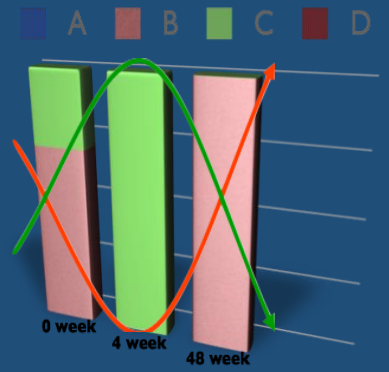


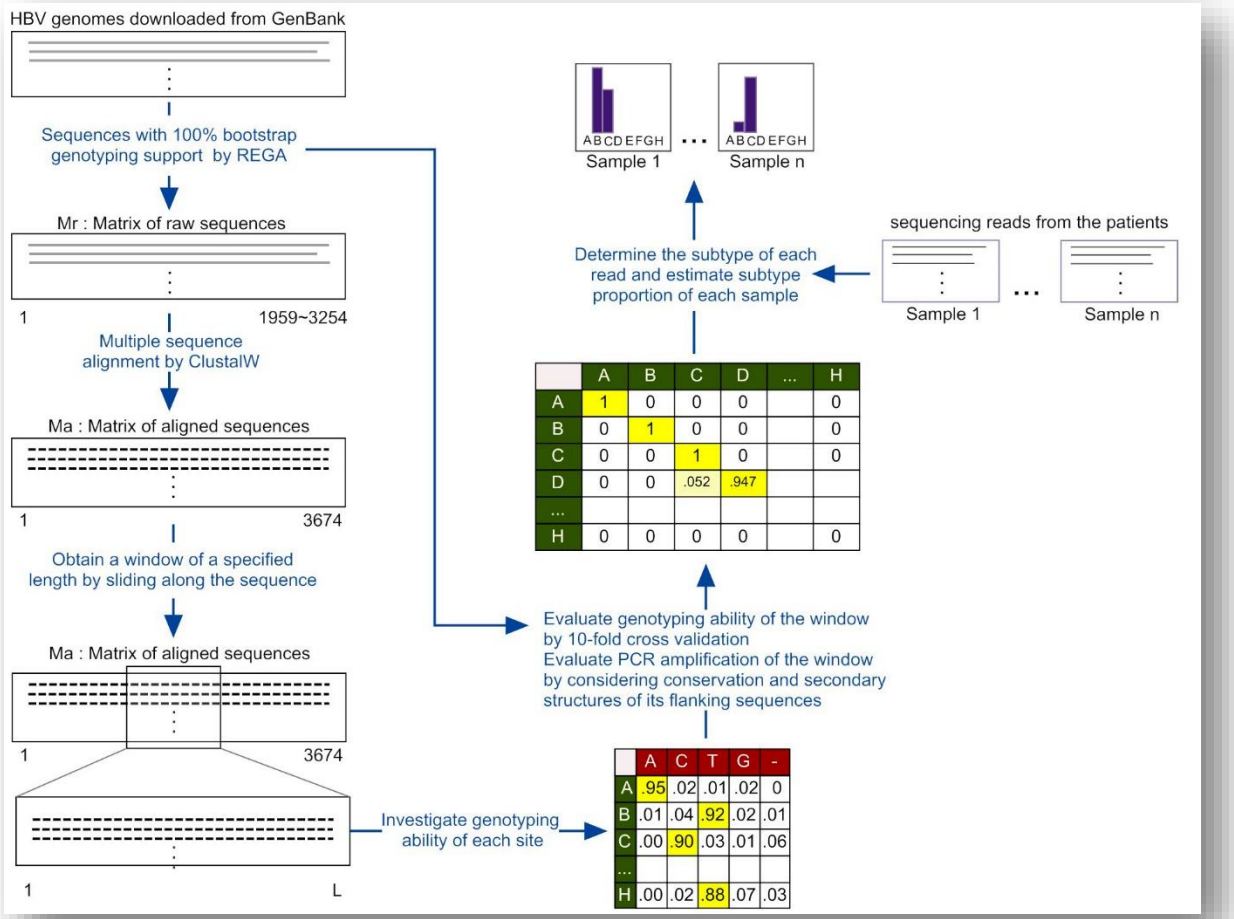
From local to global : a new perspective of Hepatitis B Virus Genotyping framework

指导老师 ：吴家睿 梁 治

答辩人 ：周 鑫 SA11008910

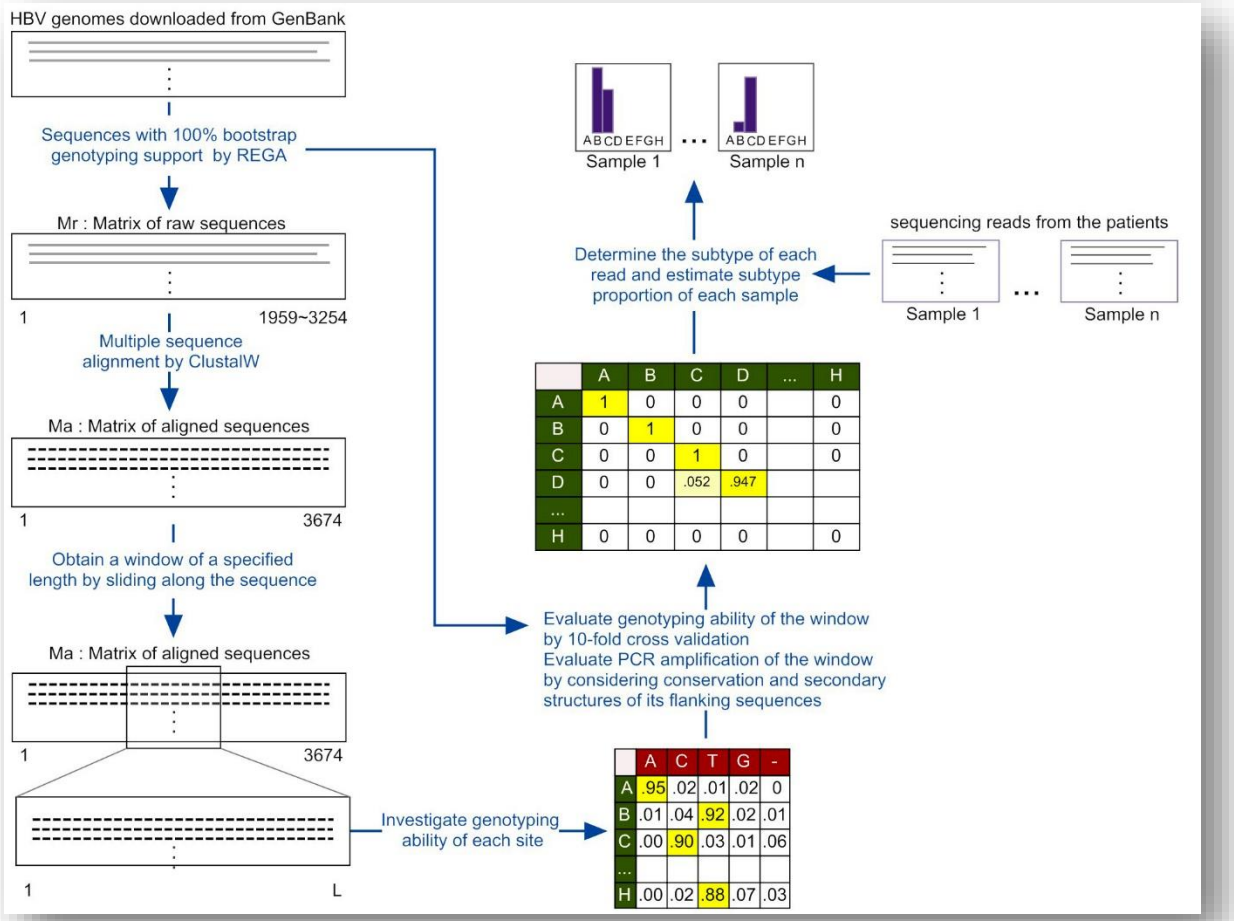
Outline of ShwinGen

1. Background and Motivation
2. Training Set Retrieve …
    1. NCBI Data Retrieve
    2. REGA Genotype
3. Correlation Position Determination Via ICA …
    1. ICA among Different Loci of Genome
    2. ICA among Different Sequences
4. Genotyping Short Windows Selection
5. Barcode Design
6. Next Generation Sequencing of Short Windows …
    1. Control Template Design
    2. Noise Removal
7. HBV Subtype Inference
8. Analysis between Subtype Shift and Drug Therapy



Outline of ShwinGen

1. **Background and Motivation**
2. Training Set Retrieve …
    1. NCBI Data Retrieve
    2. REGA Genotype
3. Correlation Position Determination Via ICA …
    1. ICA among Different Loci of Genome
    2. ICA among Different Sequences
4. Genotyping Short Windows Selection
5. Barcode Design
6. Next Generation Sequencing of Short Windows …
    1. Control Template Design
    2. Noise Removal
7. HBV Subtype Inference
8. Analysis between Subtype Shift and Drug Therapy

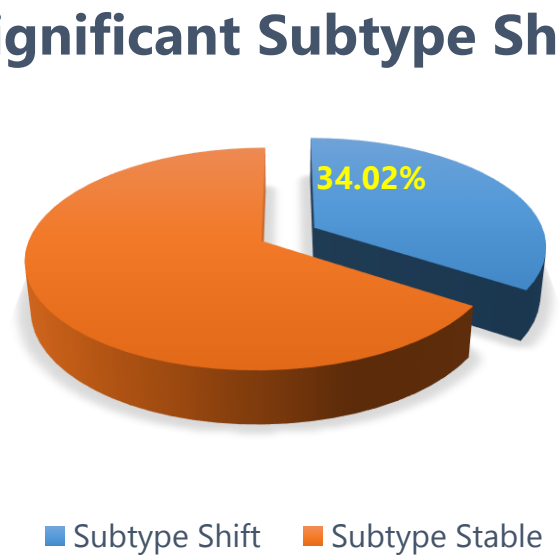


Background and Motivation

Features	Description	p-value
Age	30.3 (19,52)	
Gender(male/female)	32 / 6	
Alanine aminotransferase(IU/mL)		0.026 *
Pre-treatment	161.8(45-611)	
After-treatment	65.1 (16-451)	
HBV DNA (Log(copies/mL))		0.046 *
Pre-treatment	9.5 (7.3-11.1)	
After-treatment	7.3 (3.0-10.6)	
Hepatitis B e antigen (+/-)		0.024 *
Pre-treatment	38 / 0	
After treatment	20 /18	

Characteristics of the 38 ADV-treated chronic hepatitis B patients

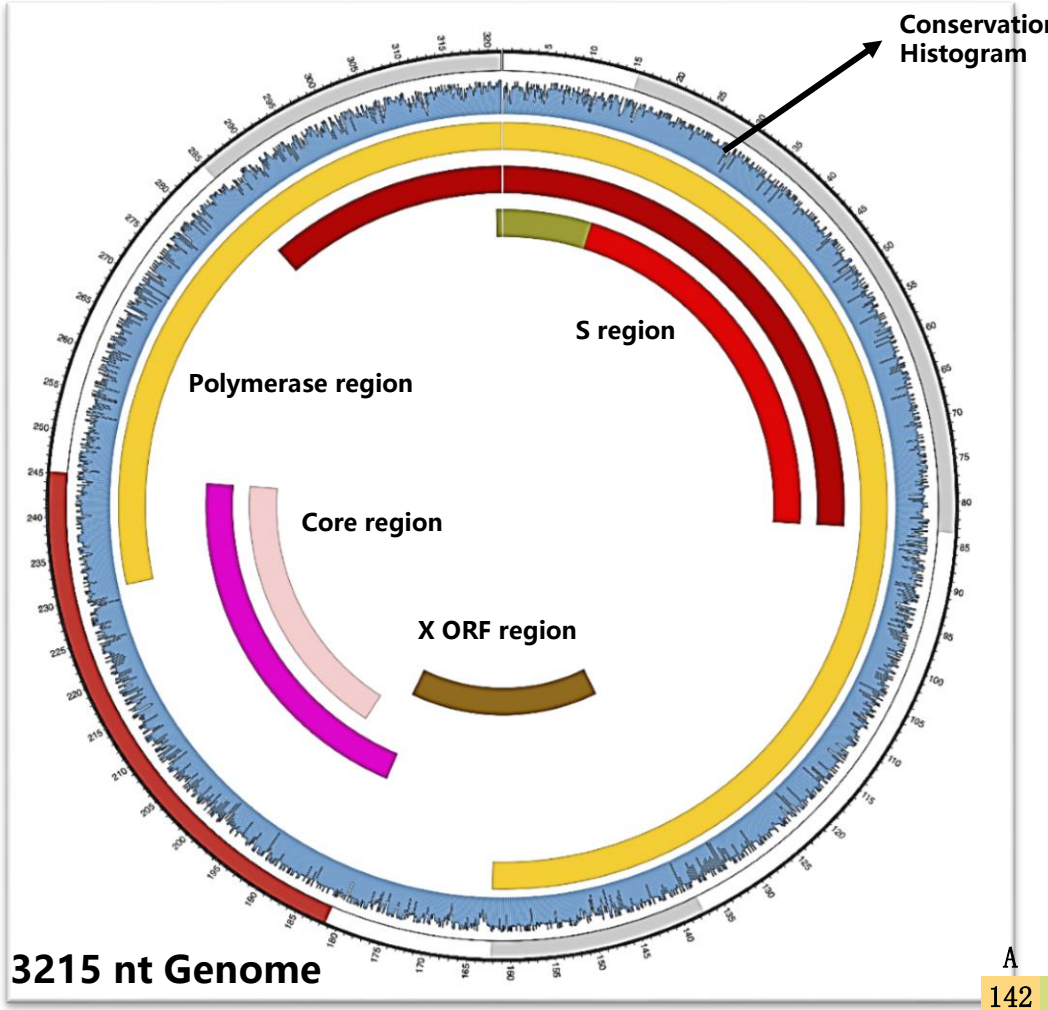
200 samples



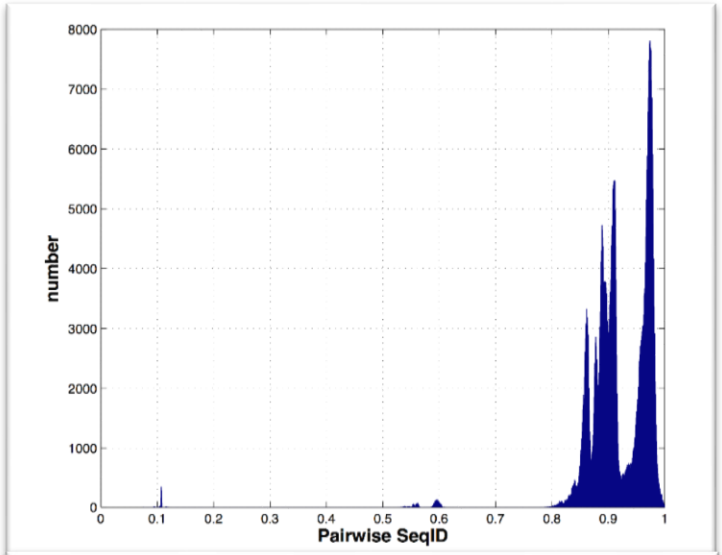
Significant Subtype shift have been found after AVD treatment

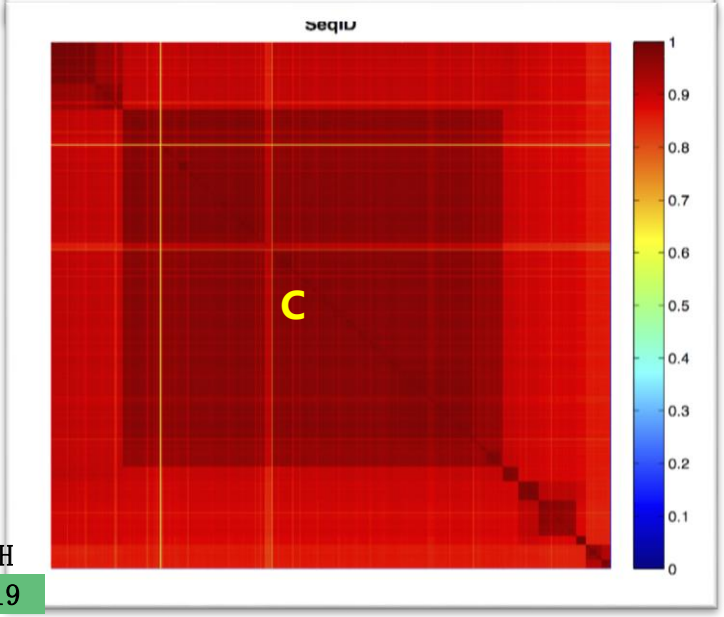
Background and Motivation

HBV Genome Structure … Totally overlapped and Conservation



Inter-Subtype Pattern:

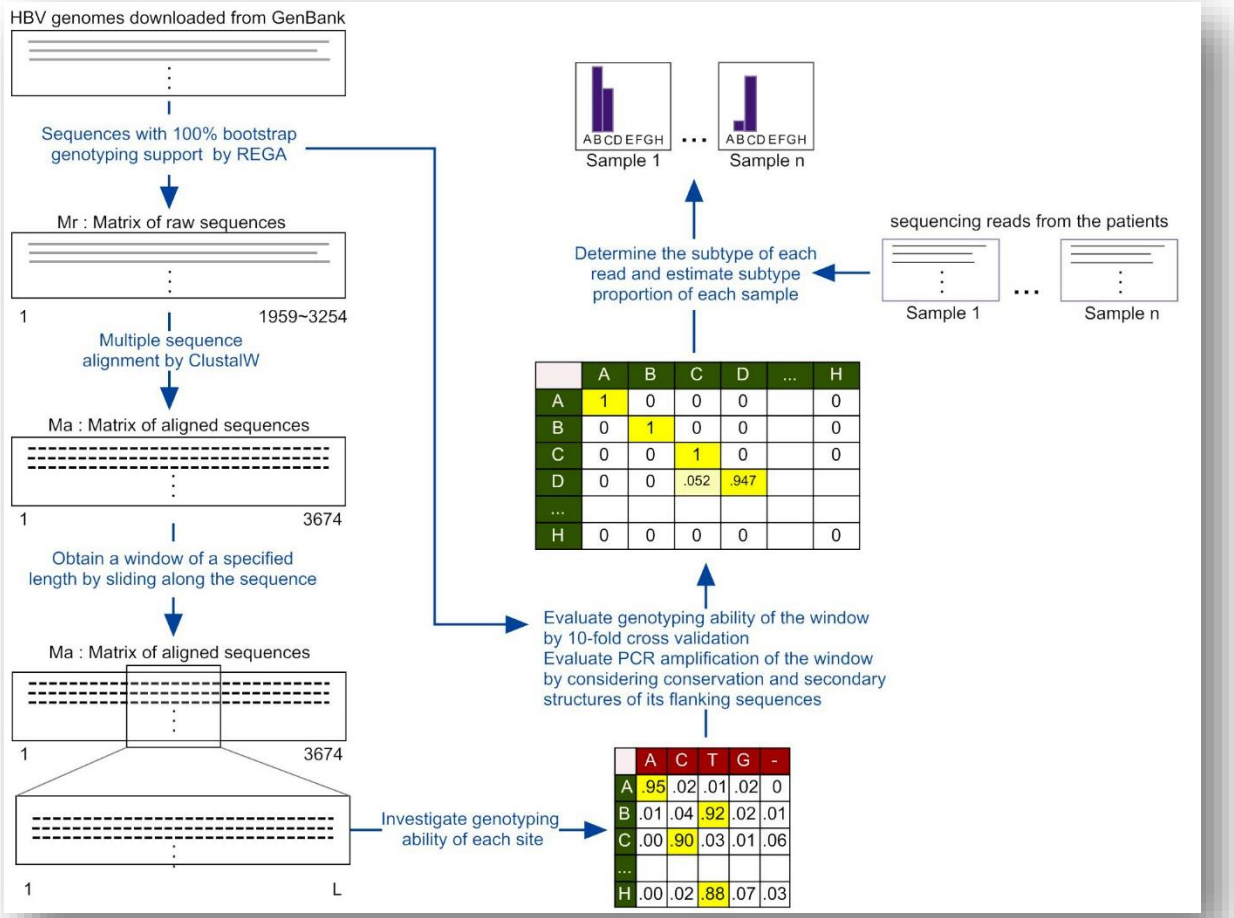




A	B	C	D	E	F	G	H
142	31	752	75	40	31	19	19

Outline of ShwinGen

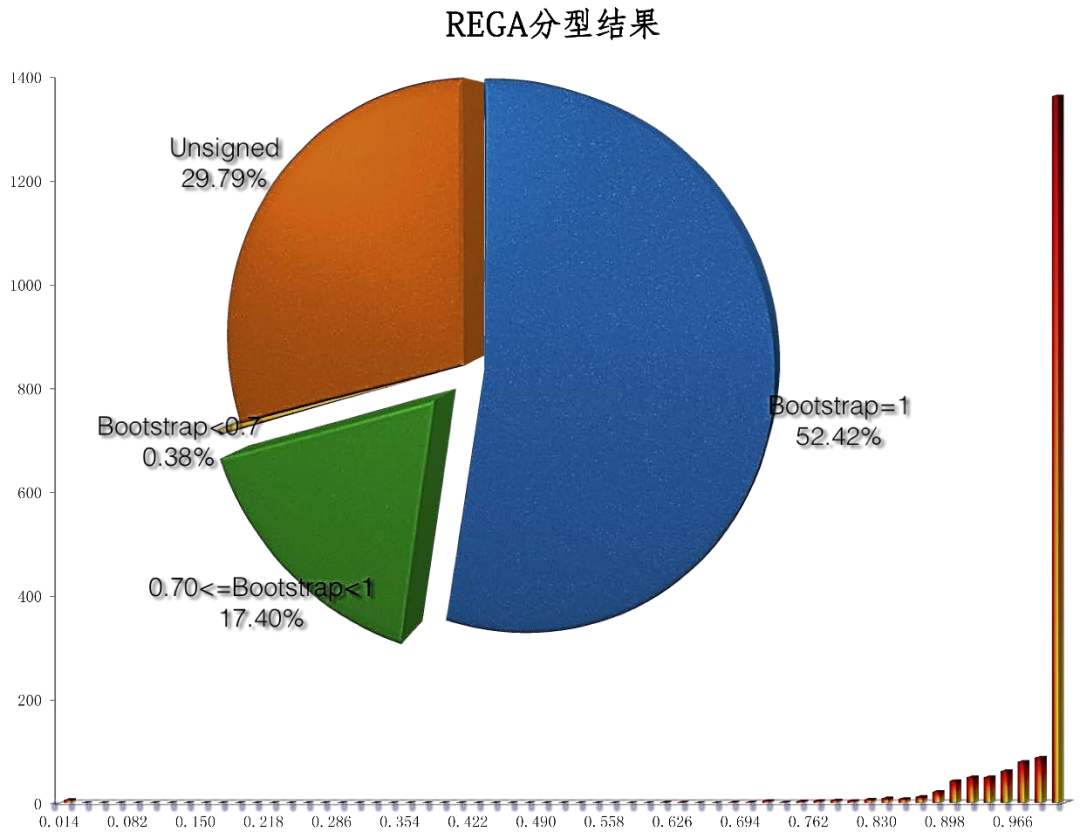
1. Background and Motivation
2. **Training Set Retrieve …**
    1. **NCBI Data Retrieve**
    2. **REGA Genotype**
3. **Correlation Position Determination Via ICA …**
    1. **ICA among Different Loci of Genome**
    2. **ICA among Different Sequences**
4. Genotyping Short Windows Selection
5. Barcode Design
6. Next Generation Sequencing of Short Windows …
    1. Control Template Design
    2. Noise Removal
7. HBV Subtype Inference
8. Analysis between Subtype Shift and Drug Therapy

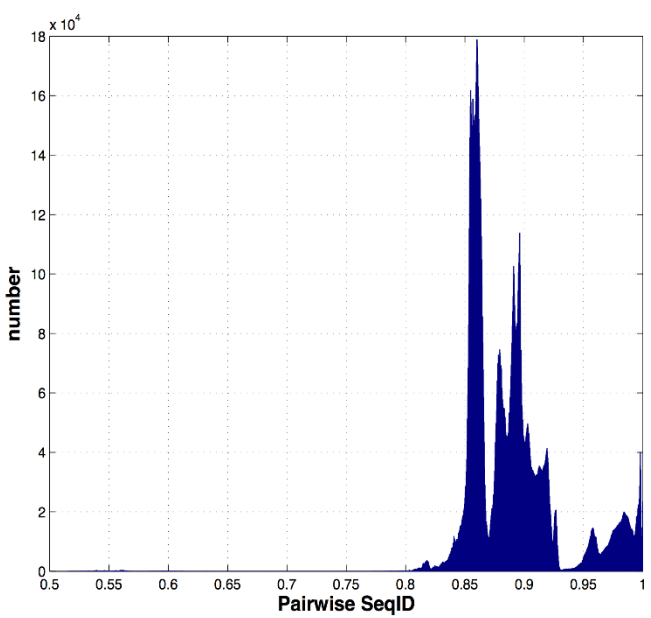


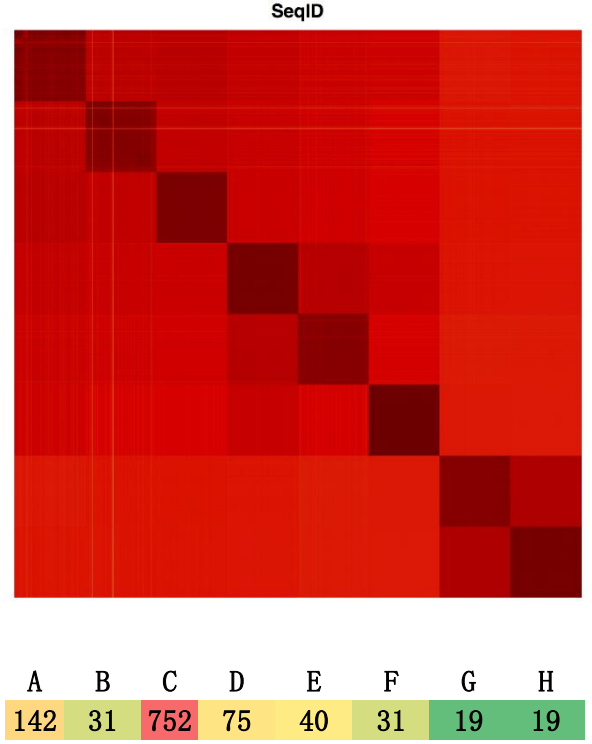
Independent Position Subsets

NCBI and REGA Retrieve …
1109 Standard Subtypable Sequences

SMOTE of Subtype Unbalanced Dataset …
To Balanced Dataset





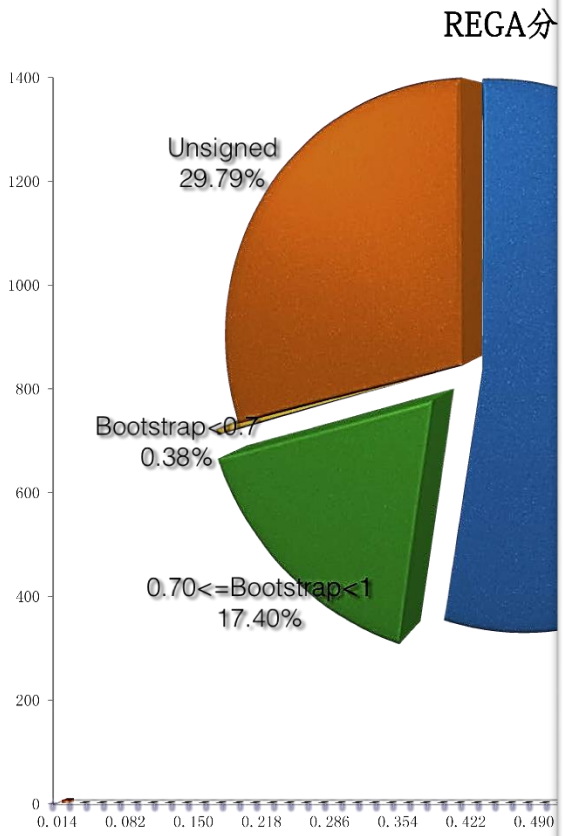


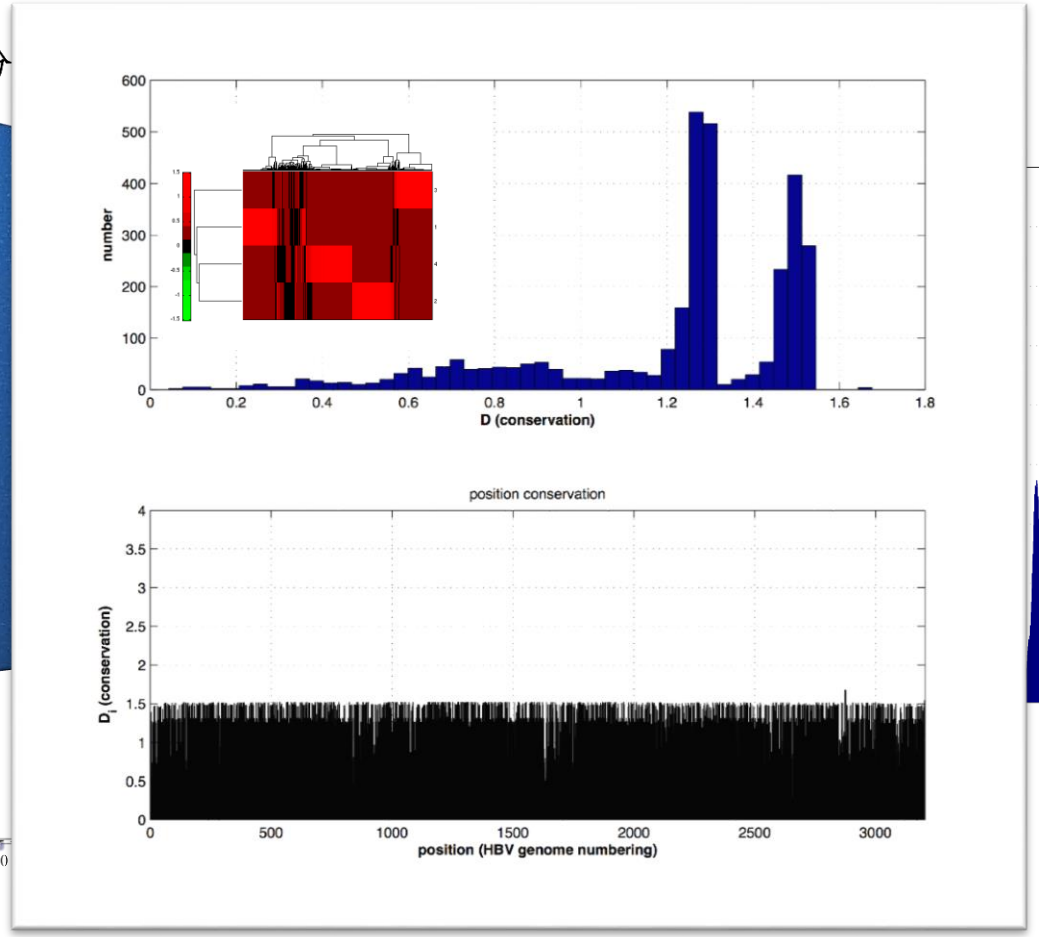
A	B	C	D	E	F	G	H
142	31	752	75	40	31	19	19

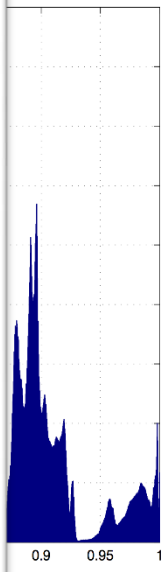
Independent Position Subsets

NCBI and REGA Retrieve …
1109 Standard Subtypable Sequences

SMOTE of Subtype Unbalanced Dataset …
To Balanced Dataset







Loci Conservation

inter-Position Correlation

$$C_{ij}^{ab} = f_{ij}^{ab} - f_i^{(a)} f_j^{(b)}$$

$$\hat{C}_{ij}^{ab} = \phi(D_i^{(a)})\phi(D_j^{(b)})\, C_{ij}^{ab} \qquad \phi(D) = \frac{\partial D}{\partial f}$$

$$\widehat{C_{ij}} = \phi_i \phi_j (\langle X3d_{si} X3d_{sj} \rangle_s - \langle X3d_{si} \rangle_s \langle X3d_{sj} \rangle_s)$$

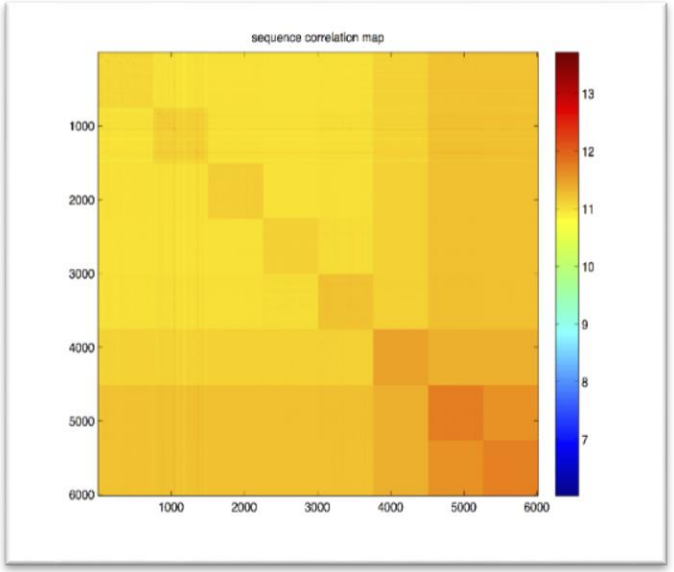
$\widehat{C_{ij}}$ means the correlation between each pair of loci i and loci j

Inter-positional correlation

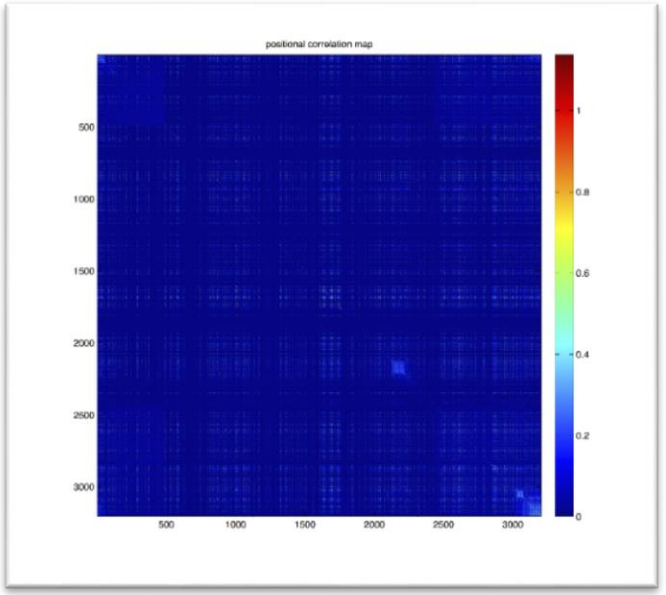
Considering KL entropy as change tradeoff

Correlation average of C project the 3D tensor Matrix to M x L (sequence and loci position) ***X***

$$\tilde{C} = \frac{\widetilde{X^T}\tilde{X}}{M} \qquad \text{and} \qquad \tilde{S} = \frac{\tilde{X}\widetilde{X^T}}{M}$$

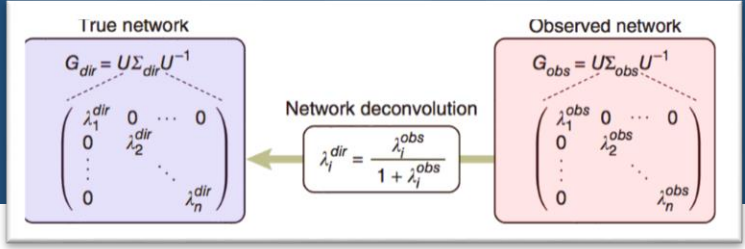


Inter-loci correlation



Inter-sequential correlation

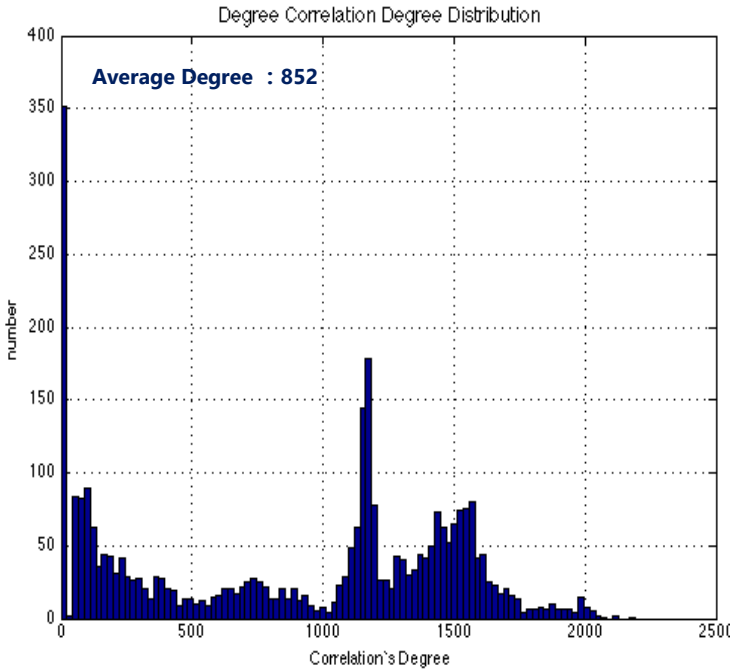
Correlation Network Cleanup

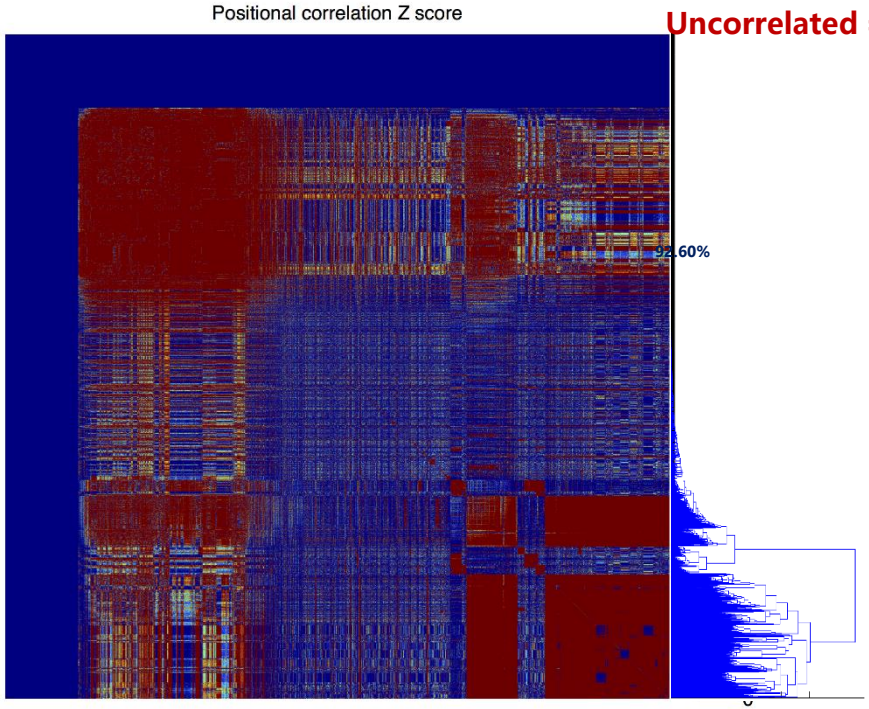


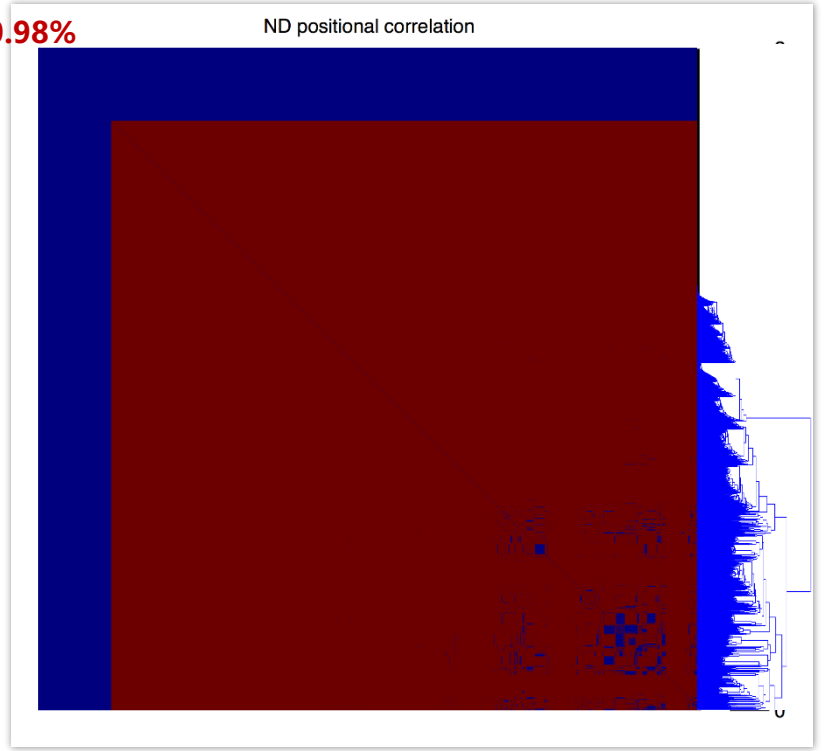
Correlation observed among different loci always contain many in-direct trans-effect

$$G_{dir} = G_{obs}(I + G_{obs})^{-1}$$

$$G_{obs} = G_{dir} + G_{dir}^2 + G_{dir}^3 \ldots$$

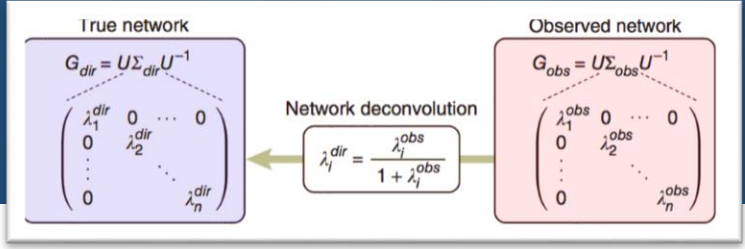






Inter-Positional Correlation is almost direct between each pairs of position 89.02% loci are directly correlated with each other

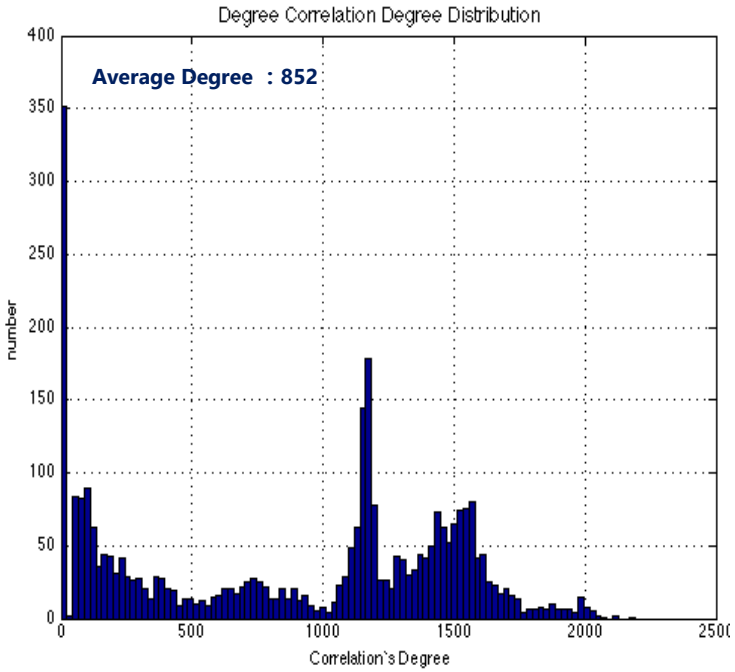
Correlation Network Cleanup

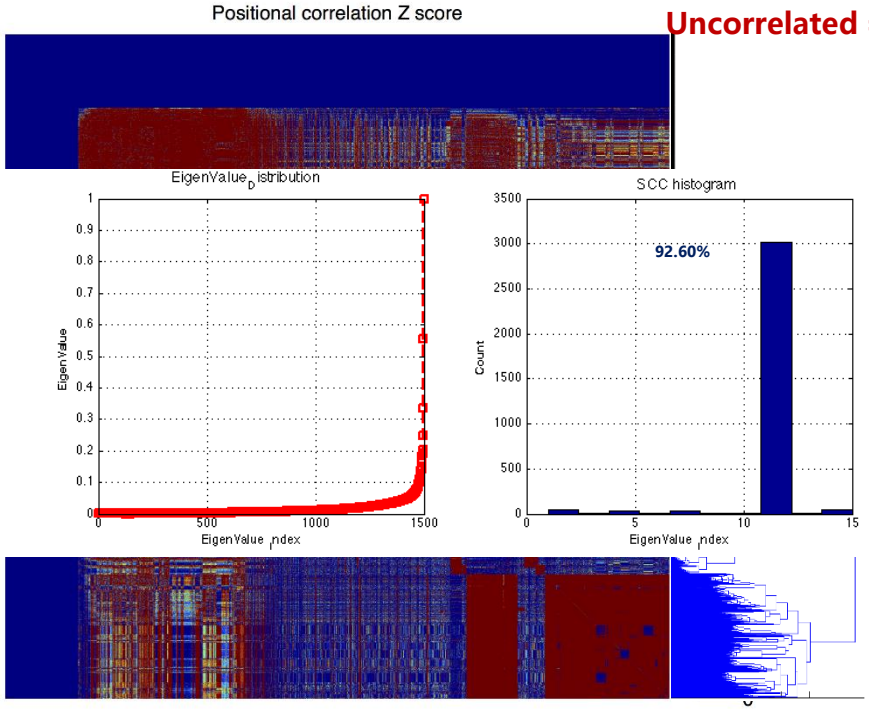


Correlation observed among different loci always contain many in-direct trans-effect

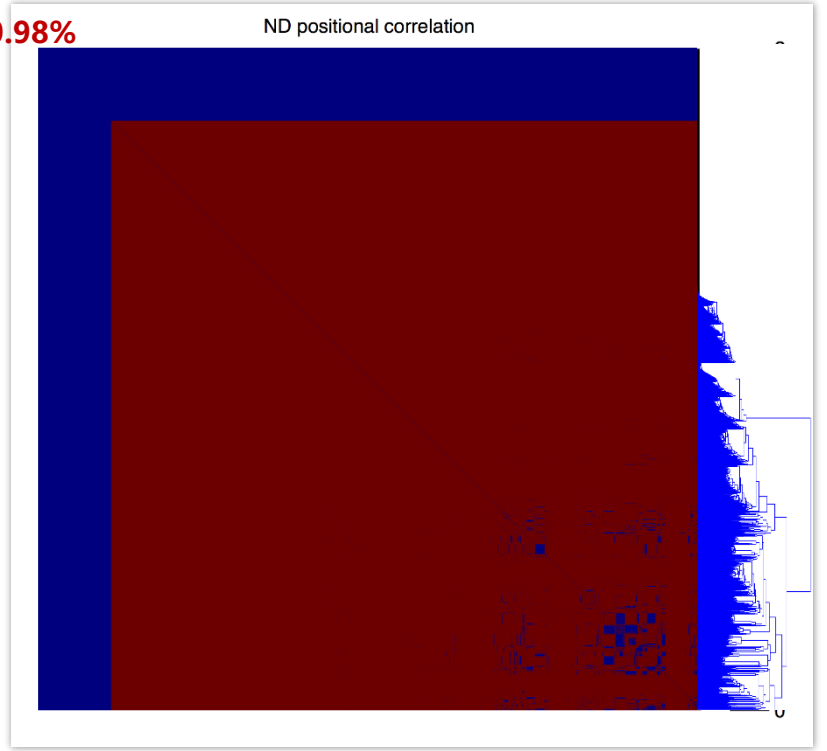
$$G_{dir} = G_{obs}(I + G_{obs})^{-1}$$

$$G_{obs} = G_{dir} + G_{dir}^2 + G_{dir}^3 \ldots$$





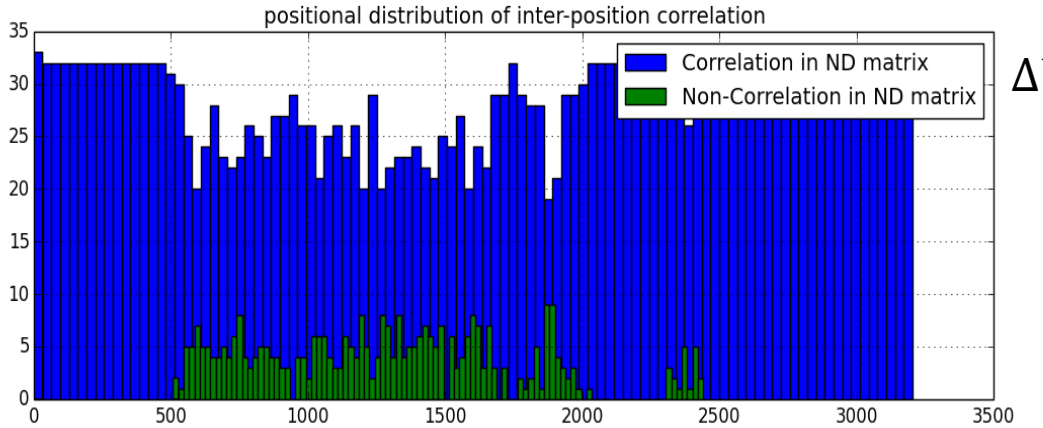
Uncorrelated == 10.98%



Inter-Positional Correlation is almost direct between each pairs of position 89.02% loci are directly correlated with each other

Blind Source of Correlation Matrix

352 uncorrelated loci histogram on HBV genome $S: Loci\ dependent$ $WZ=S\ From\ Z \rightarrow S$

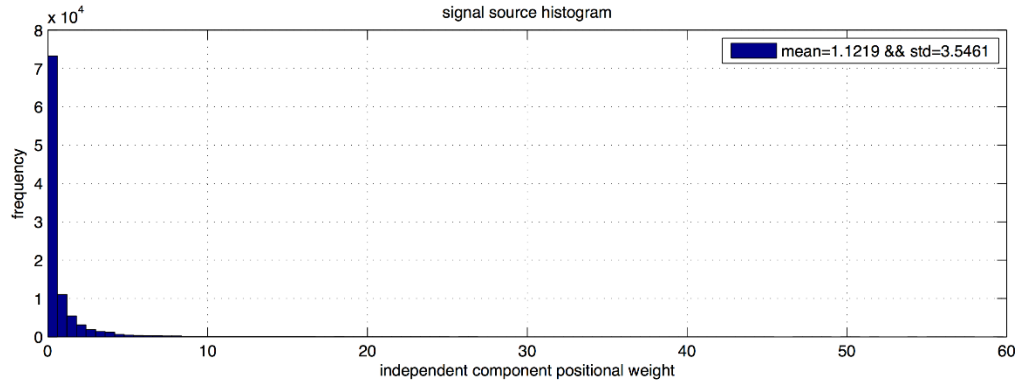


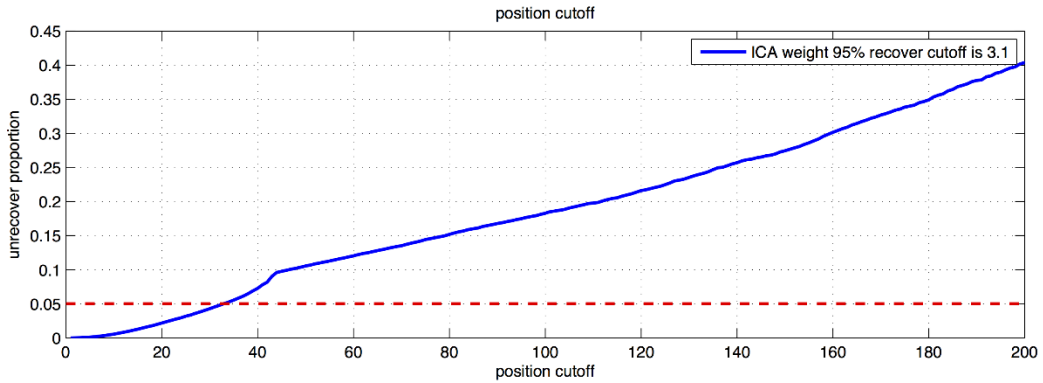
$$\Delta W = \varepsilon(I_k + (1 - f(-WZ)(WZ)^T)W \quad :$$

Blind Source of Correlation Matrix

352 uncorrelated loci histogram on HBV genome

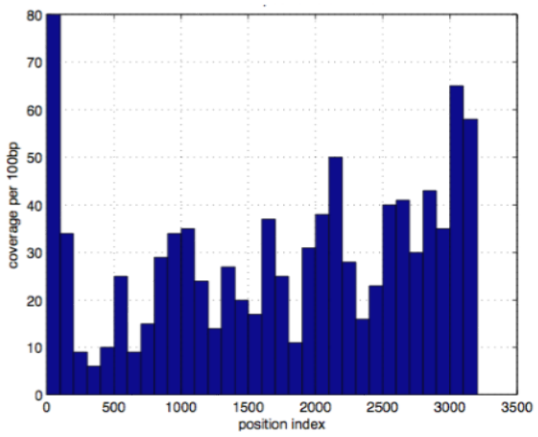
$S : Loci\ dependent$ $WZ=S\ From\ Z \rightarrow S$



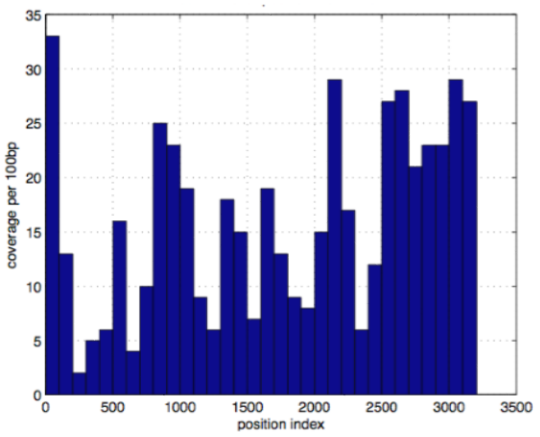


$$\Delta W = \varepsilon(I_k + (1 - f(-WZ)(WZ)^T)W \quad :$$

891 loci kept for 95% recovery

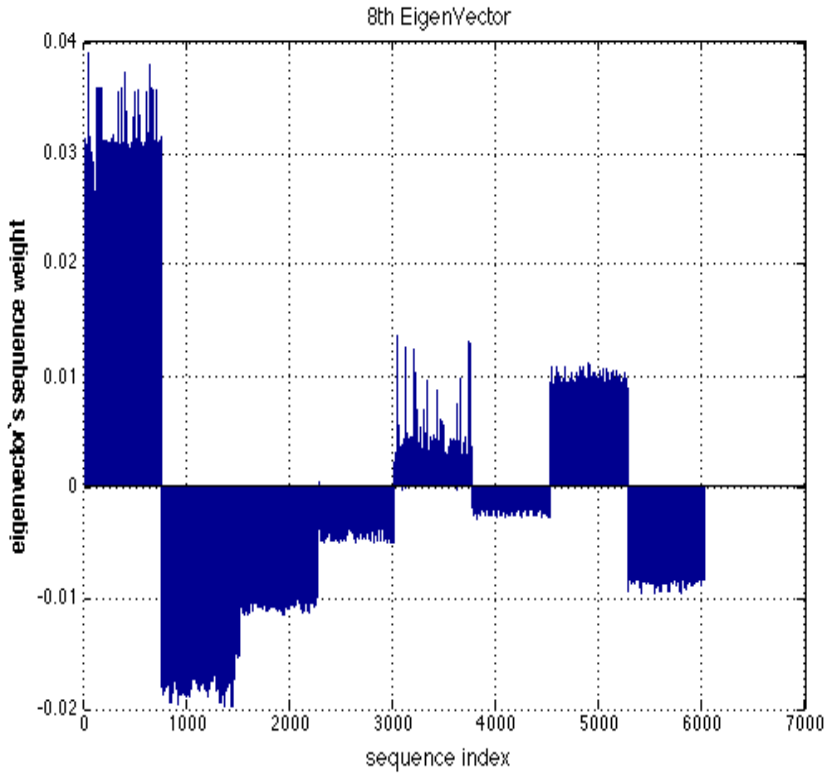


517 loci kept for 85% recovery



Subtype and Independent Component $\tilde{S} = \frac{\tilde{X}\tilde{X^T}}{M}$

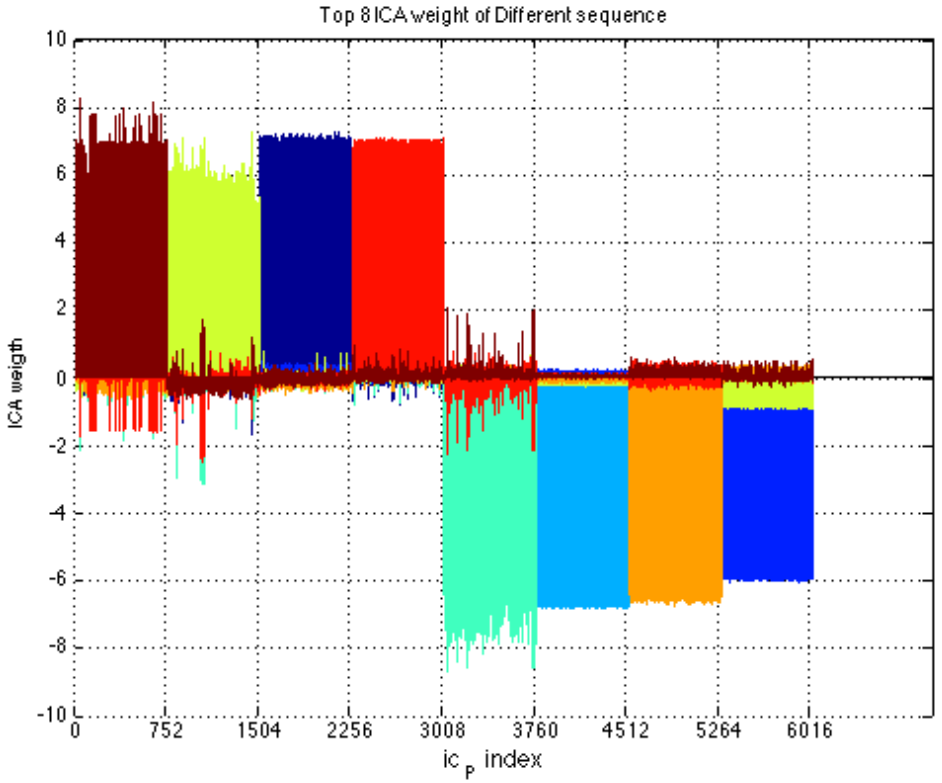
Ranking KL-diversity of each loci into sequence correlation matrix increase the performance of HBV subtype Category in ICA



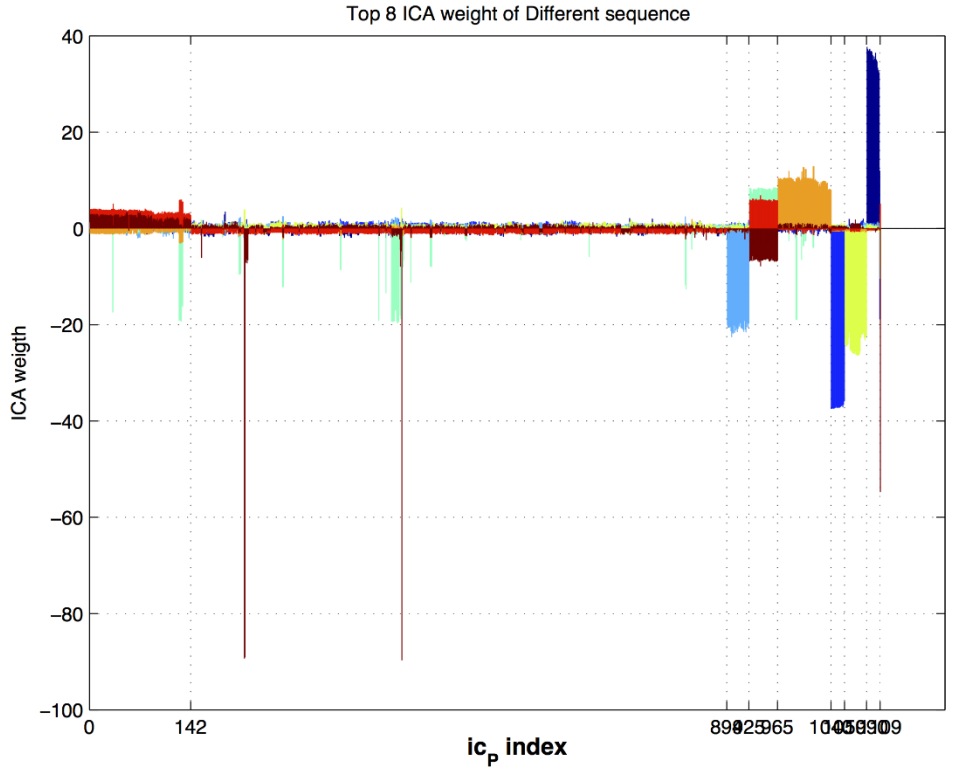
Sequence Eigen-vector

Subtype and Independent Component $\tilde{S} = \frac{\tilde{X}\tilde{X^T}}{M}$

Ranking KL-diversity of each loci into sequence correlation matrix increase the performance of HBV subtype Category in ICA



Standard HBV training set

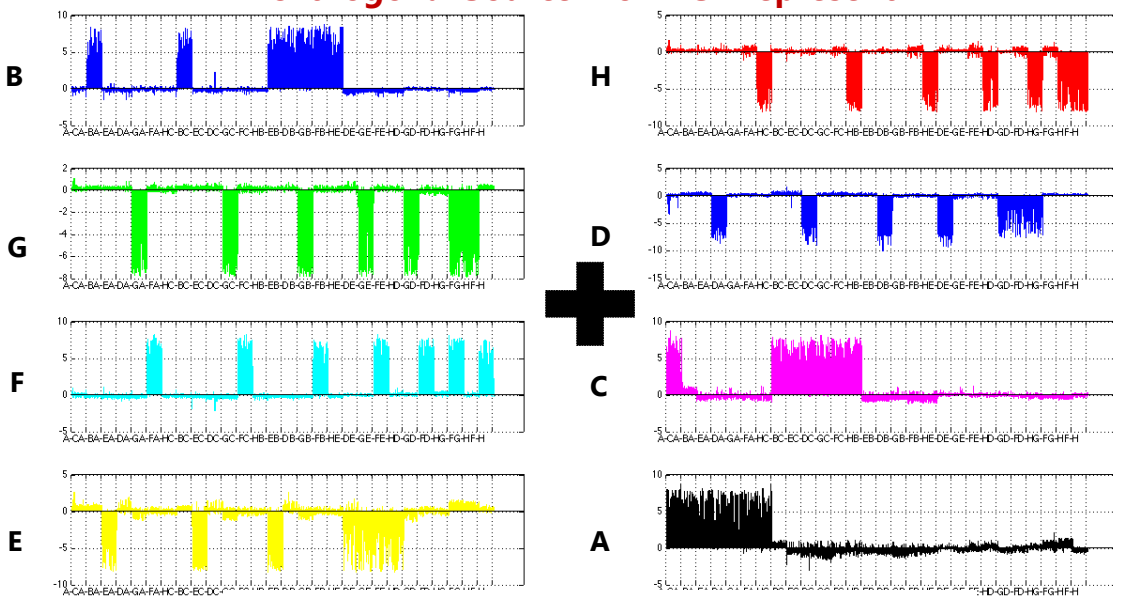


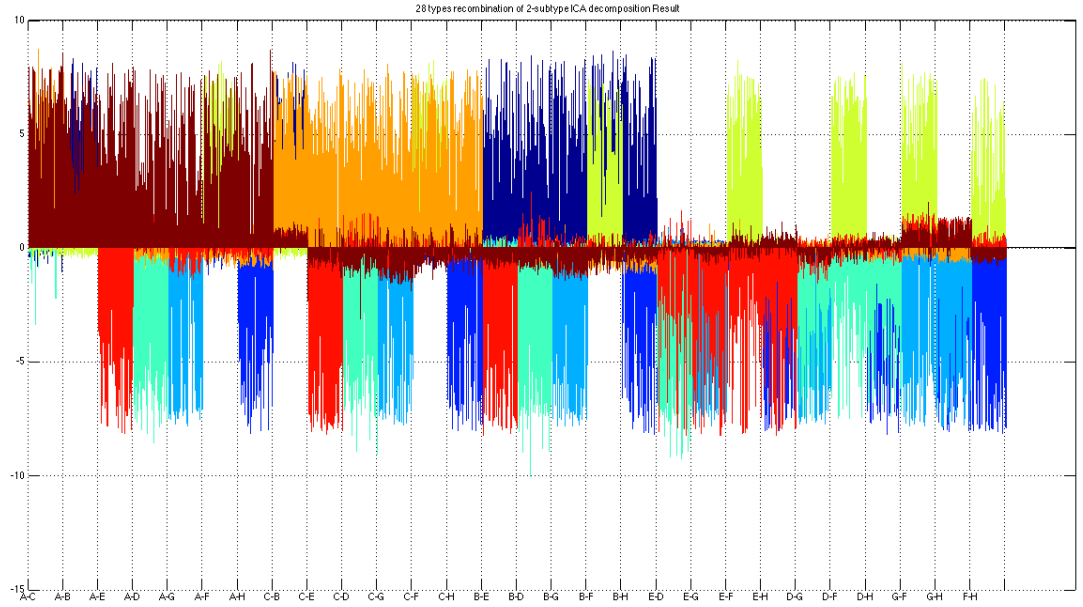
Unbalanced subtype combination(pure)

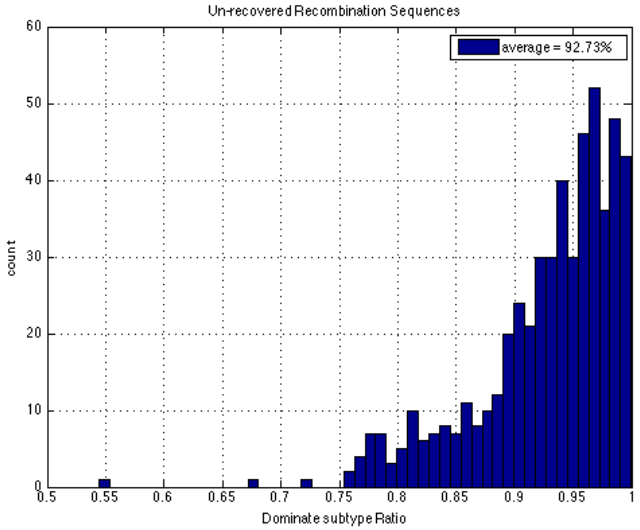
De-noise out orthogonal source of subtype

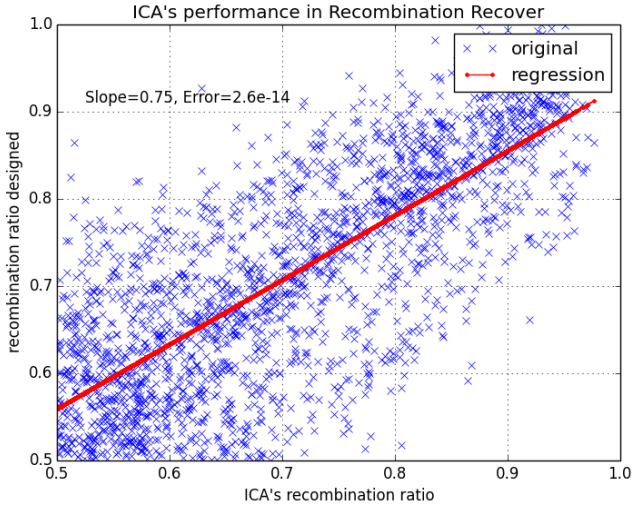
Recombination Test Sequence Subset Validation

Orthogonal Source From ICA represent



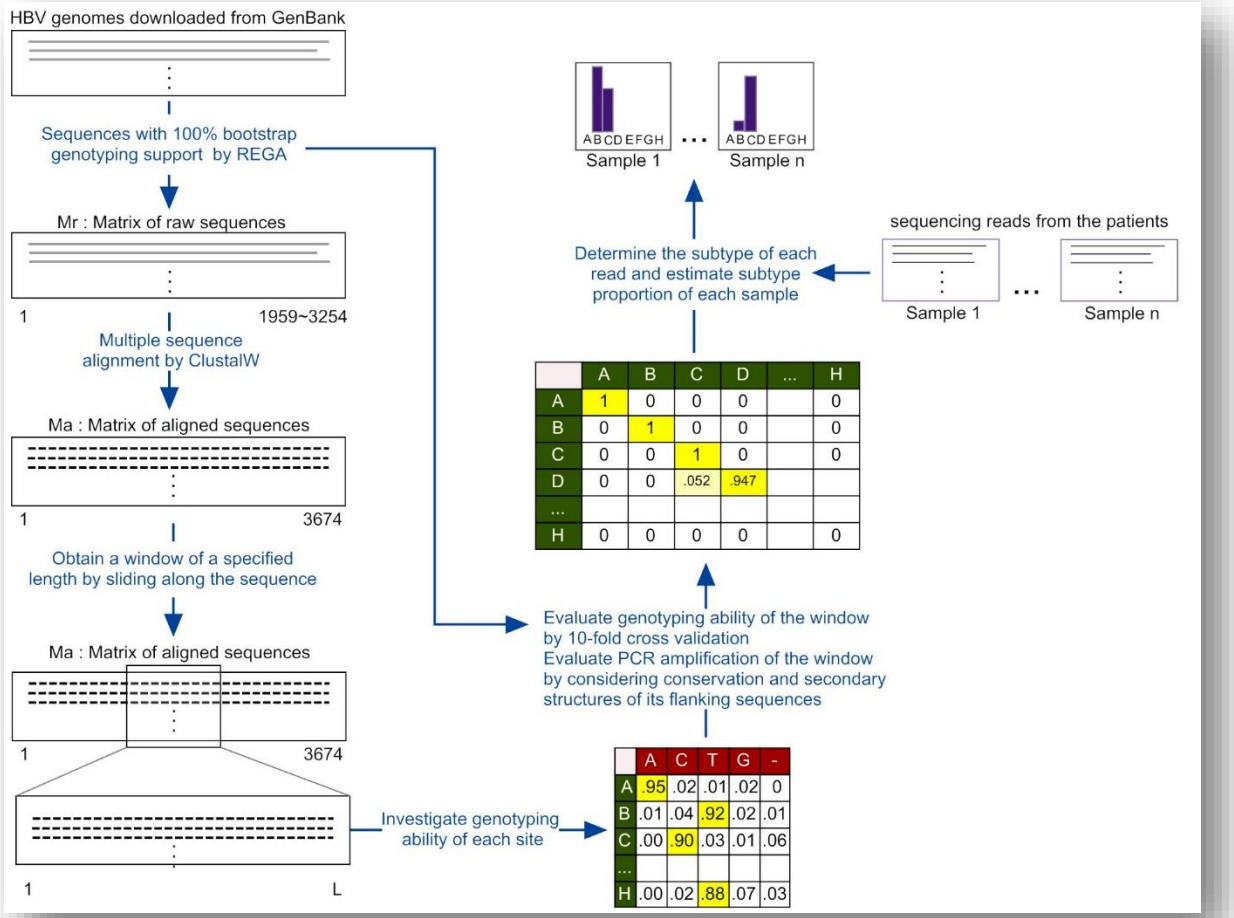






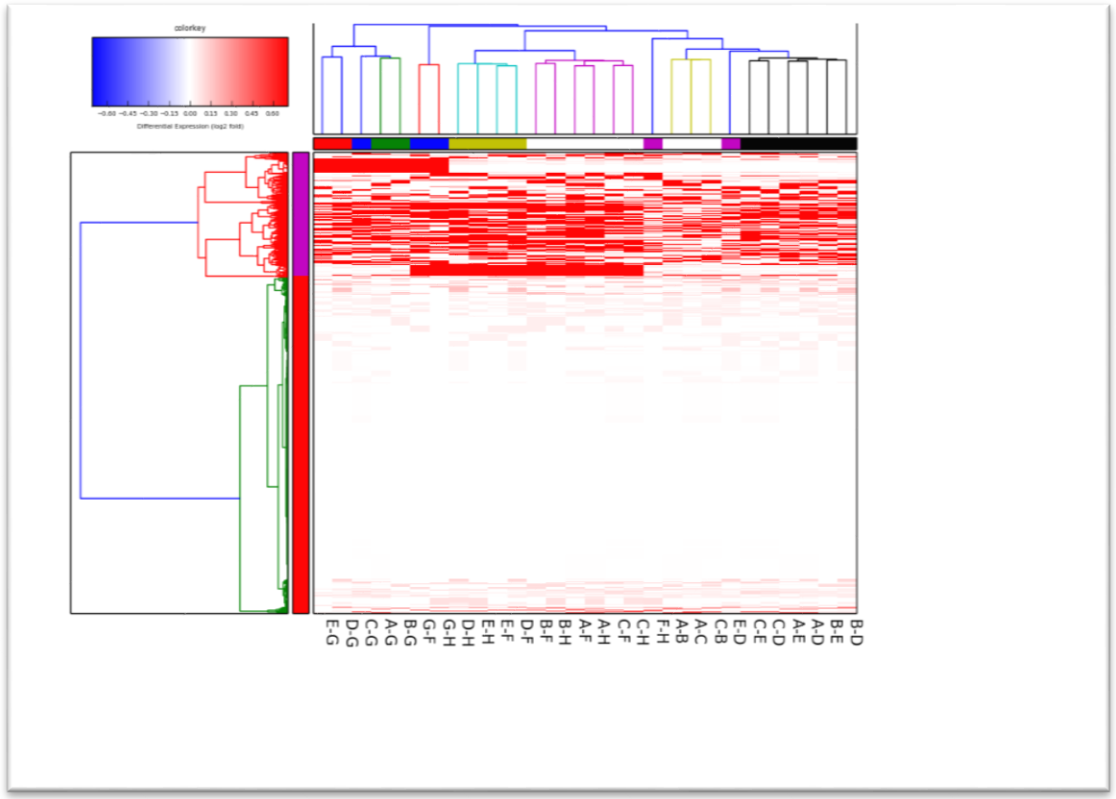
Outline of ShwinGen

1. Background and Motivation
2. Training Set Retrieve …
    1. NCBI Data Retrieve
    2. REGA Genotype
3. Correlation Position Determination Via ICA …
    1. ICA among Different Loci of Genome
    2. ICA among Different Sequences
4. **Genotyping Short Windows Selection**
5. **Barcode Design**
6. Next Generation Sequencing of Short Windows …
    1. Control Template Design
    2. Noise Removal
7. HBV Subtype Inference
8. Analysis between Subtype Shift and Drug Therapy



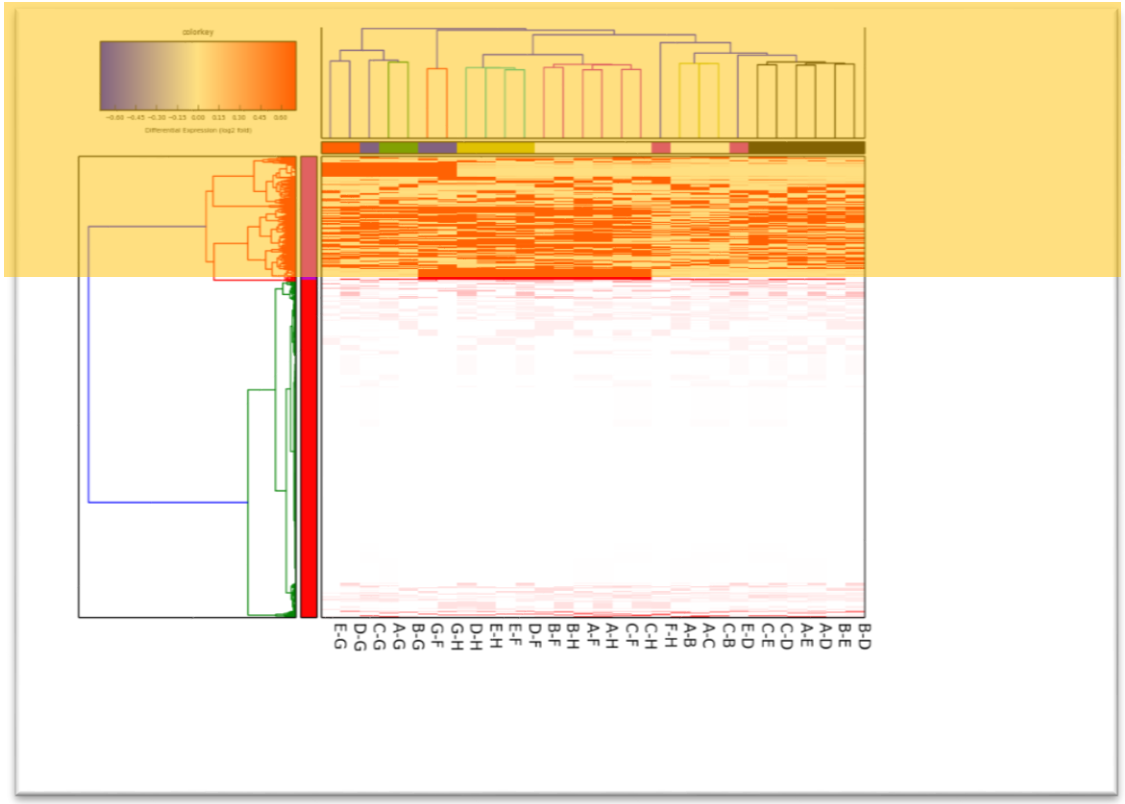
Short Window For Classification

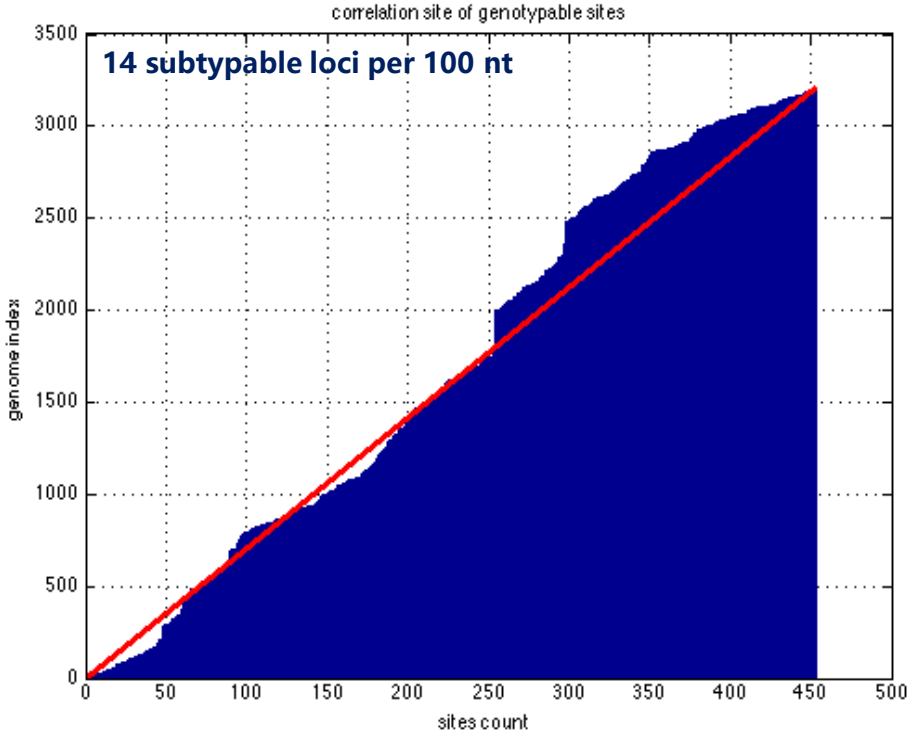
Non-Correlated Loci are sparse and they are conserved at same time
On the other hand, Since sequence of different subtype can be clustered well by ICA, we can describe each subtype with PSSM center



Short Window For Classification

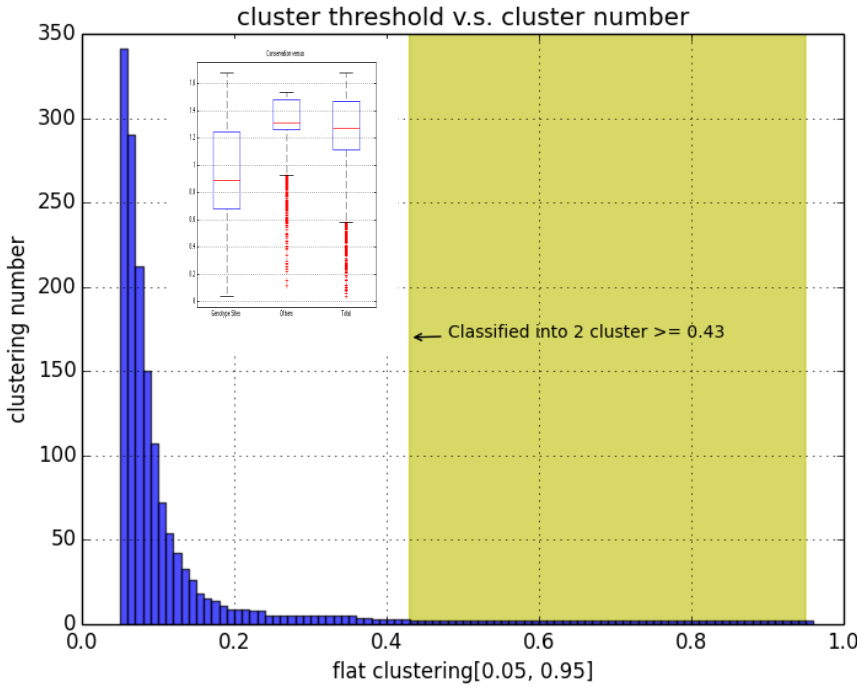
Non-Correlated Loci are sparse and they are conserved at same time
On the other hand, Since sequence of different subtype can be clustered well by ICA,
we can describe each subtype with PSSM center

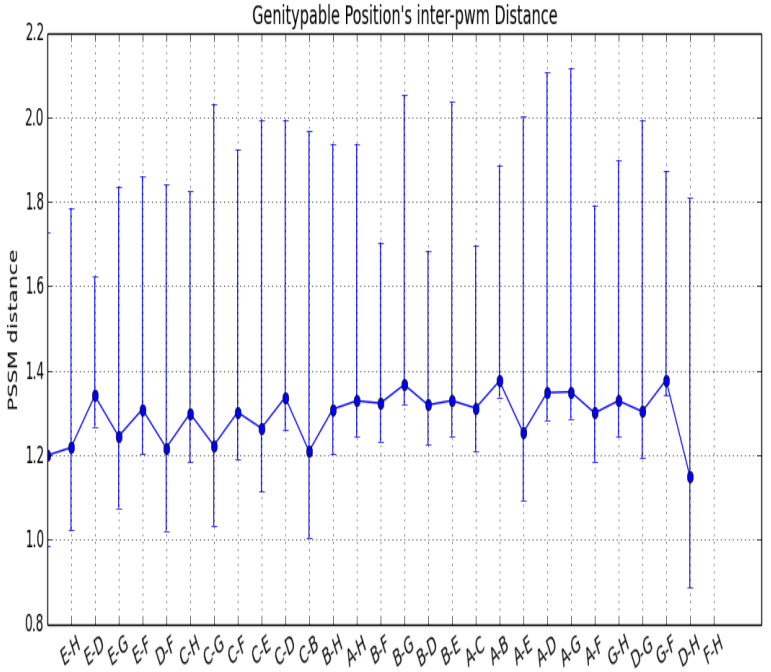




Compressed sense of HBV subtype

Flat Clustering Of Hierarchical Structure is a hard problem



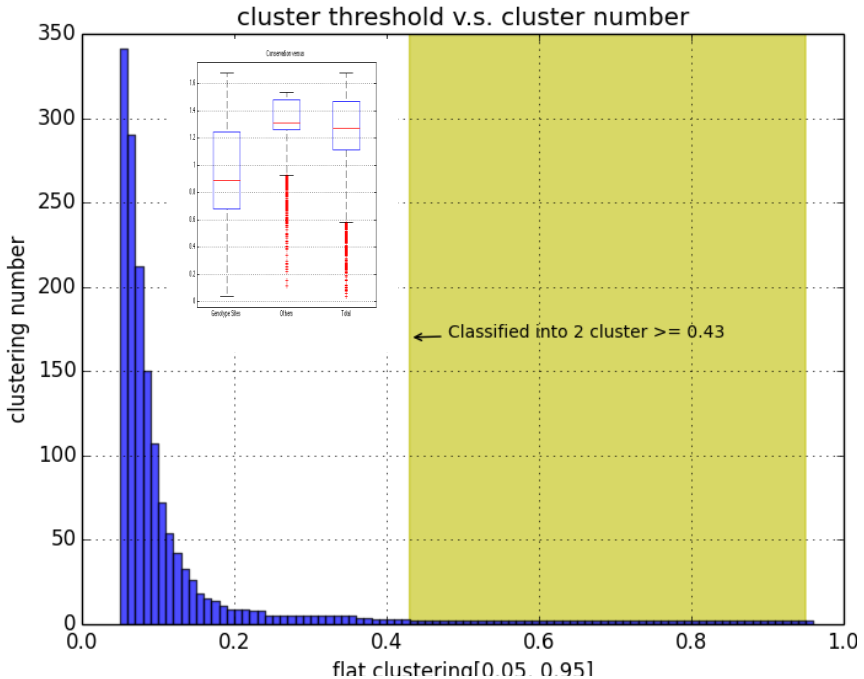


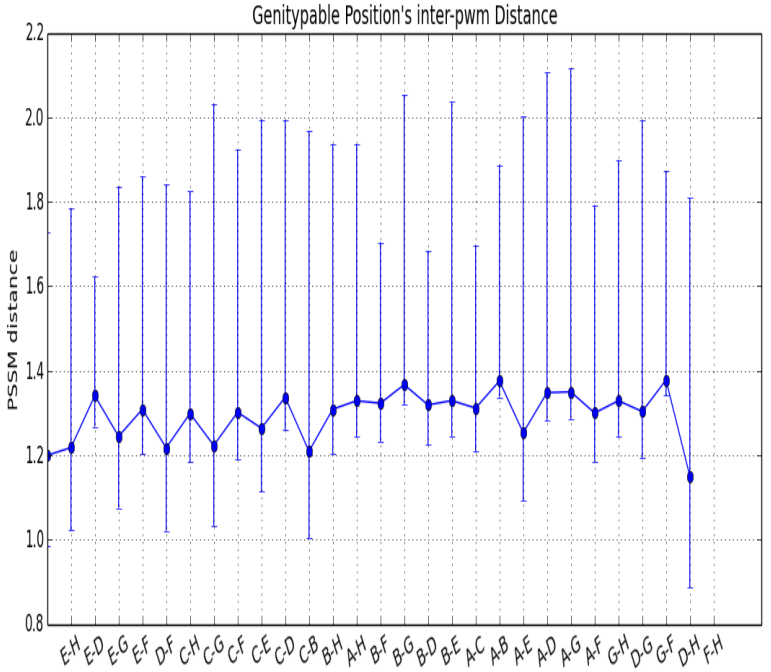
$$Ax = B$$

$$x = (0,0, \dots, 1,1, \dots 1,0)$$

Compressed sense of HBV subtype

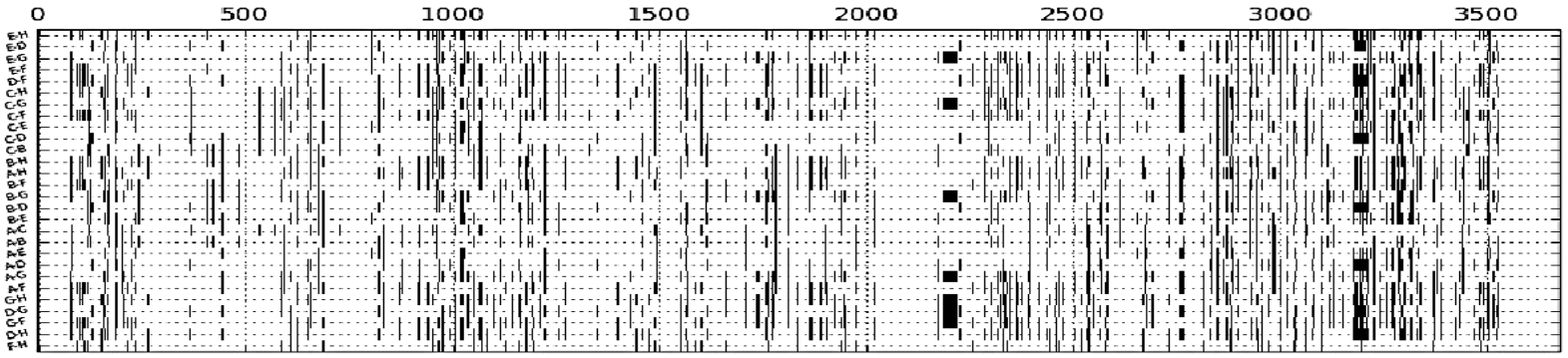
Flat Clustering Of Hierarchical Structure is a hard problem

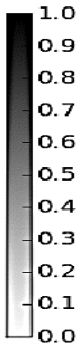


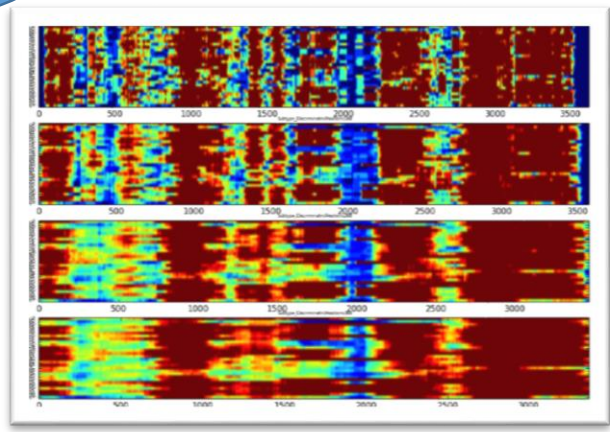


$$Ax = B$$

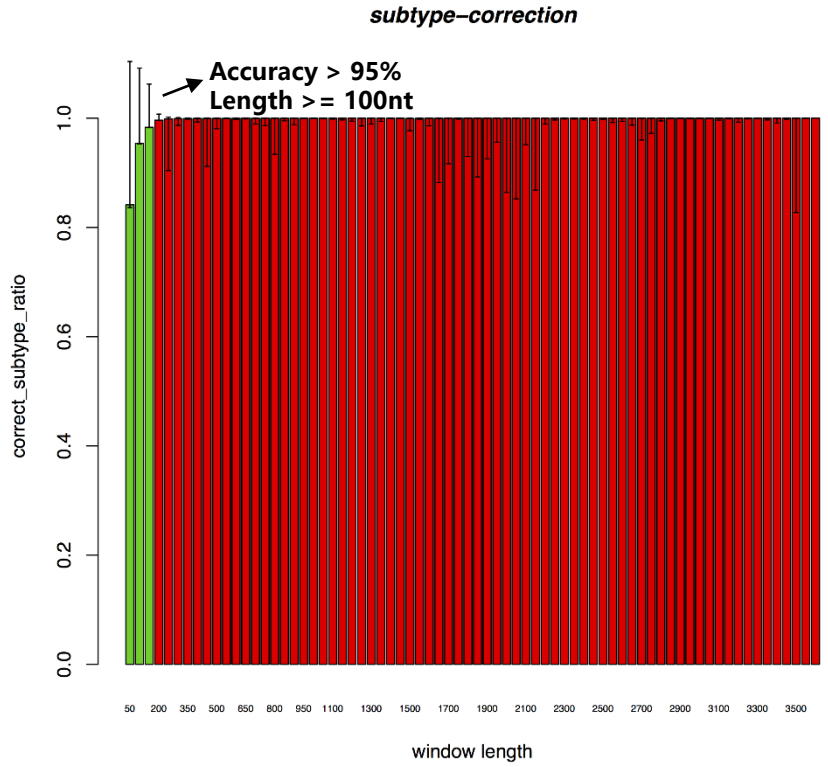
$$x = (0,0,\dots,1,1,..1,0)$$

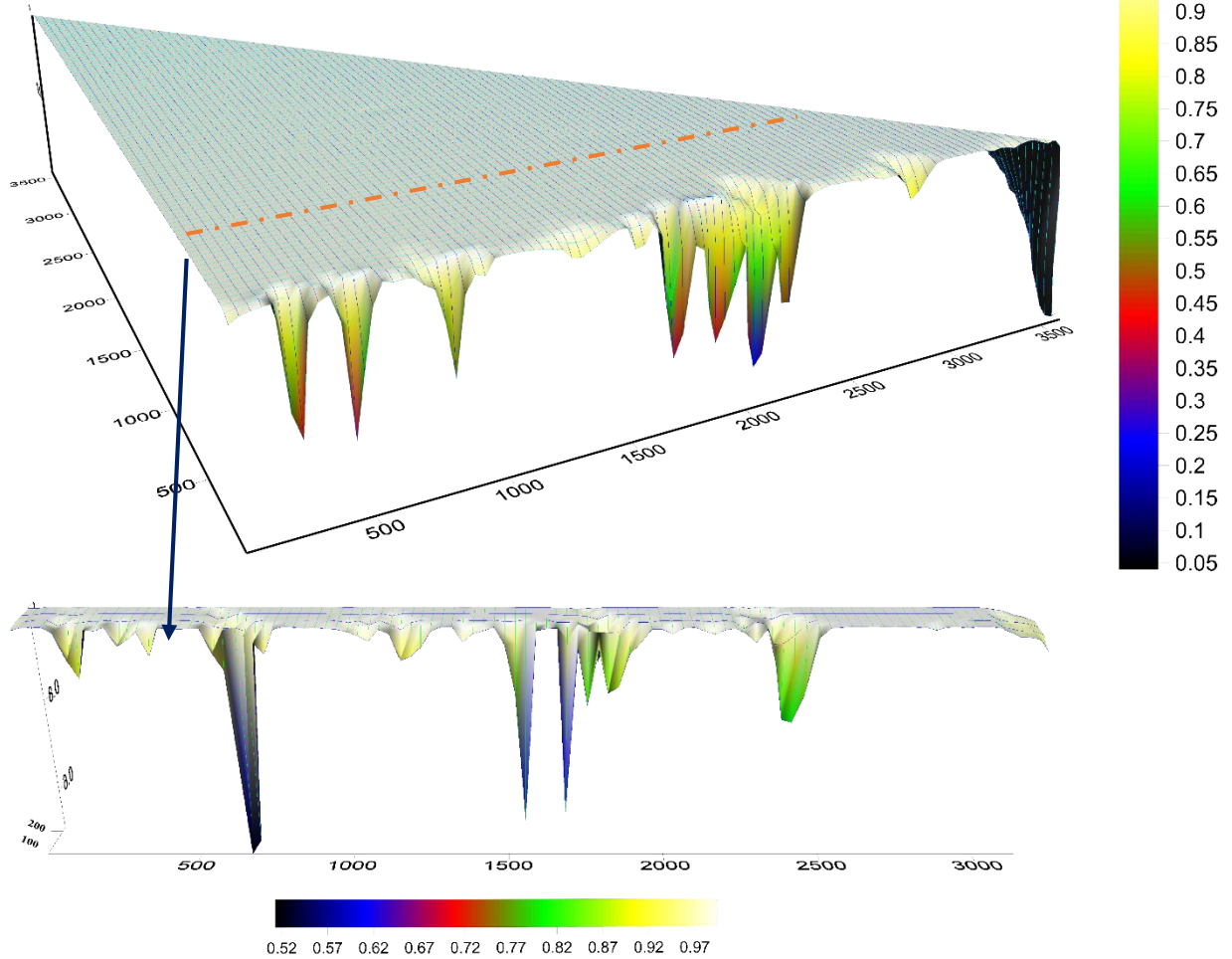






Short window genotype ability





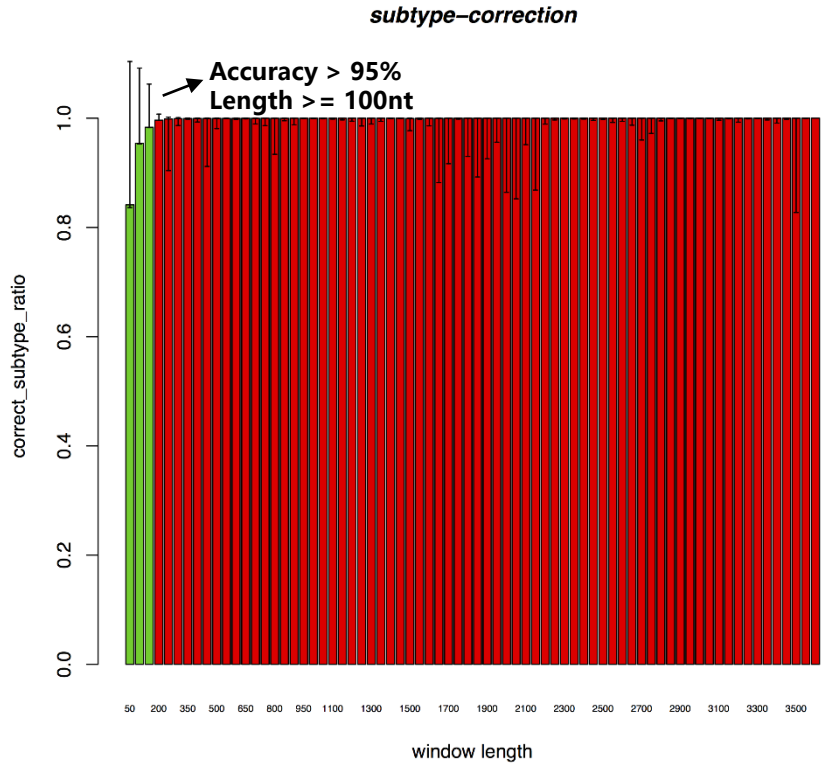
RT region for pyro-sequencing : window of nt [619,879] (261bp) + primer(40bp)

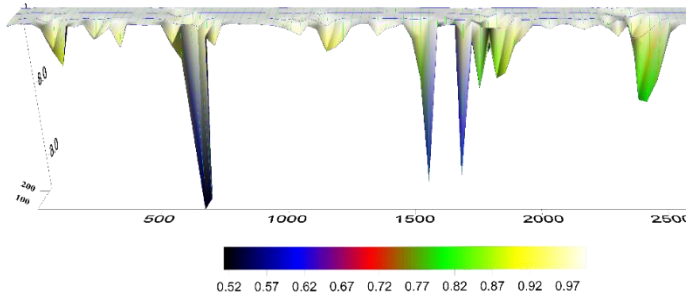
Most suitable region for solexa sequencing : window of nt [1446,1544] (99bp) + primer (40bp)

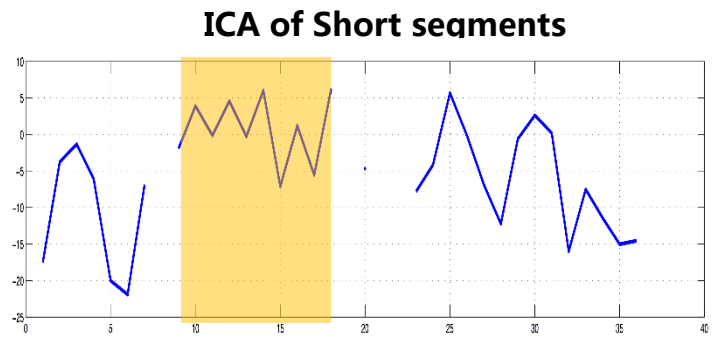
Primer constrain + specificity + context constrain

.. For Barcode Design CSP

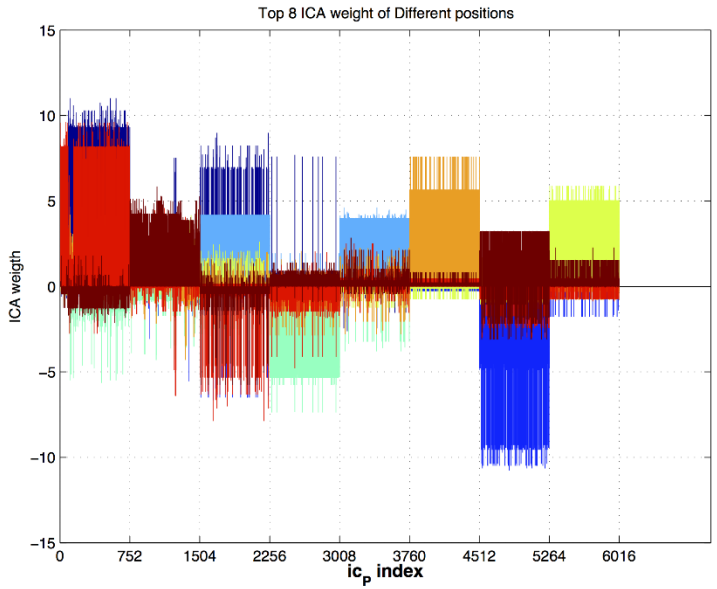
Short window genotype ability



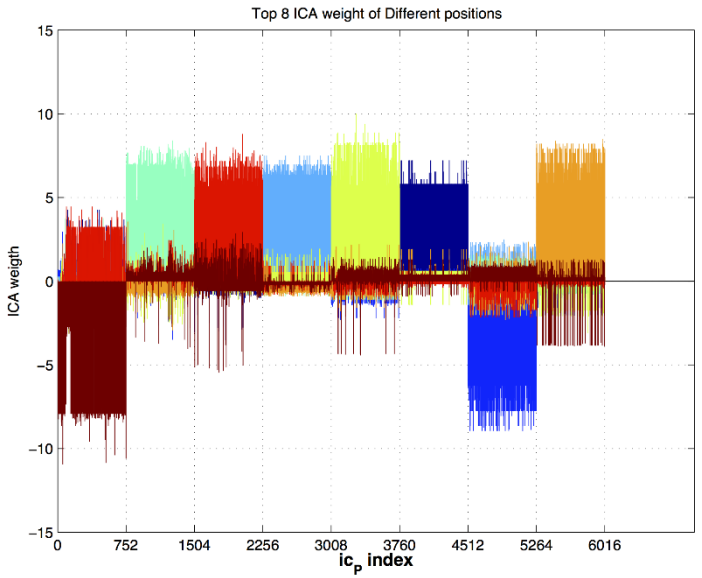




(100bp)



(300bp)



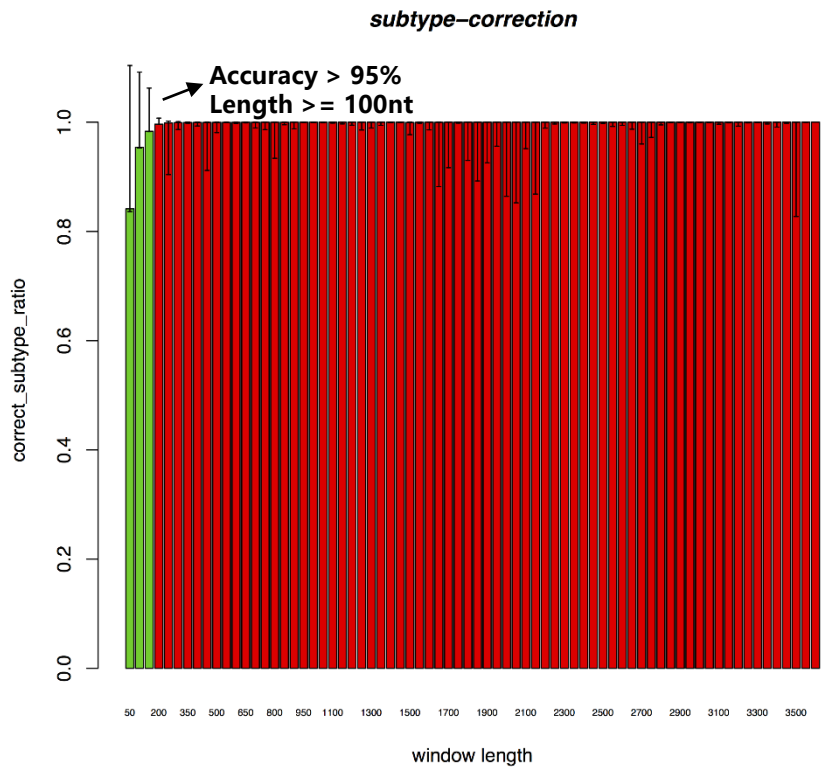
RT region for pyro-sequencing : window of nt [619,879] (261bp) + primer(40bp)

Most suitable region for solexa sequencing : window of nt [1446,1544] (99bp) + primer (40bp)

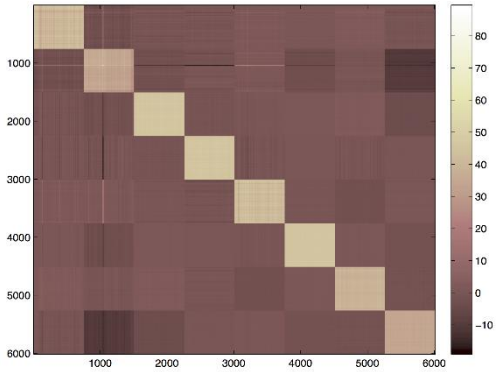
Primer constrain + specificity + context constrain

.. For Barcode Design CSP

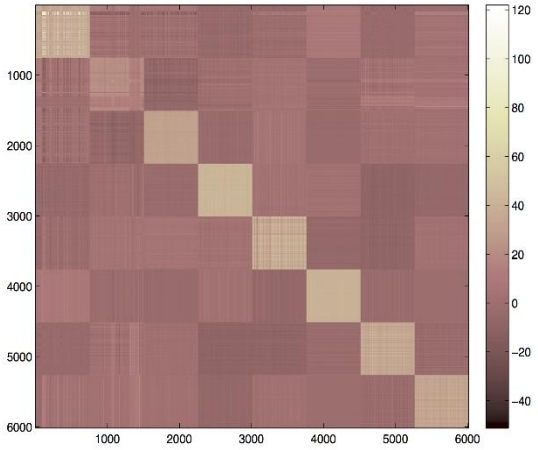
Short window genotype ability

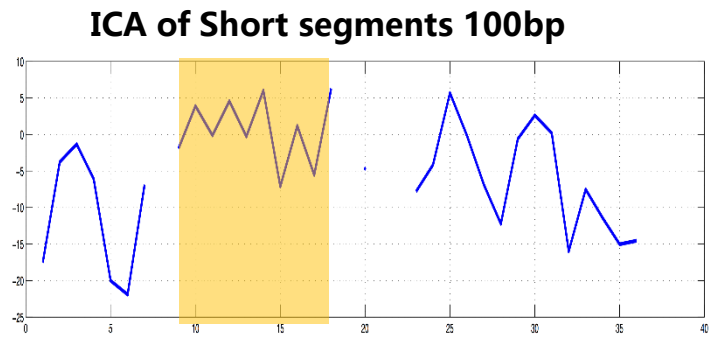


standard

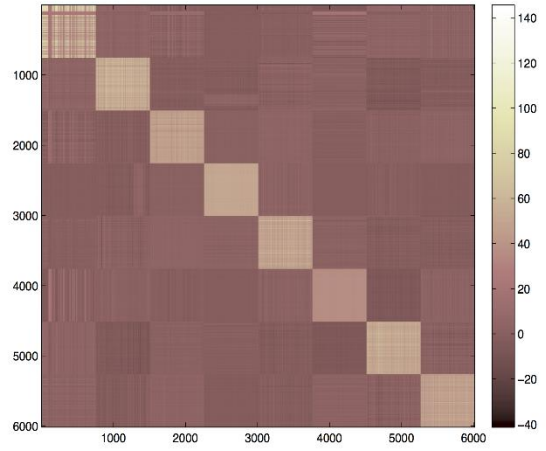


(100bp)





(300bp)



RT region for pyro-sequencing : window of nt [619,879] (261bp) + primer(40bp)

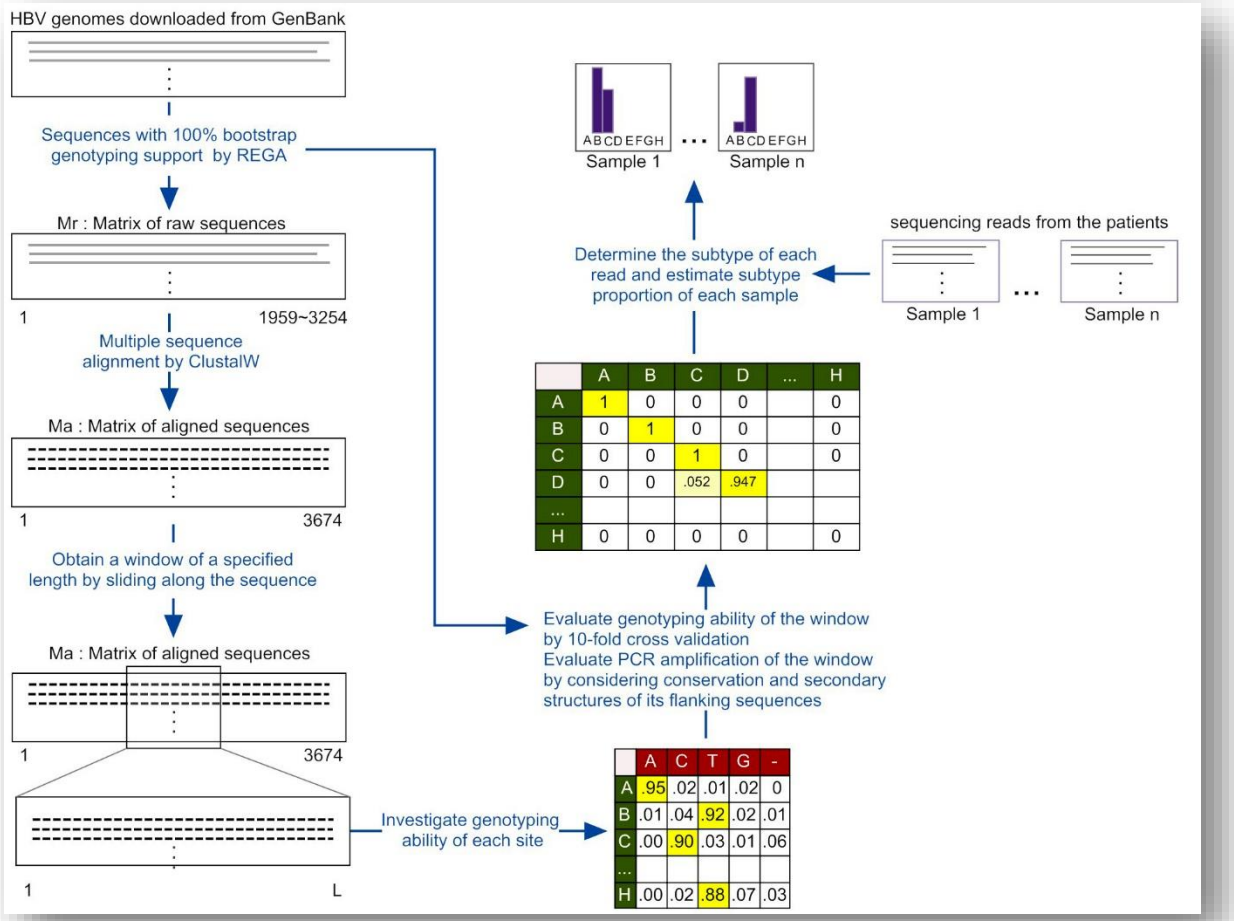
Most suitable region for solexa sequencing : window of nt [1446,1544] (99bp) + primer (40bp)

Primer constrain + specificity + context constrain

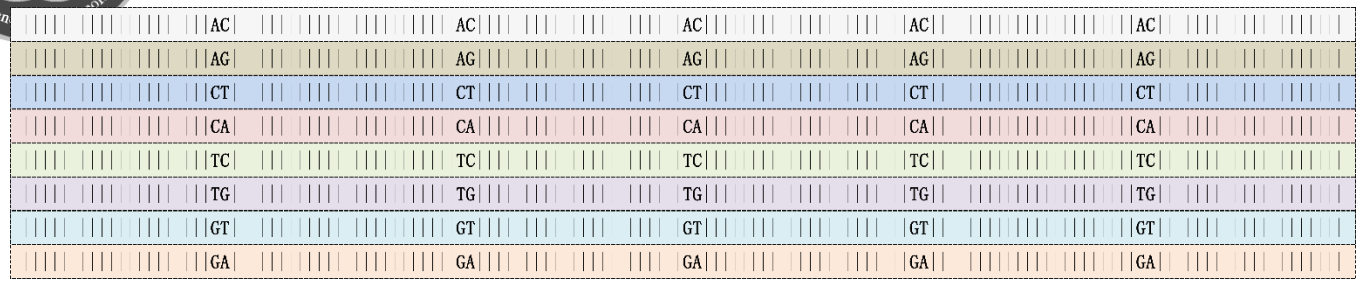
.. For Barcode Design CSP

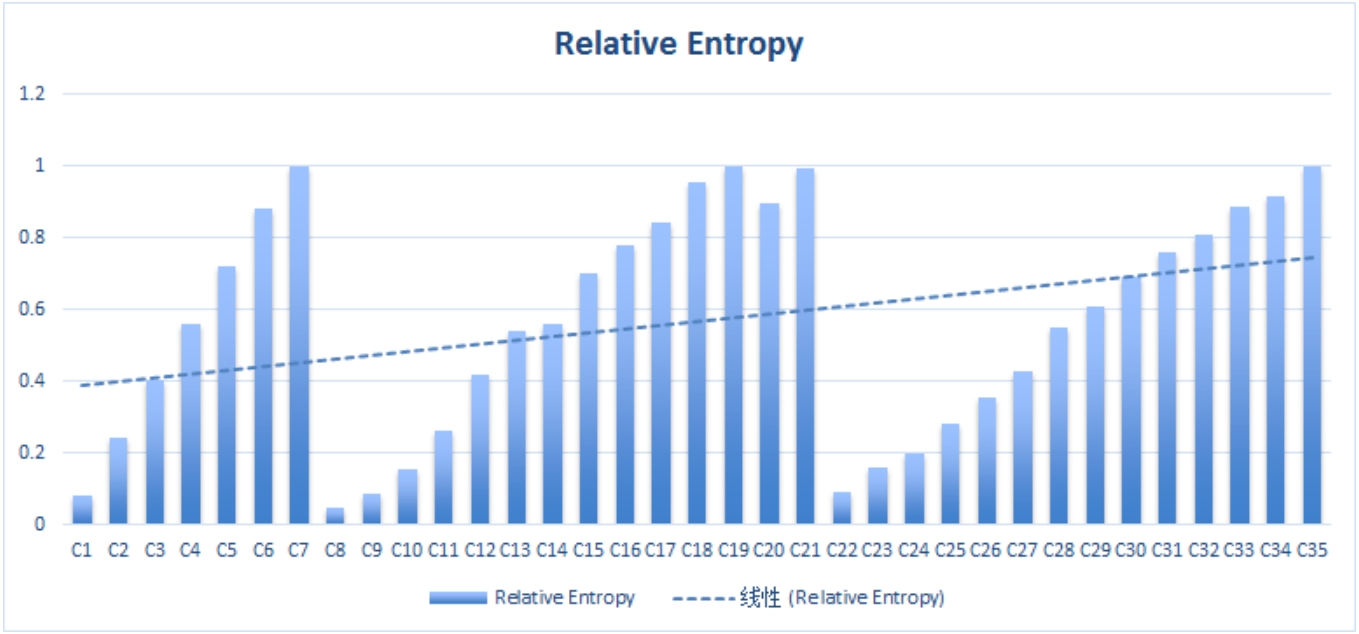
Outline of ShwinGen

1. Background and Motivation
2. Training Set Retrieve …
    1. NCBI Data Retrieve
    2. REGA Genotype
3. Correlation Position Determination Via ICA …
    1. ICA among Different Loci of Genome
    2. ICA among Different Sequences
4. Genotyping Short Windows Selection
5. Barcode Design
6. **Next Generation Sequencing of Short Windows …**
    1. **Control Template Design**
    2. **Noise Removal**
7. **HBV Subtype Inference**
8. Analysis between Subtype Shift and Drug Therapy

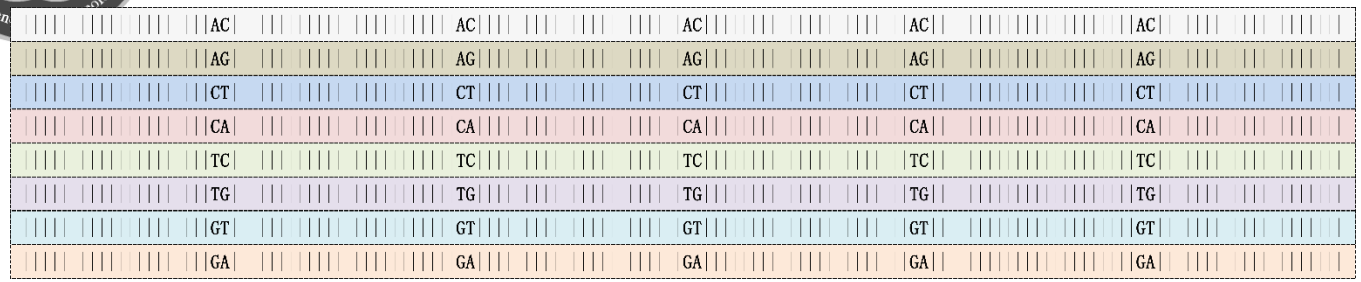


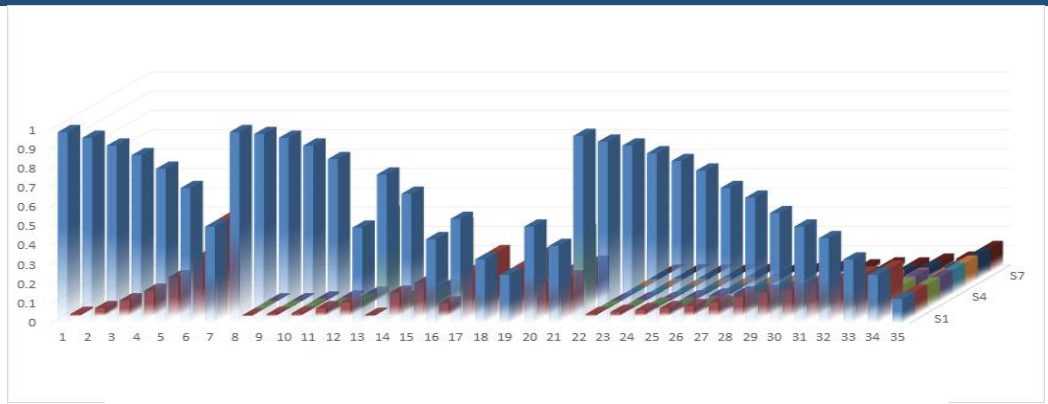
Standard Build

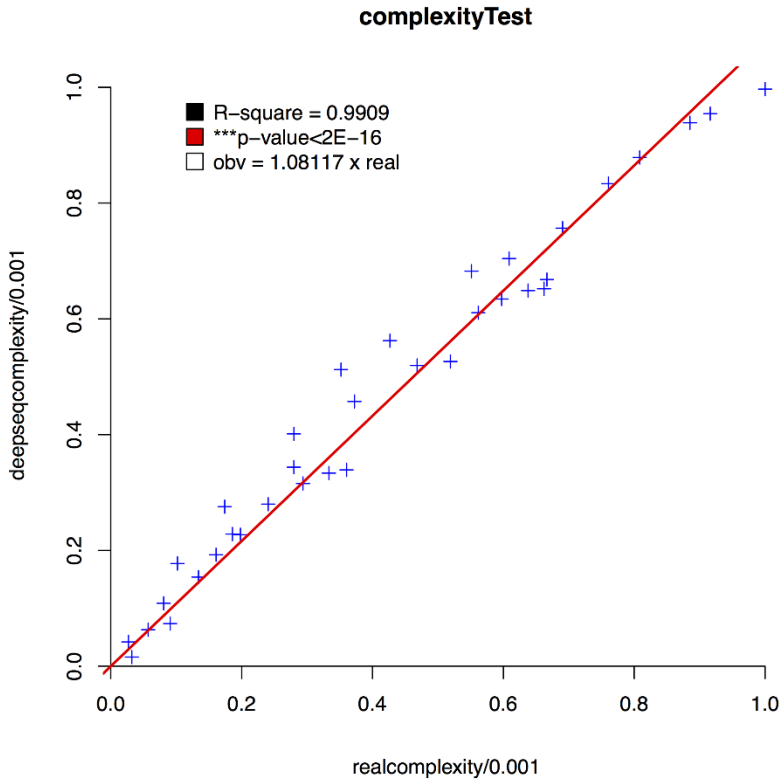


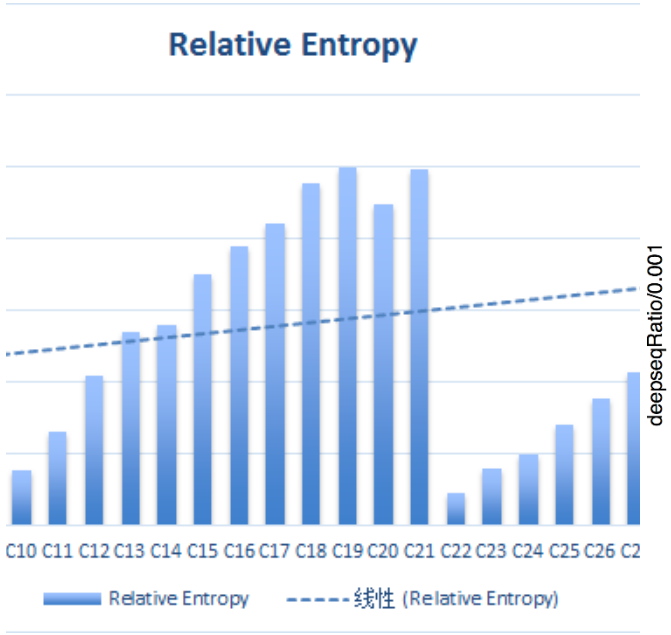


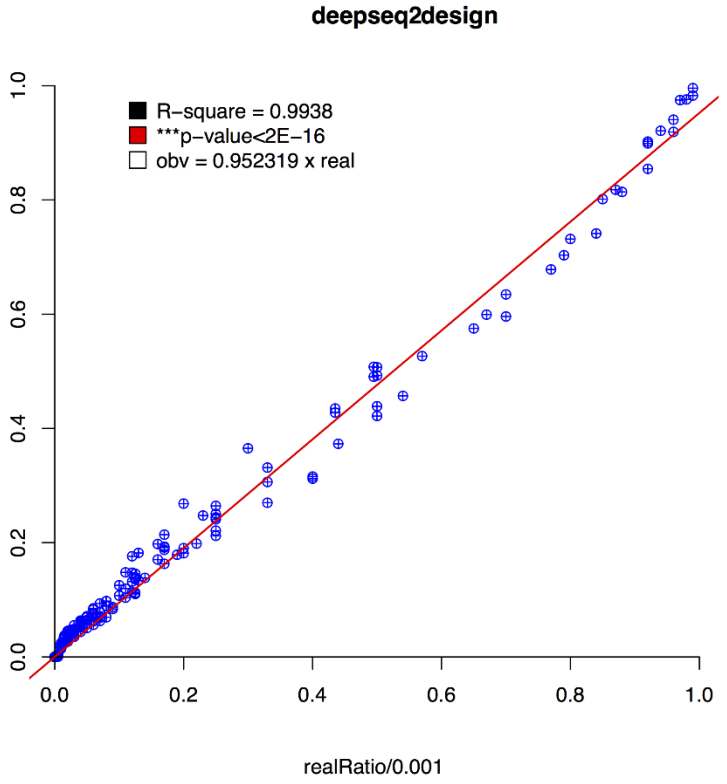
Standard Build







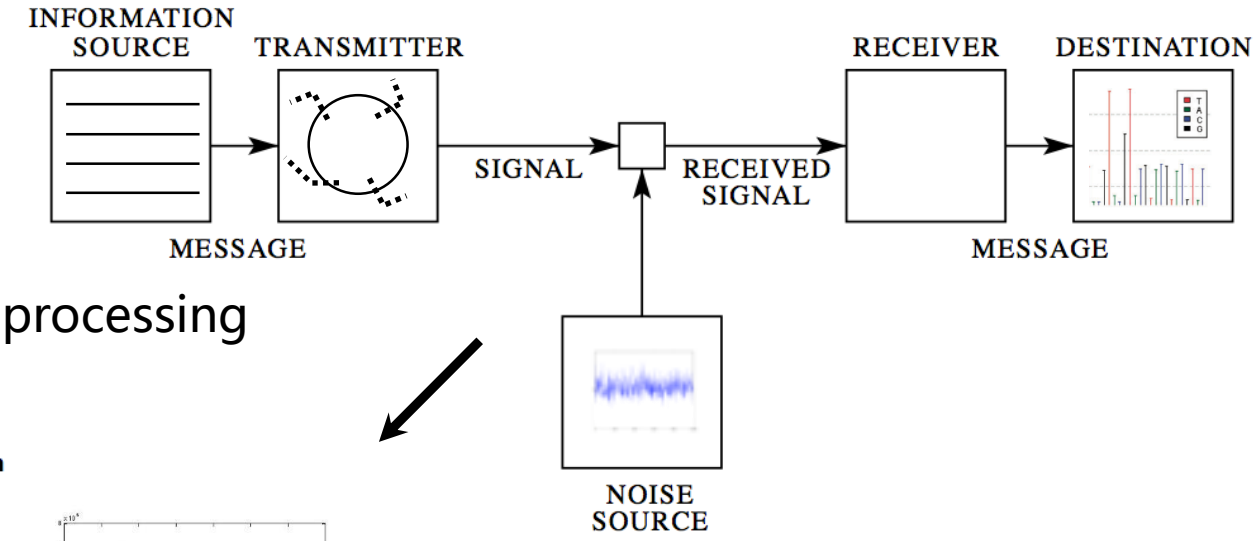




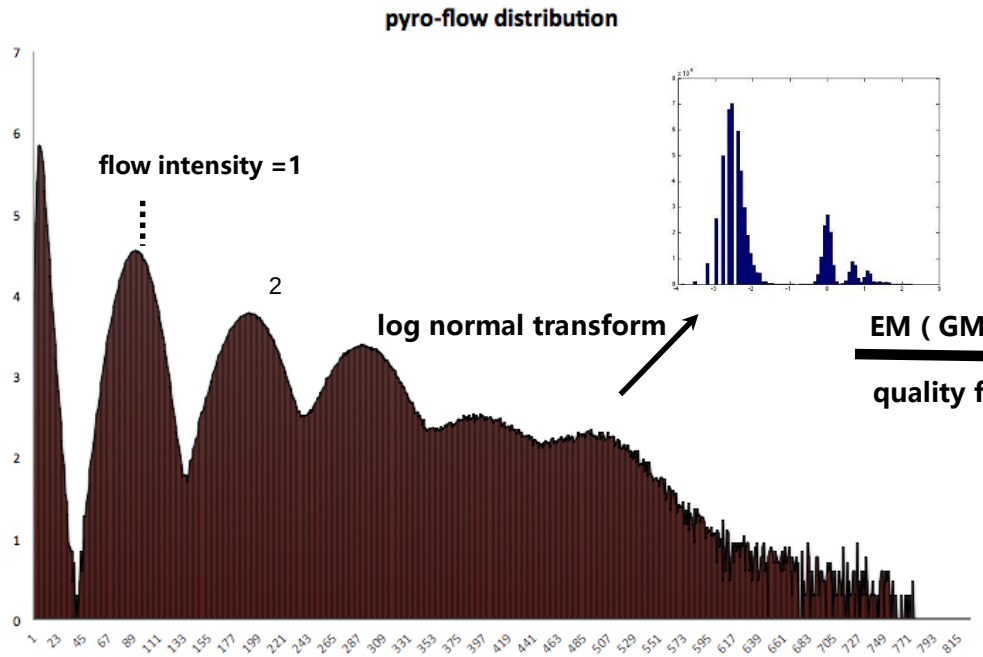
And recover ratio's Infimum is 0.02%

Noise Removal

Primer constrain + specificity
+ context constrain
.. For Barcode Design CSP
= [Shwigen]



Deep sequencing data processing

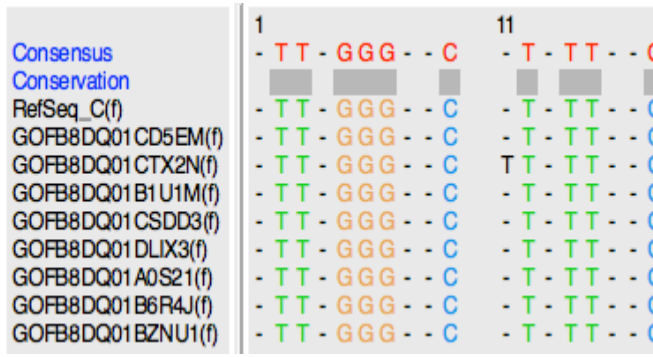


EM (GMM)

quality file

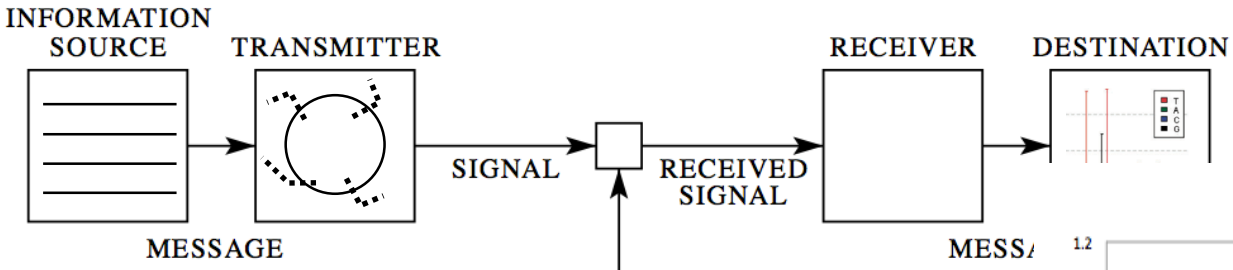
FlowIr	E	S	Prob
3.62	4	0.021393	0.998211
4.4	4	0.021393	0.98641
1.52	2	0.078356	0.998669
0.82	1	0.326043	1
6.23	5	0.000449	0.987954
6.06	5	0.000449	0.992621
4.41	4	0.021393	0.985788
5.79	5	0.000449	0.984649
3.44	4	0.021393	0.862544

iterate profile
MSA

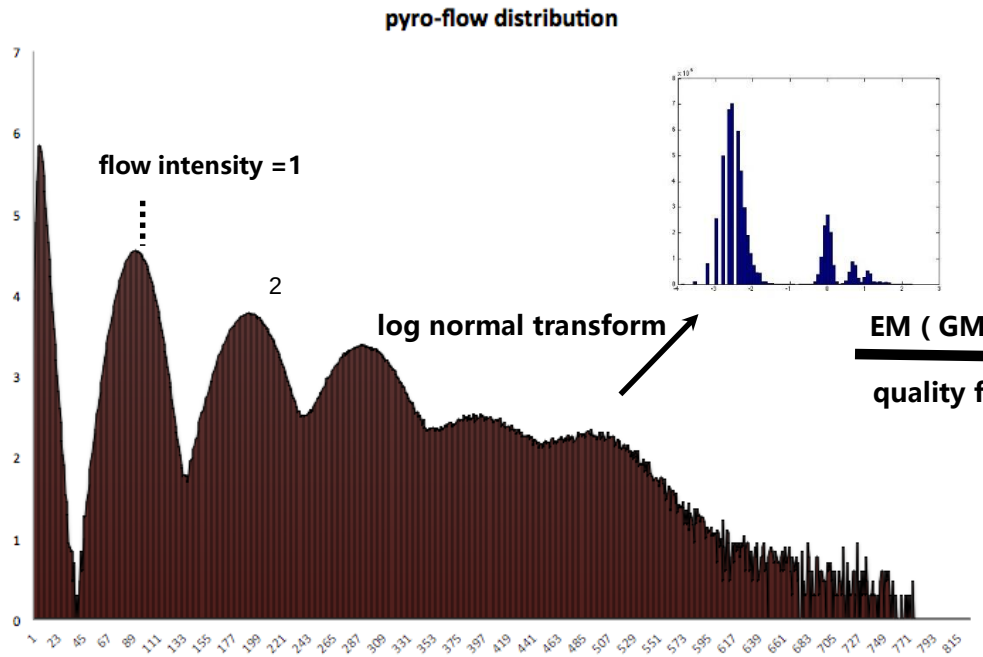


Noise Removal

Primer constrain + specificity
+ context constrain
.. For Barcode Design CSP
= Shwigen



Deep sequencing data processing

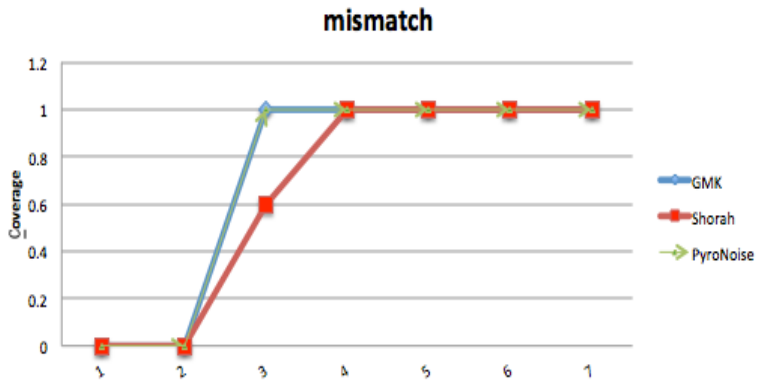


EM (GMM)

quality file

FlowIr	E	S	Prob
3.62	4	0.021393	0.998211
4.4	4	0.021393	0.98641
1.52	2	0.078356	0.998669
0.82	1	0.326043	1
6.23	5	0.000449	0.987954
6.06	5	0.000449	0.992621
4.41	4	0.021393	0.985788
5.79	5	0.000449	0.984649
3.44	4	0.021393	0.862544

iterate profile
MSA

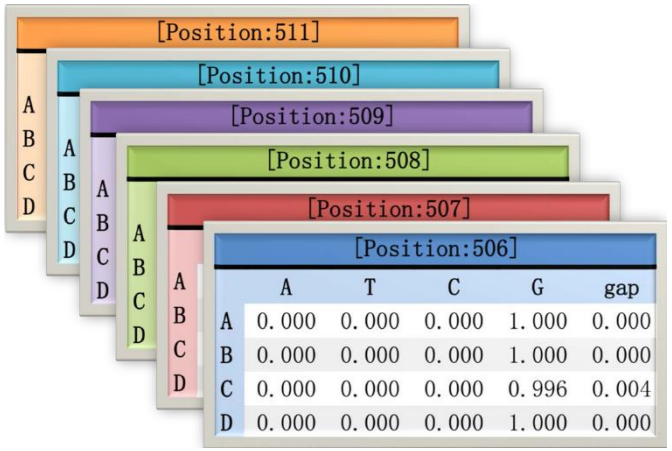


HBV Subtype Inference

http://netalign.ustc.edu.cn/ShwinGen/

Reads Genotype assignment via MAP from PSSM

$$\widehat{Genotype}_{\mathrm{MLE}}('atcg....') = \mathrm{argmax}_{\mathrm{Genotype}} \sum_{n=1}^{L} \log\left(P(\mathrm{Genotype})P(S_n|\mathrm{Genotype})\right)$$

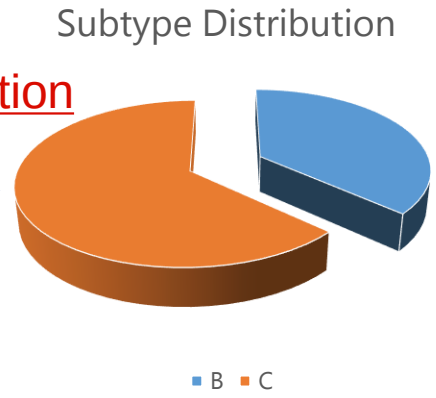


MAP

SeqID/subtype	Likelihood	Log-Liklihood
A	2.34E-80	-79.63167652
B	1.33E-62	-61.87583337
C	3.26E-56	-55.4870609
D	2.26E-110	-109.645238
E	8.91E-124	-123.0499411
F	4.96E-192	-191.304381
G	2.35E-110	-109.629293
H	8.36E-182	-181.0780014

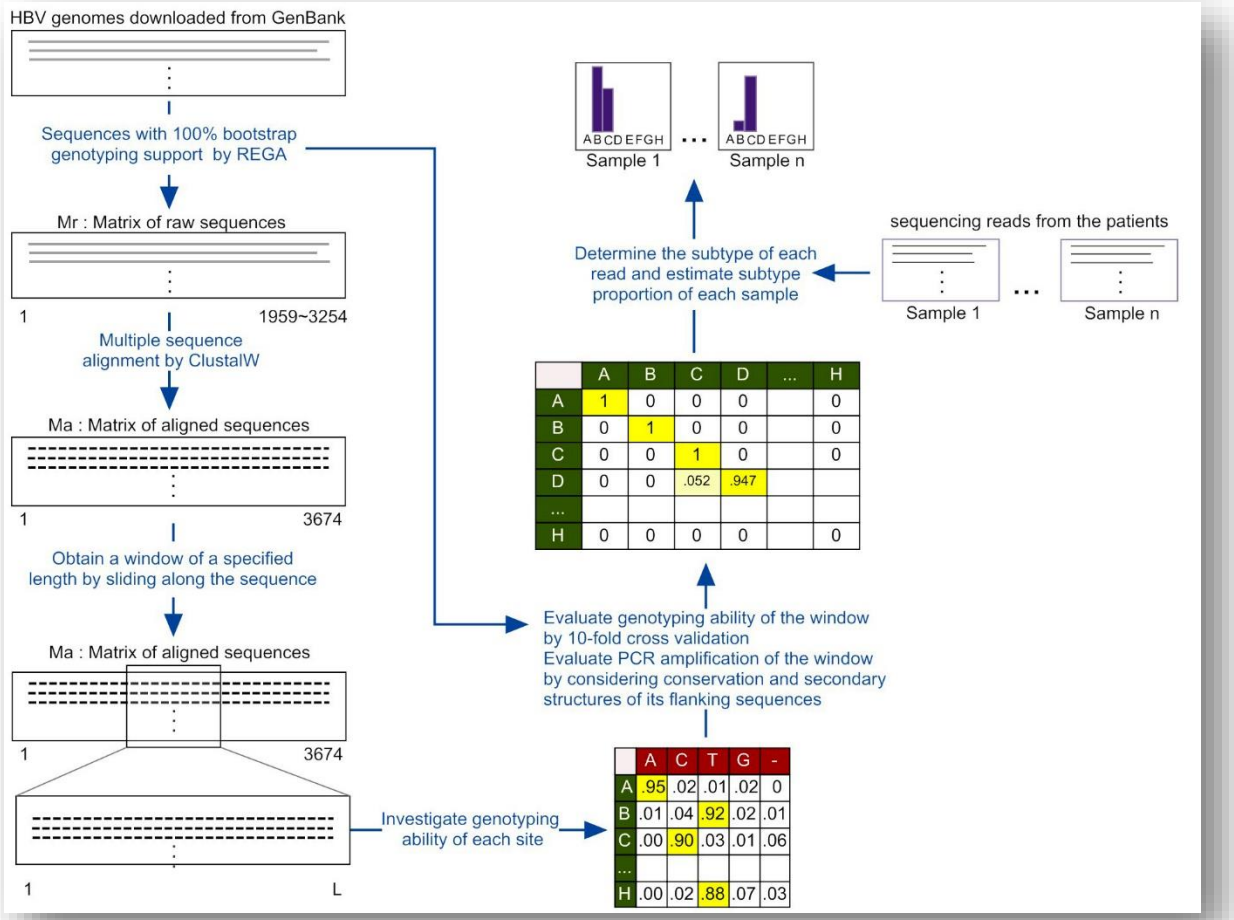
C is most in mix infection

/SUBTYP	A	B	C	D
Fraction	0	0.366667	0.633333	0
Frequency	0	11	19	0



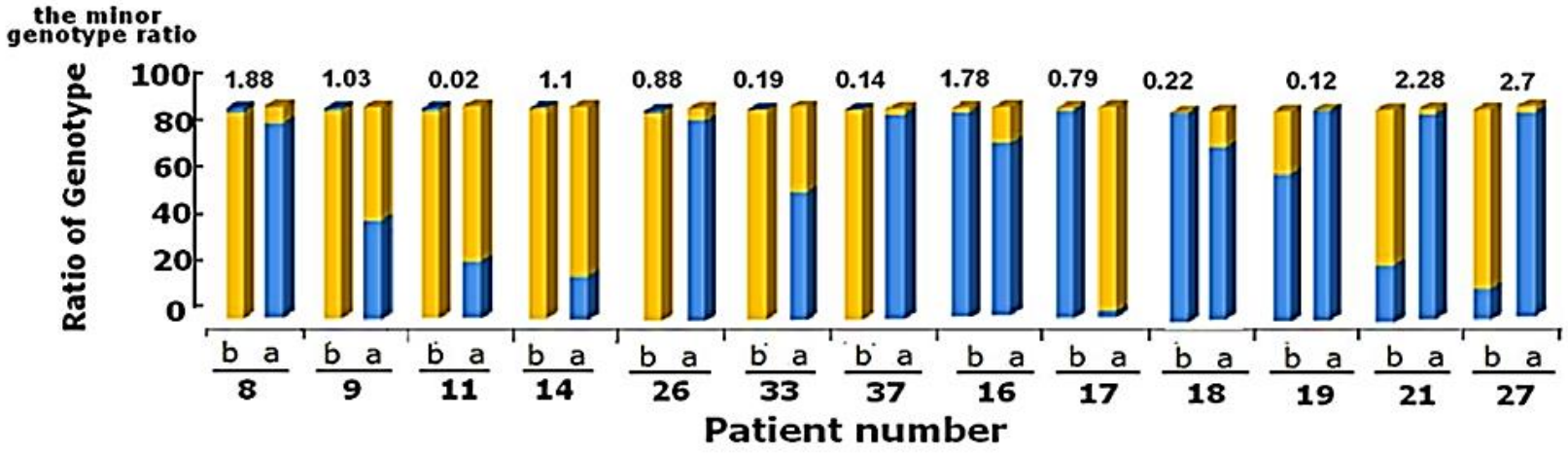
Outline of ShwinGen

1. Background and Motivation
2. Training Set Retrieve …
    1. NCBI Data Retrieve
    2. REGA Genotype
3. Correlation Position Determination Via ICA …
    1. ICA among Different Positions of Genome
    2. ICA among Different Sequences
4. Genotyping Short Windows Selection
5. Barcode Design
6. Next Generation Sequencing of Short Windows …
    1. Control Template Design
    2. Noise Removal
7. HBV Subtype Inference
8. **Analysis between Subtype Shift and Drug Therapy**

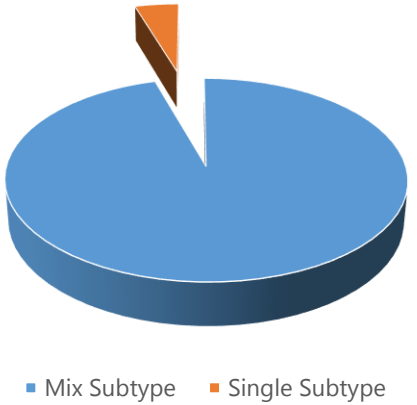


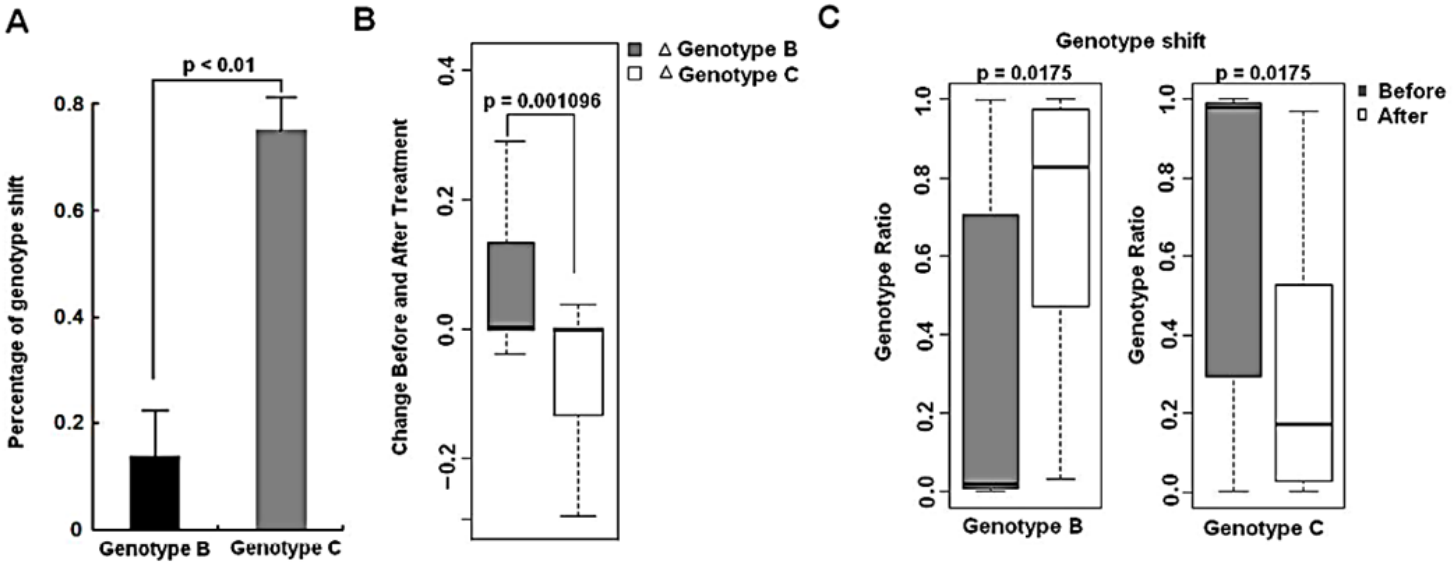
Subtype Distribution

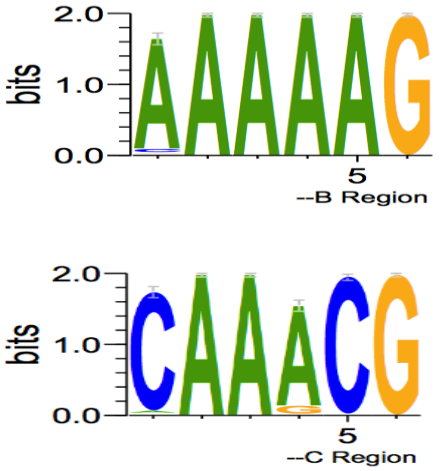
13 Subtype Shift in Sanger Detected By NGS



0.02% cutoff







CHB display B subtype preference

Conclusion

- Classify HBV subtype via short segment
- Compressed representation of HBV genome
- **First time** to retrieve HBV subtype and recombination chimera via **ICA** (Blind source recover)
- Convince the assumption that ADV resistance is from mix infection
- Pipeline construction of short window segment sequencing by 454 or Solexa
- NGS data's noise removal by different algorithms is taken into account
- First paper [Analysis of HBV genotype shift and correlation with antiviral efficiency during adefovir dipivoxil therapy by deep sequencing] has been submitted to Journal of Hepatology
- And [ICA's application in HBV subtype classification]'s manuscript is in preparation

Thanks for everyone's help very much these 3 years
Thanks for everyone's listening very much
Thanks for your attention and questions