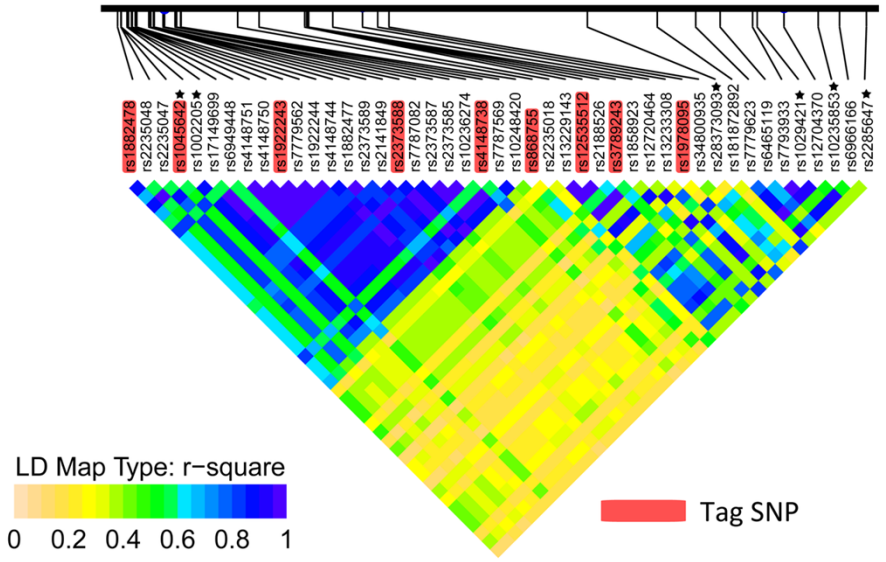
Deconvoluting the Most Clinically Relevant Region of the Human Genome

Dimitri Monos Ph.D.

Immunogenetics Laboratory
The Children's Hospital of Philadelphia
Department of Pathology and Lab Medicine
Perelman School of Medicine, University of Pennsylvania

ARUP LABORATORIES, Pathology Grand Rounds, September 20, 2018

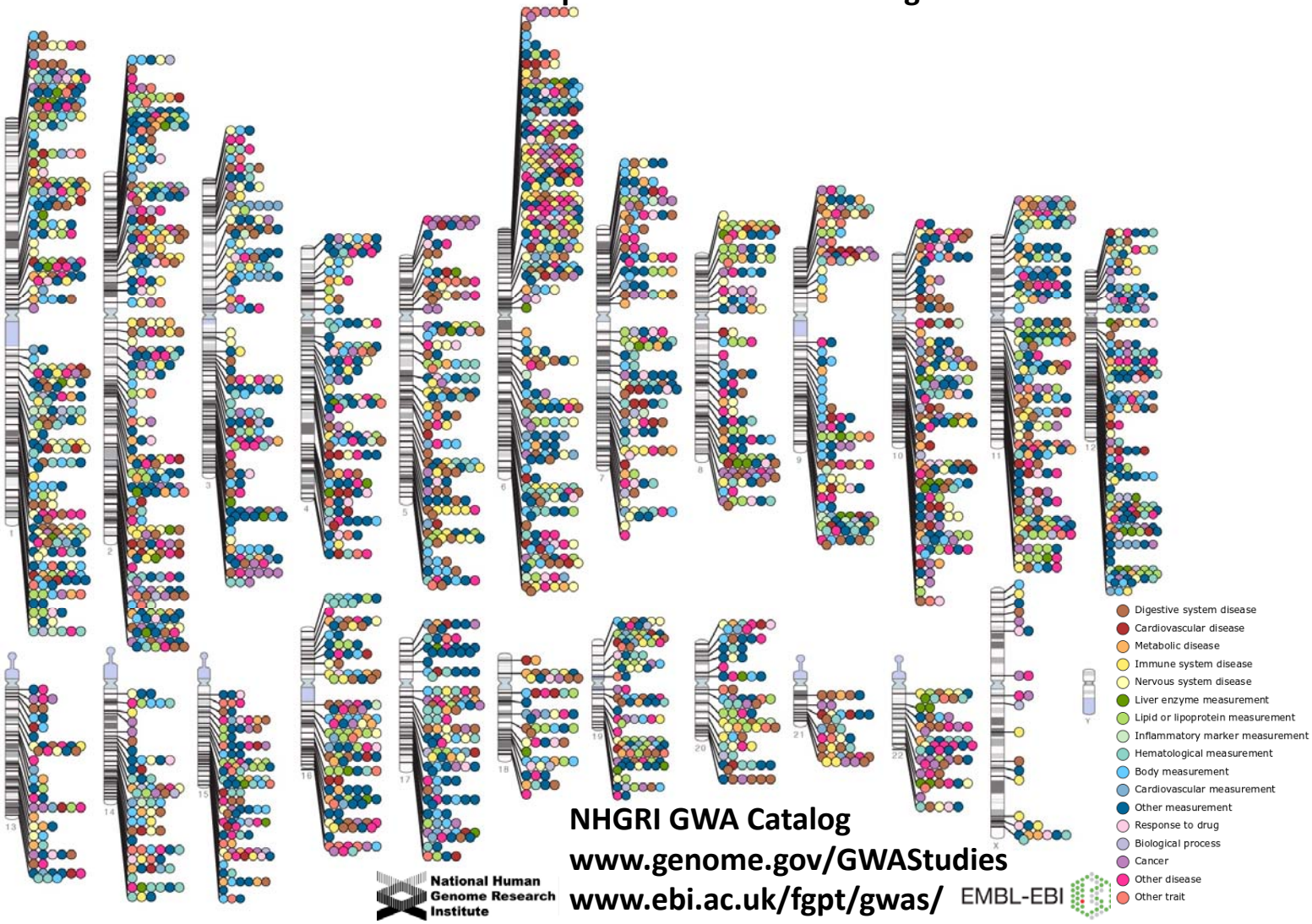
GWAS Interpretation - Tag SNPs are Markers of LD blocks



*Concept of LD is population specific

http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0046295

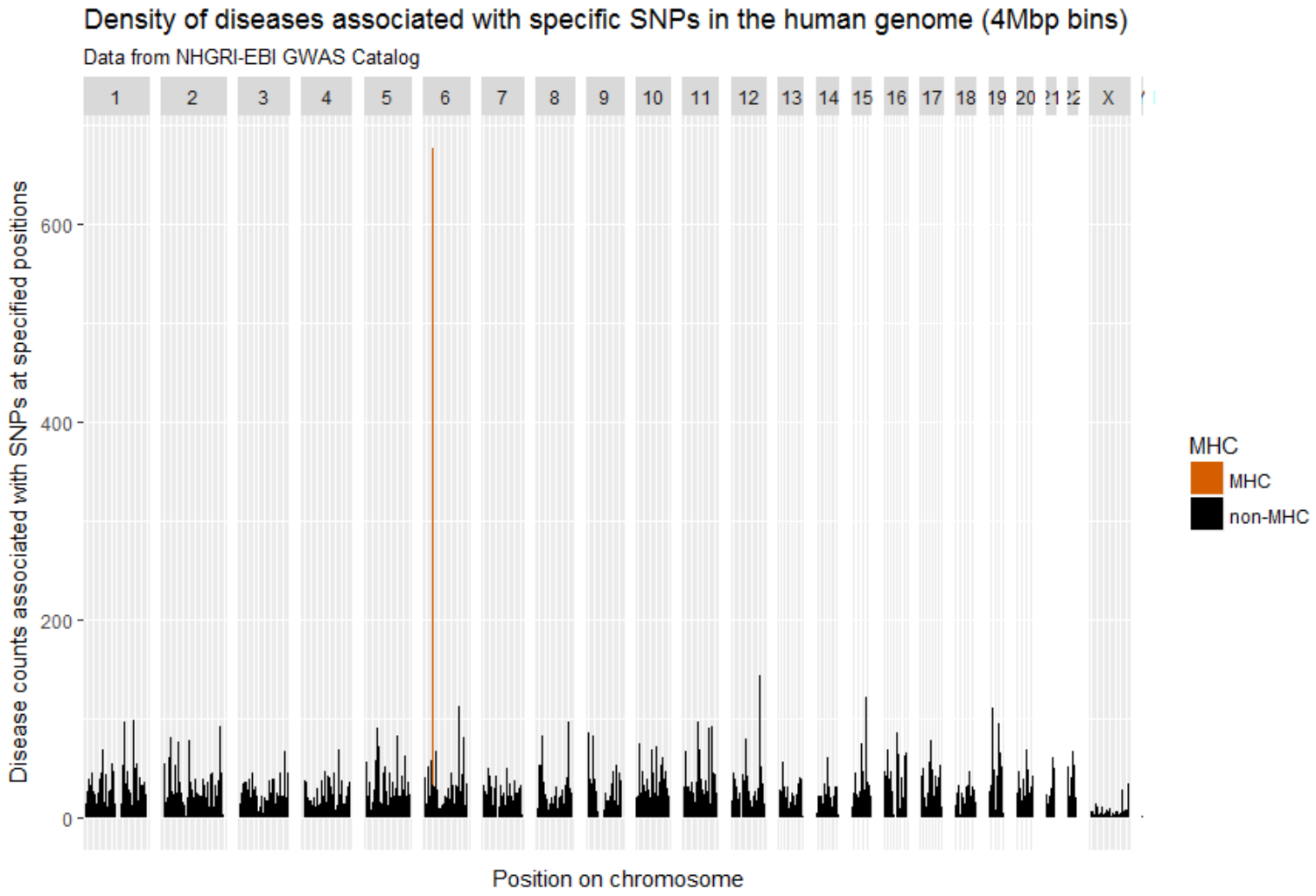
Published Genome-Wide Associations through 12/2012
Published GWA at p≤5X10^{-8} for 17 trait categories



NHGRI GWA Catalog
www.genome.gov/GWAStudies
www.ebi.ac.uk/fgpt/gwas/

National Human Genome Research Institute

EMBL-EBI



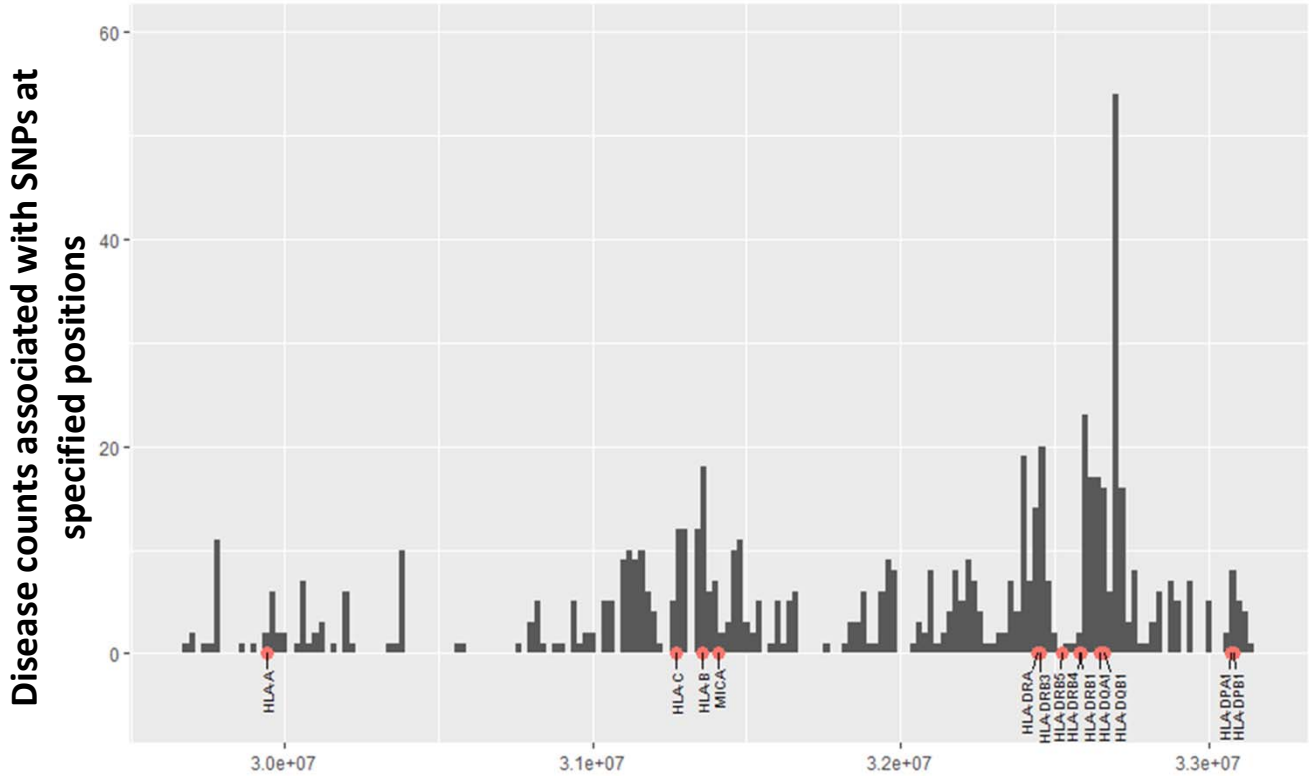
A SNP may appear twice if it has been associated with more than one disease

***Clark* et al. The Dichotomy Between Disease Phenotype Databases and the Implications for Understanding Complex Diseases Involving the Major Histocompatibility Complex. Intern. J. of Immunogenetics 42:413-422, 2015**

Genome-wide Density of SNPs Associated with Diseases

- The MHC (chr6:29-33Mb = 4Mb) includes ~260 genes, about half of which are involved in the immune response
- 884 unique loci associated with 479 unique traits/diseases; 112 unique disease phenotypes
- The MHC is recognized as the most important region of the human genome in relation to disease susceptibility

Density of diseases associated with specific SNPs in MHC region (20Kbp bins)
Data from NGHRI-EBI GWAS Catalog



Position on chromosome; HLA gene locations indicated

Approaches

1. Sequencing characterization of the MHC: Complete and accurate sequencing of the 4Mb of heterozygote samples using long sequencing reads (3-10kb) and de novo (not reference-based) assembly. Eventual objective is the generation of MHC haplotypes if possible for the whole MHC or any other sizeable segment of interest.

2. Identify MHC genomic elements, like miRNAs, long non-coding RNAs, pseudogenes, methylation sites and possibly new elements with functional roles.

3. Use alternative approaches combining NGS/Genetics and Complexity Theory/Physics that provide totally new insights in the relationships of genomic sequences and their possible interdependences by computational means.

Dapprich *et al. BMC Genomics* (2016) 17:486
DOI 10.1186/s12864-016-2836-6

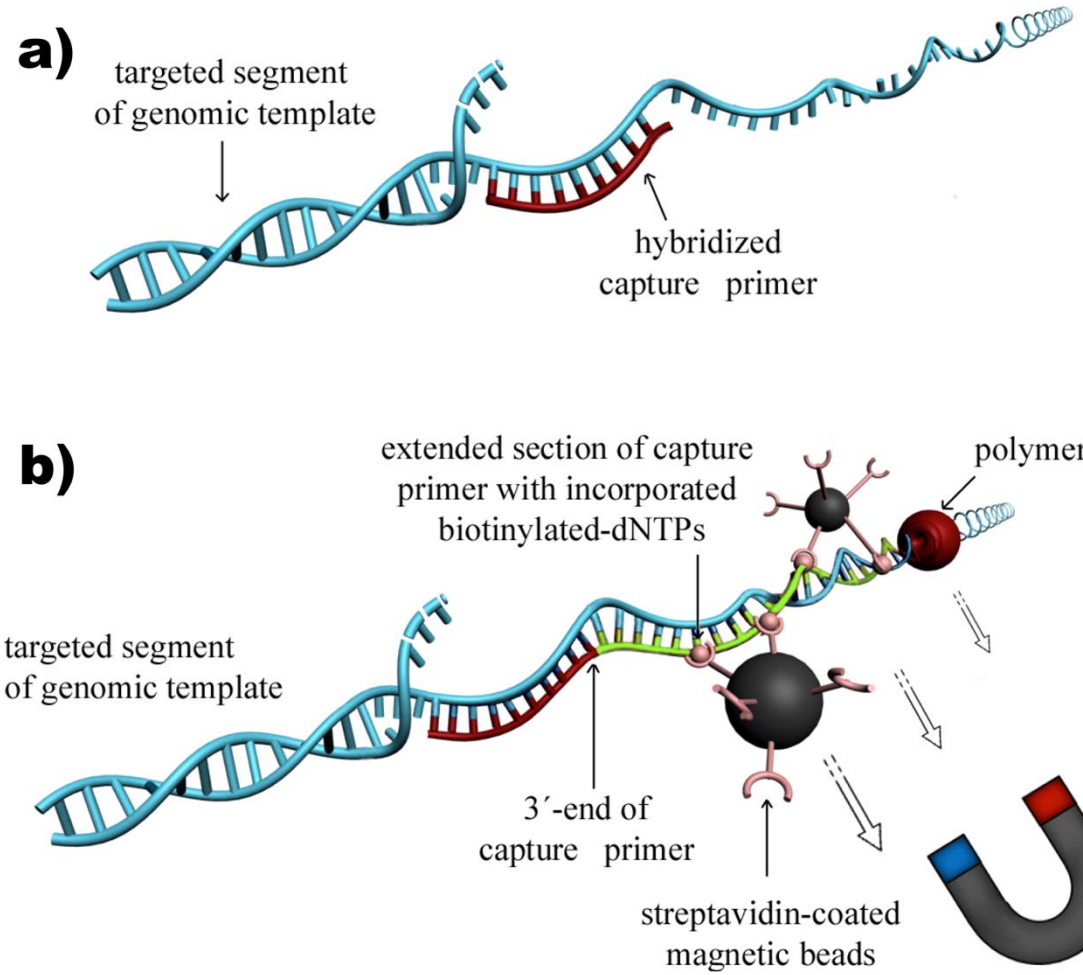
BMC Genomics

METHODOLOGY ARTICLE Open Access

The next generation of target capture technologies - large DNA fragment enrichment and sequencing determines regional genomic variation of high complexity

Johannes Dapprich[1*], Deborah Ferriola[1,2], Kate Mackiewicz[2], Peter M. Clark[2], Eric Rappaport[3], Monica D'Arcy[4], Ariella Sasson[4], Xiaowu Gai[4], Jonathan Schug[5], Klaus H. Kaestner[5] and Dimitri Monos[2,6*]

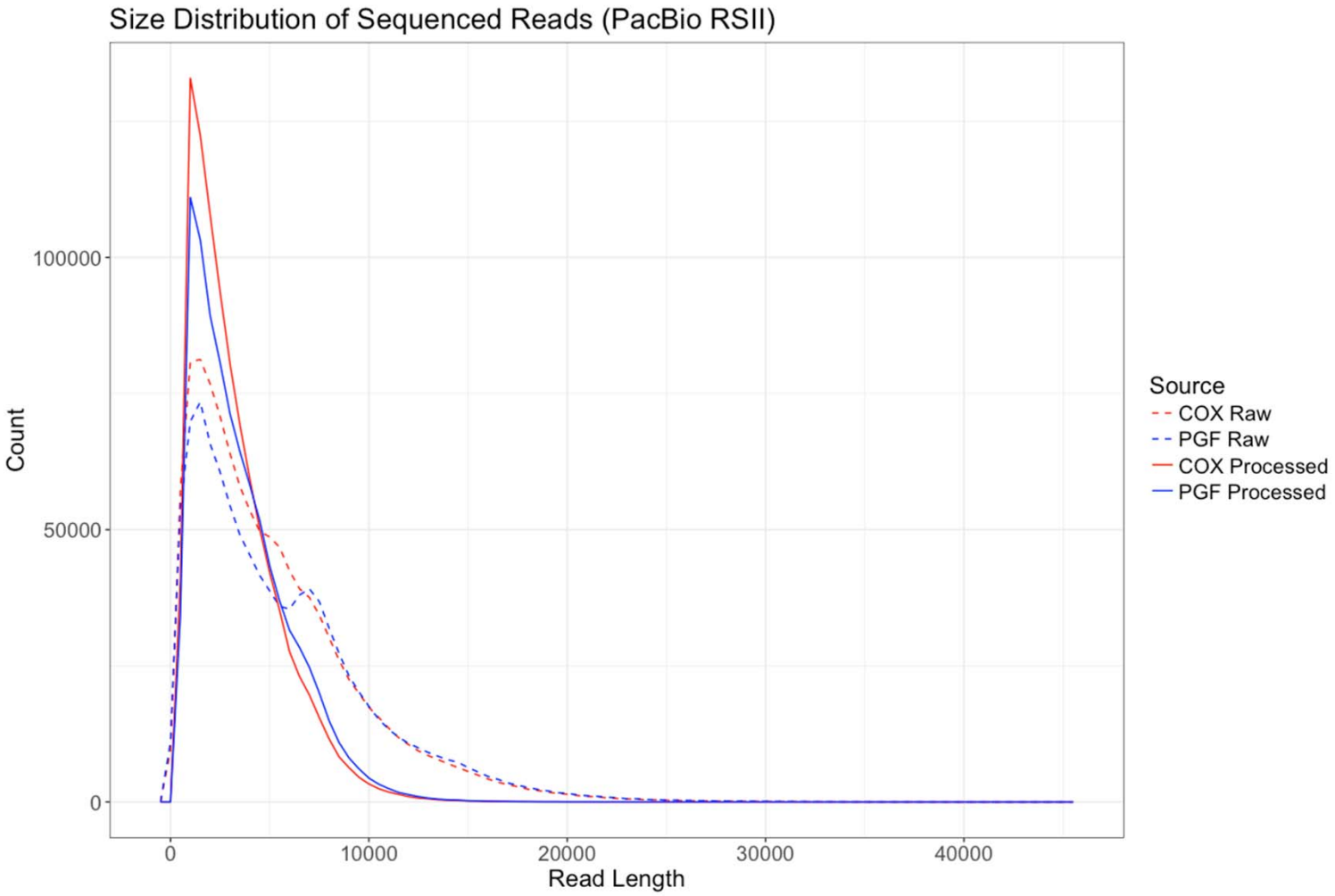
Region Specific DNA Extraction (RSE)



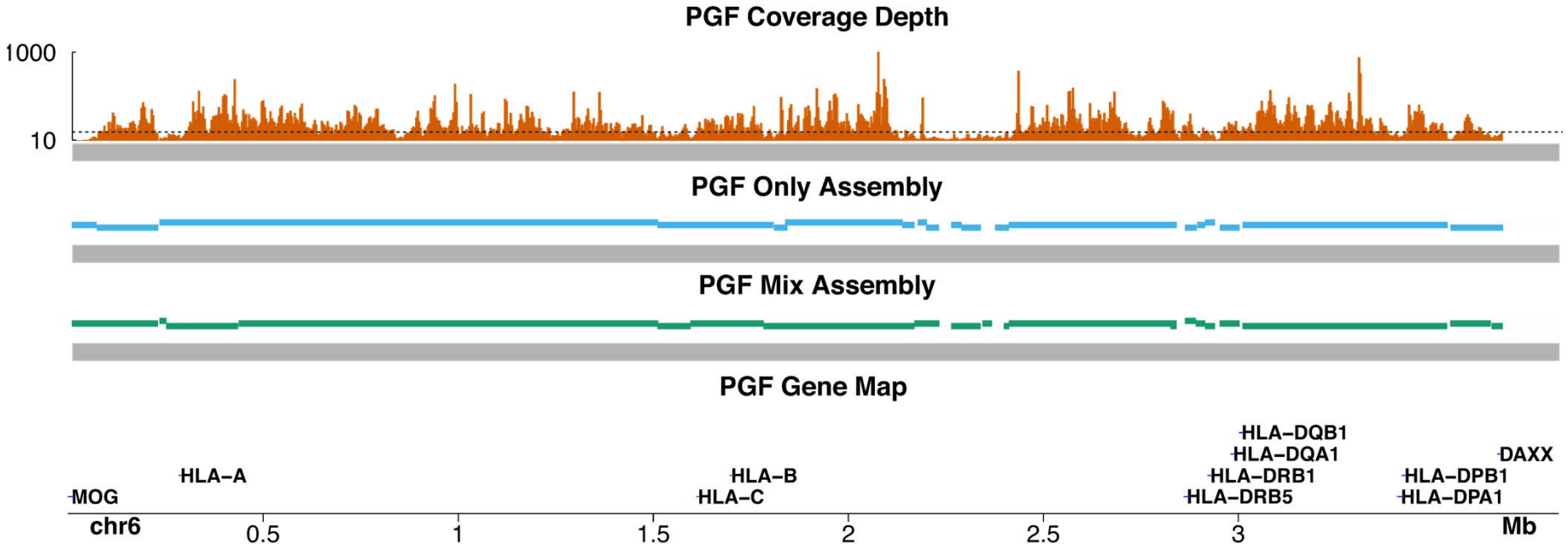
- **36 different genomic sequences of a single DNA sample, have been targeted captured and sequenced, totaling 25Mb**

- **Using a software program we have developed (Antholigo) 500 oligos were designed for the capture of the 4Mb of the MHC, i.e one oligo every 8Kb.**

PacBio Sequencing for *de novo* Assembly of the MHC

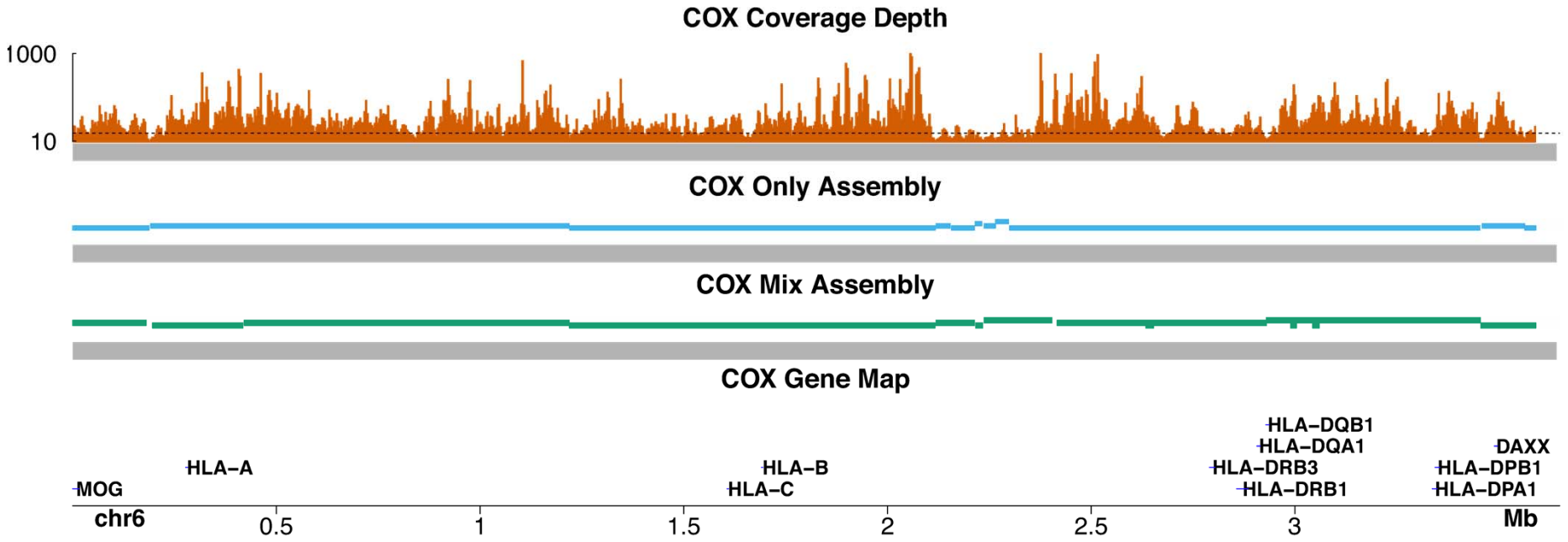


PGF Assembly



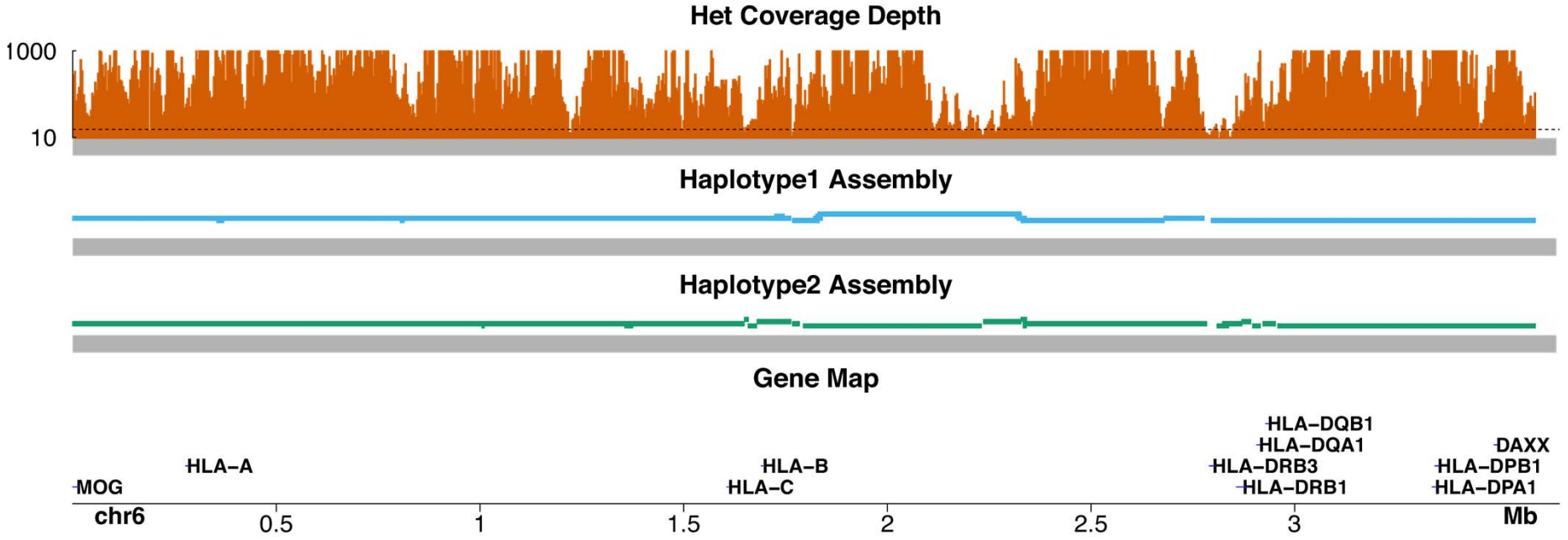
- **PGF Alignment: Mean depth of coverage: 176X, 93.8% of positions >20x**
- **PGF Assembly using only PGF reads**
  - **21 contigs >10 Kb. 96% coverage of targeted region with 99.95% accuracy**
  - **Longest contig 1.2 Mb**
- **PGF Assembly using mixed reads**
  - **20 contigs >10 kb. 96% coverage of targeted region with 99.69% accuracy.**
  - **Longest contig 1.0 Mb**

COX Assembly



- **COX Alignment: Mean depth of coverage 253X, 99.6% of positions >20X**
- **<u>COX Assembly using only COX reads</u>**
  - **11 contigs >10 Kb. 99% coverage of targeted region with 99.97% accuracy.**
  - **Lonest Contig 1.1 Mb**
- **<u>COX assembly using mixed (PGF+COX) reads</u>**
  - **13 contigs >10 Kb. 99% coverage of targeted region with 99.95% accuracy.**
  - **Longest Contig 900 Kb**

Heterozygous Sample Assembly



- **Haplotype 1**
  - **13 contigs >10 Kb**
  - **Longest contig 1.8 Mb**
- **Haplotype 2**
  - **18 contigs >10 Kb**
  - **Longest contig 1.7 Mb**
- **Accuracy**
  - **The HLA haplotypes derived from family tree analysis was the same as the HLA haplotypes after sequencing and *de novo* assembly for 10 genes. The total number of bases in the 19 HLA alleles of the two haplotypes were 105,098 with an accuracy 99.95%.**
  - **96.6% (4816/4988) of expected OmniExpress-24 SNPs found in contigs, with 99.2% (4777/4816) accuracy.**

Genetic and epigenetic fine mapping of causal autoimmune disease variants

Kyle Kai-How Farh[1,2]*, Alexander Marson[3]*, Jiang Zhu[1,4,5,6], Markus Kleinewietfeld[1,7]†, William J. Housley[7], Samantha Beik[1], Noam Shoresh[1], Holly Whitton[1], Russell J. H. Ryan[1,5], Alexander A. Shishkin[1,8], Meital Hatan[1], Marlene J. Carrasco-Alfonso[9], Dita Mayer[9], C. John Luckey[9], Nikolaos A. Patsopoulos[1,10,11], Philip L. De Jager[1,10,11], Vijay K. Kuchroo[12], Charles B. Epstein[1], Mark J. Daly[1,2], David A. Hafler[1,7]§ & Bradley E. Bernstein[1,4,5,6]§

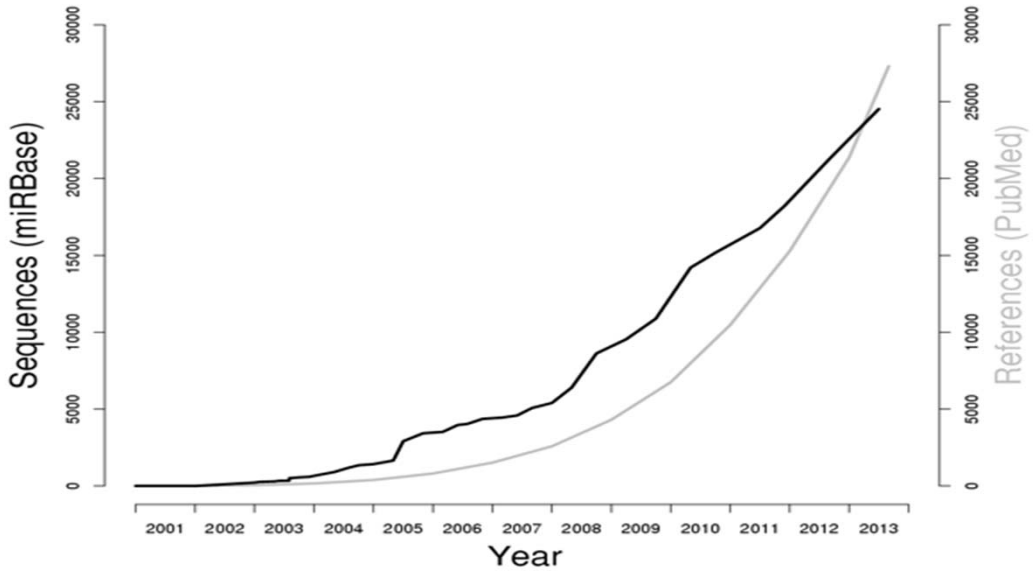
19 FEBRUARY 2015 | VOL 518 | NATURE | 337

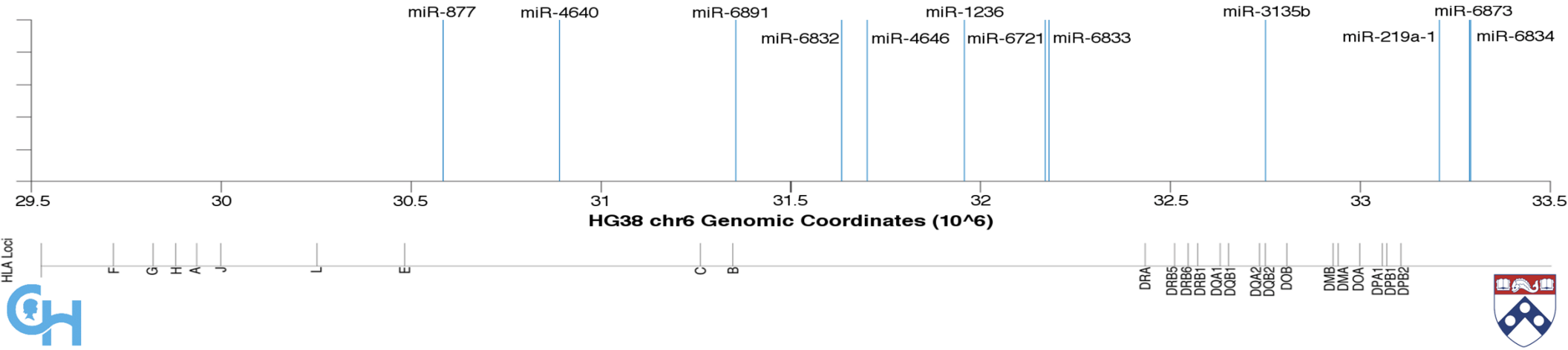
GWAS data reveal that ~90% of causal variants in autoimmune diseases are non-coding

The above statement is concordant with the major findings of the ENCODE project, whereby the majority of the genome encodes for meaningful elements of primarily regulatory nature

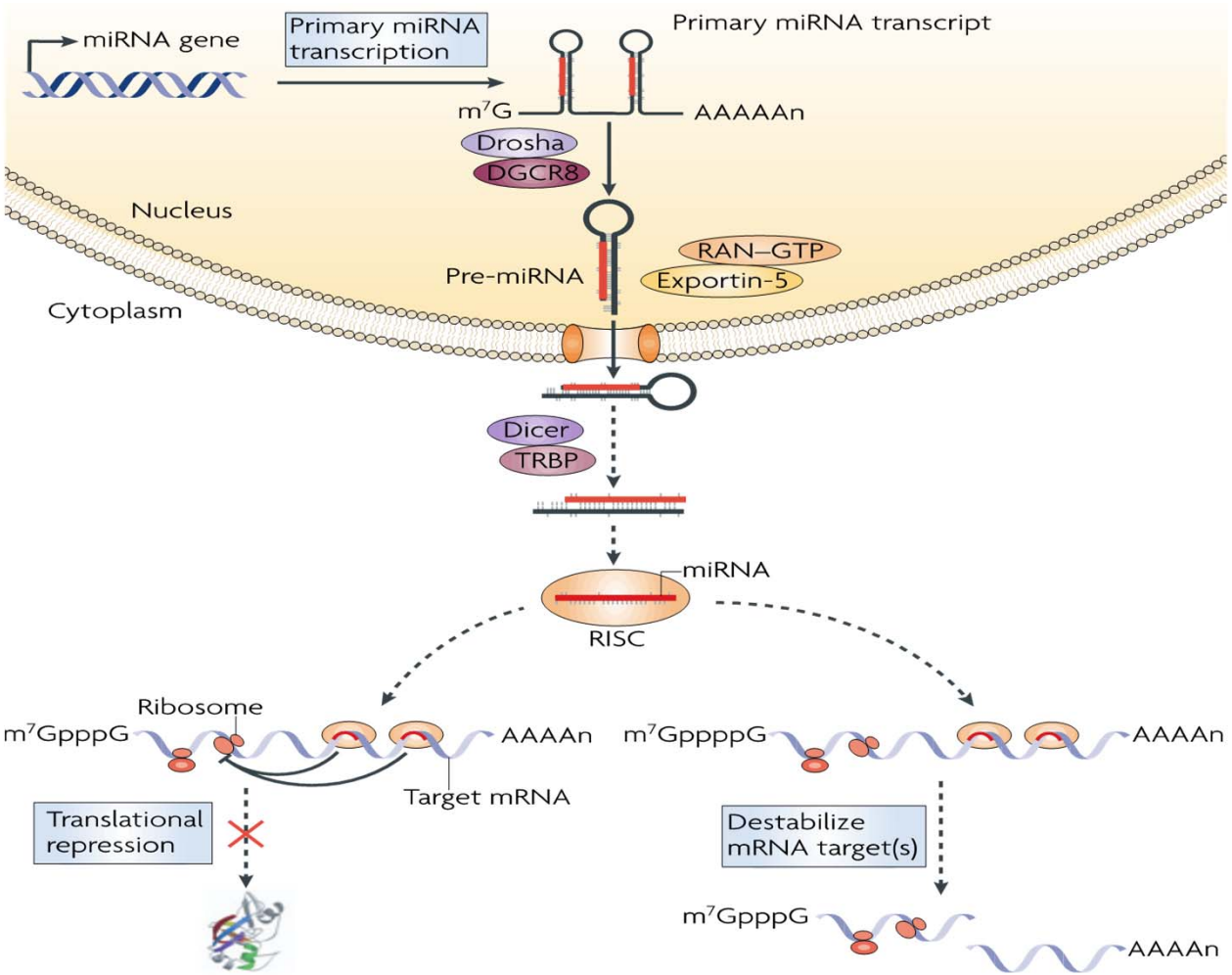
Therefore the "Junk DNA" theory is definitely a theory of the past …

Annotated miRNA – miRBase (Rel. 21)

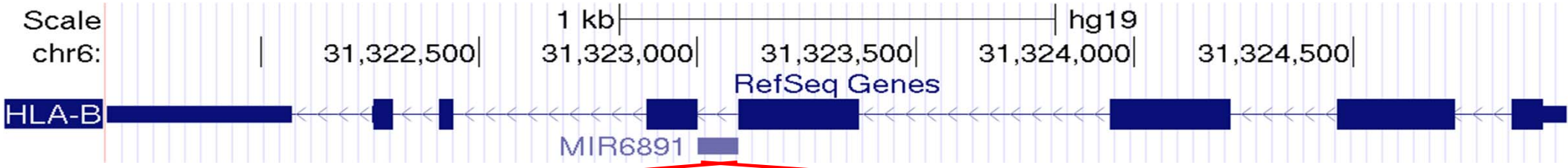




MicroRNAs: what do they do? MicroRNA biogenesis and mechanism of action



Lodish, H.F. et al. (2008). *Nat Rev Immun.*, 8, 120-130.

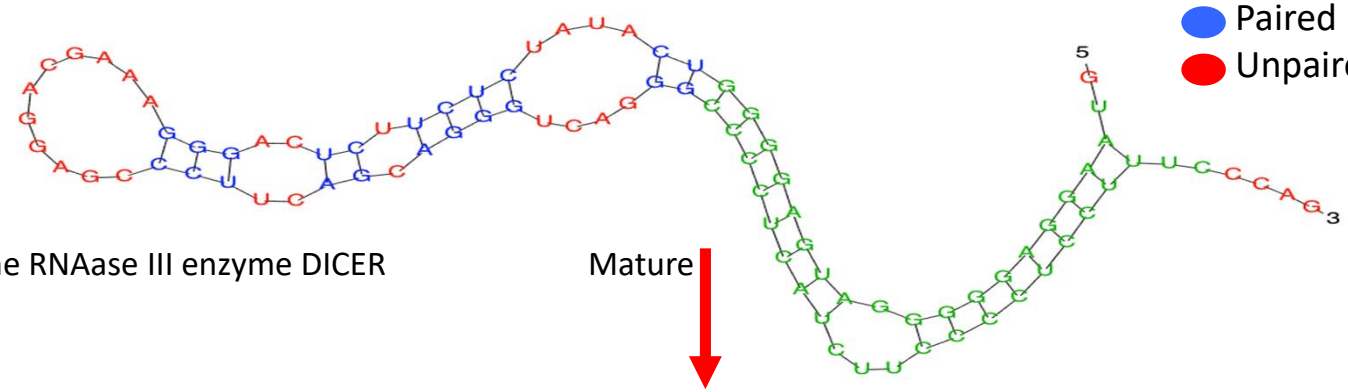


GUAAGGAGGGGGAUGAGGGGUCAUAUCUCUUCUCAGGGAAAGCAGGAGCCCUUCAGCAGGGUCAGGGCCCCUCAUCUUCCCCUCCUUUCCCAG

5’ End

3’ End

Bioenergetically stable Pre-miRNA hairpin structure formation



Cleavage by the RNAase III enzyme DICER

Mature miRNA

hsa-miR-6891-5p

UAAGGAGGGGGAUGAGGGG

hsa-miR-6891-3p

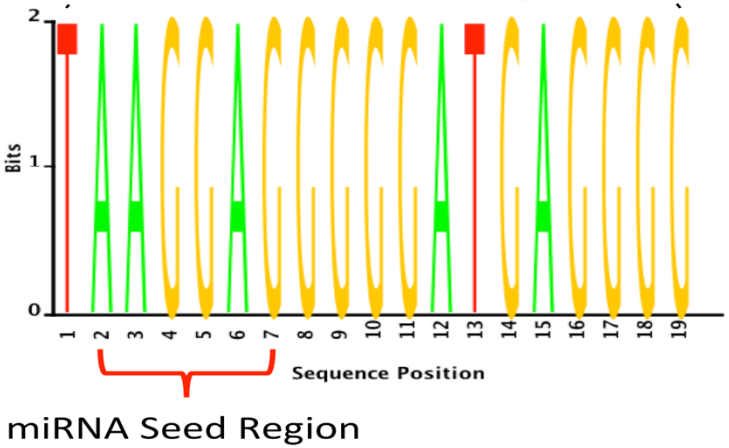
CCCUCAUCUUCCCCUCCUUUC

Translational Suppression of mRNA Targets

A)

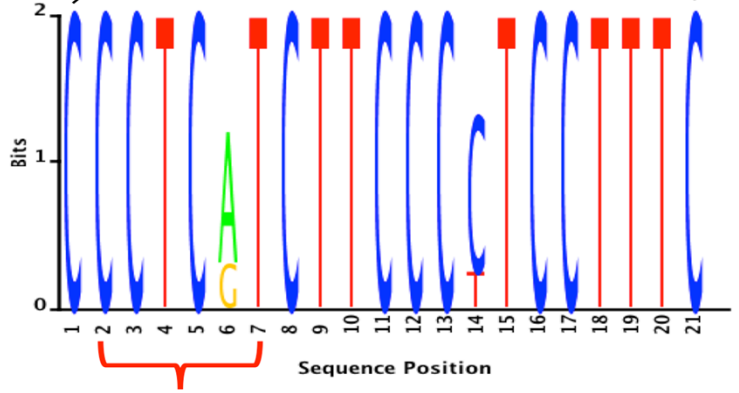
Allele	Count	HLA-B Intron 4 Sequence Alignment
	110	GTAAGGAGGGGGATGAGGGGTCATATCTCTTCTCAGGGAAAGCAGGAGCCCTTC..........AGCAGGGTCAGGGCCCCTCATCTTCCCCTCCTTTCCCAG
	88	GTAAGGAGGGGGATGAGGGGTCATATCTCTTCTCAGGGAAAGCAGGAGCCCTTCTGGAGCCCTTCAGCAGGGTCAGGGCCCCTCGTCTTCCCCTCCTTTCCCAG
	72	GTAAGGAGGGGGATGAGGGGTCATATCTGTTCTCAGGGAAAGCAGGAGCCCTTCTGGAGCCCTTCAGCAGGGTCAGGGCCCCTCATCTTCCCCTCCTTTCCCAG
	65	GTAAGGAGGGGGATGAGGGGTCATATCTCTTCTCAGGGAAAGCAGGAGCCCTTC..........AGCAGGGTCAGGGCCCCTCATCTTCCCTTCCTTTCCCAG
	40	GTAAGGAGGGGGATGAGGGGTCATATCTCTTCTCAGGGAAAGCAGGAGCCCTTCTGGAGCCCTTCAGCAGGGTCAGGGCCCCTCATCTTCCCCTCCTTTCCCAG
	7	GTAAGGAGGGGGATGAGGGGTCATATCTCTTCTCAGGGAAAGCAGGAGCCCTTCTGGAGCCCTTCAGCAGGGTCAGGACCCCTCATCTTCCCCTCCTTTCCCAG
	1	GTAAGGAGGGGGATGAGGGGTCATATCTCTTCTCAGGGAAAGCAGGAGCCCTTC..........AGCAGGGTCAGGGCCCCTCATCTTCCCTTCCTTTCCCGG
Total	1	GTAAGGAGGGGGATGAGGGGTCATATCTCTTCTCAGGGAAAGCAGGAGCCCTTCTGGAGCCCTTCAGCAGGGTCGGGGCCCCTCATCTTCCCCTCCTTTCCCAG
	384	

B)



miRNA Seed Region

C)



miRNA Seed Region

Studying the Role of miR-6891-5p
Experimental design

1. Establish appropriate cell model
   - Evaluate expression of miR-6891-5p
2. Assess the role of miR-6891-5p within a cell model

   That is: Identify putative miRNA targets through RNA expression microarray analysis (miR-6891-5p inhibition vs. control)

   For inhibition of miR-6891-5p, a construct with antisense of miR-6891-5p and a scrambled sequence as a control needed to be expressed in COX cells

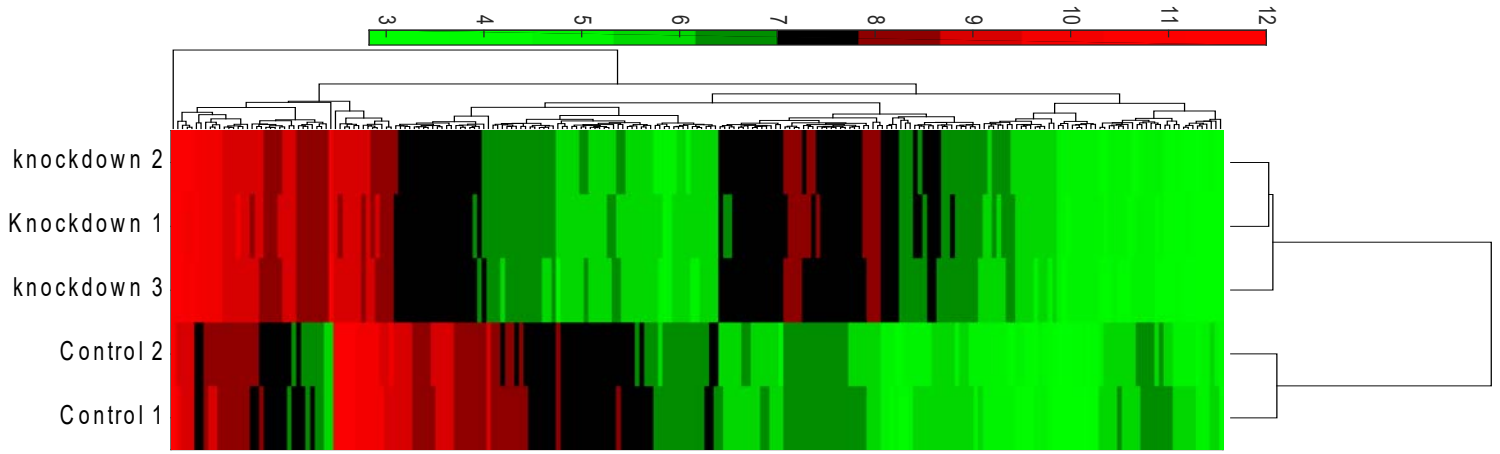
   Therefore, antisense and scrambled sequence expressing plasmids were packaged separately into lentiviruses for better delivery in COX cells

Identification of miR-6891-5p targets

All samples were hybridized onto the Affymetrix HuGene 2.0 ST array for analysis. 1.35 million probes/ ~33,500 interrogated coding transcripts/ ~11,000 interrogated long intergenic non-coding transcripts

miR-6891-5p Inhibited Samples Inhibition of miR-6891-5p within the COX B-lymphocyte cell line using a lentivirus construct engineered to express the antisense transcript of miR-6891-5p

Control Samples Scrambled antisense miR-6891-5p lentivirus expression vector was used as control



HSA miR-6891-5p differentially regulates targets in B-cell line knockdown vs. control samples (RNA microarray analysis)

104 up-regulated transcripts were identified. Only top 10 are shown. Identified genes are putative targets of HSA-miR-6891-5p.

Ensemble Gene ID	Gene Symbol	Fold Change	FDR
ENSG00000226777	KIAA0125	22.7	1.2E-02
ENSG00000211890	IGHA2	8.5	2.0E-02
ENSG00000186522	SEPT10	7.8	3.8E-03
ENSG00000229807	XIST	7.5	2.0E-03
ENSG00000133124	IRS4	6.4	4.5E-03
ENSG00000237438	CECR7	6.3	2.4E-02
ENSG00000258667	HIF1A-AS2	6.0	7.5E-04
ENSG00000079691	LRRC16A	5.9	9.8E-04
ENSG00000184258	CDR1	5.6	3.2E-02
ENSG00000073282	TP63	5.4	2.6E-03

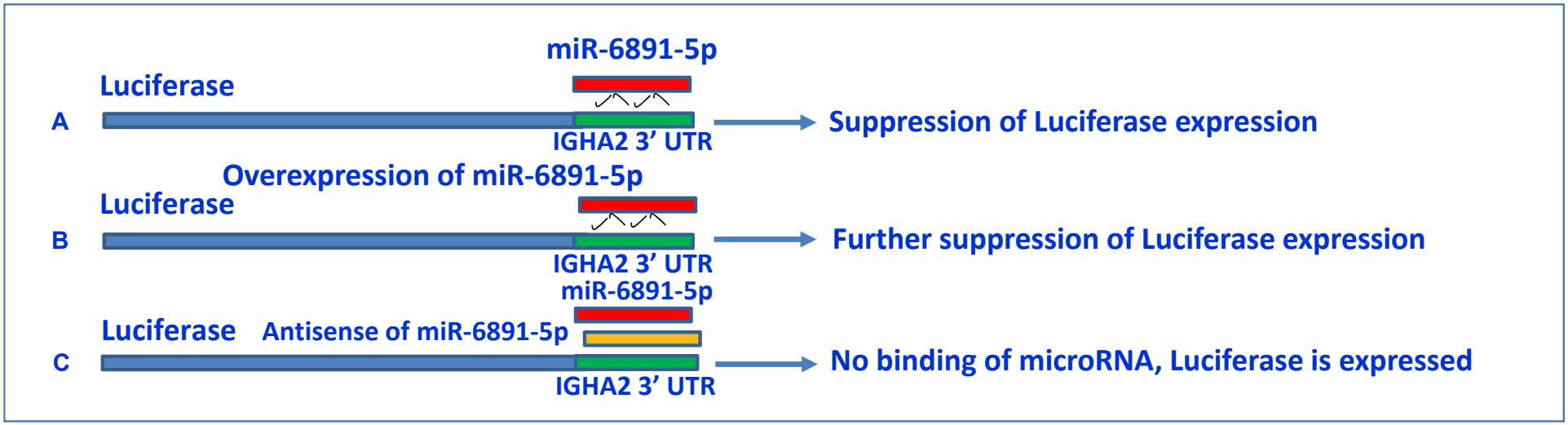
99 down-regulated differentially expressed transcripts were identified. Not shown. Identified genes are indirect targets of HSA-miR-6891-5p.

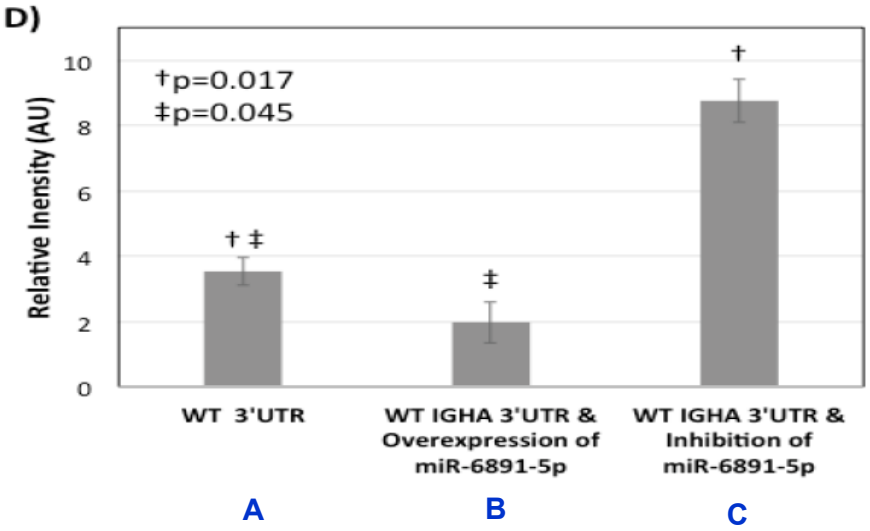
Putative mRNA Targets of miR-6891-5p – Disease Association

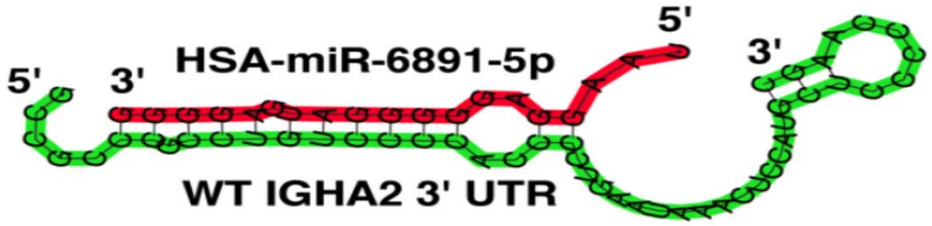
DISEASE (17/52)	Targeted Genes
Crohn's disease ulcerative colitis	IRAK3, FCRL3
rheumatoid arthritis	CXCR3, FCRL3
asthma	IRAK3, CXCR3
thyroid disease, autoimmune	FCRL3
multiple sclerosis	FCRL3
hepatitis, autoimmune	FCRL3
Addison's disease	FCRL3
diabetes, type 2	IRS4, SORBS1, IRS4
diabetes, type 1	FCRL3
bladder cancer	RGS6
Graves' disease	FCRL3
systemic lupus erythematosus	CXCR3, GMAP5
Alzheimer's Disease	FOS
lung cancer	RGS6
Sebaceous tumors, somatic	LEF1
Urinary bladder cancer	TP63
Chronic lymphocytic leukemia	GRAMD1B

Putative mRNA targets of miR-6891-5p identified by microarray analysis were found to be involved in 52 diseases (OMIM), including the subset of autoimmune and cancer related diseases listed above.

miRNA-6891-5p targets the 3’UTR of the heavy chain IgA mRNA







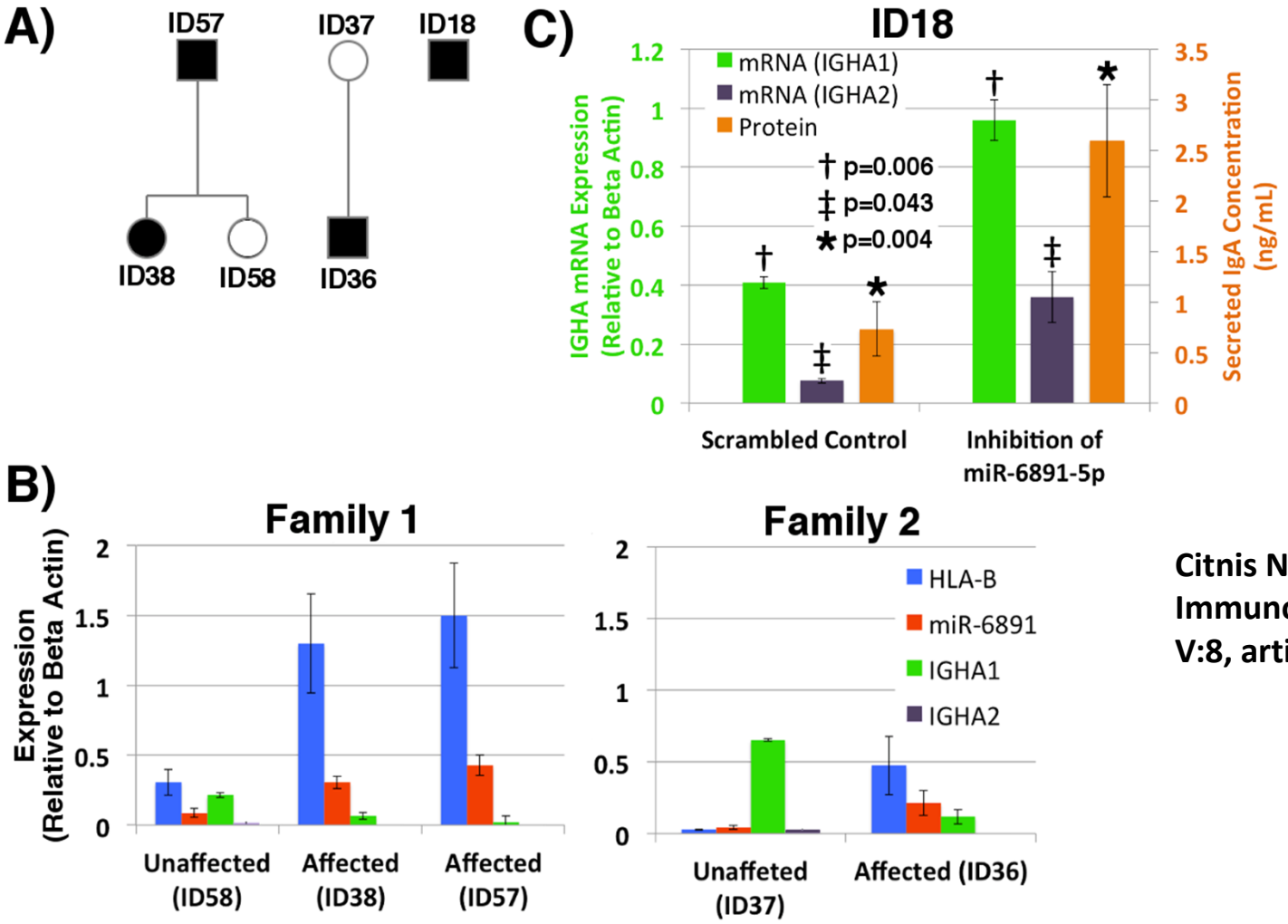
IGHA1 and IGHA2 3’UTR are highly conserved

Selective IgA deficiency

- Most common antibody deficiency (can be up to 0.6% and is population dependent)
- IgA deficiency is IgA level of 0.07 g/l after the age of four years in the absence of IgG and IgM deficiencies.
- Patients suffer from increased incidences of upper respiratory tract infections.
- Selective IgA deficiency is believed to be the result of defects in B-cell maturation

Nature Reviews Immunology 13, pp; 519–533

Exploring the role of miR-6891-5p in selective IgA deficiency



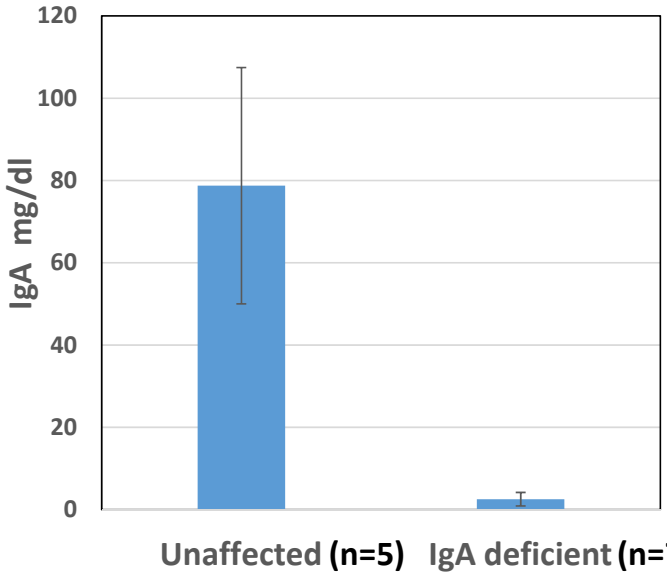
Citnis N. et al. Frontiers in Immunology, May 2017, V:8, article 583

Effect of miRNA-6891-5p suppression in *IGHA1* expression in human primary low-IgA expressing B-cells

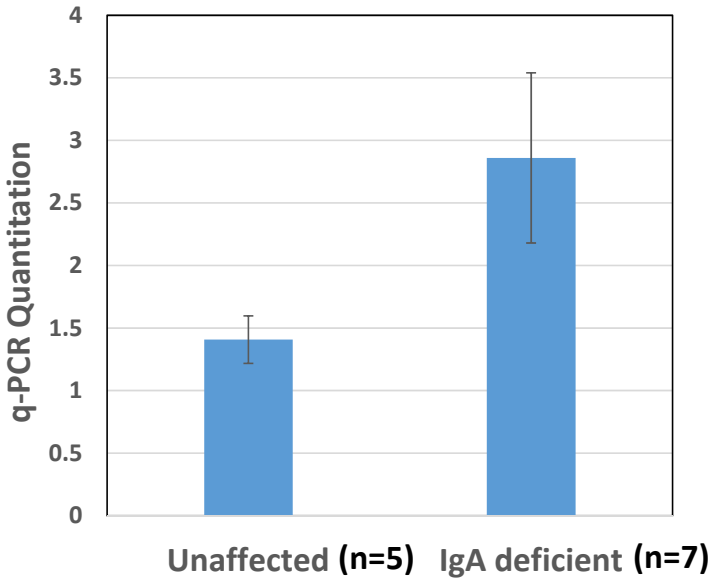
Method:

- **B-cells from families with low and normal IgA expressing siblings were purified.**
- **IgA low expressing patient cells were transduced with either the scrambled or the miR-6891-5p antisense construct expressing lentiviruses. Total RNA was purified from these cells and qPCR was performed to analyze *IGHA1* expression.**

IgA secretion by primary B cells (mg/dl)



Upregulation of *IGHA1* in antisense treated primary B cells



Conclusion: Suppression of miR-6891-5p can upregulate *IGHA1* expression in primary human IgA-low expressing B-cells

Significance of the MHC encoded miRNAs

- **HLA/MHC encoded microRNAs can influence/regulate many cellular functions and therefore many diseases**
- **We have studied only one of the targets of miR-6891-5p while expression analysis data indicate that the particular miRNA controls about 200 transcripts and therefore a wide spectrum of processes**
- **If MHC encodes more miRNAs and each miRNA can target hundreds of transcripts, MHC noncoding regions certainly play a critical role in human biology**

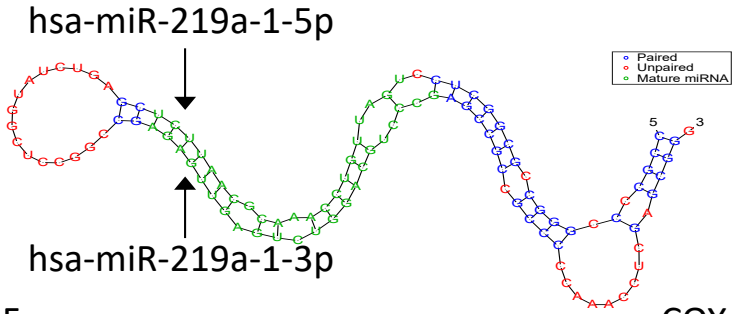
Identifying Novel miRNA of the MHC
Experimental Design

1. **Perform deep sequencing of the miRNA transcriptome on two homozygous lymphoblastoid cell lines with completely characterized MHC haplotypes, PGF and COX**

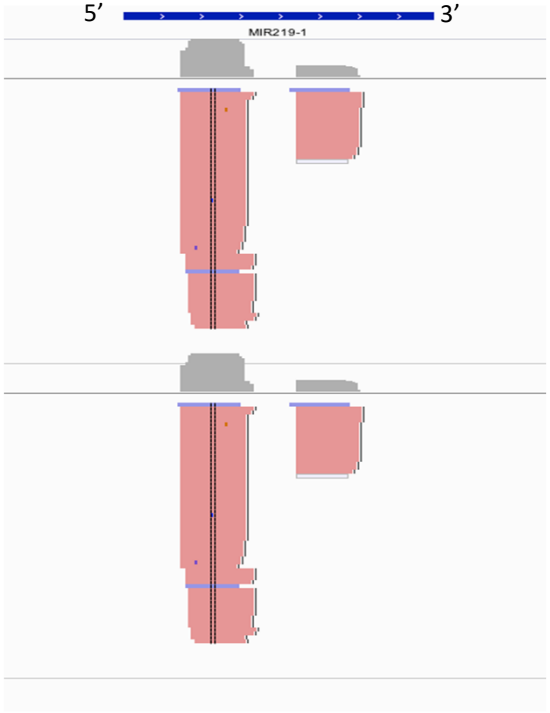
1. **Genome wide deep sequencing of the miRNA transcriptome reveals the expression of over 800 known miRNA with an average depth of coverage > 20X**

Expression Signature (HSA-miR-219a-1)

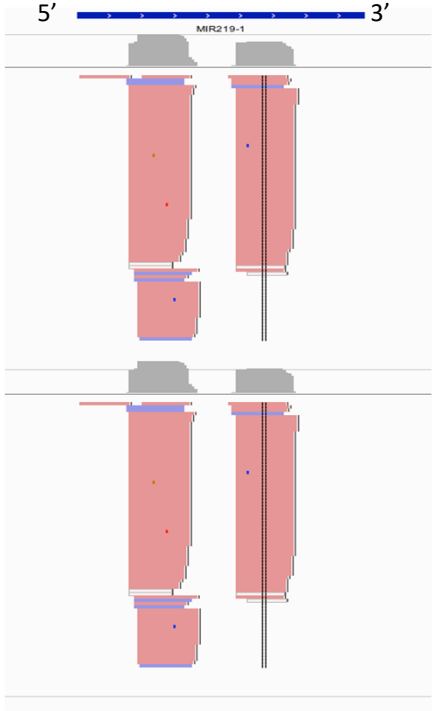
Pre-miRNA Hairpin Structure



PGF



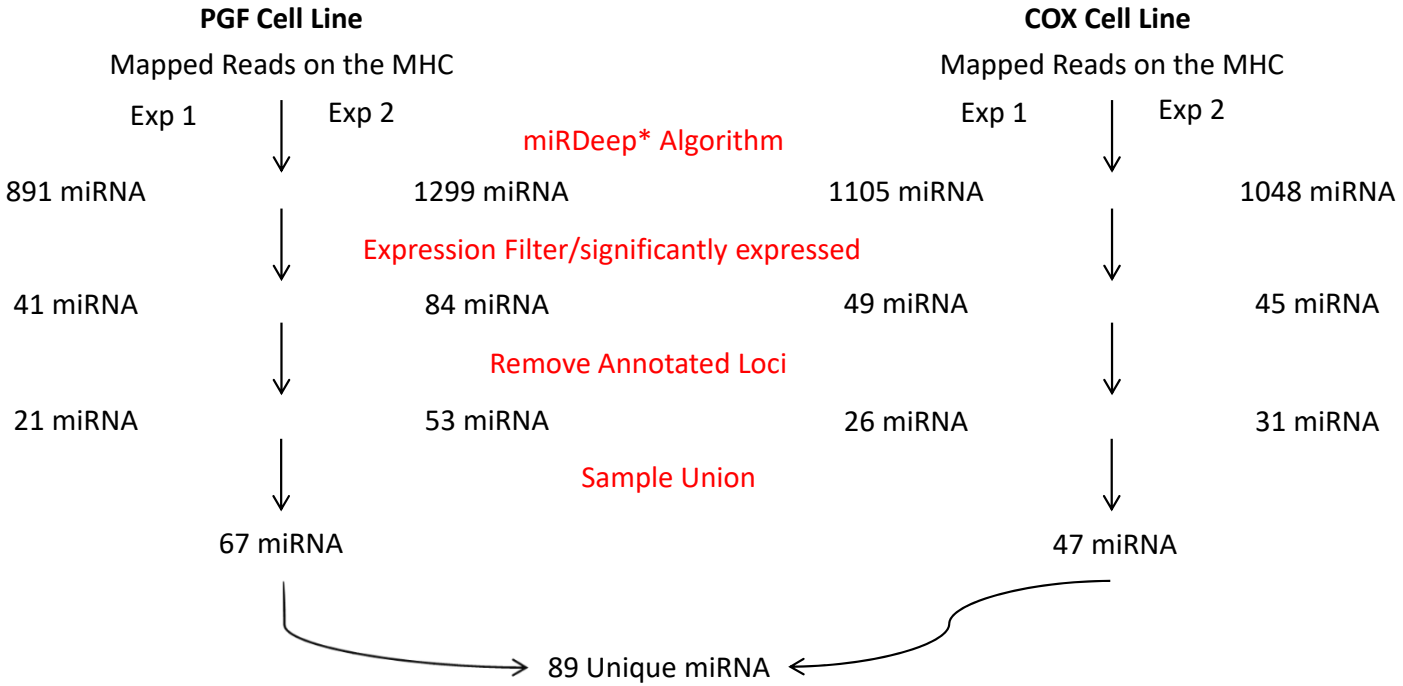
COX



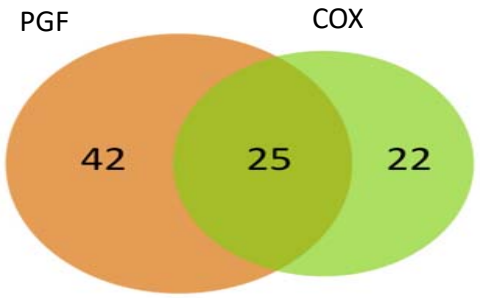
Biological Replicate 1

Biological Replicate 2

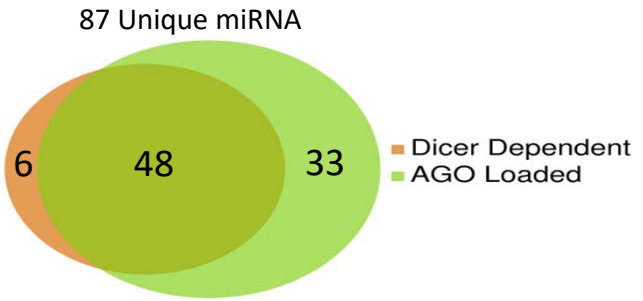
Identification of Novel miRNAs from the MHC



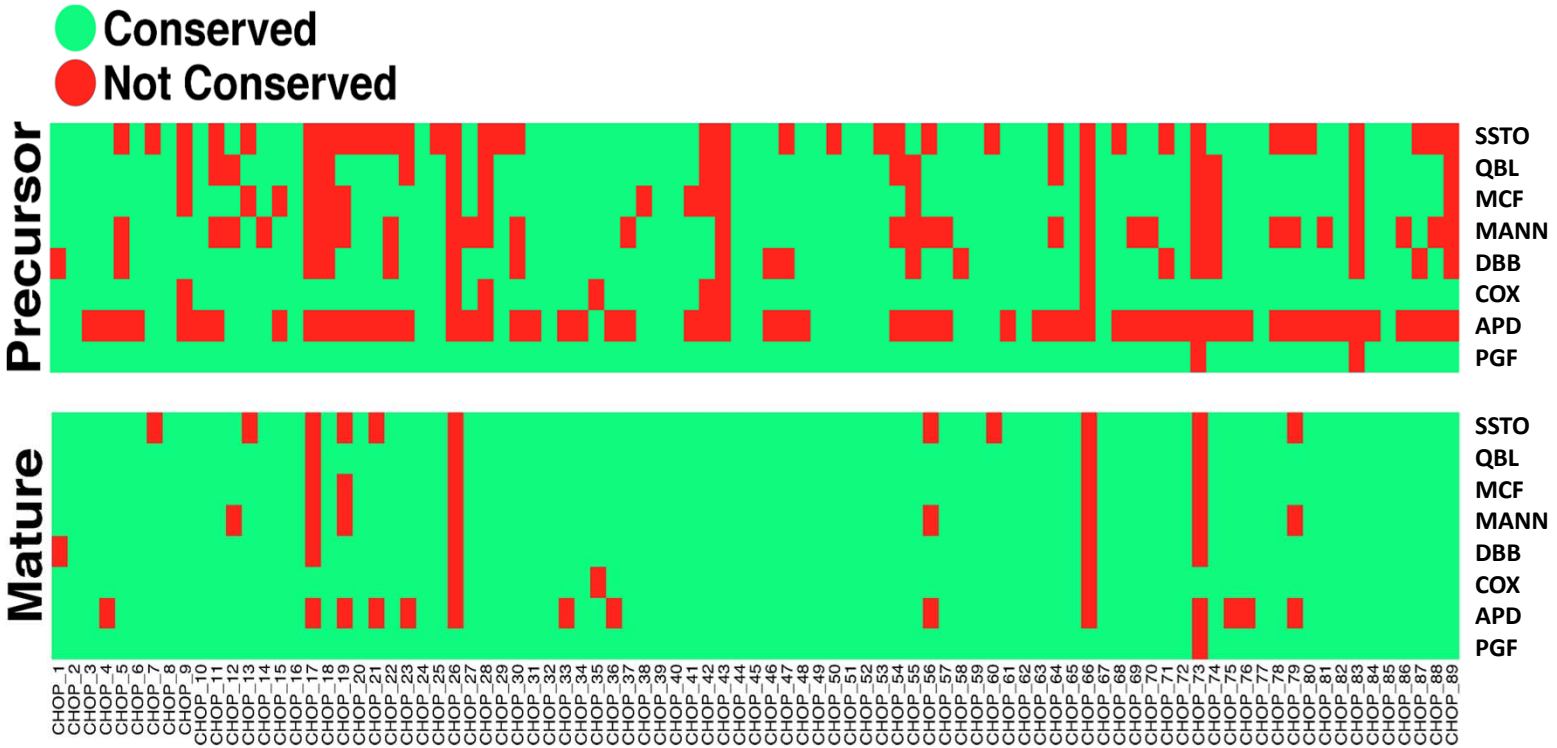
Haplotype Specific Expression



Additional Supporting Evidence



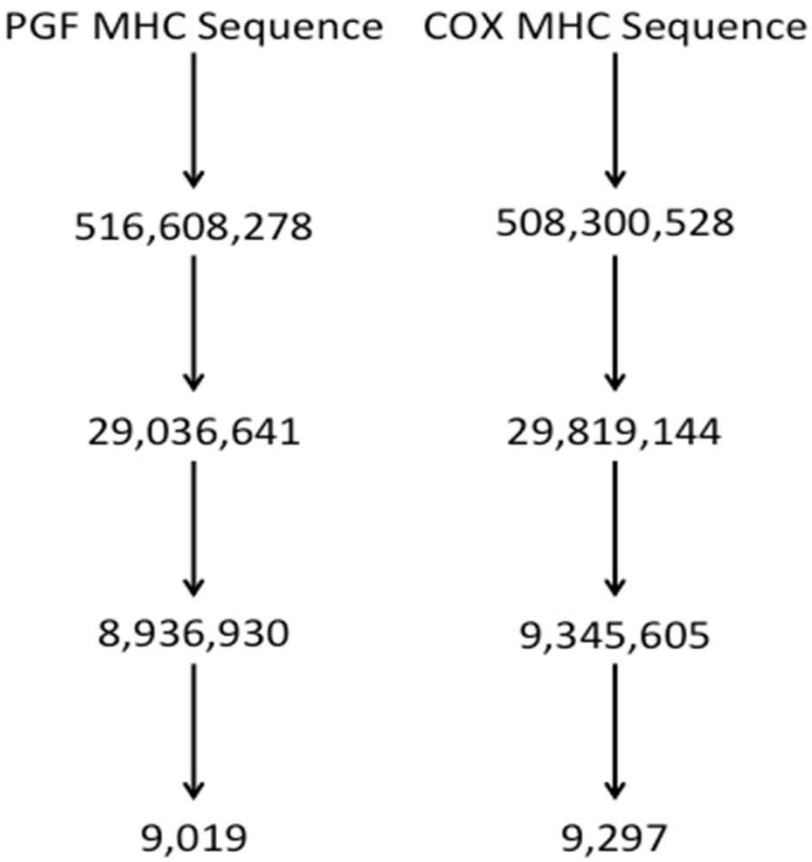
Patterns of Novel miRNAs Encoded by the MHC are Haplotype Specific



Associated Disease / Trait	Disease Associated SNP	Genomic Context	Novel miRNA ID
Age-related hearing impairment	rs6904029	non coding transcript exon variant	CHOP_34
Age-related macular degeneration	rs12153855	intron variant	CHOP_11
			CHOP_64
	rs2071277	intron variant	CHOP_46
Antinuclear antibody levels	rs2395185	intron variant	CHOP_66
Arthritis (juvenile idiopathic)	rs2395148	intron variant	CHOP_52
Asthma	rs3129943	intron variant	CHOP_52
Atopic dermatitis	rs12153855	intron variant	CHOP_11
			CHOP_64
Autism spectrum disorder	rs3132581	intron variant	CHOP_19
			CHOP_25
			CHOP_42
			CHOP_47
			CHOP_9
Bipolar disorder and schizophrenia	rs2524005	upstream gene variant	CHOP_34
	rs886424	non coding transcript exon variant	CHOP_42
			CHOP_47
			CHOP_9
Blood metabolite ratios	rs1046080	missense variant	CHOP_23
			CHOP_4
			CHOP_61
Cholesterol, total	rs3177928	3' UTR variant	CHOP_43
Complement C3 and C4 levels	rs11575839	synonymous variant	CHOP_2
			CHOP_33
			CHOP_39
	rs2071278	intron variant	CHOP_11
			CHOP_27
			CHOP_46
			CHOP_60
			CHOP_64
			CHOP_7
Crohns disease	rs1799964	upstream gene variant	CHOP_32
	rs9258260	upstream gene variant	CHOP_53
	rs9271366	intergenic variant	CHOP_26
			CHOP_43
			CHOP_52
			CHOP_66

Each novel miRNA (89) that is in LD with a disease associated SNP as annotated by GWAS Catalog is reported along with the associated trait/disease (64), SNP ID and genomic context of each variant (as annotated by GWAS catalog).

Computational prediction pipeline and putative miRNA loci identified from the MHC haplotype sequences of PGF and COX lymphoblastoid cell lines



1. Define Pre-miRNA Loci (58 ≤ Length ≥ 110)
2. *In silico* pre-miRNA Hairpin Folding
3. Minimum Free Energy Filter (-20Kcal/mol)
4. Linear Hairpin Filter
5. Machine Learning Filter (miRBoost)
6. Remove Entries Overlapping Annotated Genomic Features (GENCODE rel24)
7. Merge Overlapping Loci into Single Locus

	PGF-MHC	COX-MHC	Region 1 chr2:90248739-93848739	Region 2 chr14:21214050-24809251
Total Length	3,666,036	3,591,053	3,600,000	3,595,201
Exonic Base Count	408,172	394,882	4,427	338,354
Intronic Base Count	1,233,387	1,251,155	26,964	1,343,178
Intergenic Base Count	2,024,477	1,945,016	3,568,609	1,913,669
Computationally predicted pre miRNAs	9,019	9,297	1,828	6,996

A Multi-Disciplinary Approach for Understanding the Organization of DNA Sequences : Where Genetics Meets Physics and the High Throughput Sequencing Technologies Meet Complexity Theory

George Pavlos, Ph.D
Democritus University of Thrace, Department of Electrical and Computing Engineering Section of Telecommunications and Space Science
Research Team of Chaos & Complexity
Xanthi, Greece

Dimitri Monos, Ph.D
Immunogenetics Laboratory, The Children's Hospital of Philadelphia
Department of Pathology and Lab Medicine
University of Pennsylvania, School of Medicine

Structures

The world is full of Complexity!

- The weather
- The oceans
- The winds
- A tree structure

Complex Systems

A system is complex when it is composed of many parts that interconnect in intricate ways. The degree and nature of the relationships is imperfectly known. Metric for intricateness (or complexity) is the amount of information contained in the system. The overall emergent behavior is difficult to predict, even when subsystem behavior is readily predictable. Small changes in inputs or parameters may produce large changes in behavior.

Examples demonstrating the relevance of Complexity Theory in diverse physical systems

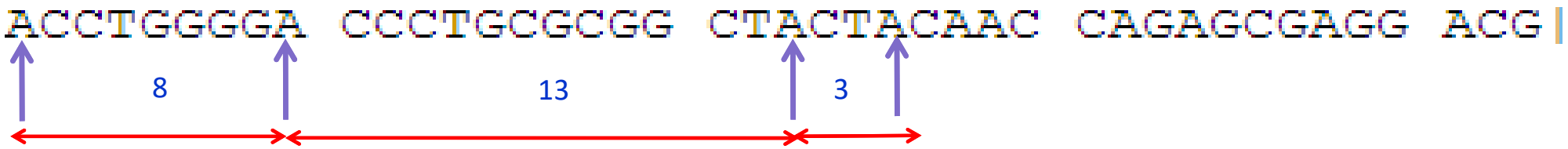
Space plasmas, atmospheric dynamics and seismicity (earthquakes) as well as brain functions, cardiac activity and most recently studying DNA sequence dynamics.

A remarkable agreement of theoretical predictions by Complexity Theory and experimental estimations was found in all cases.

The Problem

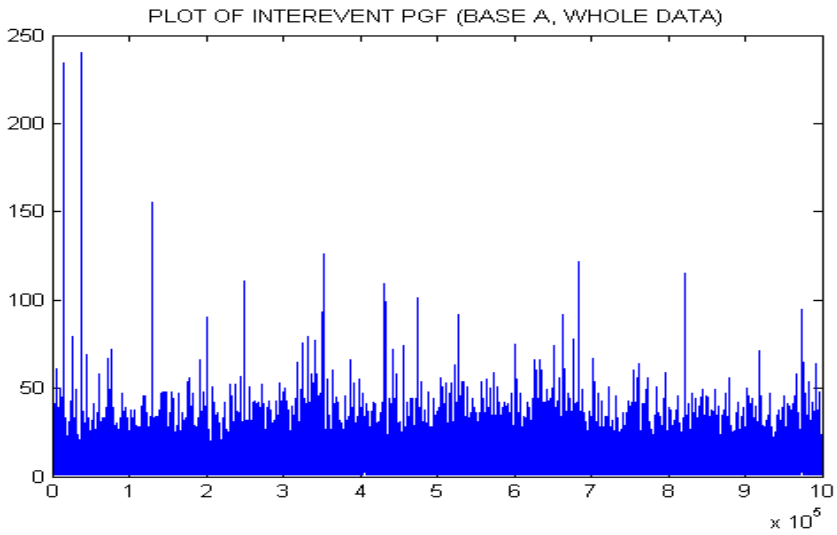
To understand the hidden dynamics (patterns/ information) encrypted within DNA sequences by tools from Complexity Theory and Non-Linear Dynamics

From Sequences of Bases to Arithmetic Data



Interevent Data: 8, 13, 3,…

We constructed the arithmetic data from the DNA sequences by counting the number of intervening bases from a specific base (A) to the next one and so on (Interevent data)



Contents lists available at ScienceDirect

Physica A

journal homepage: www.elsevier.com/locate/physa

Measuring complexity, nonextensivity and chaos in the DNA sequence of the Major Histocompatibility Complex

G.P. Pavlos [a], L.P. Karakatsanis [a,*], A.C. Iliopoulos [a], E.G. Pavlos [a], M.N. Xenakis [a], Peter Clark [b], Jamie Duke [b], D.S. Monos [b]

[a] Department of Electrical and Computer Engineering, Democritus University of Thrace, 67100 Xanthi, Greece
[b] Department of Pathology and Laboratory Medicine, Perelman School of Medicine, University of Pennsylvania and The Children's Hospital of Philadelphia, Philadelphia, PA 19104, USA

G.P. Pavlos et al. / Physica A 438 (2015) 188–209

The basic finding of this work was that the continuum of the MHC sequences are characterized by a dynamical process whereby long-range correlations are observed. Otherwise there is no discontinuity in the information included in different segments of the DNA.

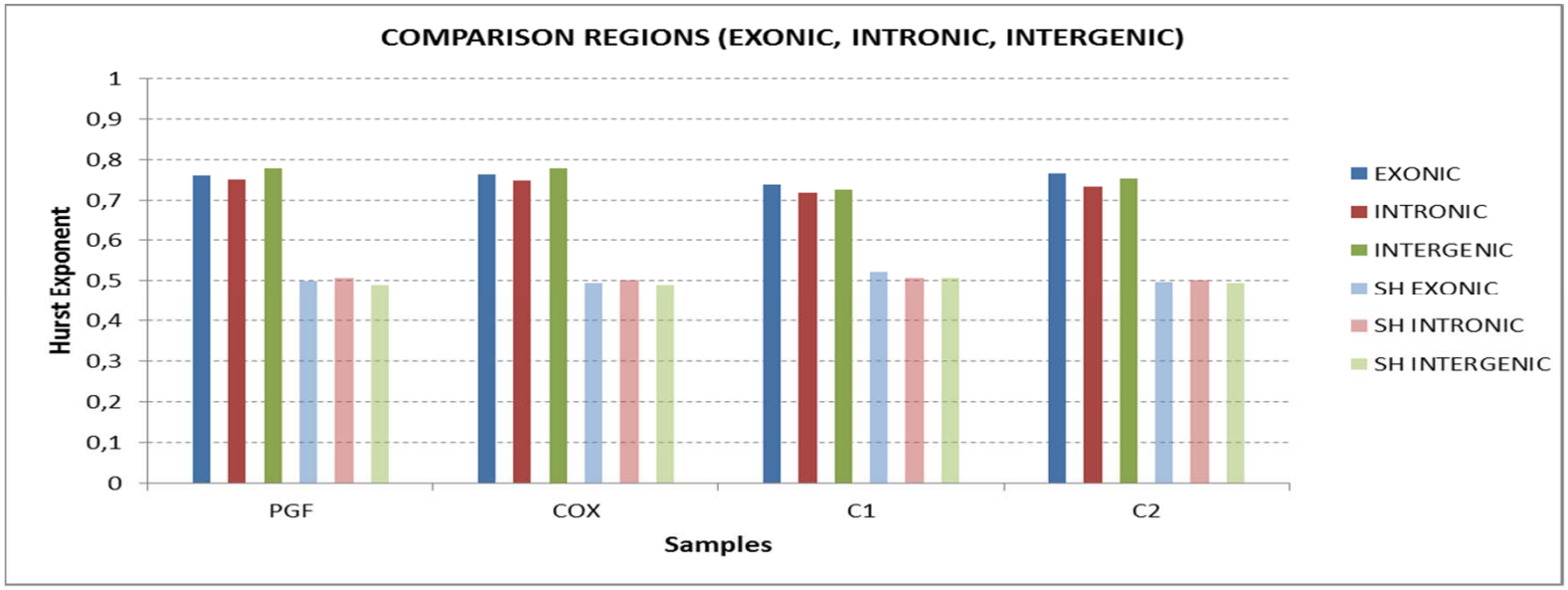
Experimental Design and Data

	PGF-MHC	COX-MHC	Region 1 chr2:90248739-93848739	Region 2 chr14:21214050-24809251
Total Length	3,666,036	3,591,053	3,600,000	3,595,201
Exonic Base Count	408,172	394,882	4,427	338,354
Intronic Base Count	1,233,387	1,251,155	26,964	1,343,178
Intergenic Base Count	2,024,477	1,945,016	3,568,609	1,913,669

PGF and COX complete MHC sequences were used. Region 1 was selected to have very few genes, with the overwhelming percentage of bases derived from intergenic space. Region 2 was selected to have the relative same ratio of base composition as well as gene density, making this region very similar to the ~4MB MHC. The individual sequences that comprise a specific genomic region (exonic, intronic and intergenic) for each assembly were concatenated into a single contiguous sequence, giving rise to four sequence files for each assembly. These four concatenated sequences were then randomly shuffled while preserving the underlying base composition, giving rise to four additional random sequences for each assembly.

The Hurst Exponent

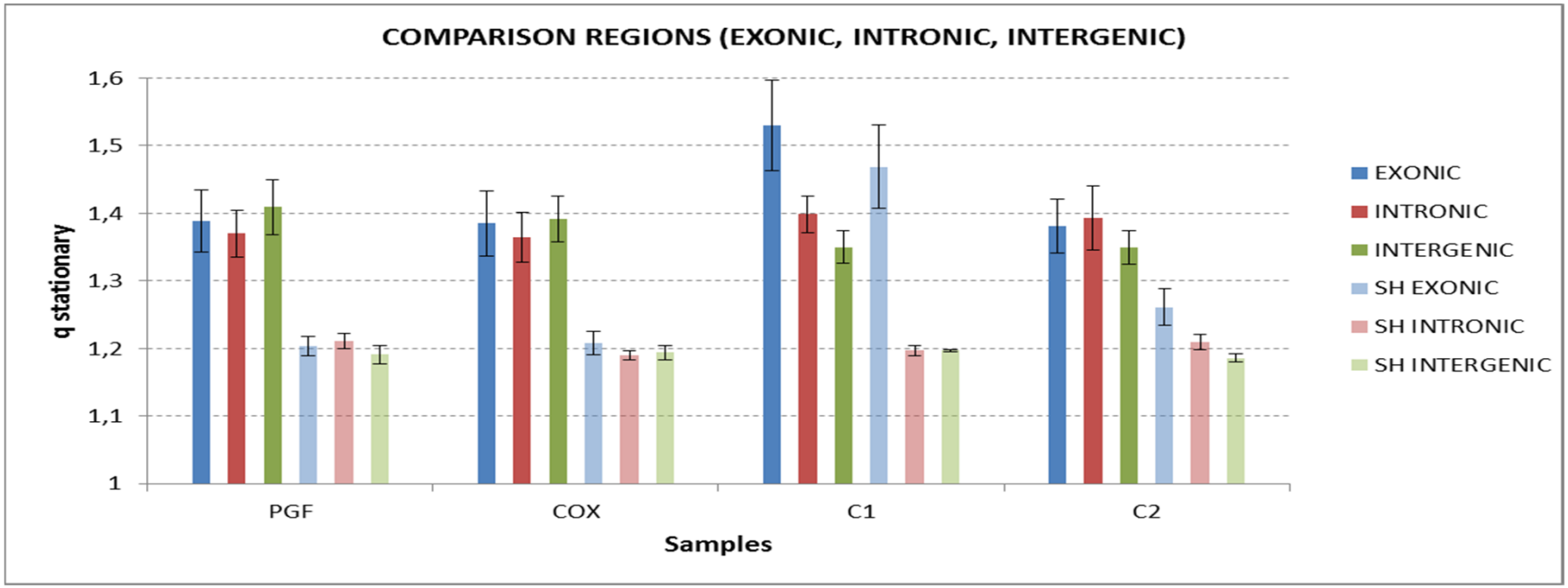
The Hurst exponent is a statistical index used for describing long range correlations in the data series. Its values range between 0 and 1. A value of 0.5 indicates a true random process (e.g. a Brownian series). A Hurst exponent value, 0.5 < H < 1 indicates "persistent behavior" or pattern.



Exonic, intergenic and intronic sequences are characterized by long range correlations and “memory character” or “persistent behavior” or patterns of DNA sequences. Every next step is influenced/determined by previous steps. Shuffled data are clearly distinguishable from the physiological sequences.

Tsallis q triplet estimation – (q stationary)

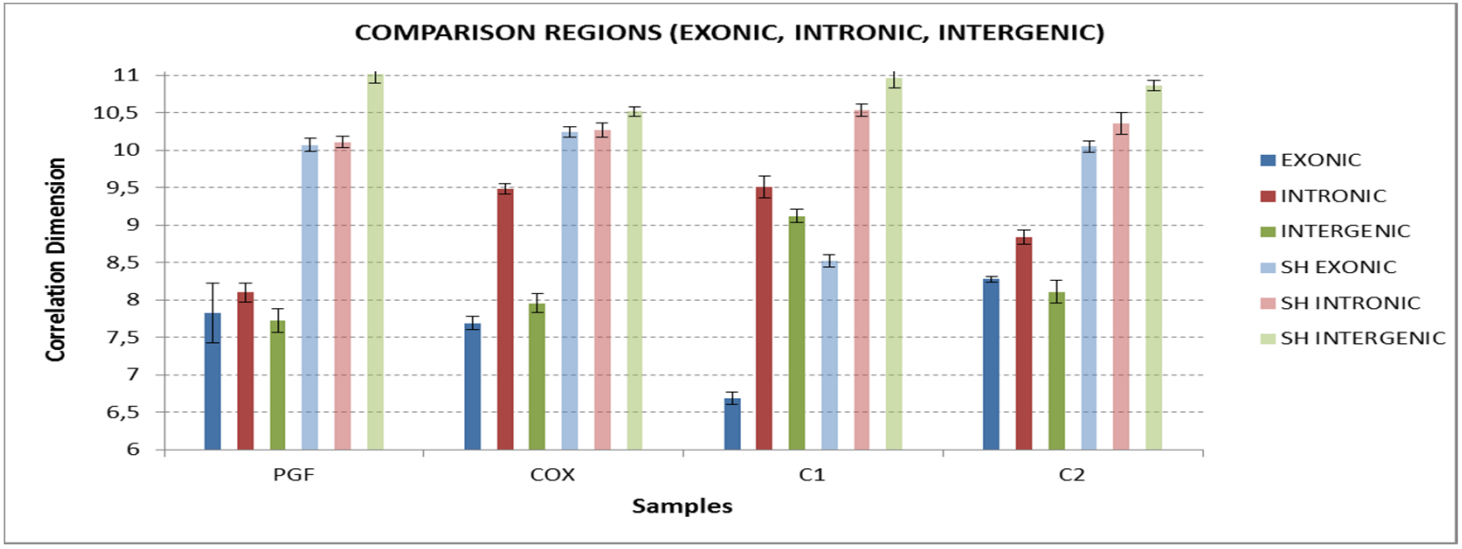
Tsallis q-Gaussians describe far from equilibrium, metastable stationary states. When the dynamic of a system is attracted in a confined subset of the phase space, then long-range correlations can develop



Within each of the exonic, intergenic and intronic sequences long range correlations are observed. q=1 means totally random.

Correlation Dimension

This tool is a measure of the dimensionality of the state space occupied by a set of random points and is directly connected with the efficient degrees of freedom needed to fully describe the system's dynamics. This tool reveals the degree of self-organization of a particular system. Higher values lower degree of self-organization (noise).



Parameters:
- Embedding Dimension (m=12)
- Delay Time (τ=2)
- Theiler Parameter (w=10)

The lower number of genes reveals a lower-dimensional fractal set in the DNA dynamical phase space. On the contrary, the higher number of genes reveals a higher-dimensional fractal set with a more stochastic behavior. In the other two regions intronic and intergenic, this ratio became inversely proportional. More specifically in the intergenic region, the higher number of genes of the exonic region reveals a lower-dimensional fractal set. Furthermore we saw an interaction among the number of genes in exonic region and the intergenic region regarding the profile of a fractal set in the phase space.

Finally the above result indicates the presence of nonlinear and low-dimensional DNA dynamics underlying the DNA structure and it evolves on a low-dimensional fractal set in the DNA dynamical phase space.

The lower the number, the higher the degree of self organization

Conclusions

1. Clear discrimination was observed between the two sets of data (real DNA sequences and shuffled sequences). This observation provides the first evidence that the approach is credible.
2. We observed significant dynamical behavior for the specific regions of the DNA sequences (EXONIC, INTRONIC, INTERGENIC) suggesting complexity behavior. This would imply that the system contains patterns and information. Considering that exonic regions are indeed regions with patterns and information, then the present statistical/computational tools reveal that intronic and intergenic regions are also DNA segments including patterns and information at a comparable level with exonic. It is the ENCODE project today that makes this finding a rather credible assertion.
3. The approach indicates that exonic, intergenic and intronic sequences are characterized by long range correlations (LD concept?). Not of specific loci but rather as an overarching concept.
4. An interactive relationship has been identified among exonic, intronic and intergenic sequences.
5. An important next step is to examine whether these internal degree of organization of intronic and intergenic sequences have an identifiable character.

Conclusions overall

- Using GWAS data we find that the MHC is the only region in the whole genome characterized by such a high density of SNPs associated with traits and diseases. The high throughput technologies can now enable the complete and accurate characterization of the MHC.
- This development will promote the study of other genomic elements within the MHC, like miRNAs, and their role in regulating gene expression. It is not unlikely that other genomic elements currently unknown may be discovered. The miRNAs control a large number of transcripts and the MHC, most likely harbors thousands of miRNAs.
- The merging of distinct disciplines like genetics and complexity theory from Physics can provide insights previously unattainable. This most likely is relevant not only to the MHC but to the rest of the genome as well.
  - MHC sequences, whether exonic, intronic or intergenic contain comparable levels of information.
  - These regions interact with each other.
  - The regions within the MHC are characterized by long range correlations (reminiscent of Linkage Disequilibrium?).

Acknowledgements

Immunogenetics Lab/CHOP- PENN
Deborah Ferriola
Katarzyna Mackiewicz
Hilary Melher
Jamie Duke
Peter Clark
Malek Kamoun-HUP
Brad Johnson-HUP

Generation Biotech
Johannes Dapprich

University Of Thrace-Greece, Dept. of Electrical Engineering and Computer Sciences
George Pavlos
Leonidas Karakatsanis
Aggelos Iliopoulos
Evgenios Pavlos

The Children's Hospital *of* Philadelphia

Immunogenetics Laboratory

Deborah Ferriola

Allison Gasiewski

Jamie Duke
Hilary Mehler

Larissa Slavich

Timothy Mosbruger

Steven Heron

Michael Mignogno
Ryan Morlen

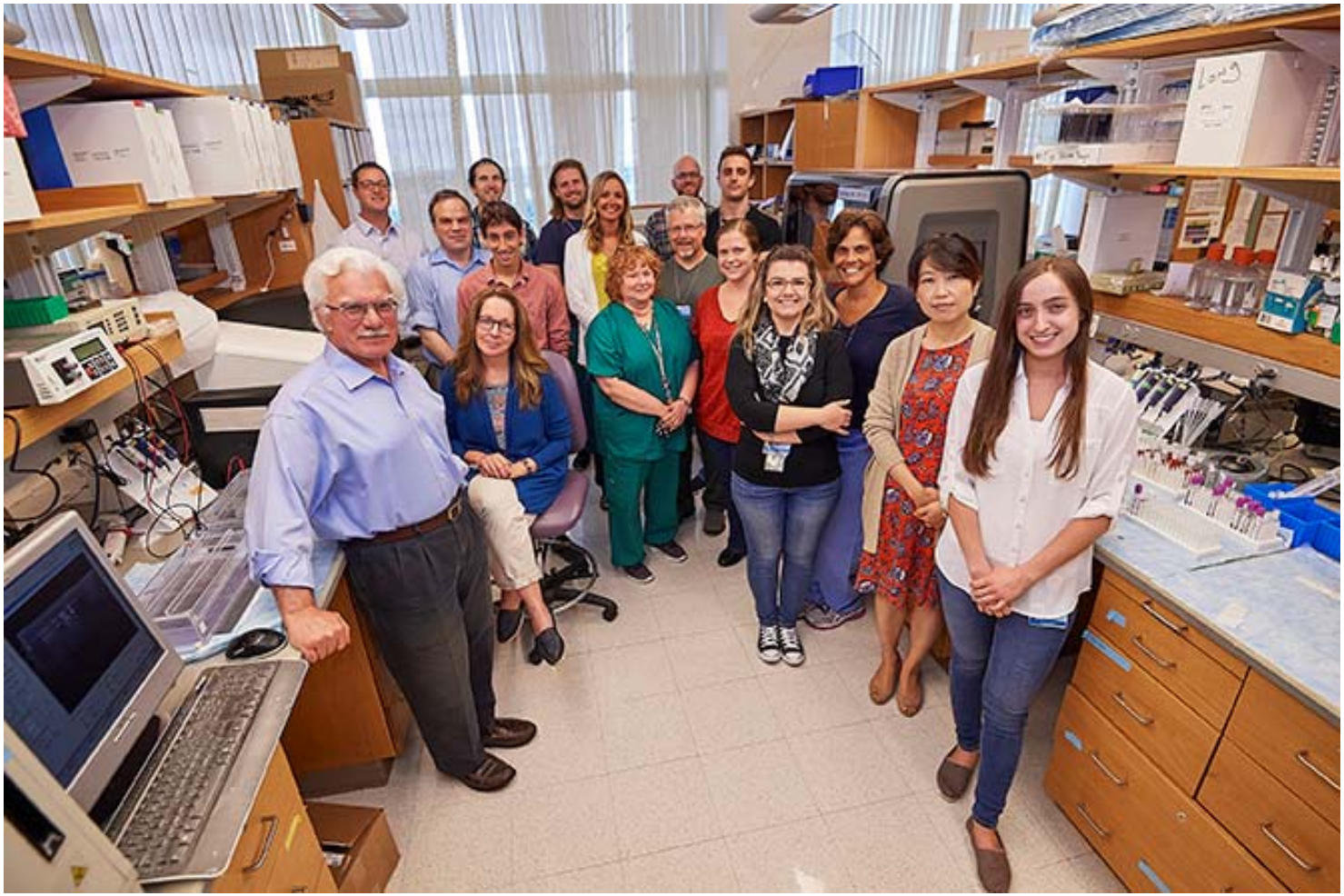
Rita Walker

Ed Frackelton

Yanping Huang

Jenna Wasserman
Ethan Kentzel

Carolina Kneib



Thank you