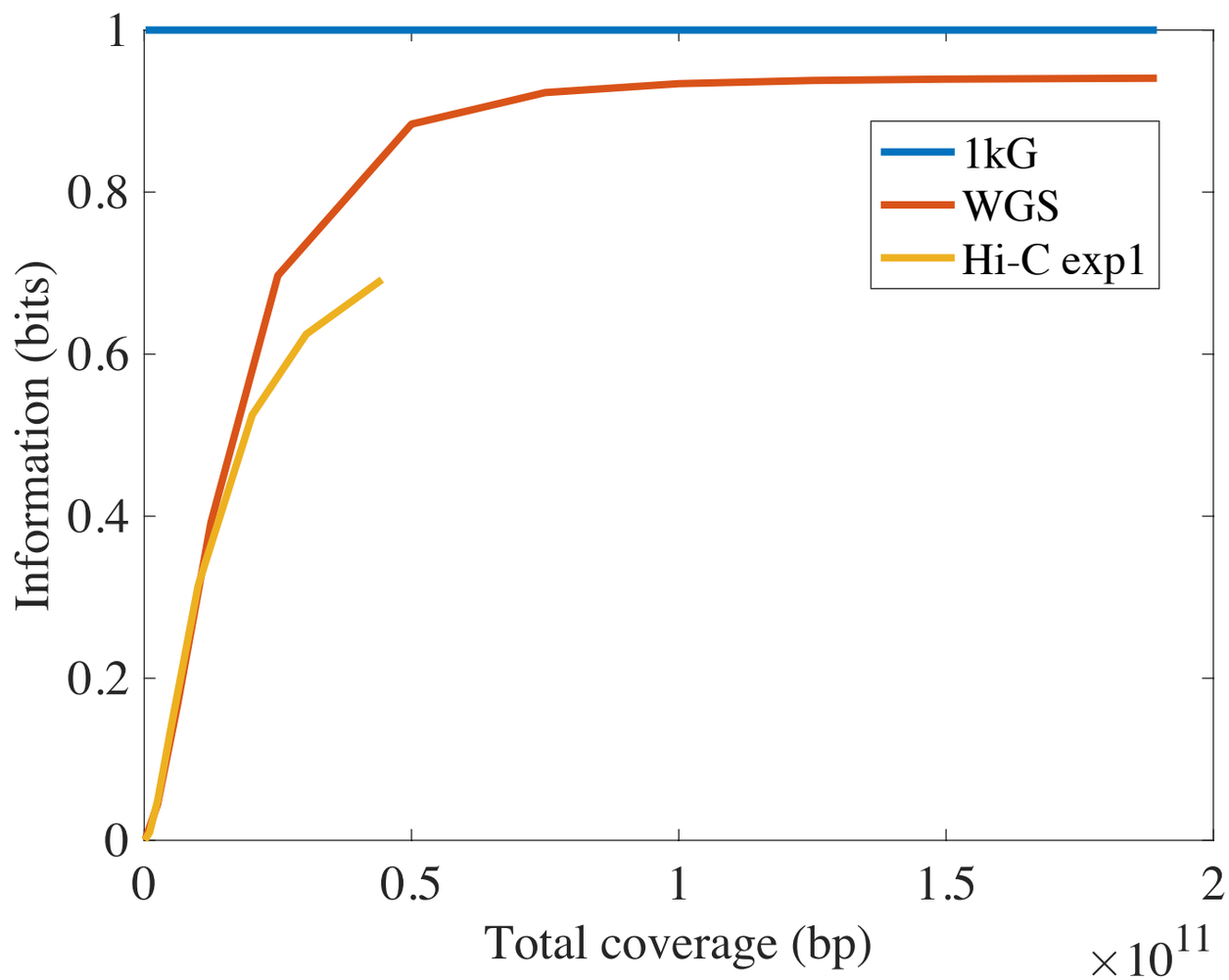
$$p(s_i) = \frac{f(s_i)}{n_T}$$

$f(s_i)$: number of individuals with SNV s_i
n_T : total number of individuals in the population

Further normalization with the gold standard

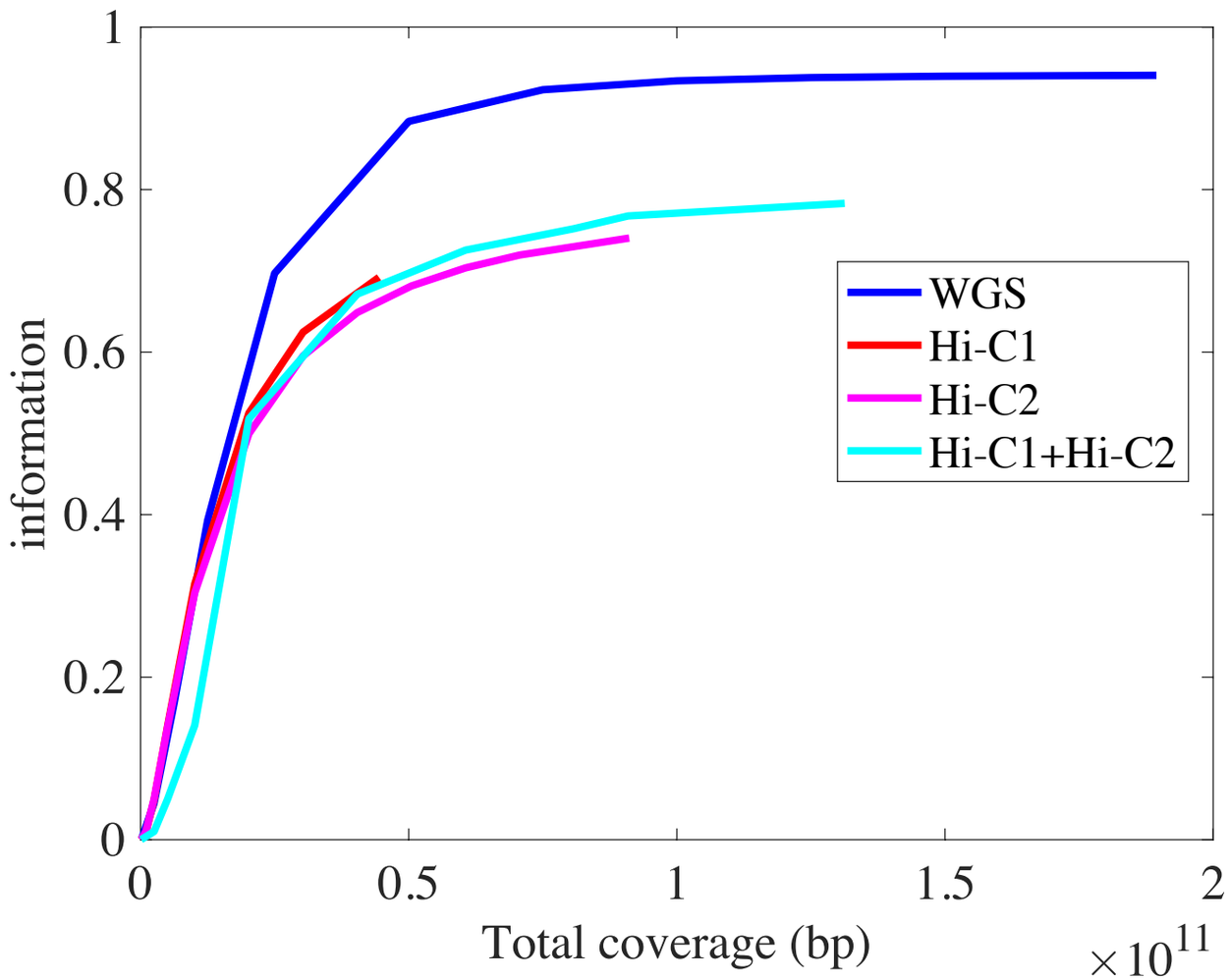
%of the gold standard information = pmi(S^{FGE};NA12878) / S^{GS}(NA12878)

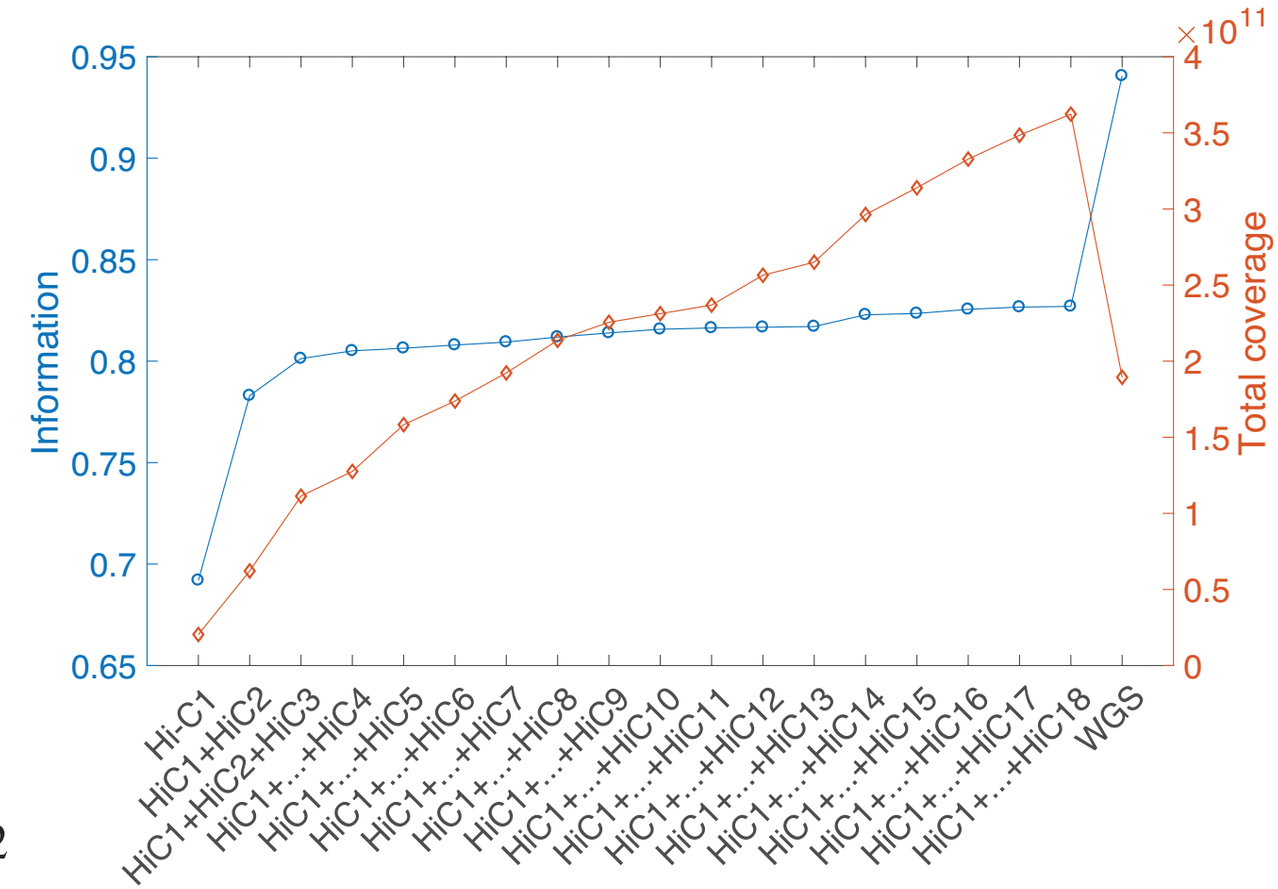
How much information a typical Hi-C experiment contain?



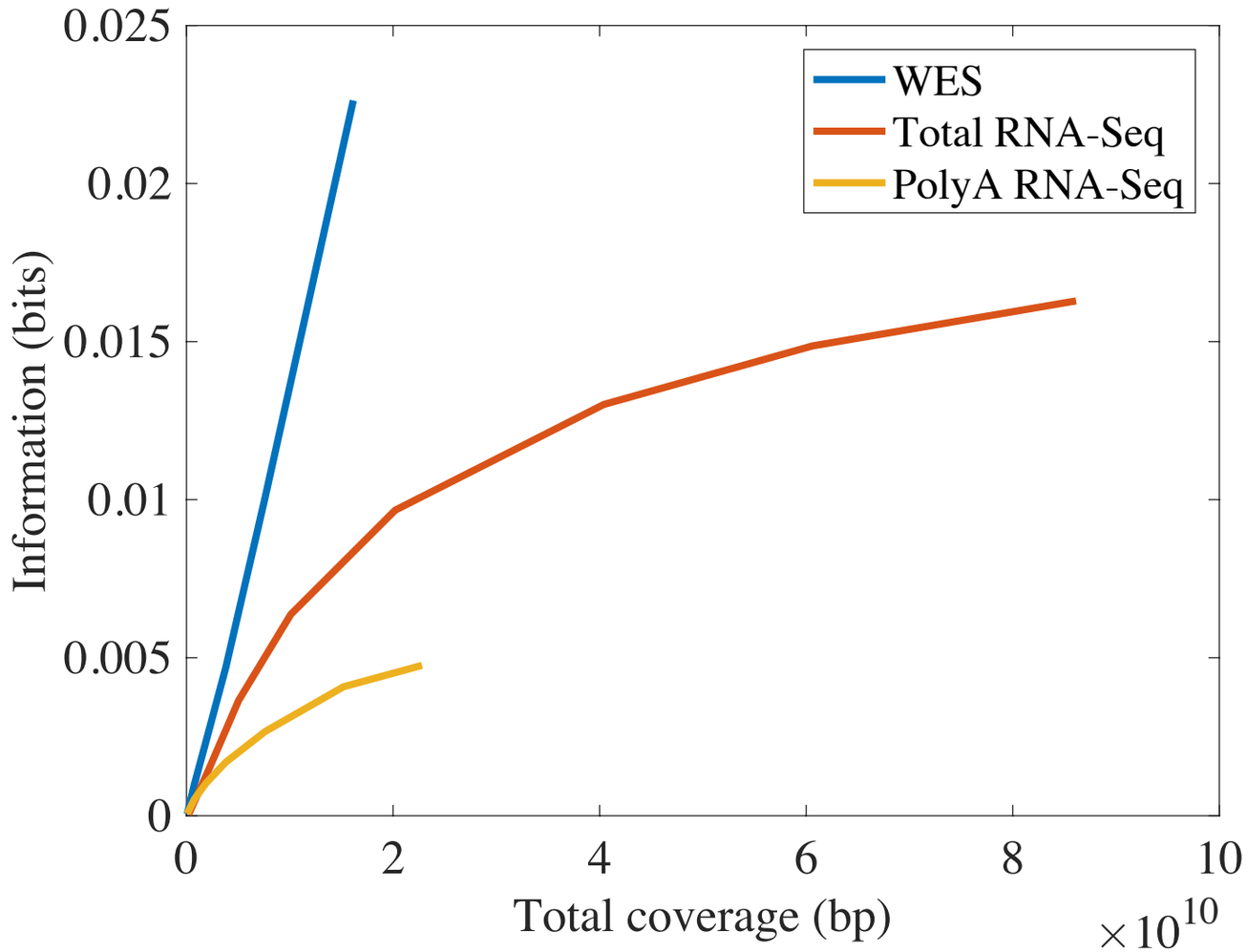
However, this is just one experiment out of 18 that was used to create the whole Hi-C library. Can we get more information by adding these experiments together?

Putting Hi-C experiments together does not change the outcome

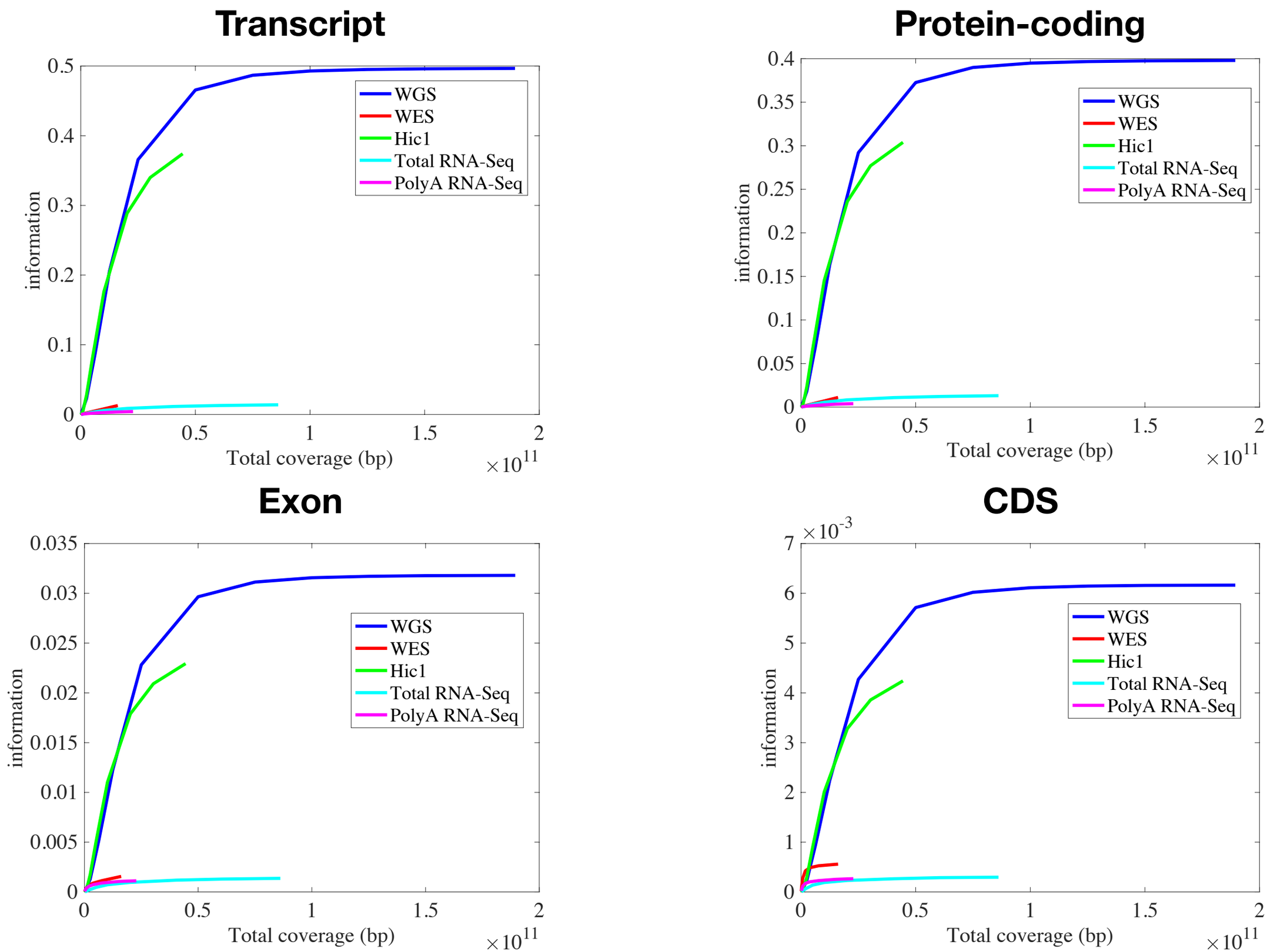




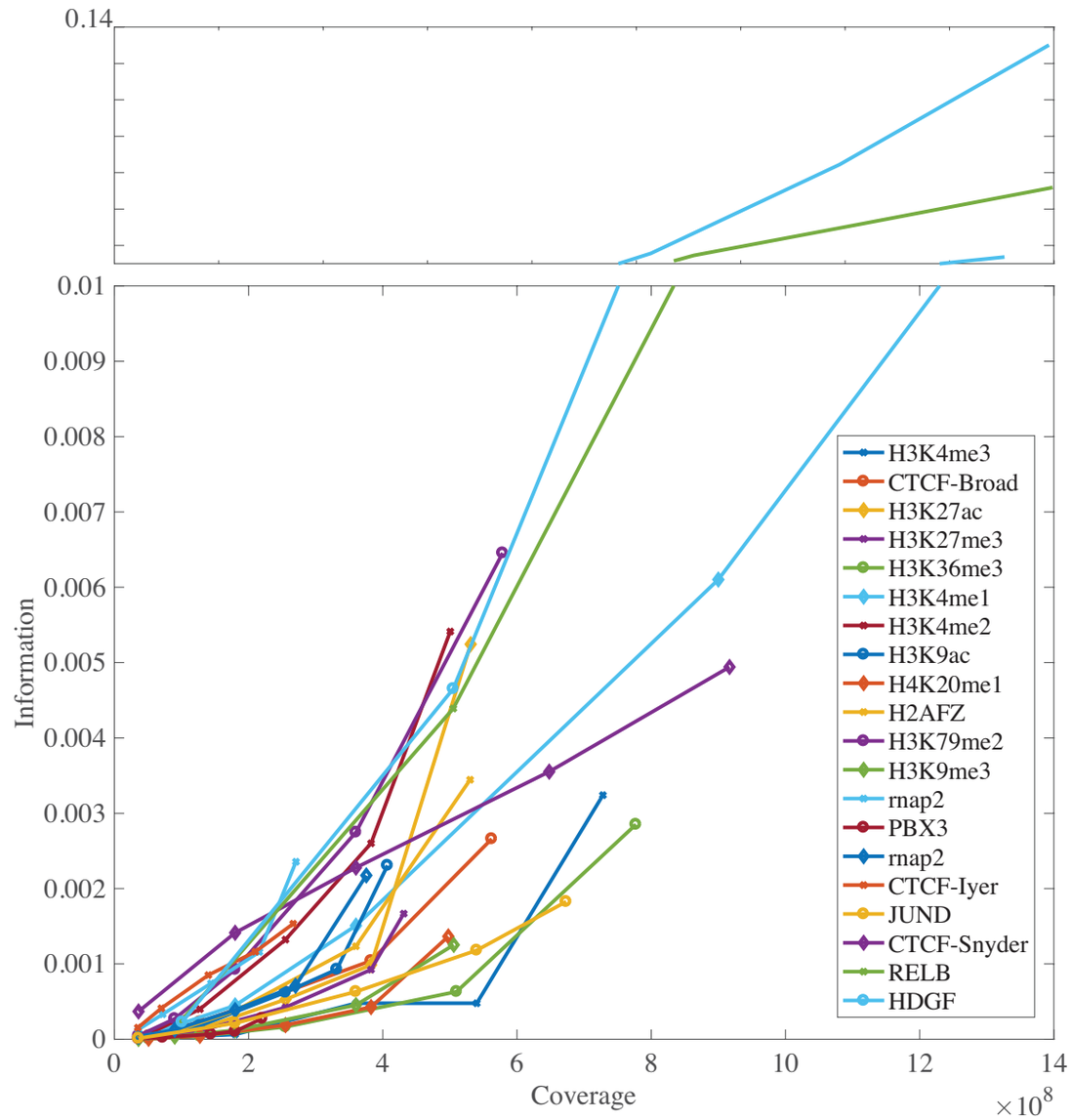
How much information a typical RNA-Seq experiment contain?



Hi-C reveals more information at transcript and coding regions compared to RNA-Seq

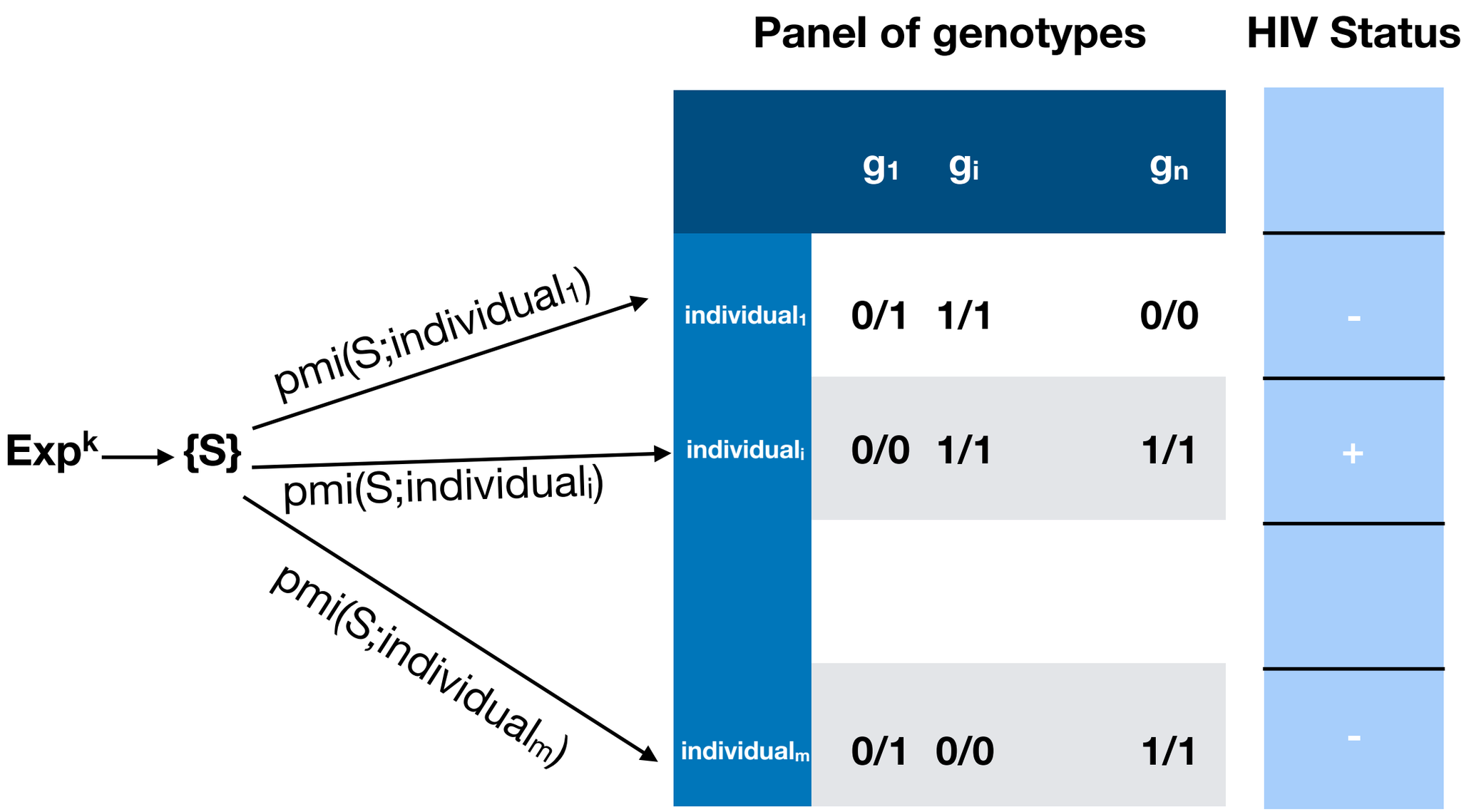


How much information ChIP-Seq contain?



Seems like both ChIP-Seq and RNA-Seq leak only a small proportion of private information. Now, the question becomes “does this leakage enough to link the individuals to a panel of genotypes?

Linking attack with the publicly available fastq files



Quantification of Linking Accuracy

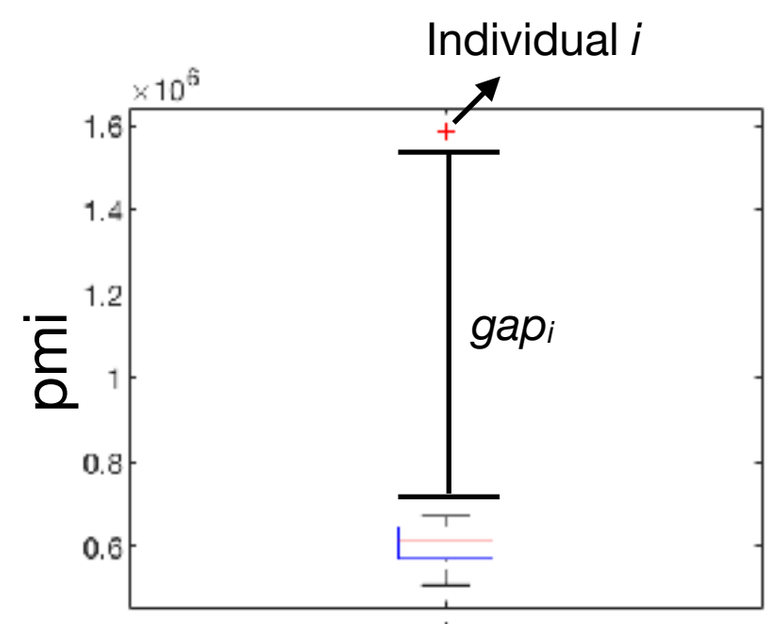
1. Amount of information we have for the target individual +
2. Amount of information we have for other individuals in the panel -

- Rank of all the *pmi(S;i)* values, where *S* is the set of genotypes called from experiment, and *i* is the genotypes of individual *i* in the panel.
- Calculate gap_i for each individual as

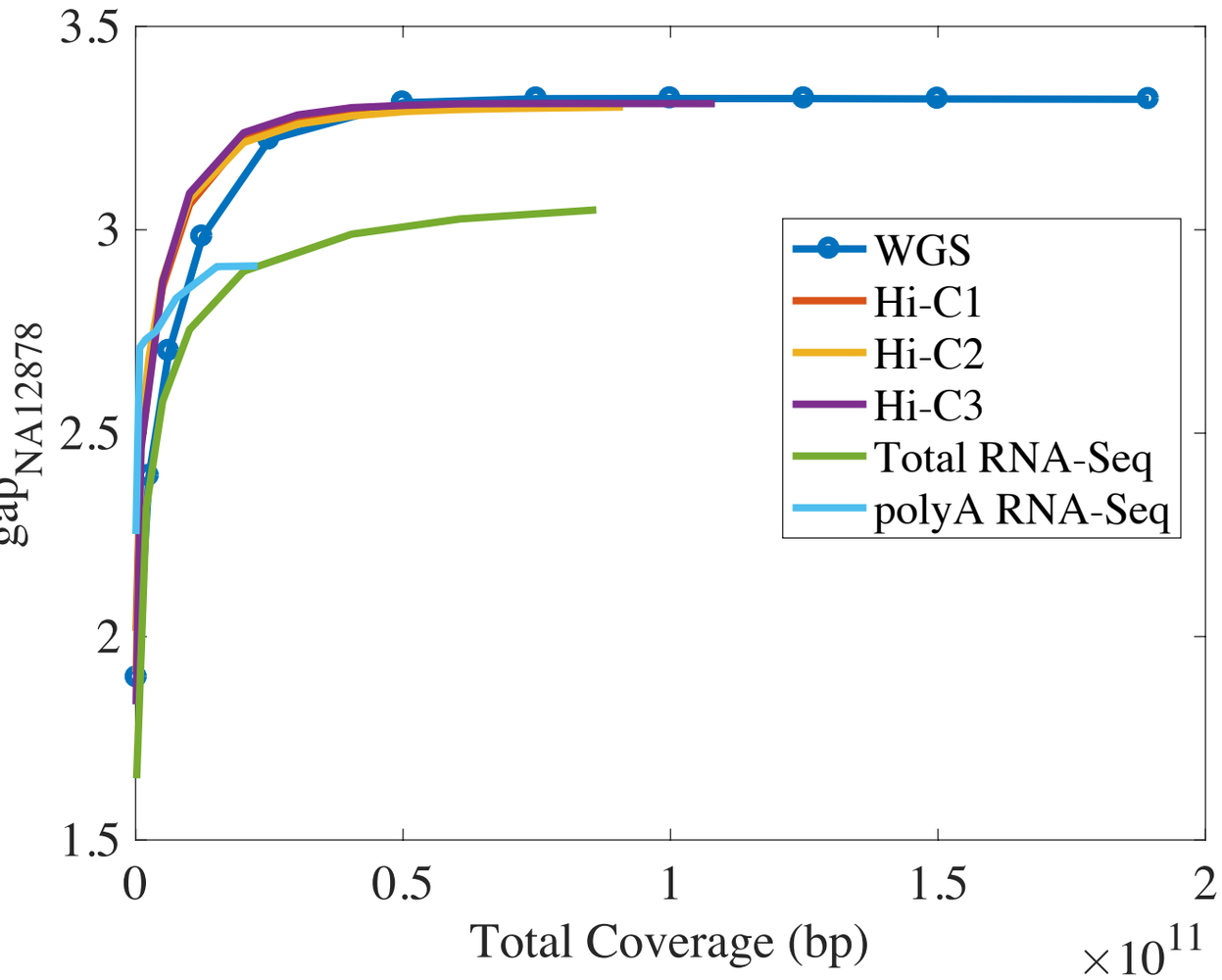
$$if\ rank\big(pmi(S;i)\big) \leq 5\ then\ gap_i = \frac{pmi(S;i)}{pmi(S;j)}\ where\ rank\big(pmi(S;j)\big) = 2$$

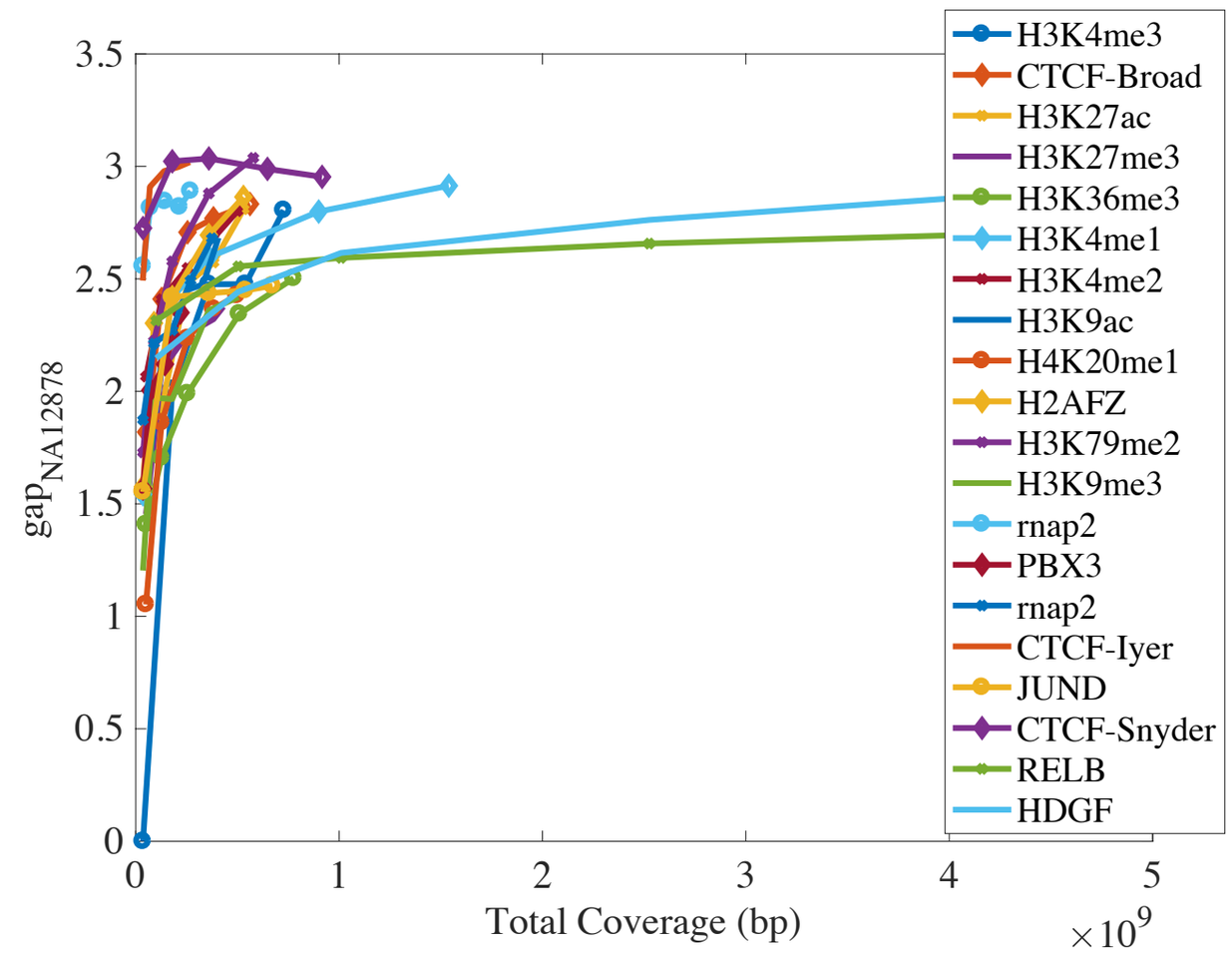
$$otherwise, \qquad gap_i = 0$$

$gap_i \geq 2$	Individual *i* is extremely vulnerable
$1 < gap_i < 2$	Individual *i* is vulnerable
$0 < gap_i \leq 1$	Individual *i* can be vulnerable with auxiliary information
$gap_i = 0$	Individual *i* cannot be identified



ChIP-Seq reveals way less information compared to other assays. However, it provides comparable linking accuracy to WGS and Hi-C!



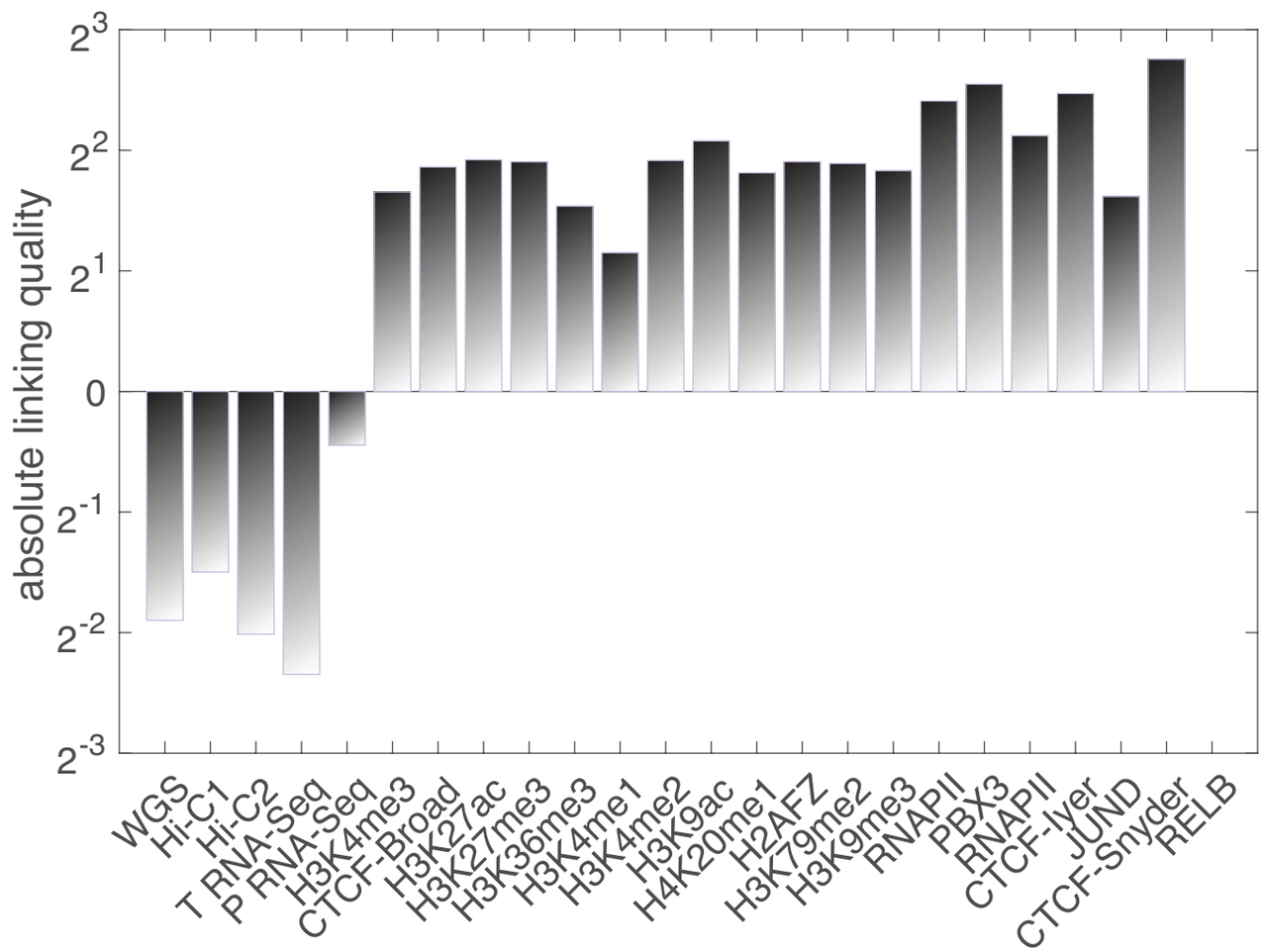


ChIP-Seq reveals way less information compared to other assays. However, it provides highest linking accuracy!

A better comparison can be made by normalizing the gap by the coverage

absolute linking quality (exp^k) = $\log_2$ [max(gap)/{coverage at max(gap)}]

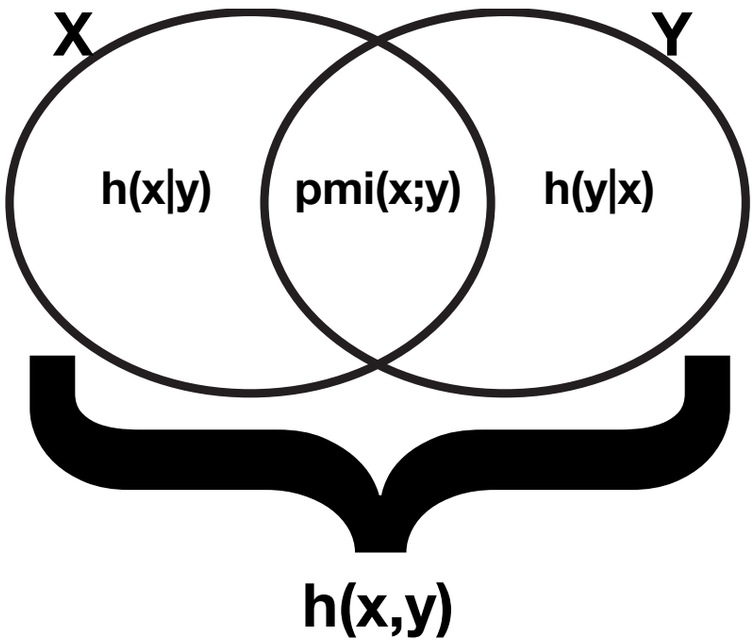
coverage = total coverage / haploid genome size



Genotyping Accuracy

For a high quality linking, we needed

1. High information overlap with the target individual
2. Less information overlap with the rest of the panel



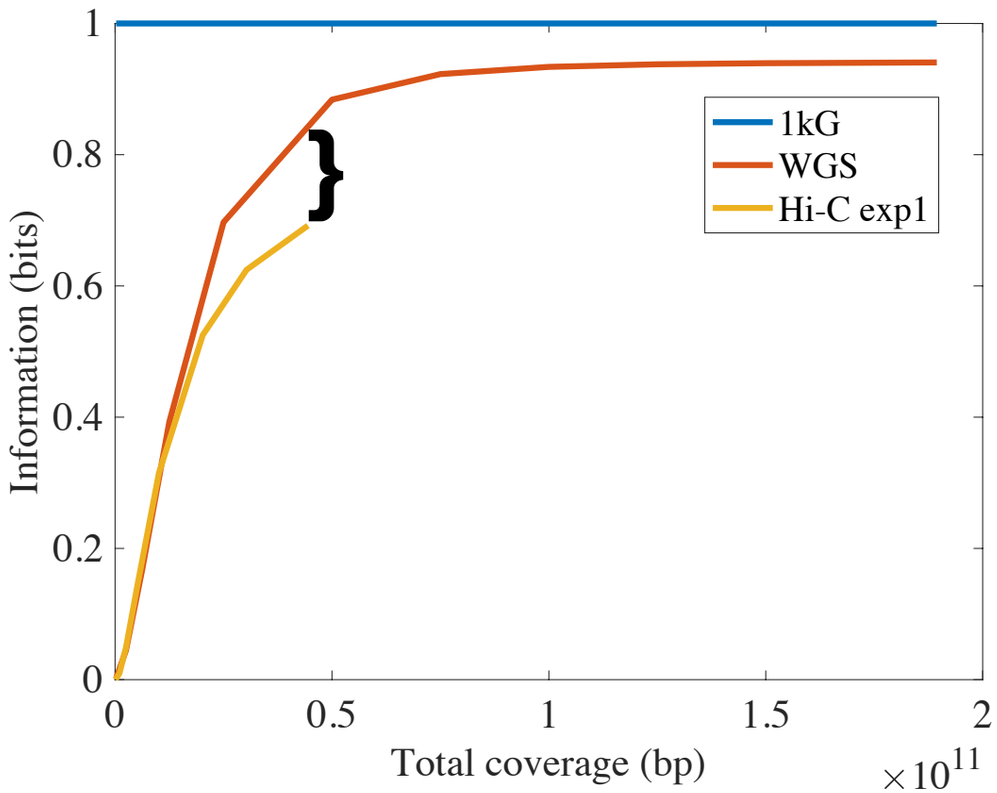
npmi(x;y) = pmi(x;y)/h(x,y)

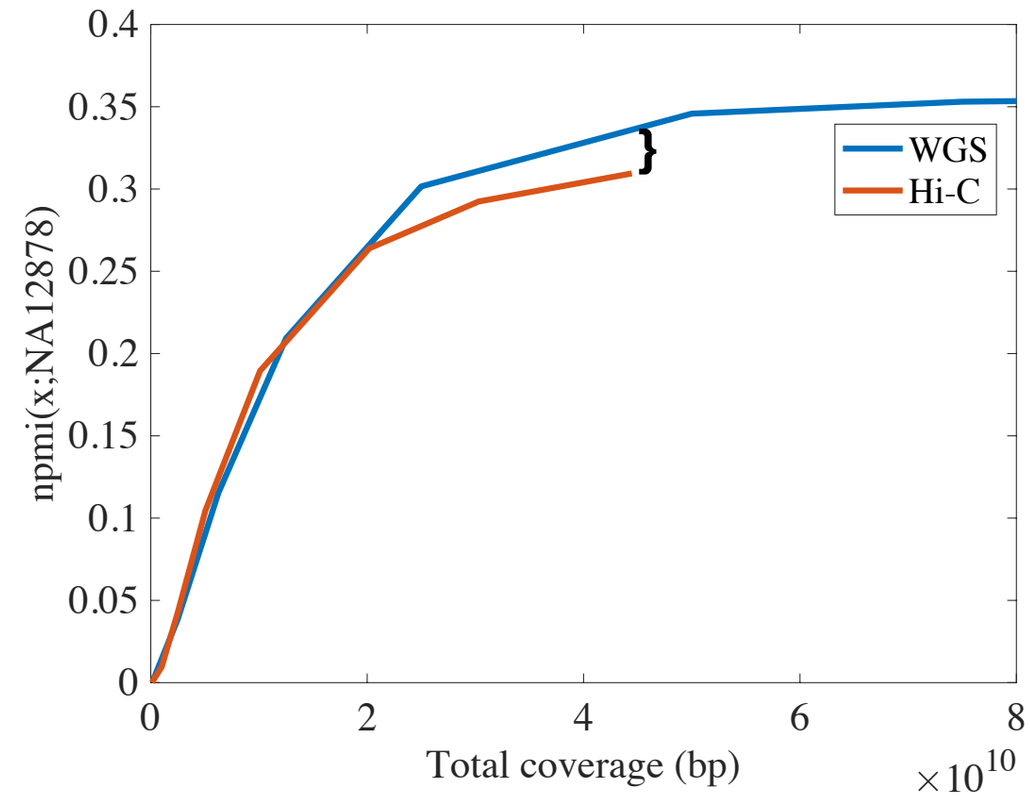
If npmi(x;y) = -1, x and y never occurring together
npmi(x;y) = 0, x and y are independent
npmi(x;y) = 1, x and y are completely co-occuring

Normalized point wise mutual information - npmi(x;y)

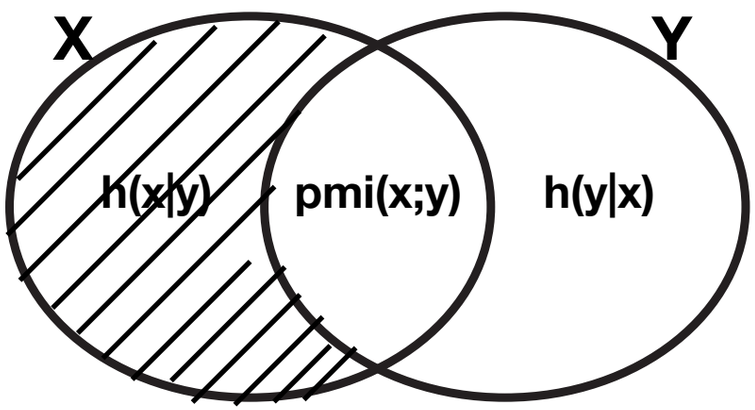
- If X is the genotype set called from experiment, Y is the gold standard, then npmi(x;y) can be used as a metric for genotyping accuracy.
- Not only missed genotypes (h(y|x)) but also the noise (or the False Positives or h(x|y)) will affect this metric.

The difference between npmi of Hi-C and WGS is smaller compared to difference between pmi





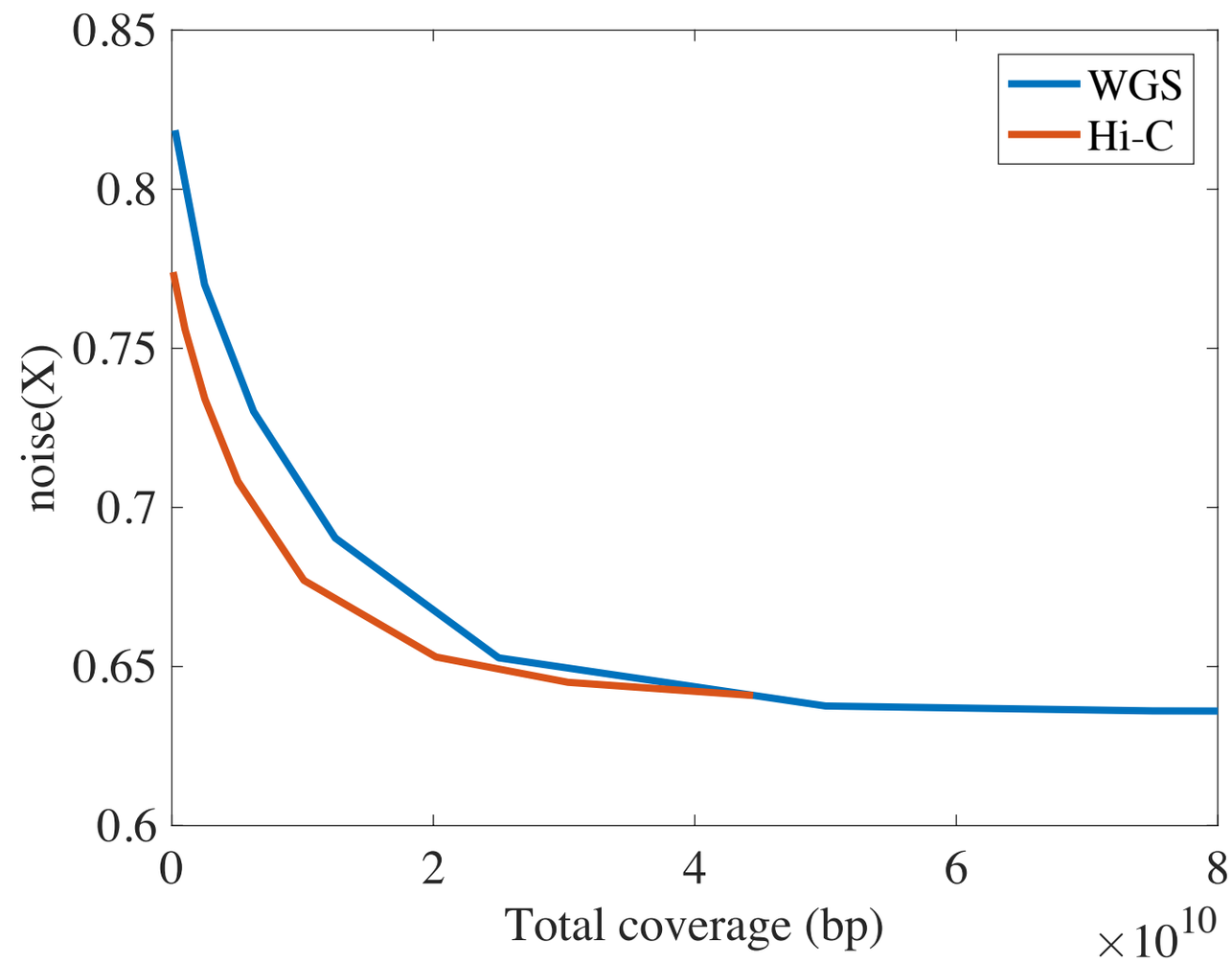
Variant calling from Hi-C is less noisy compared to WGS?



noise(X)=h(x|y)/h(x)

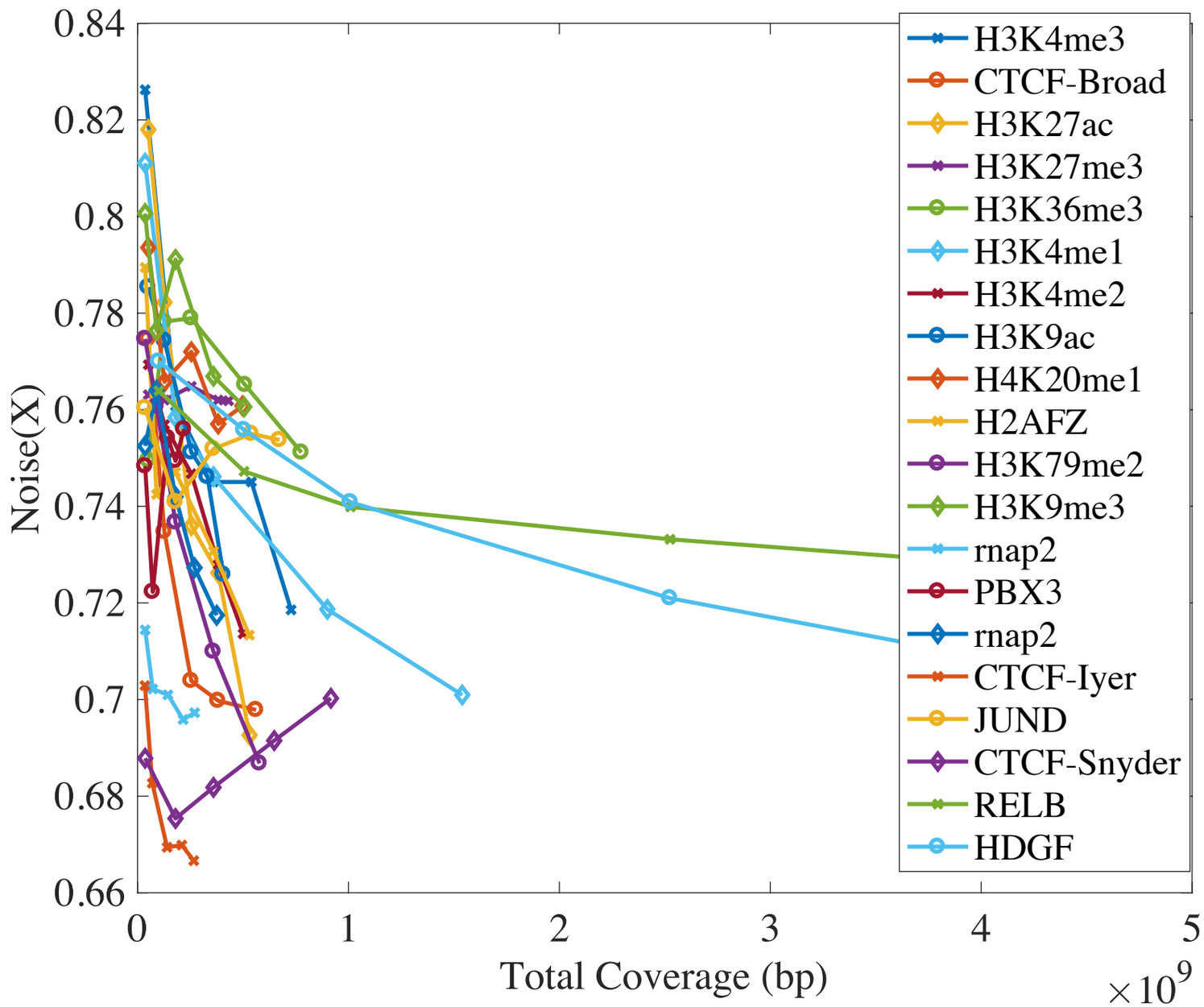
Self information of X that does not overlap with self information of Y

Self information of X



Yes! At low coverages …
Still somewhat better at high coverages

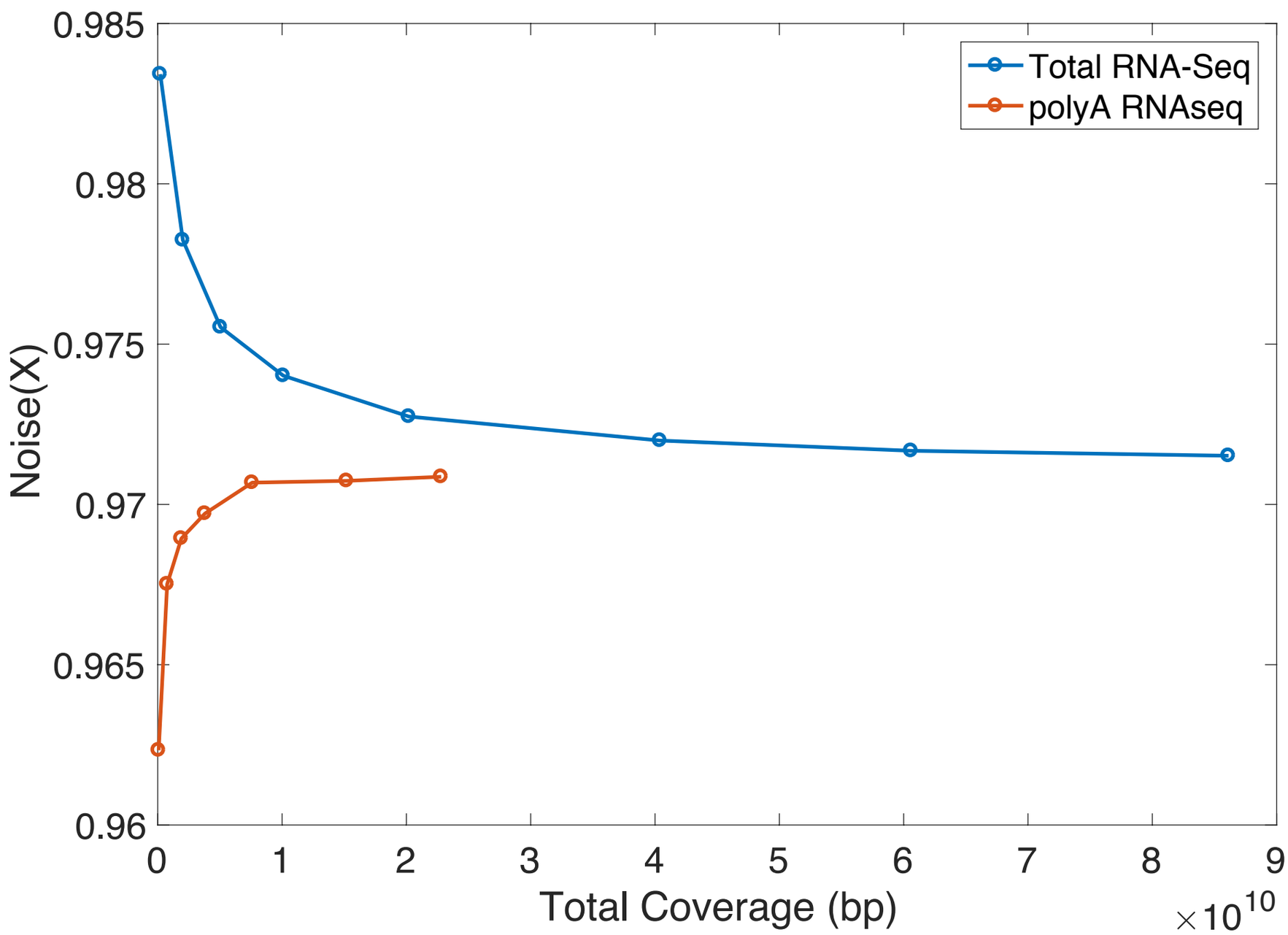
Variant calling from ChIP-Seq is comparable to Hi-C and WGS in terms of noise - maybe a bit more noisy



Variant calling from RNA-Seq has the highest amount of noise

Potential Reasons:

* Split Reads
* RNA editing



Imputation of more SNVs

- Use IMPUTE2 and 1000genomes panel to impute new SNVs using LD blocks
- Imputed SNVs are further filtered based on <0.3 certainty threshold
- pmi score is adjusted

$$S = \{s_1, ..., s_i, ..., s_N\}$$

$$h(S) = \sum_{i=1}^{i=N} -log(p(s_i))$$

$$p(s_i) = \frac{f(s_i)}{n_T}$$

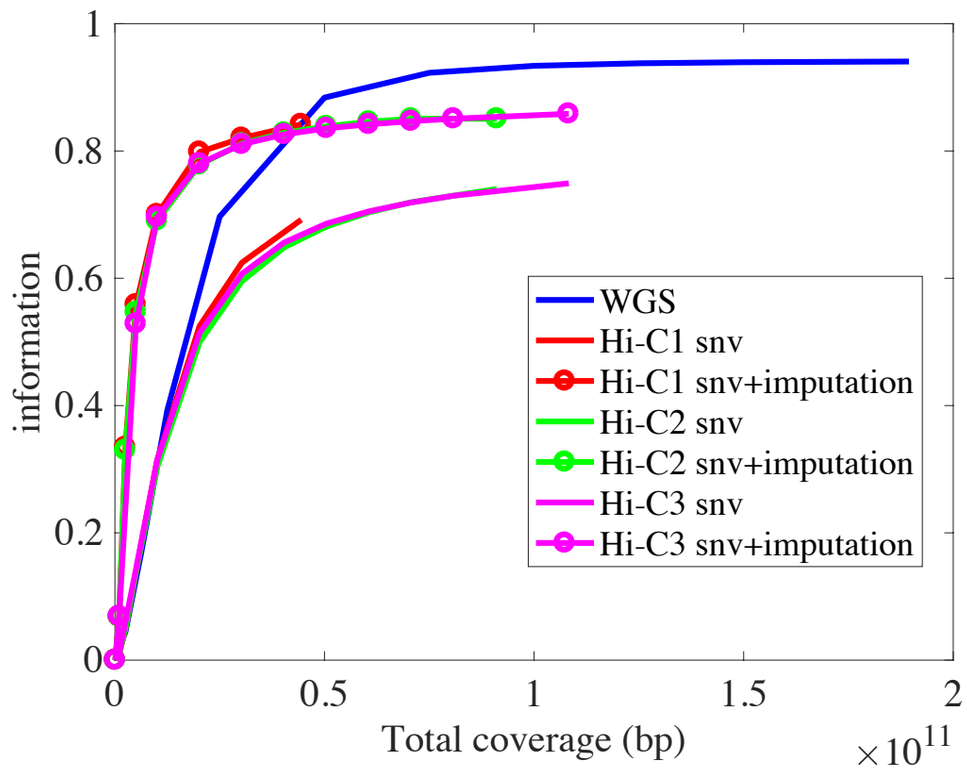
$$S^{im} = \{s_1^{im}, .., s_i^{im}, ..., s^{im}\}$$

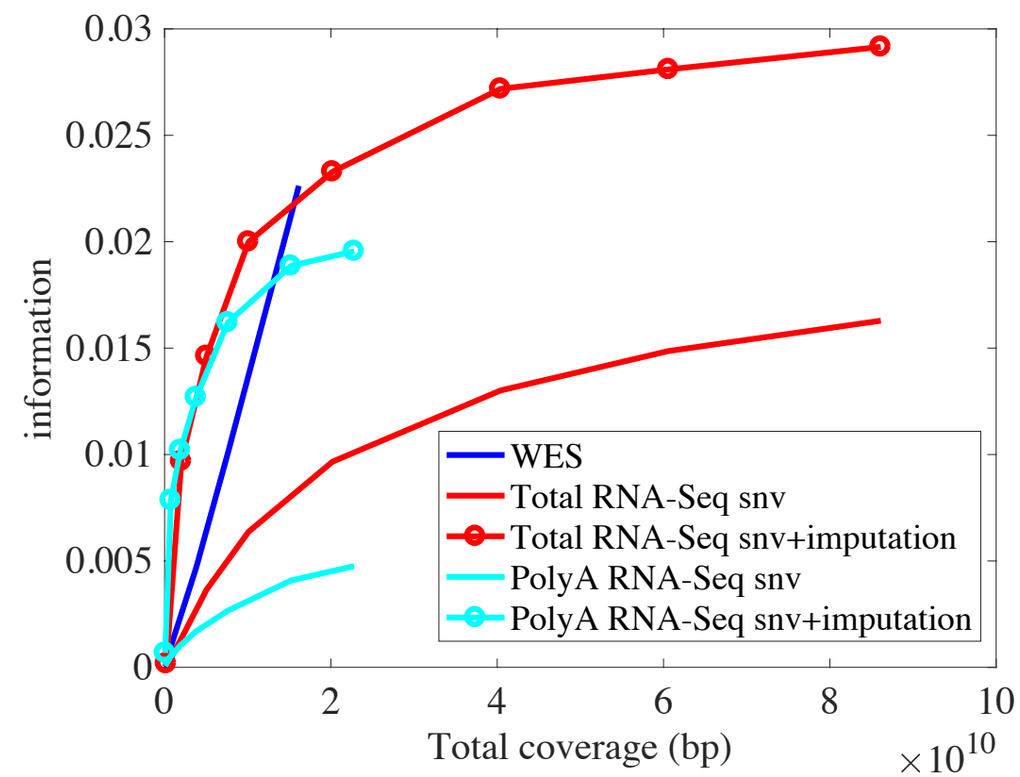
$$h(S^{im}) = \sum_{i=1}^{i=M} -log(p(S_i^{im}))$$

$$p(s_i^{im}) = \frac{\alpha_i f(S_i^{im})}{n_T}$$

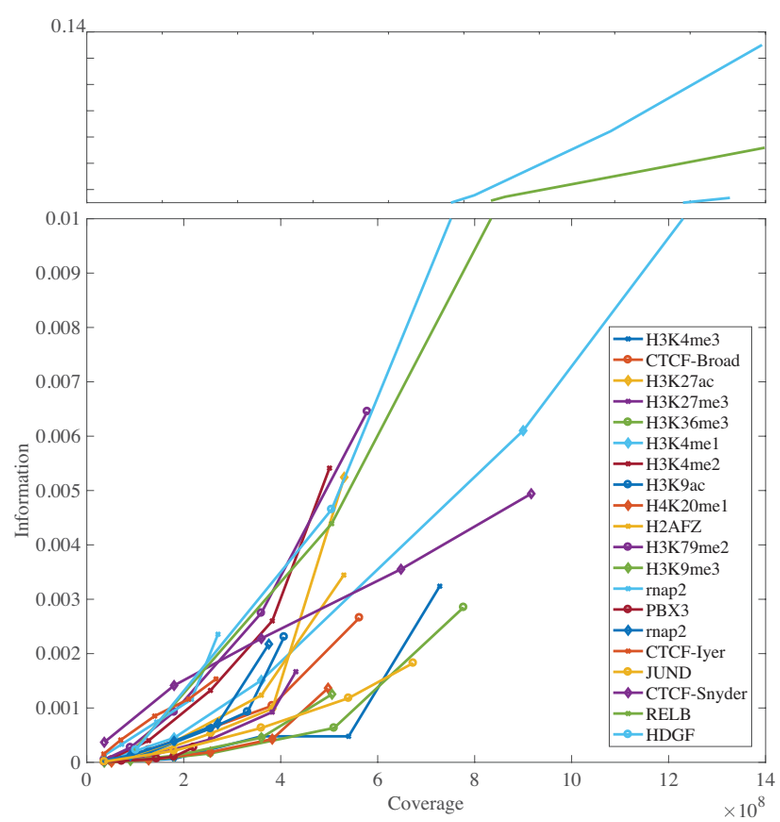
$$h(S) + h(S_i^{im})$$

Imputation substantially increases the information

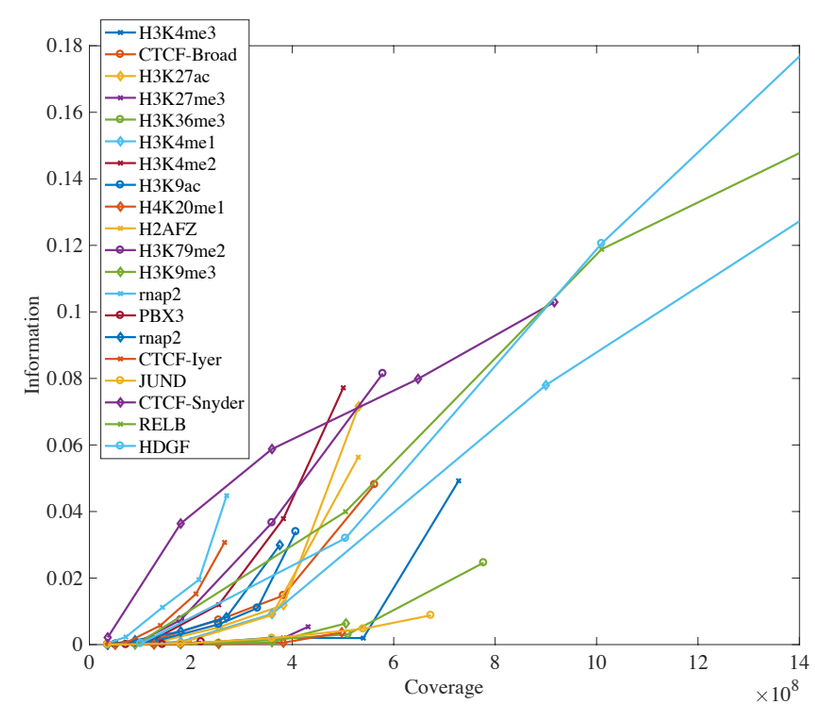




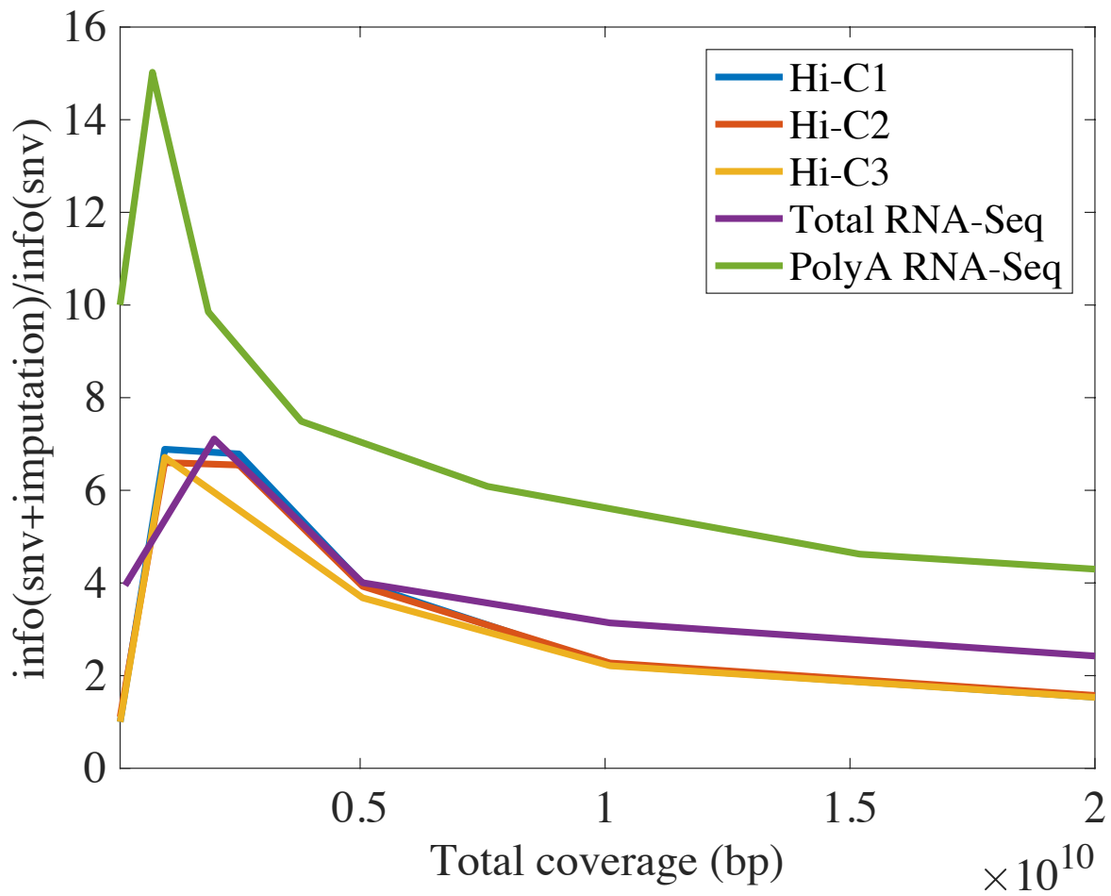
ChIP-Seq before imputation

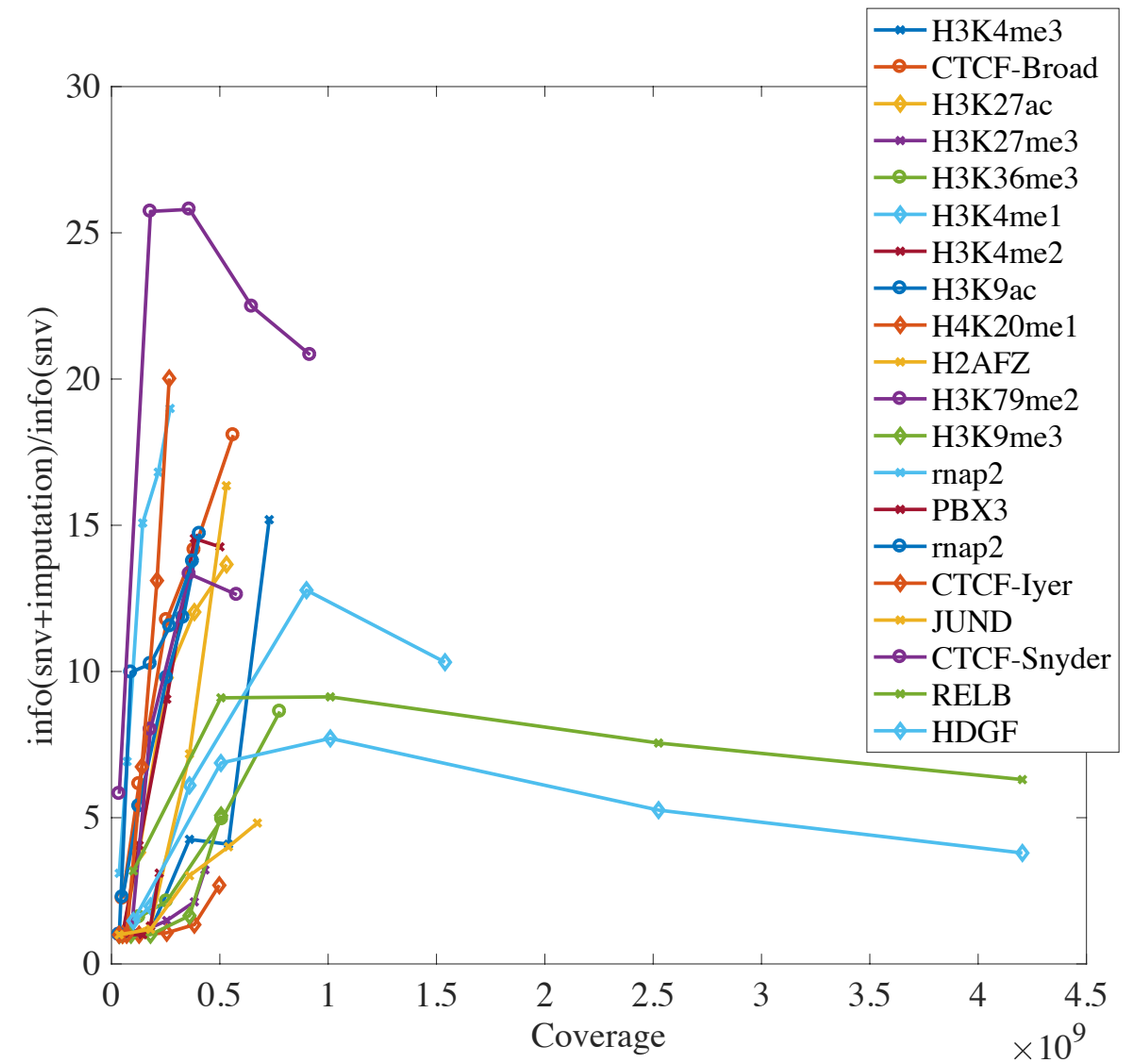


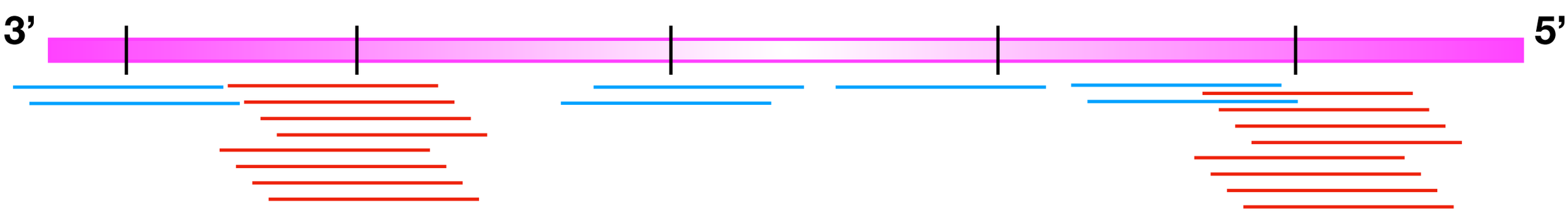
ChIP-Seq after imputation



More information can be imputed from ChIP-Seq&RNA-Seq compared to Hi-C



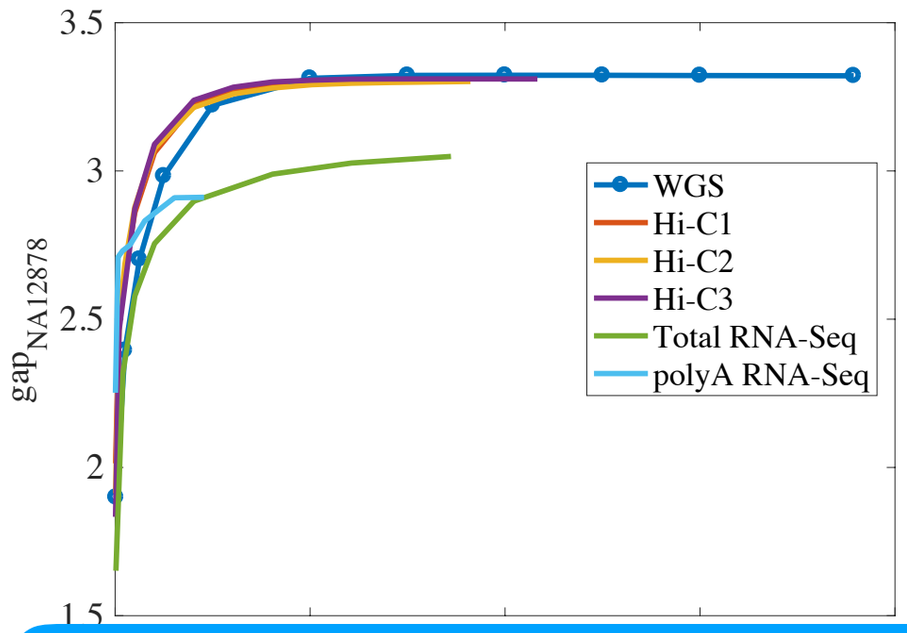




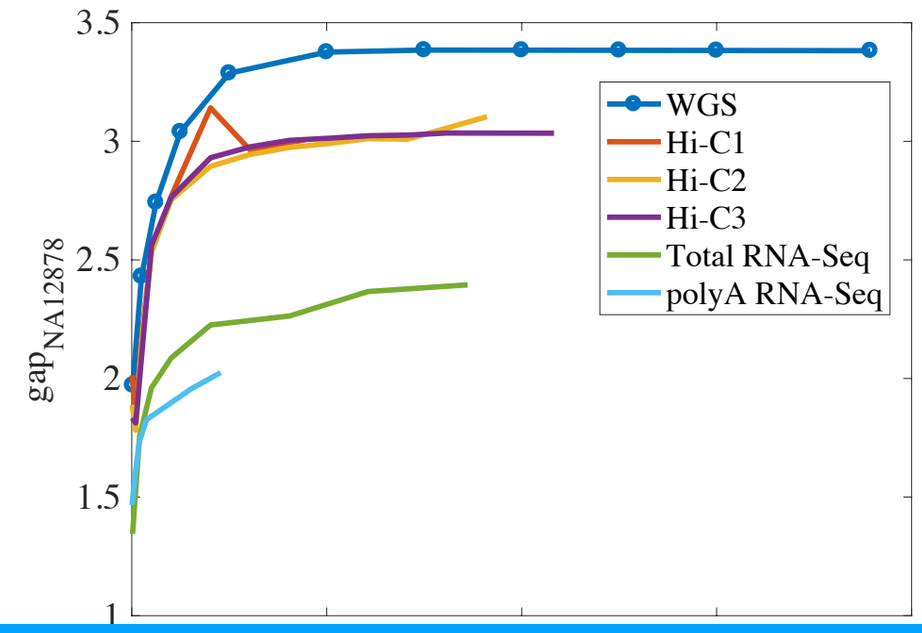
- **Having more depth in a concentrated area** vs. **having shallow depth but sampling the genome in many LD blocks**
- More room for imputation - data is sparse

However, imputation decreases linking accuracy

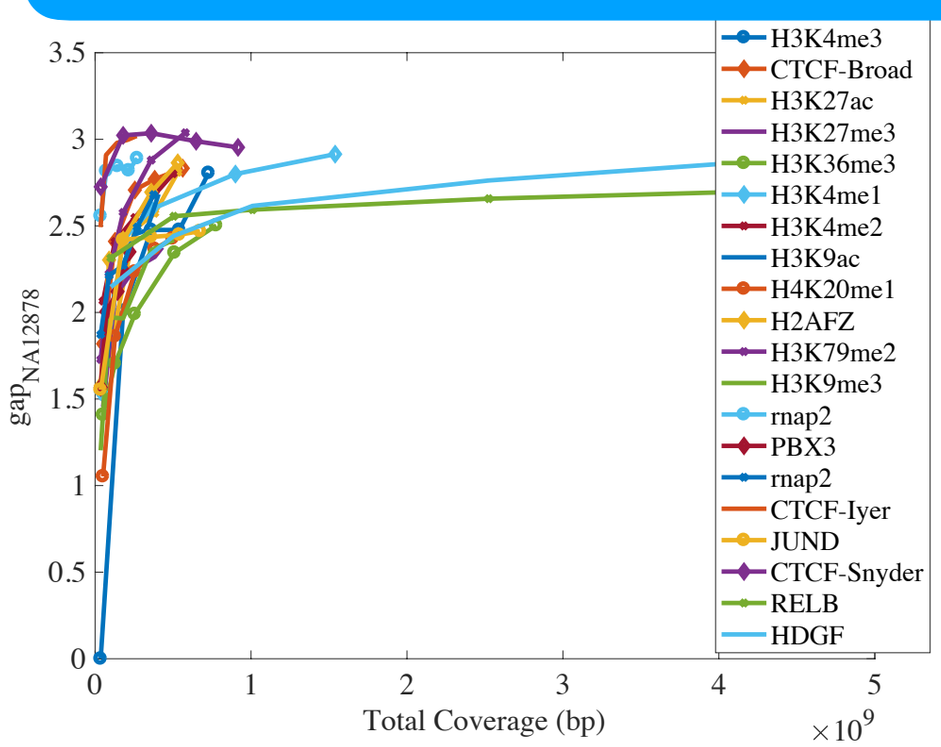
Hi-C&RNA-Seq before imputation

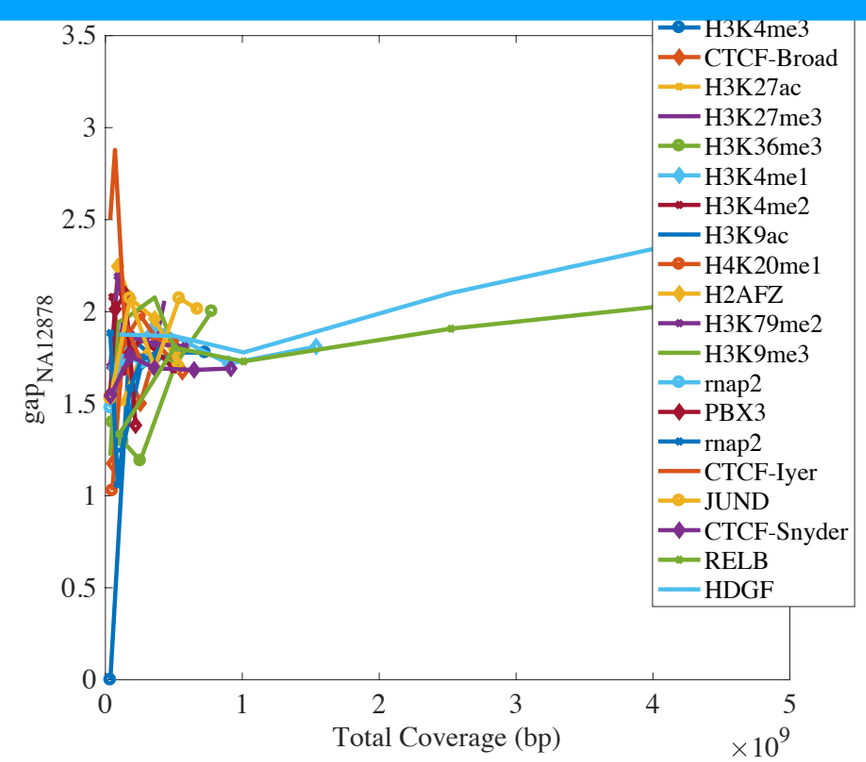


Hi-C&RNA-Seq after imputation

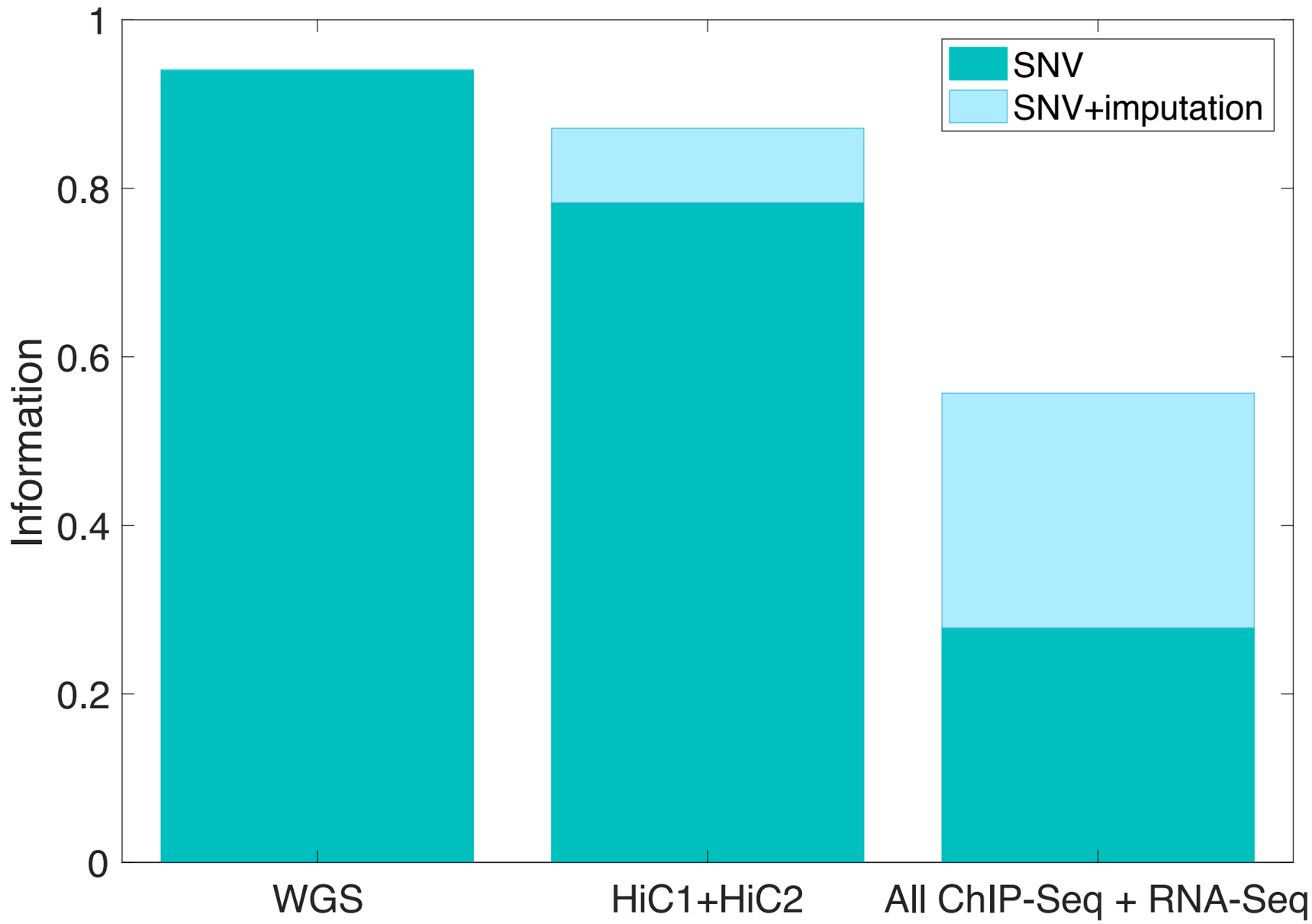


Because we impute SNVs that have high MAFs in the population and end up having common SNVs to the individuals in the panel

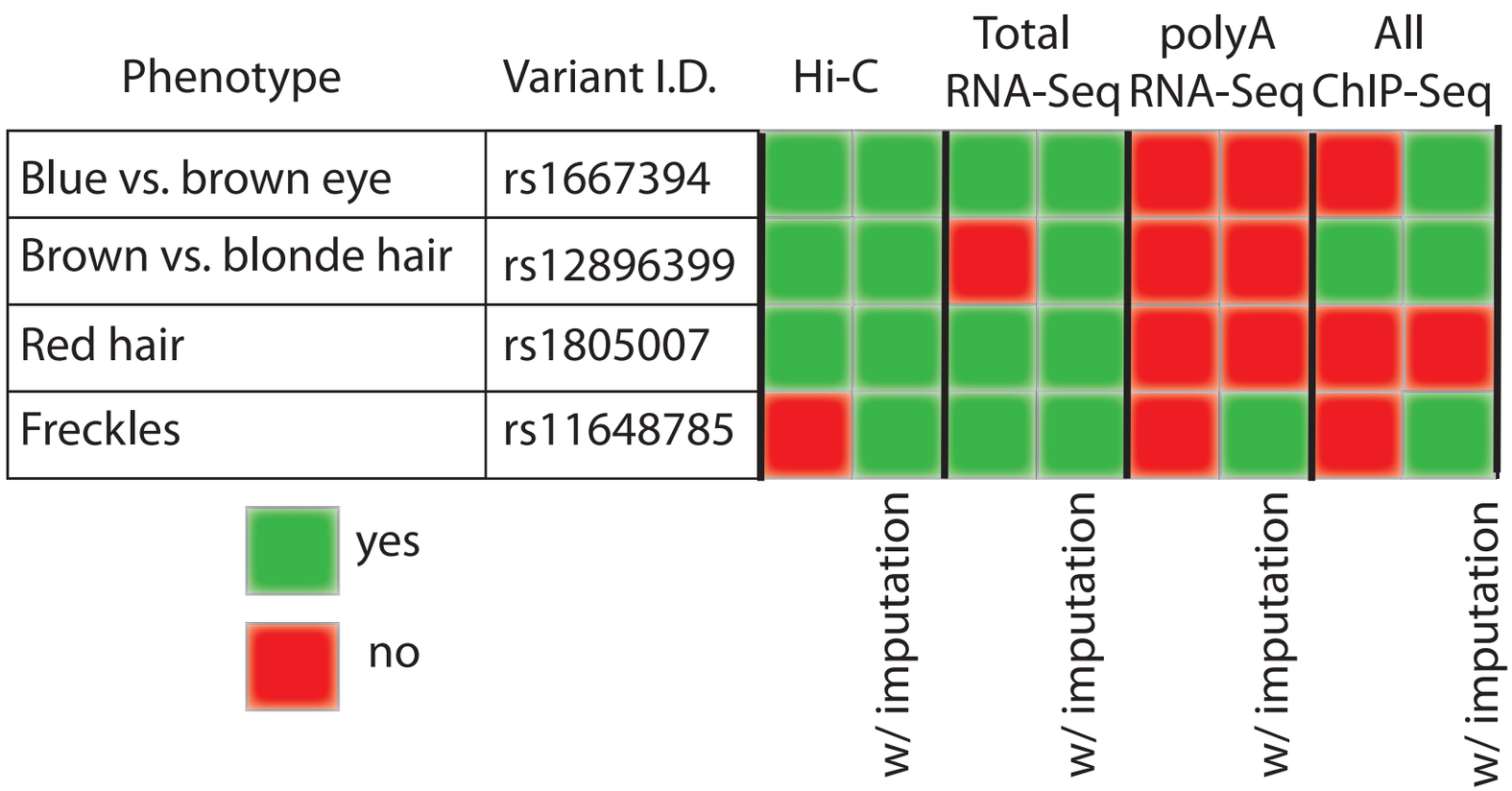




We can further put all ChIP-Seq together and gain a wealth of information about the individual



Phenotypes can be inferred using noisy&incomplete sequencing data from RNA-Seq and ChIP-Seq



A theoretical Framework

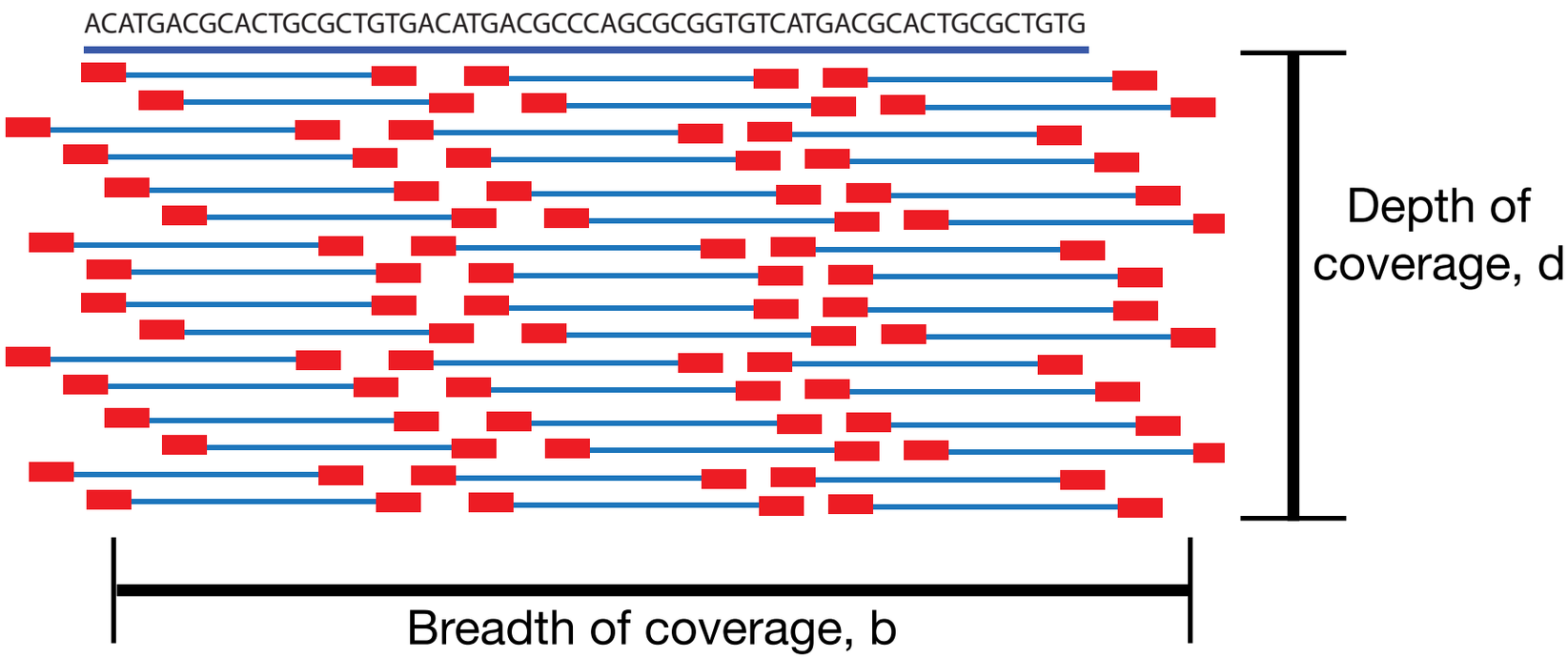
Given properties of an experiment with sequencing product, can we predict approximately how much information will be leaked without calling variants?

- As the total coverage increases, leaked information increases - trivial but is it linear?
- Can depth of the coverage alone predict the leaked information?
- Every experiments comes with biases (i.e transcription factor binding site distribution, non-coding genome, just protein coding genome, etc)
  - Can we quantify the bias? Does that help us to quantify the leaked information?

$$y \sim f(x_1, x_2, .., x_N)$$

If y is the leaked information, is there an f? If so, what are x?

Lander-Waterman Statistics



Lander-Waterman Equation

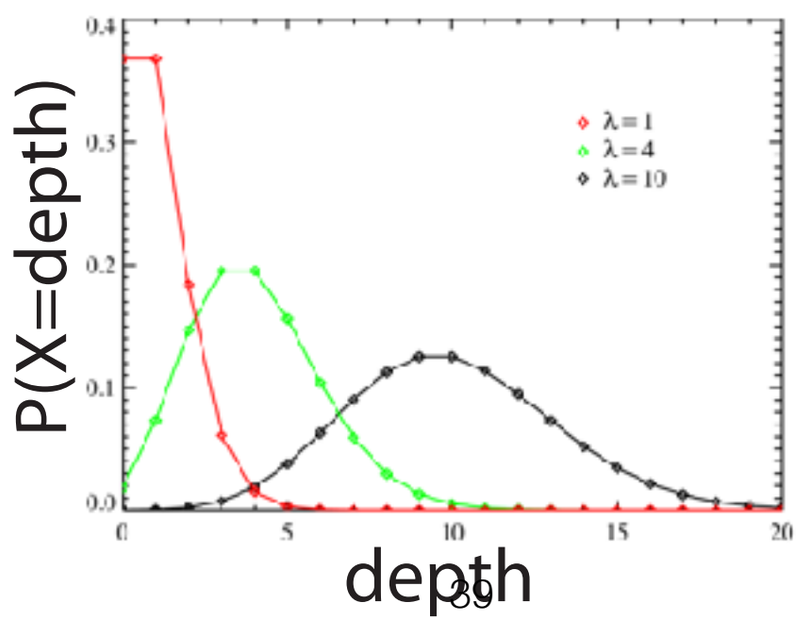
$d = N \times L / G$

N = number of reads

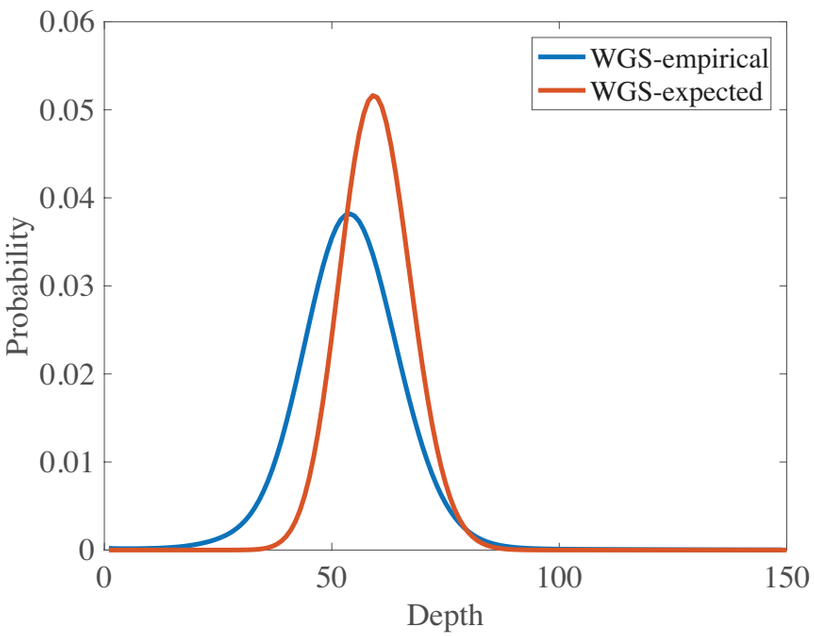
L = length of the reads

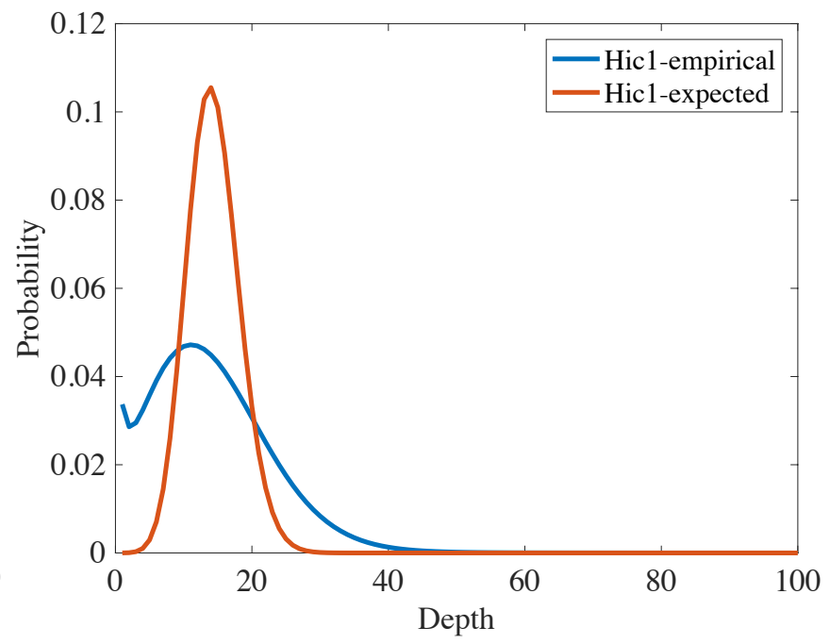
G = haploid genome length

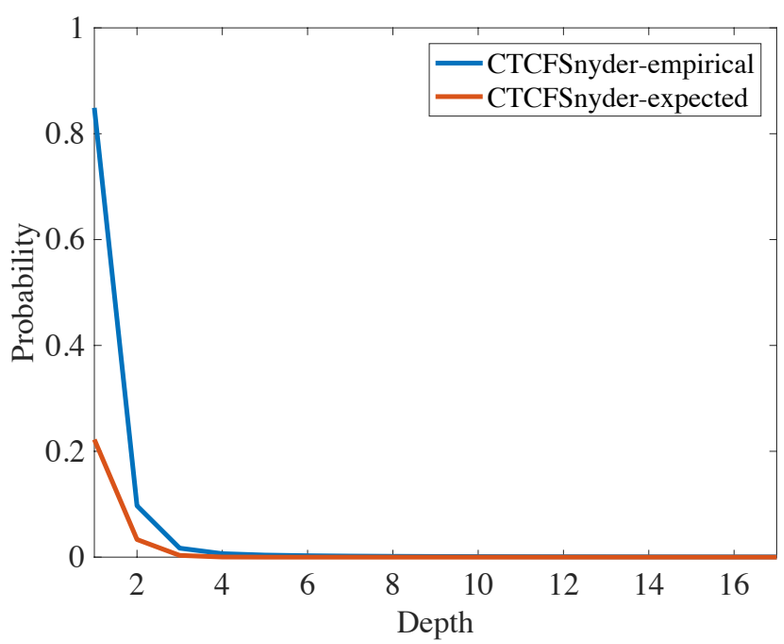
- Since G >> L, end effects are ignored
- Left-hand ends of the fragments are independently distributed with uniform distribution over [0, G]
- Any left hand falls in an interval (x , x+L) with a probability of N/(G-1), G is large, so N/G
- Number of fragments that fall in to this interval has a binomial distribution with a mean of NxL/G = d
- Since N>>L, this distribution approximates to Poisson with a mean and std of d

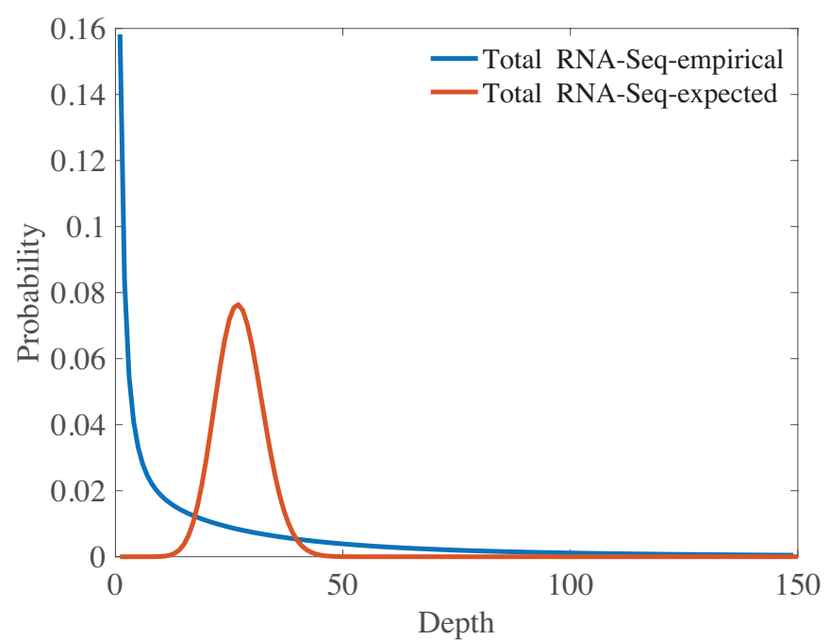


Lander-Waterman statistics can be used to estimate the bias









- Expected distributions are derived using mean = NL/G for each experiment
- Bias ~ divergence from Poisson

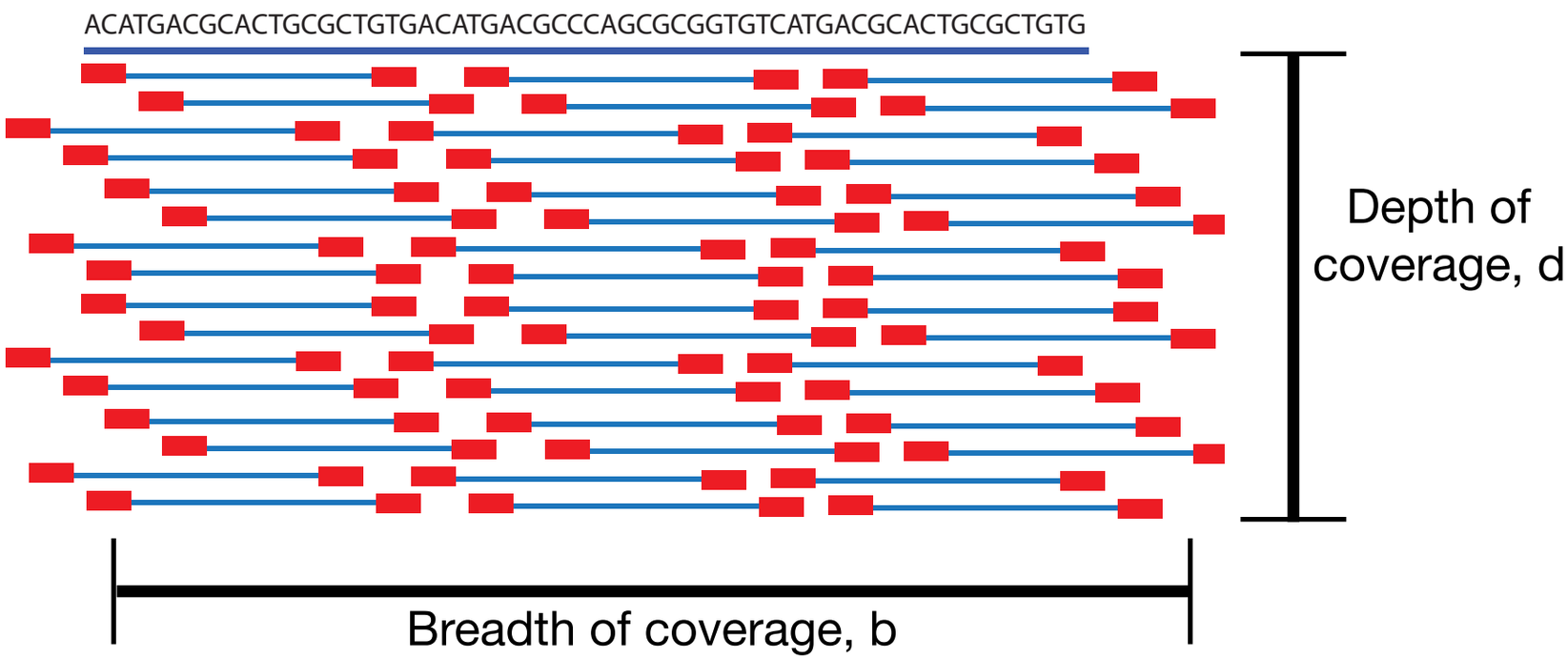
Kullback–Leibler divergence

$$D_{\mathrm{KL}}(P\|Q) = \sum_i P(i)\,\log\frac{P(i)}{Q(i)}$$

D_{KL} is the expectation of the logarithmic difference between the depth distribution of an experiment and its expected Poissonian behavior if it were to be a WGS experiment

	D_{KL}
WGS	0.235717159
Hi-C	0.447761206
RNA-Seq	1.725381274
CTCFSnyder	0.296116866

A Theoretical Framework



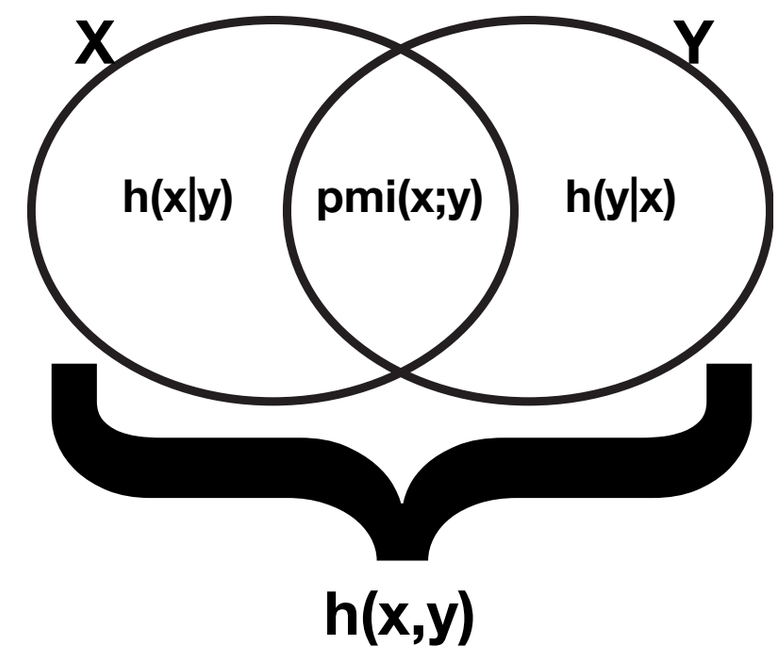
Lander-Waterman Equation

d=NxL/G

N = number of reads

L = length of the reads

G = haploid genome length



npmi(x;y) = pmi(x;y)/h(x,y)

$$y \sim f(x_1, x_2, .., x_N)$$

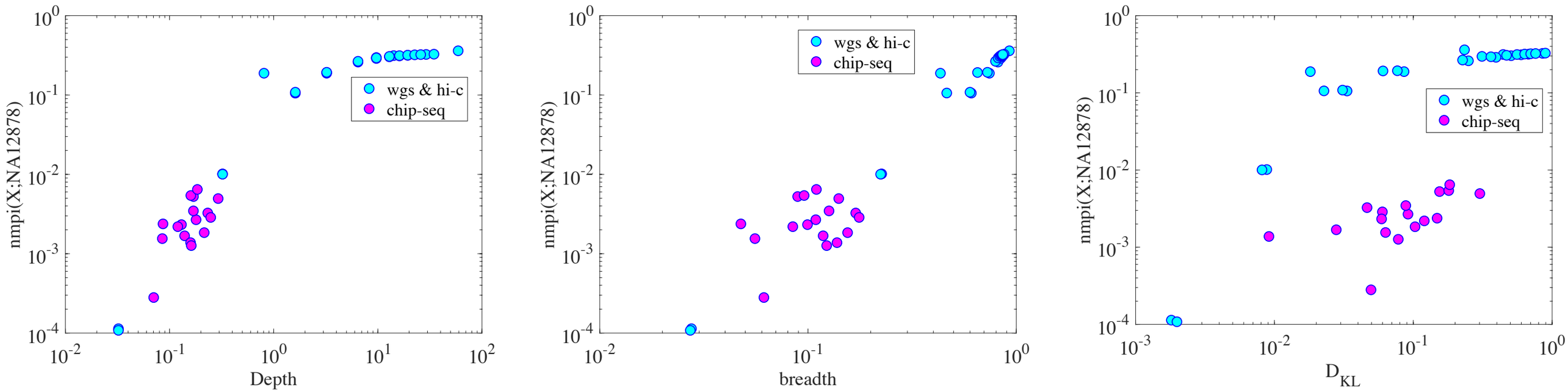
y = nmpi(X;NA12878) (genotyping accuracy)

x_1=breadth of the coverage

x_2=depth of the coverage

x_3=D_{KL} (divergence from expected distribution)

Nonlinear relationship between the features and npmi



Regression Learning

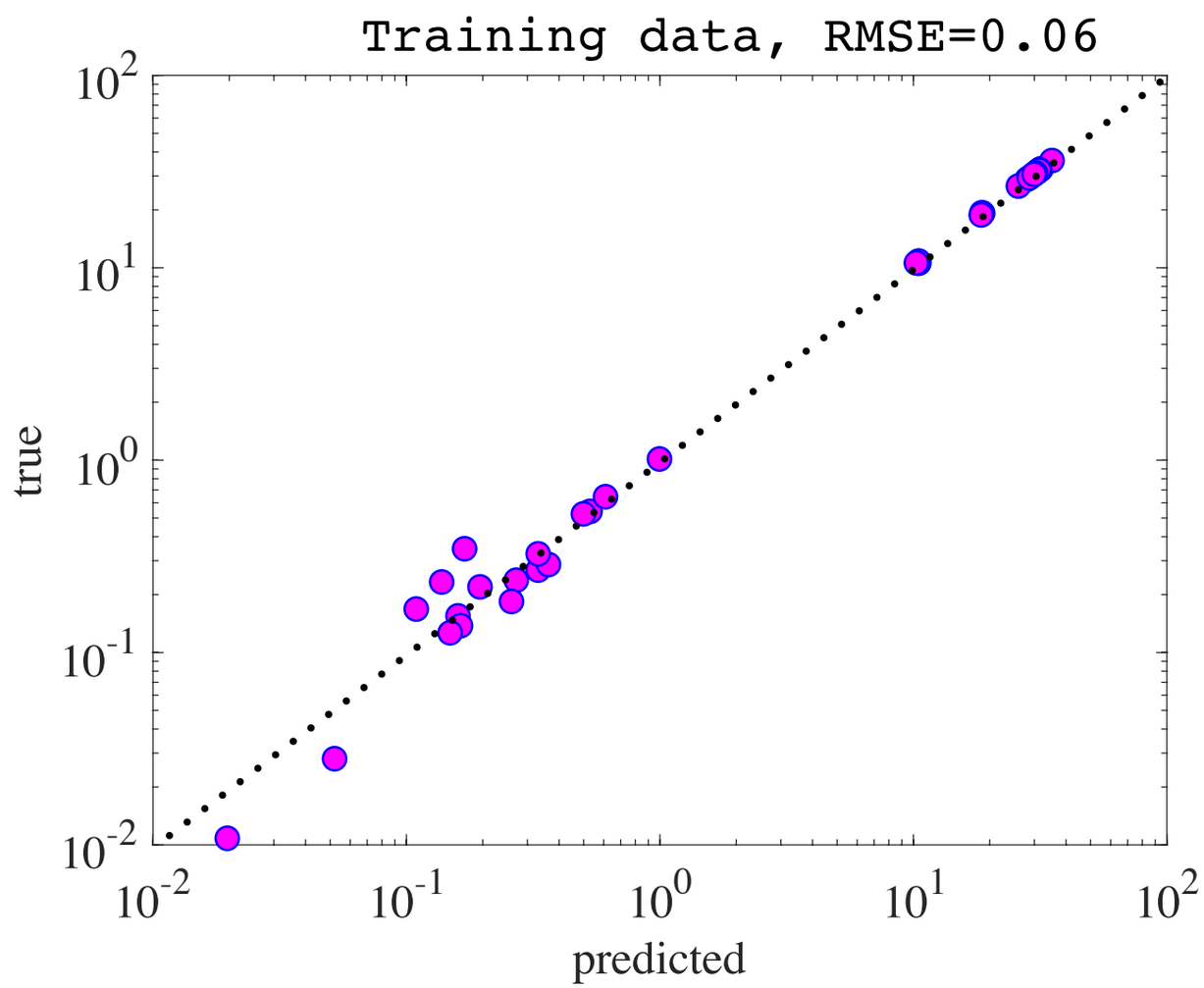
- 19 different learners (linear regression, different trees, SVM with different kernels, Gaussian process regression)
- Best fit = Gaussian Process Regression with exponential kernel

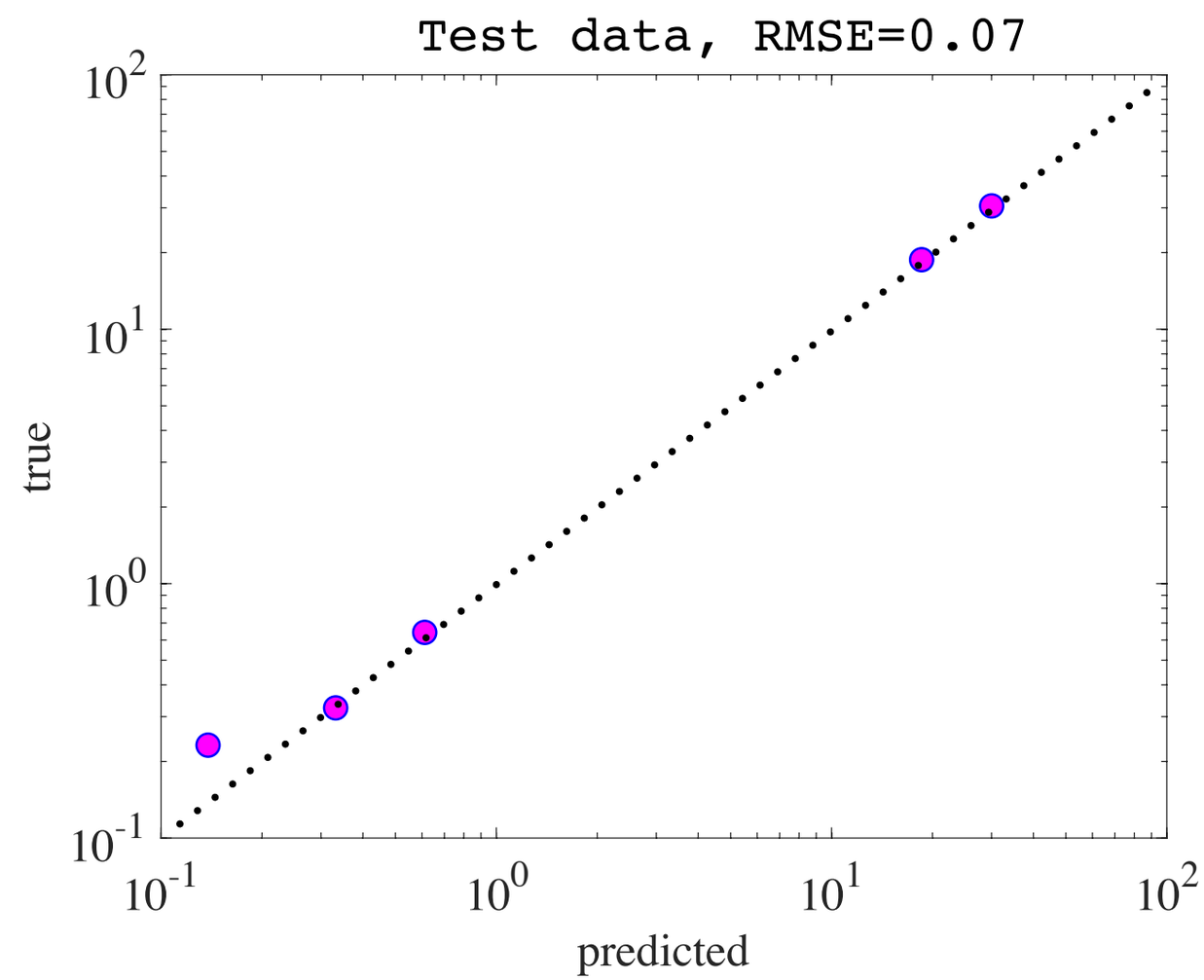
$$y = x^T\beta + \varepsilon,$$

$$\varepsilon \sim N(0, \sigma^2)$$

Regression Learning

- Total of 45 data points (values range between 0 and 35)
  - 40 is used for training, 5 is for test (randomly sampled)
  - 5-fold cross validation during training

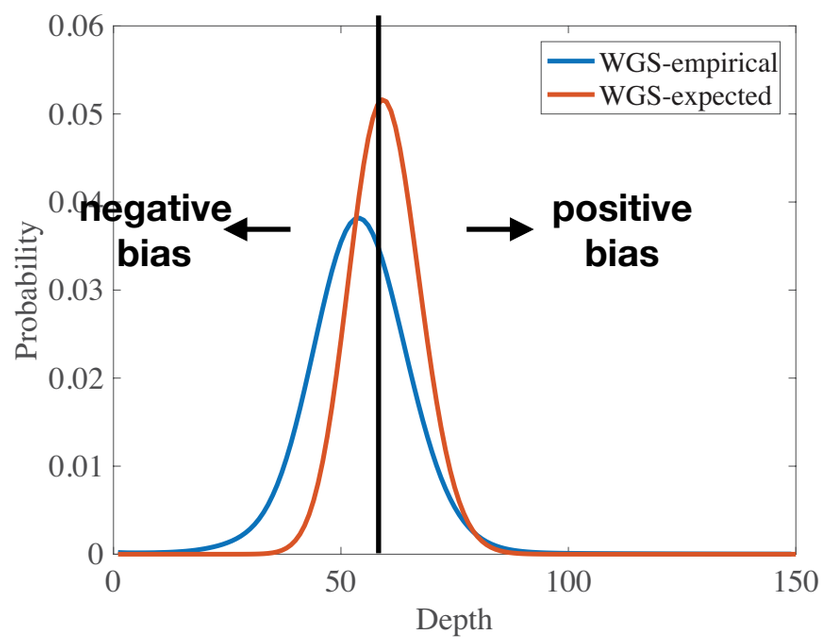




A few points to consider in the future

- We can simulate more data points to increase the sample size for the learner

- Definition of bias can be changed and it might help to build a better predicter

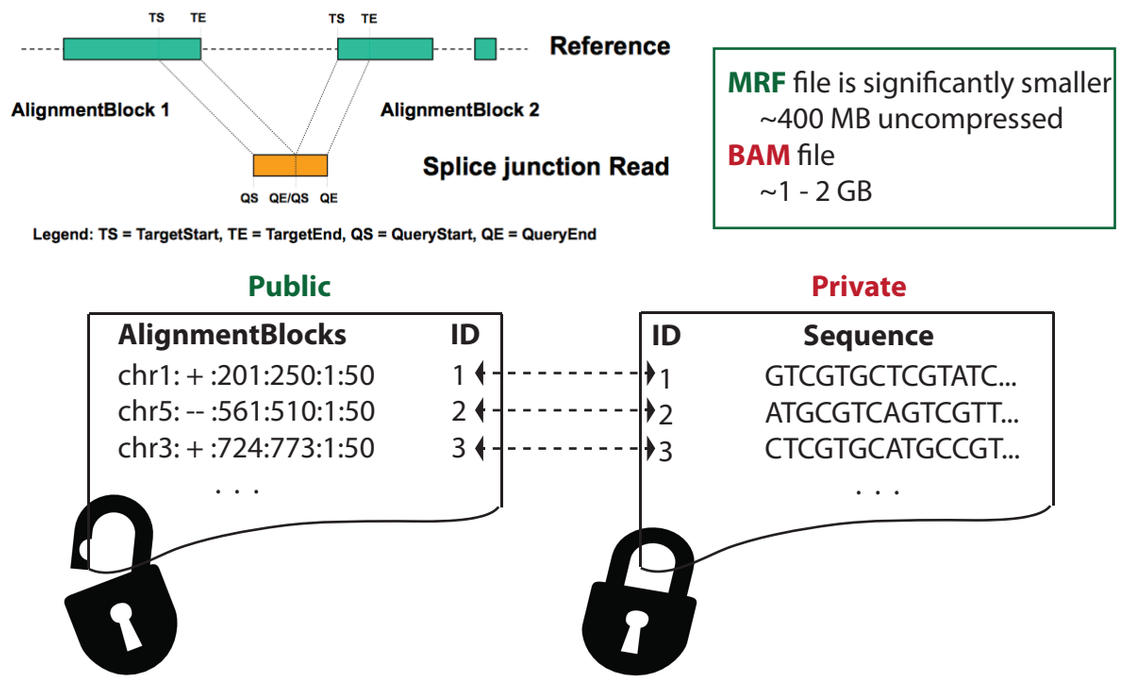


- Try to predict the information content of a different kind of functional genomic data such as ATAC-Seq or Faire-Seq

How can we secure this data?

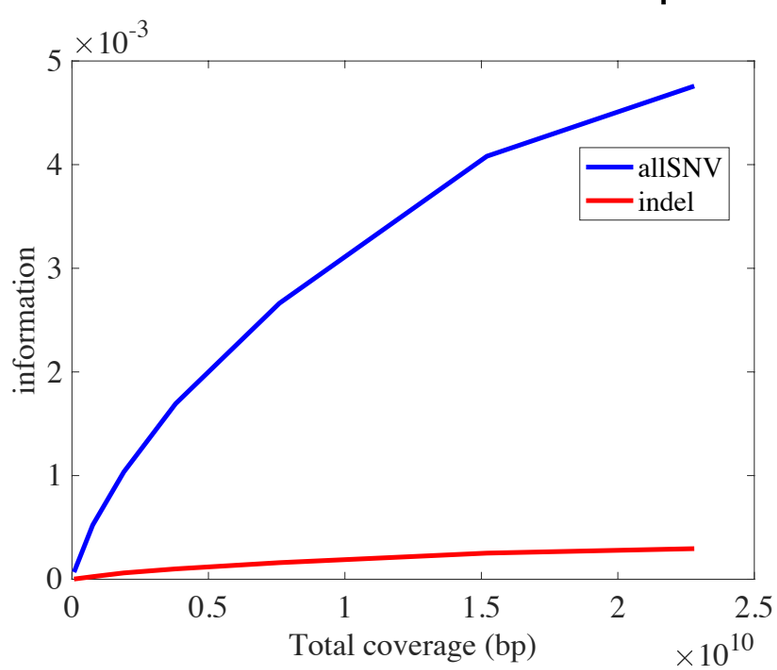
Privacy preserving file format

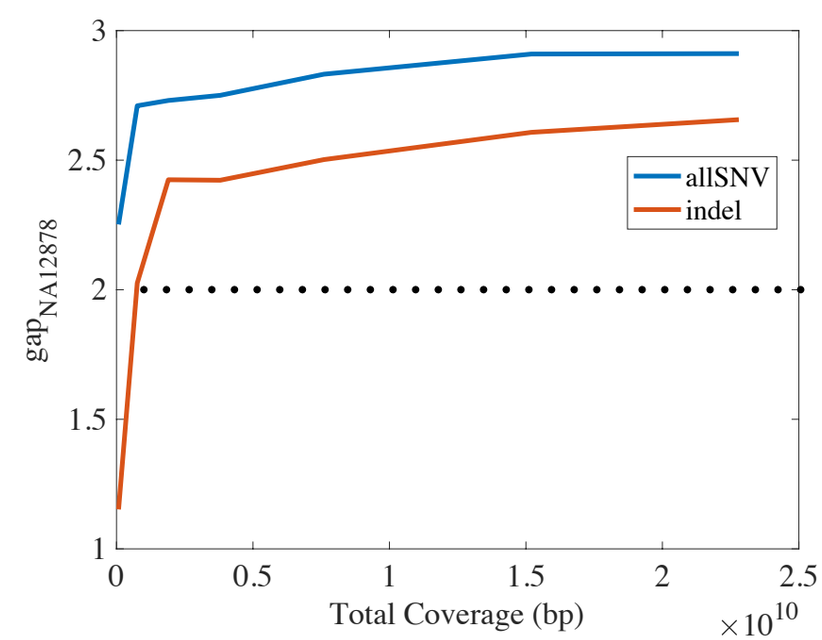
Option 1: MRF Format - For RNA-Seq



Problems:

- Indels can be inferred from the split reads





Individual is extremely vulnerable

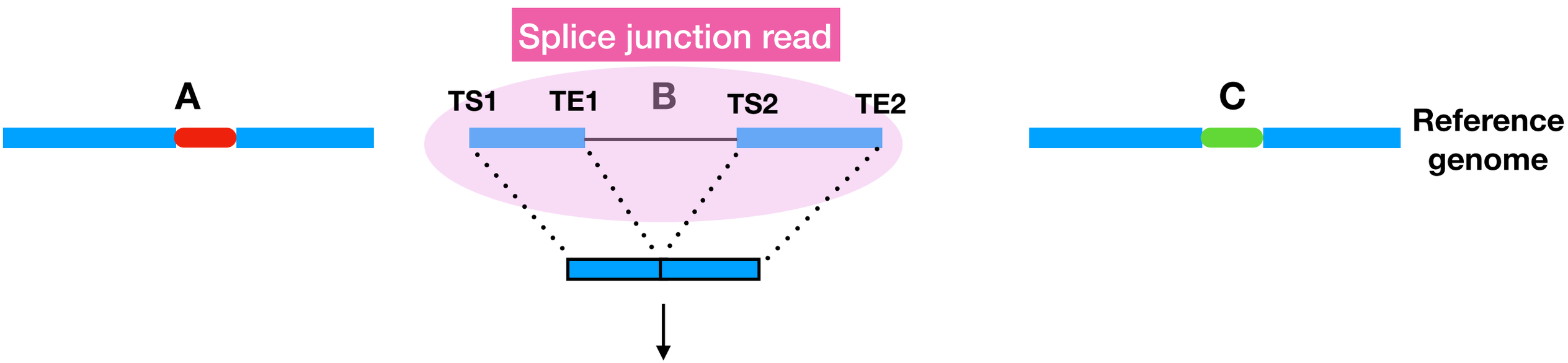
How can we secure this data?

Privacy preserving file format

Option 2: MRF Format Re-visited

As it turns out STARR prints deletions differently than split reads in a bam file

A	PRESLEY_0005:1:3:8474:3404#0	163	1	900678	255	40M4D36M
B	SINATRA_0006:7:66:10377:10458#0	147	1	708455	1	33M1063N43M
C	SINATRA_0006:7:64:17291:4457#0	83	1	567603	3	43M1I32M



Chr n: +/- : TS1—TE1:TS2—TE2

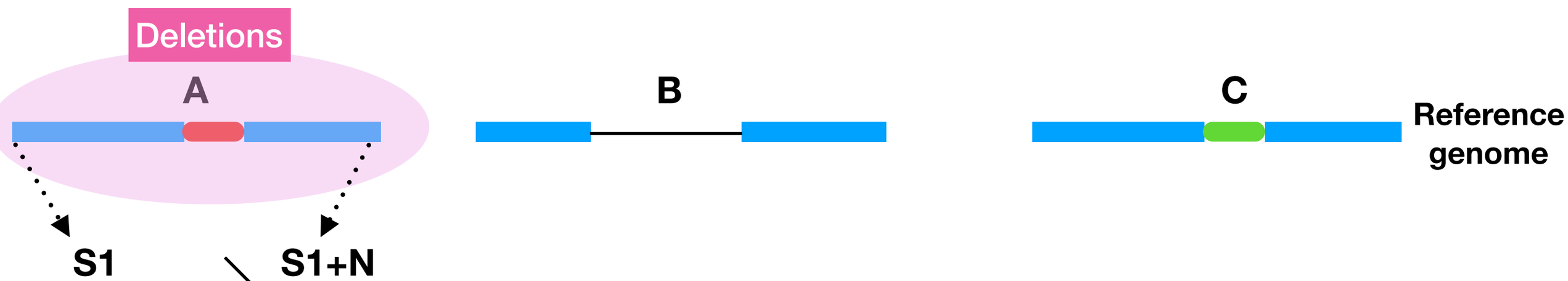
How can we secure this data?

Privacy preserving file format

Option 2: MRF Format Re-visited

As it turns out STARR prints deletions differently than split reads in a bam file

A	PRESLEY_0005:1:3:8474:3404#0	163	1	900678	255	40M4D36M
B	SINATRA_0006:7:66:10377:10458#0	147	1	708455	1	33M1063N43M
C	SINATRA_0006:7:64:17291:4457#0	83	1	567603	3	43M1I32M



N: read length

Chr n: +/- : S1—S1+N

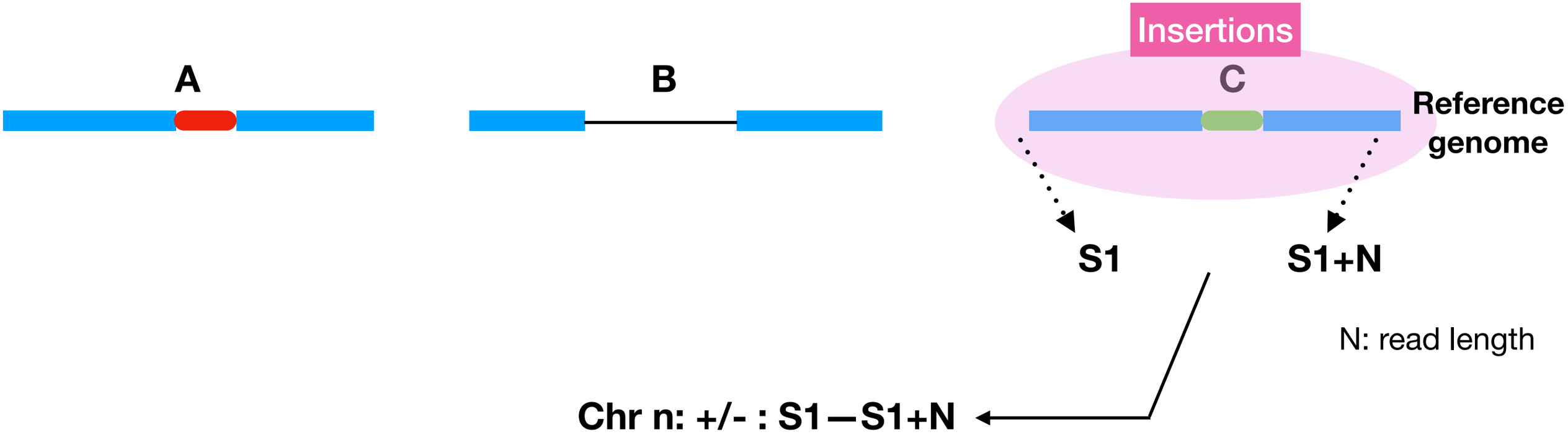
How can we secure this data?

Privacy preserving file format

Option 2: MRF Format Re-visited

As it turns out aligners prints indels differently than split reads in a bam file

A	PRESLEY_0005:1:3:8474:3404#0	163	1	900678	255	40M4D36M
B	SINATRA_0006:7:66:10377:10458#0	147	1	708455	1	33M1063N43M
C	SINATRA_0006:7:64:17291:4457#0	83	1	567603	3	43M1I32M



How can we secure this data?

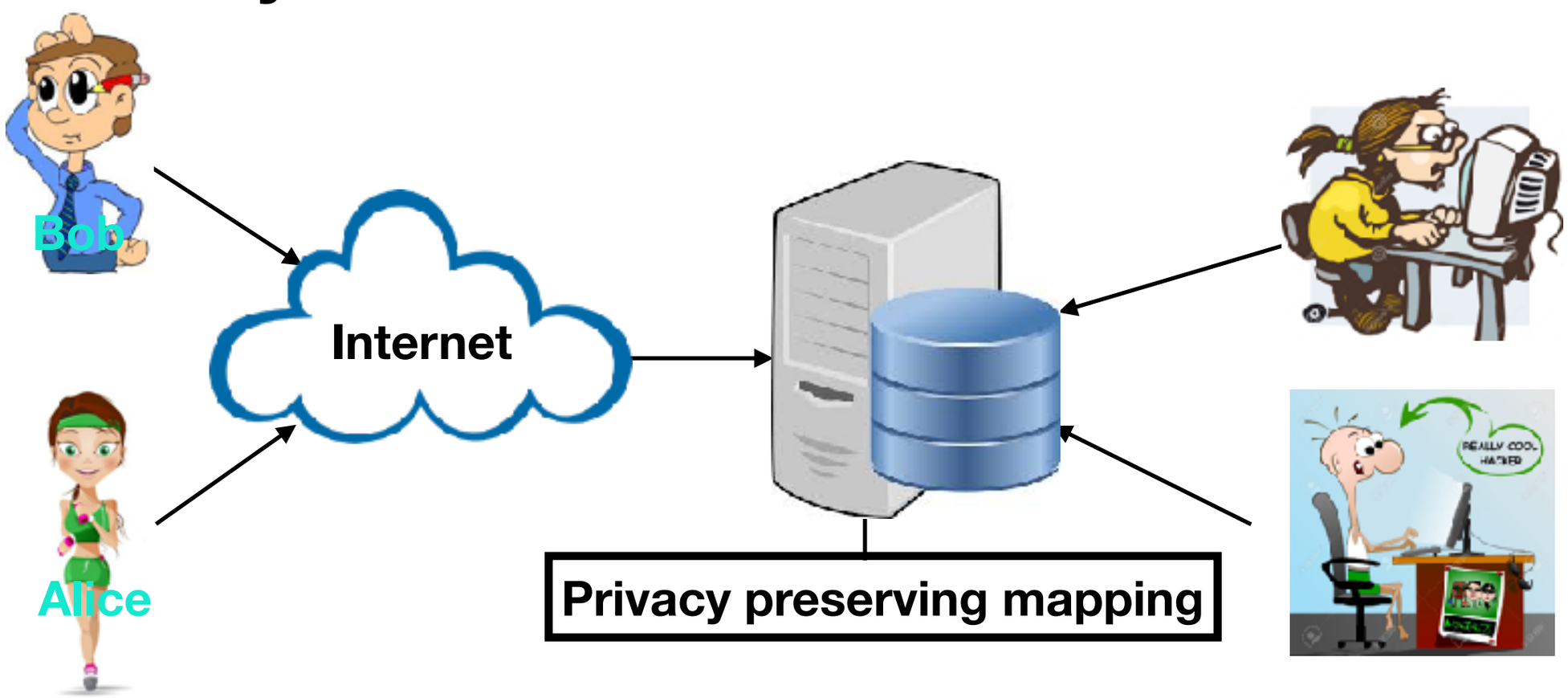
Privacy preserving file format

Option 3: GMZ Format (MRF+privateKey)

Objectives:

- Keep the public data light (small file size)
- Keep the private data light
- Minimize the information leakage
- Maximize the utility

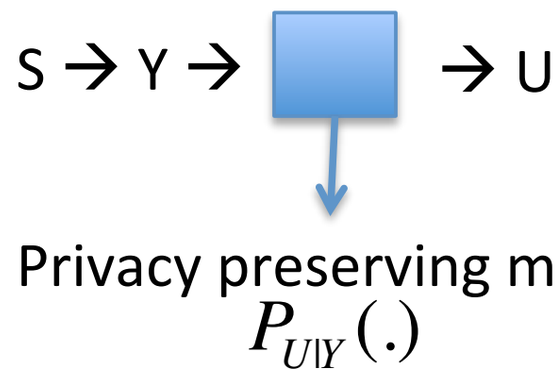
Privacy in traditional data science sense…



Privacy in traditional data science sense…

<u>Privacy preserving mapping</u>

S → set of variables that should remain private
Y → set of measurements that S can be inferred
U → distorted version of Y



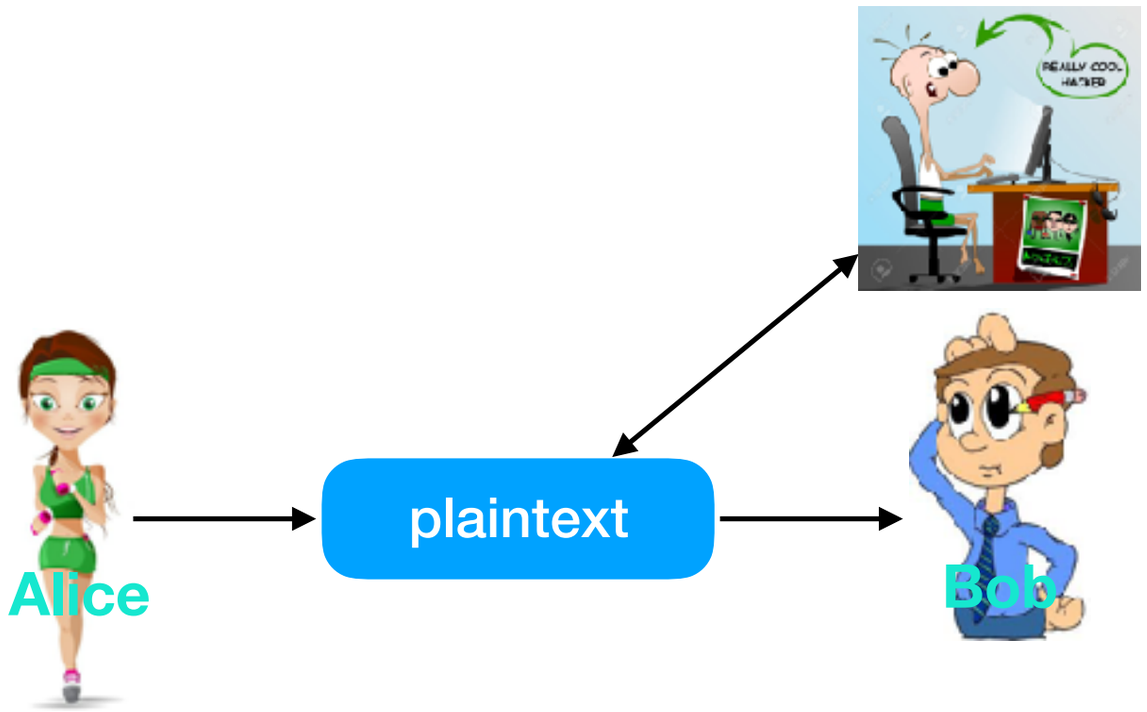
Calmon and Fawaz,2012

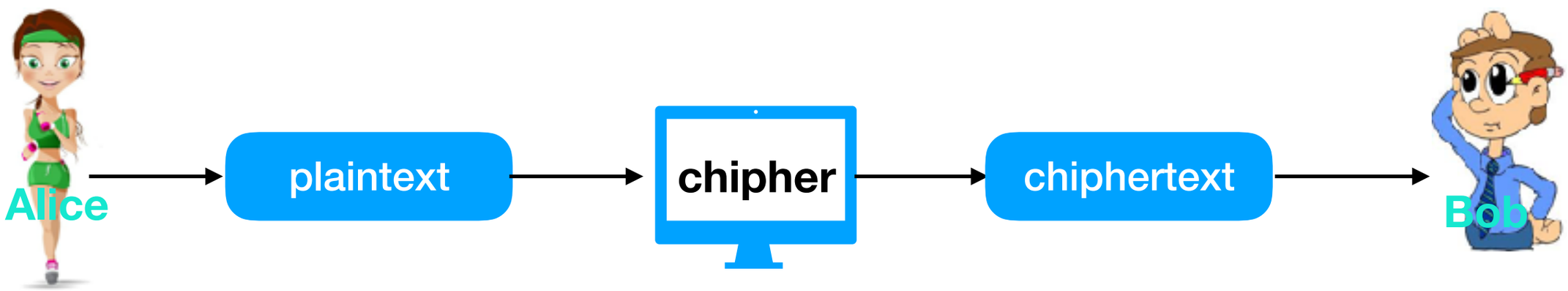
- **increase utility**
- **decrease privacy risk**

How can we secure this data?

Privacy preserving file format

Option 3: GMZ Format (MRF+private-key)

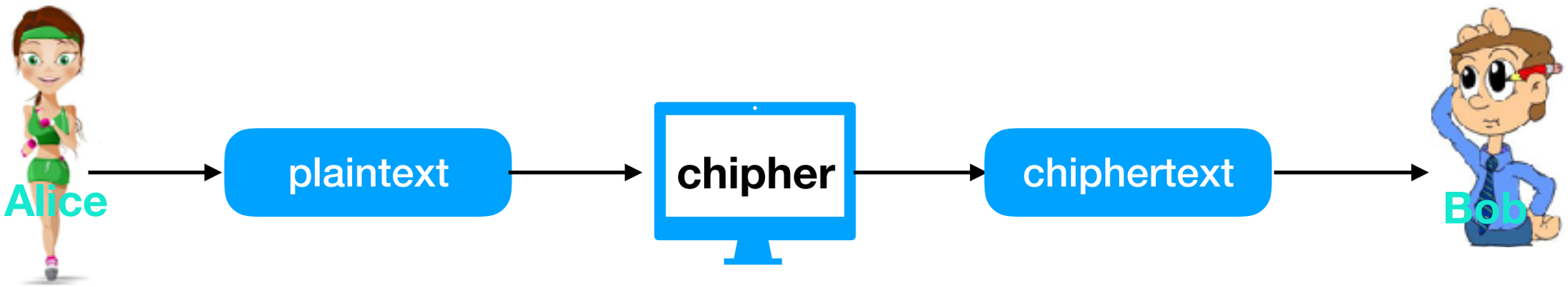




How can we secure this data?

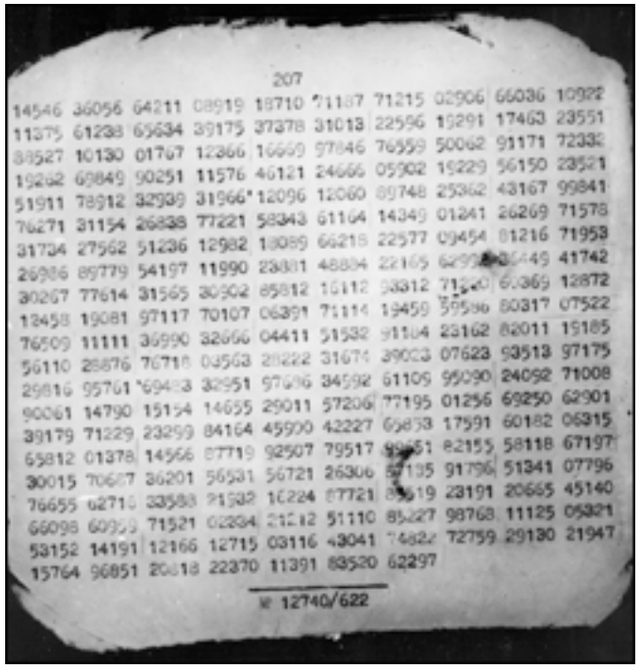
Privacy preserving file format

Option 3: GMZ Format (MRF+private-key)



m = plaintext
c = chipper text
E_k = encryption chipper, k is a cyrptographic key
$D_k = E_k^{-1}$ = decryption chipper

$c = E_k(m)$
$D_k(c) = D_k(E_k(m)) = m$



KGB ciphertext found in a hollow nickel in Brooklyn in 1953

How can we secure this data?

Privacy preserving file format

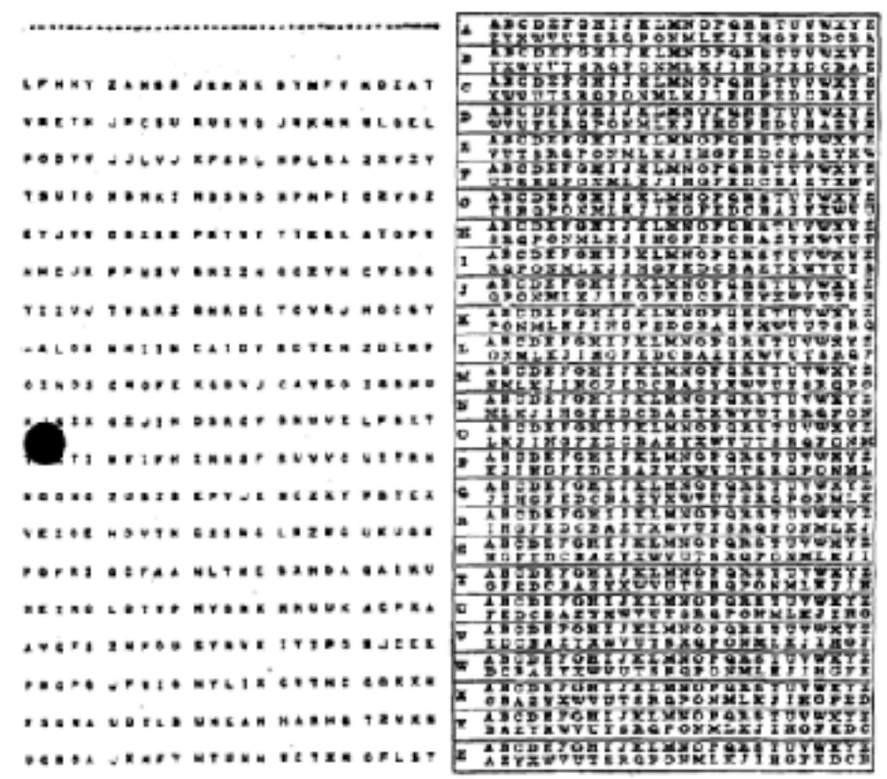
Option 3: GMZ Format (MRF+private-key)

An example encryption technique: Vernam Chipper (one-time pad)

- Information-theoretically secure - impossible to crack
  - Used to store highly sensitive data - NSA uses one-time pad

- Based on exclusive-or (XOR, $\oplus$)
  - $x \oplus y$ is true when exactly one of x and y is true
  - $x \oplus y$ is false when x and y are both true or both false
- $c = m \oplus k$ and $m = c \oplus k$

$$
\begin{aligned}
D_k(E_k(m)) &= c \oplus k \\
&= (m \oplus k) \oplus k \\
&= m \oplus (k \oplus k) \\
&= m \oplus 0 \\
&= m
\end{aligned}
$$



How can we secure this data?

An example encryption technique: Vernam Chipper (one-time pad)

$_k$ = sum of numerical values of alphabet
(A=0, B=1, C=2, …, Z=26 -> $_k$=26)
Let's say Alice wants to say "HELLO" to Bob

```
 H        E        L        L        O  message
   7 (H)    4 (E)   11 (L)   11 (L)   14 (O) message
+ 23 (X)   12 (M)    2 (C)   10 (K)   11 (L) key
= 30       16       13       21       25     message + key
=  4 (E)   16 (Q)   13 (N)   21 (V)   25 (Z) (message + key) mod 26
       E        Q        N        V        Z  → ciphertext
```

Bob converts chipher text "EQNVZ"

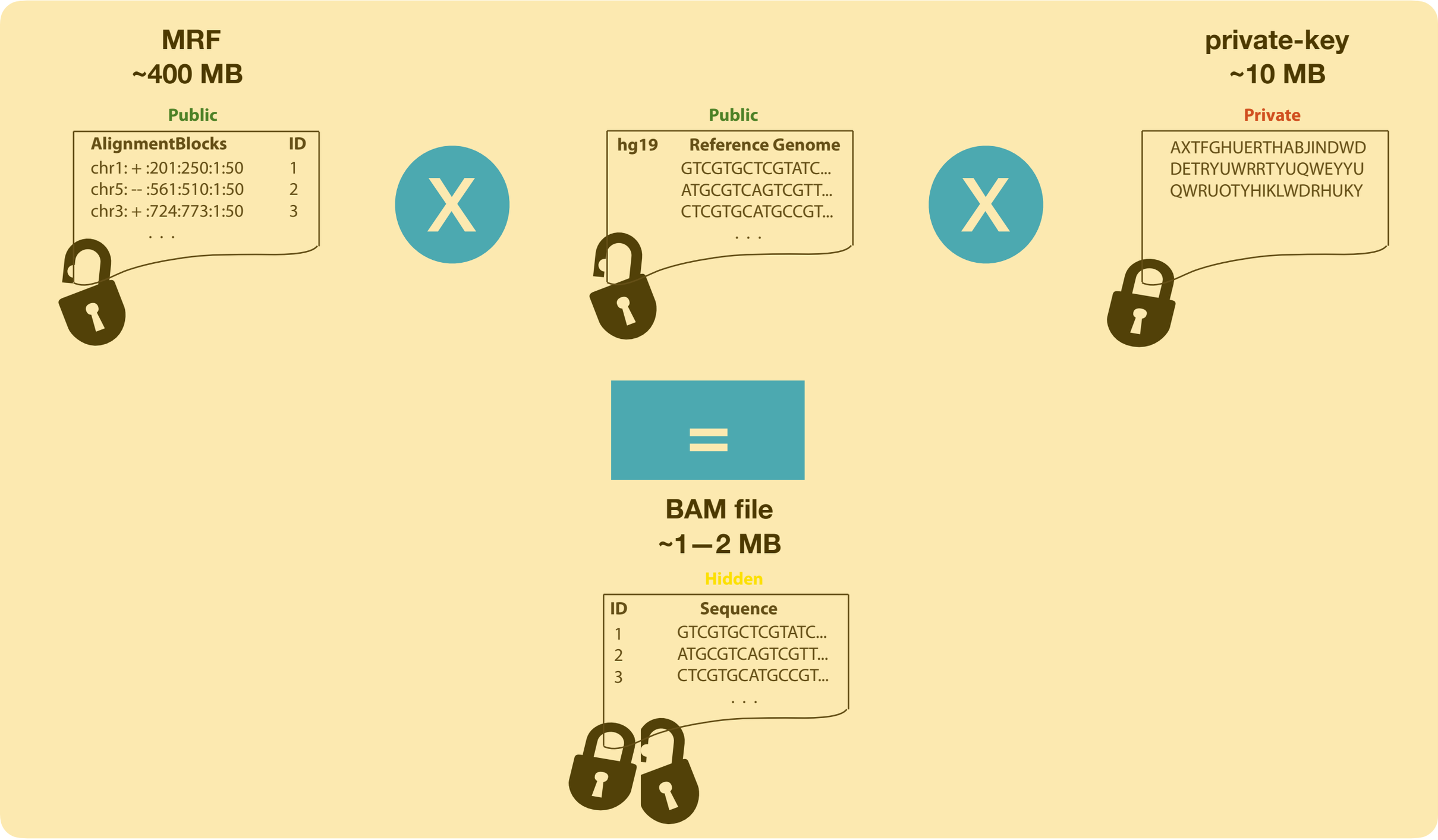
```
   E        Q        N        V        Z  ciphertext
    4 (E)   16 (Q)   13 (N)   21 (V)   25 (Z) ciphertext
-  23 (X)   12 (M)    2 (C)   10 (K)   11 (L) key
= -19        4       11       11       14     ciphertext – key
=   7 (H)    4 (E)   11 (L)   11 (L)   14 (O) ciphertext – key (mod 26)
        H        E        L        L        O  → message
```

Problem: Key has to be same size as the message
For a 2GB bam file, we need to create 2GB of key file
We might as well lock the bam file
BUT

How can we secure this data?

Privacy preserving file format

Option 3: GMZ Format (MRF+private-key)



Summary

- There is information leakage even in low coverage functional genomics data
  - Enough to identify individual in a panel
  - Sensitive phenotype information can be inferred from the leaked information
- We have developed an information theoretic framework to quantify the information leakage
- We can predict the leaked information by using the depth, breadth and the bias of the sequencing experiment
- We have improved our existing privacy preserving file formats to prevent the private information leakage

Future Directions

- Working on a linking attack using gene expression extremity + loss of function mutation from 1000genomes
- Accurate genotyping/somatic mutation load using Hi-C data for samples we don't have WGS data - especially for tissue samples from one individual or cells in different developmental stages from same donor
    - EN-TEx data can be utilized
- Private information leakage in functional genomics data in terms of SVs. I ran CNVnator on Hi-C data for deletions (NA12878)
    - Which in turn can lead to SV calling from Hi-C data (deletions and using diagonal for tandem repeats)
    - Might also lead to better understanding of SV mechanism with the underlying 3D genome architecture

WGS	
Total number of deletions	227
Total bp of deletions	26,474,150

Hi-C	
Total number of deletions	804
Total bp of deletions	42,710,150

Intersect	
Total number of deletions	82
Total bp of deletions	250,155,00

True Positive	95%
False Positive	41%

Acknowledgments

Mark

NA12878

Arif

Joel

Fabio

Timur

Sarah

Lilly

Sushant

Prashant

Jing

Donghoon

Jason