ENCODE:
"Evolution" of Approaches to Annotate the Human Genome & Interpret its Variants

Mark Gerstein, Yale

Slides freely downloadable from

Lectures.GersteinLab.org

& "tweetable" (via @markgerstein). See last slide for more info.

What is Annotation? (For Written Texts?)

No. 4356 April 25, 1953 NATURE

MOLECULAR STRUCTURE OF NUCLEIC ACIDS

A Structure for Deoxyribose Nucleic Acid

WE wish to suggest a structure for the salt of deoxyribose nucleic acid (D.N.A.). This structure has novel features which are of considerable biological interest.

A structure for nucleic acid has already been proposed by Pauling and Corey[1]. They kindly made their manuscript available to us in advance of publication. Their model consists of three inter-twined chains, with the phosphates near the fibre axis, and the bases on the outside. In our opinion, this structure is unsatisfactory for two reasons: (1) We believe that the material which gives the X-ray diagrams is the salt, not the free acid. Without the acidic hydrogen atoms it is not clear what forces would hold the structure together, especially as the negatively charged phosphates near the axis will repel each other. (2) Some of the van der Waals distances appear to be too small.

Another three-chain structure has also been suggested by Fraser (in the press). In his model the phosphates are on the outside and the bases on the inside, linked together by hydrogen bonds. This structure as described is rather ill-defined, and for this reason we shall not comment on it.

We wish to put forward a radically different structure for the salt of deoxyribose nucleic acid. This structure has two helical chains each coiled round the same axis (see diagram). We have made the usual chemical assumptions, namely, that each chain consists of phosphate di-ester groups joining β-D-deoxy-ribofuranose residues with 3′,5′ linkages. The two chains (but not their bases) are related by a dyad perpendicular to the fibre

NATURE | VOL 409 | 15 FEBRUARY 2001 |

Initial sequencing and analysis of the human genome

International Human Genome Sequencing Consortium*

** A partial list of authors appears on the opposite page. Affiliations are listed at the end of the paper.*

The human genome holds an extraordinary trove of information about human development, physiology, medicine and evolution. Here we report the results of an international collaboration to produce and make freely available a draft sequence of the human genome. We also present an initial analysis of the data, describing some of the insights that can be gleaned from the sequence.

The rediscovery of Mendel's laws of heredity in the opening weeks of the 20th century[1–3] sparked a scientific quest to understand the nature and content of genetic information that has propelled biology for the last hundred years. The scientific progress made falls naturally into four main phases, corresponding roughly to the four quarters of the century. The first established the cellular basis of heredity: the chromosomes. The second defined the molecular basis of heredity: the DNA double helix. The third unlocked the informational basis of heredity, with the discovery of the biological mechanism by which cells read the information contained in genes and with the invention of the recombinant DNA technologies of cloning and sequencing by which scientists can do the same.

The last quarter of a century has been marked by a relentless drive to decipher first genes and then entire genomes, spawning the field of genomics. The fruits of this work already include the genome sequences of 599 viruses and viroids, 205 naturally occurring plasmids, 185 organelles, 31 eubacteria, seven archaea, one fungus, two animals and one plant.

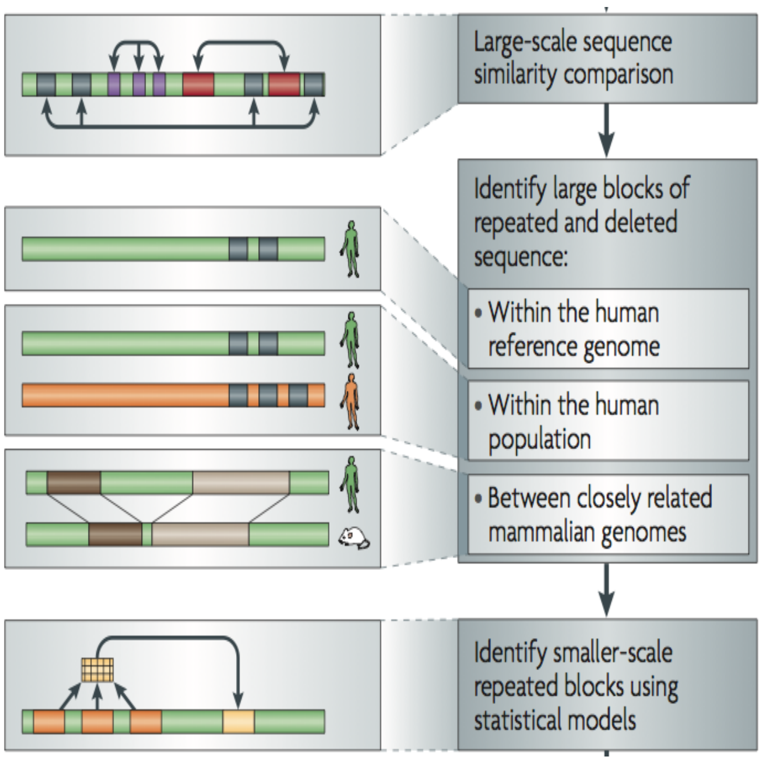
Here we report the results of a collaboration involving 20 groups from the United States, the United Kingdom, Japan, France, Germany and China to produce a draft sequence of the human genome. The draft genome sequence was generated from a physical map covering more than 96% of the euchromatic part of the human genome and, together with additional sequence in public databases, it covers about 94% of the human genome. The sequence was produced over a relatively short period, with coverage rising from about 10% to more than 90% over roughly fifteen months. The sequence data have been made available without restriction and updated daily throughout the project. The task ahead is to produce a finished sequence, by closing all gaps and resolving all ambiguities. Already about one billion bases are in final form and the task of bringing the vast majority of the sequence to this standard is now straightforward and should proceed rapidly.

coordinate regulation of the genes in the clusters.

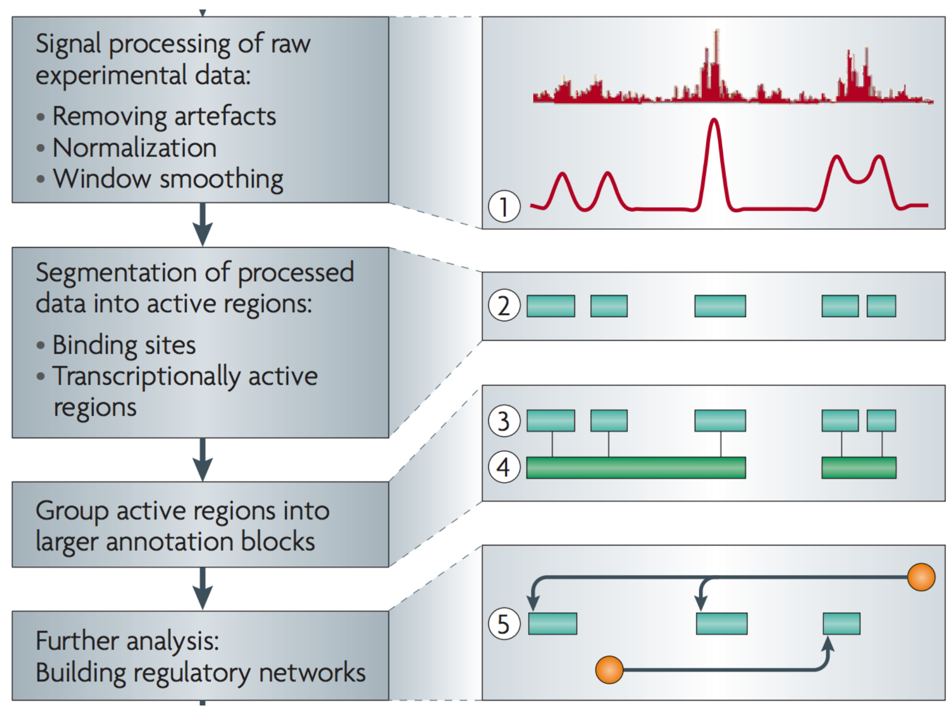
- There appear to be about 30,000–40,000 protein-coding genes in the human genome—only about twice as many as in worm or fly. However, the genes are more complex, with more alternative splicing generating a larger number of protein products.
- The full set of proteins (the 'proteome') encoded by the human genome is more complex than those of invertebrates. This is due in part to the presence of vertebrate-specific protein domains and motifs (an estimated 7% of the total), but more to the fact that vertebrates appear to have arranged pre-existing components into a richer collection of domain architectures.
- Hundreds of human genes appear likely to have resulted from horizontal transfer from bacteria at some point in the vertebrate lineage. Dozens of genes appear to have been derived from transposable elements.
- Although about half of the human genome derives from transposable elements, there has been a marked decline in the overall activity of such elements in the hominid lineage. DNA transposons appear to have become completely inactive and long-terminal repeat (LTR) retroposons may also have done so.
- The pericentromeric and subtelomeric regions of chromosomes are filled with large recent segmental duplications of sequence from elsewhere in the genome. Segmental duplication is much more frequent in humans than in yeast, fly or worm.
- Analysis of the organization of Alu elements explains the long-standing mystery of their surprising genomic distribution, and suggests that there may be strong selection in favour of preferential retention of Alu elements in GC-rich regions and that these 'selfish' elements may benefit their human hosts.
- The mutation rate is about twice as high in male as in female meiosis, showing that most mutation occurs in males.
- Cytogenetic analysis of the sequenced clones confirms suggestions that large GC-poor regions are strongly correlated with 'dark

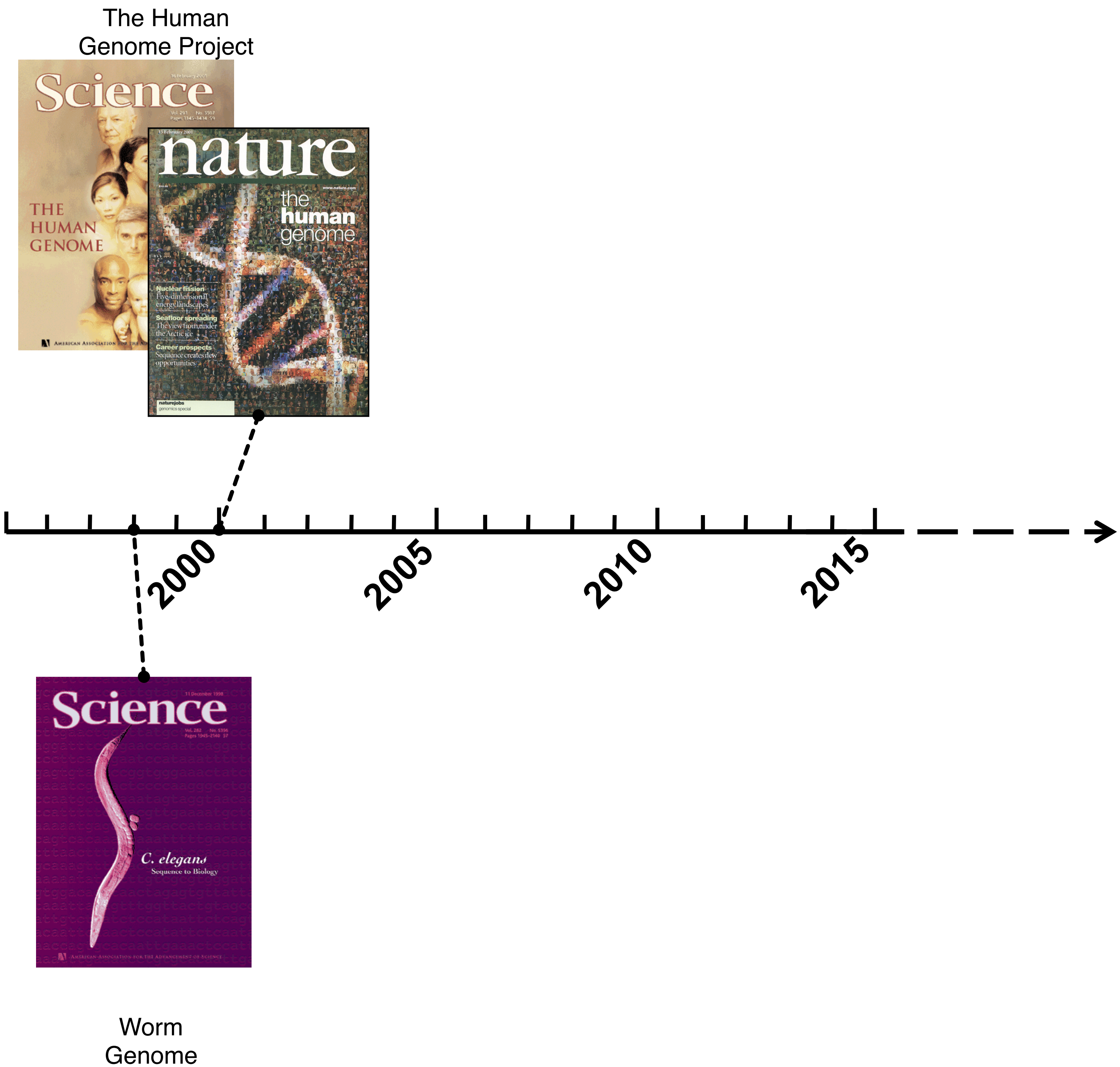
Non-coding Annotations: Overview

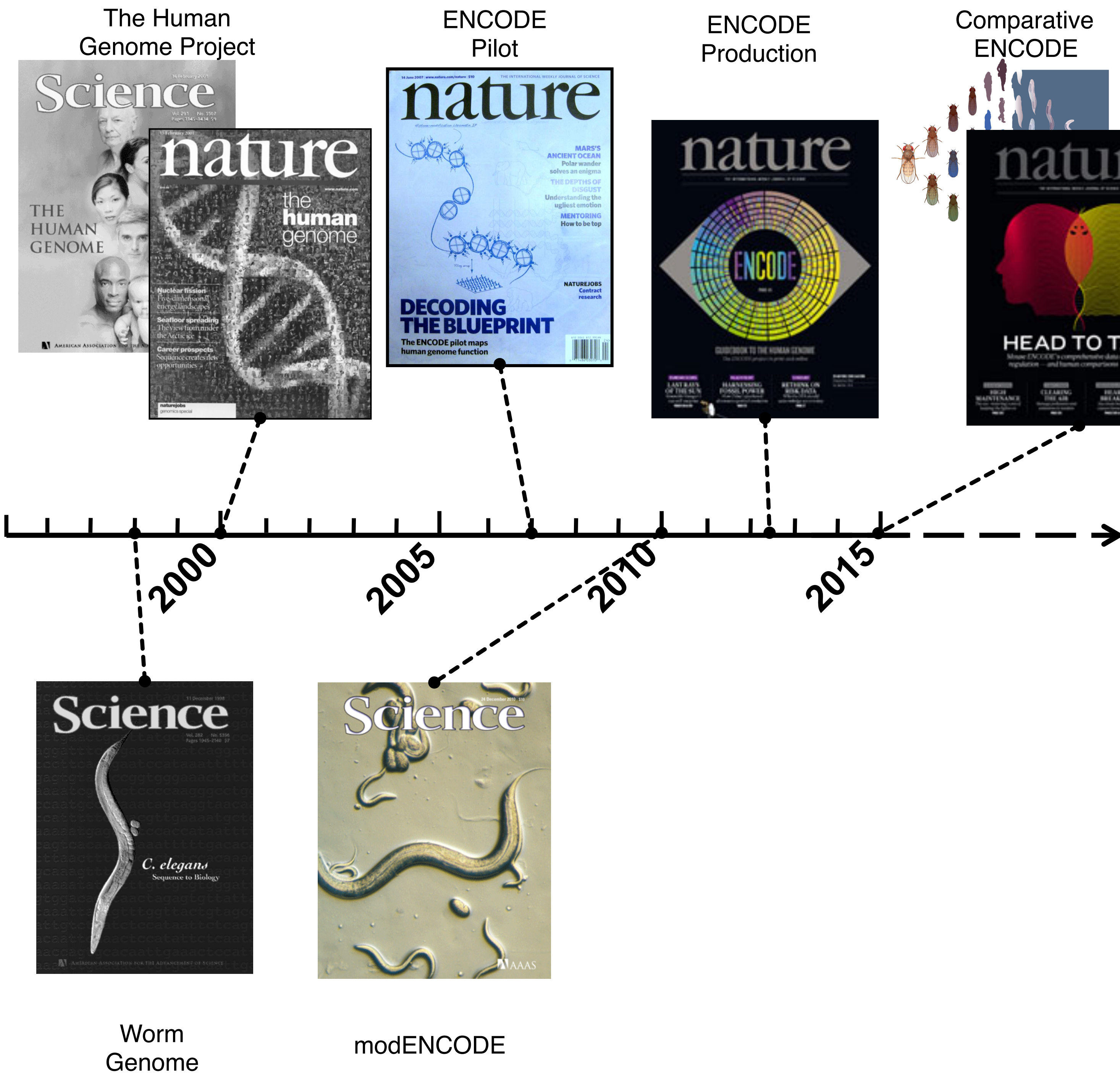
Sequence features, incl. **Conservation**



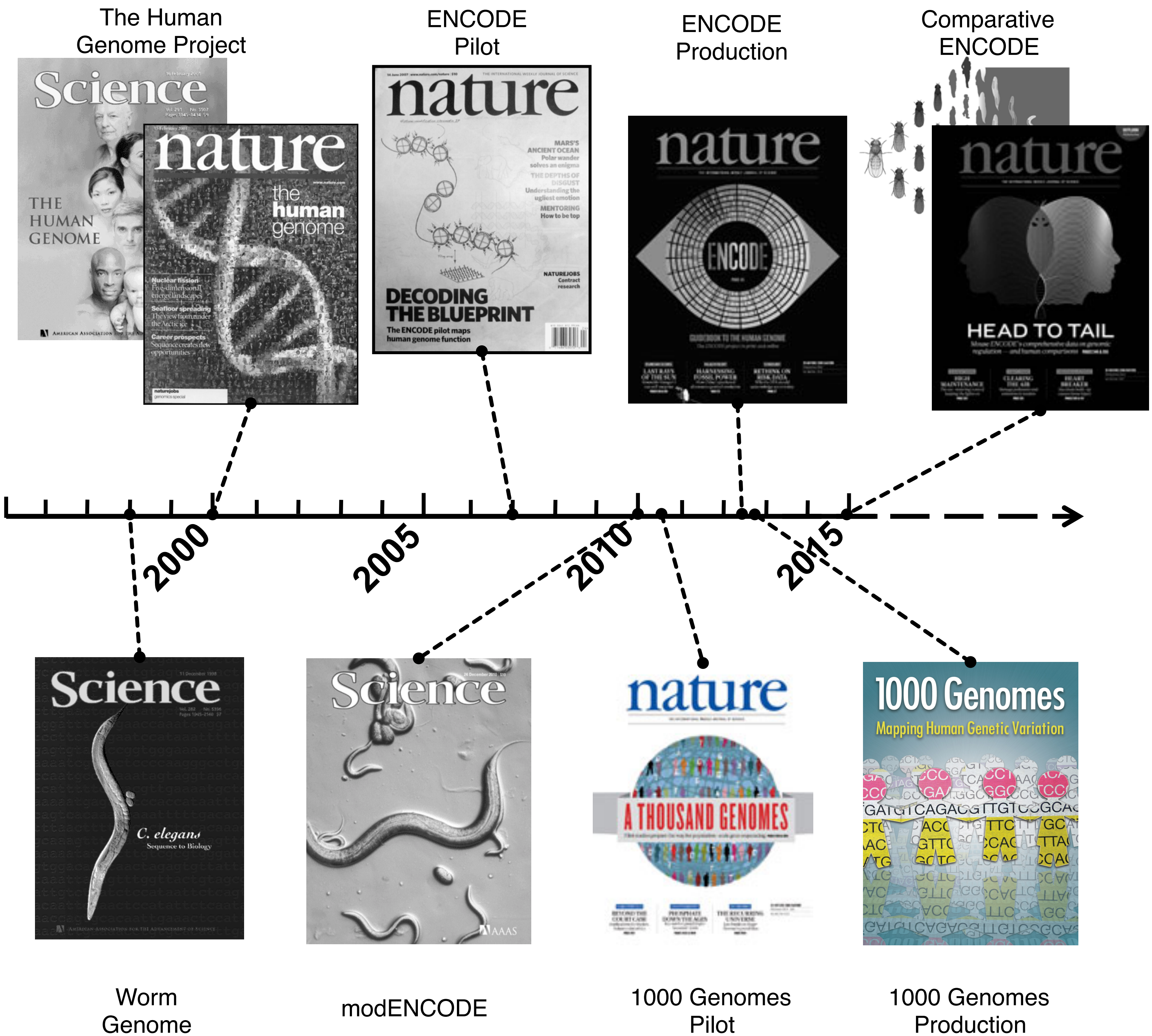
Functional Genomics
ChIP-seq (Epigenome & seq. specific TF) and ncRNA & un-annotated transcription



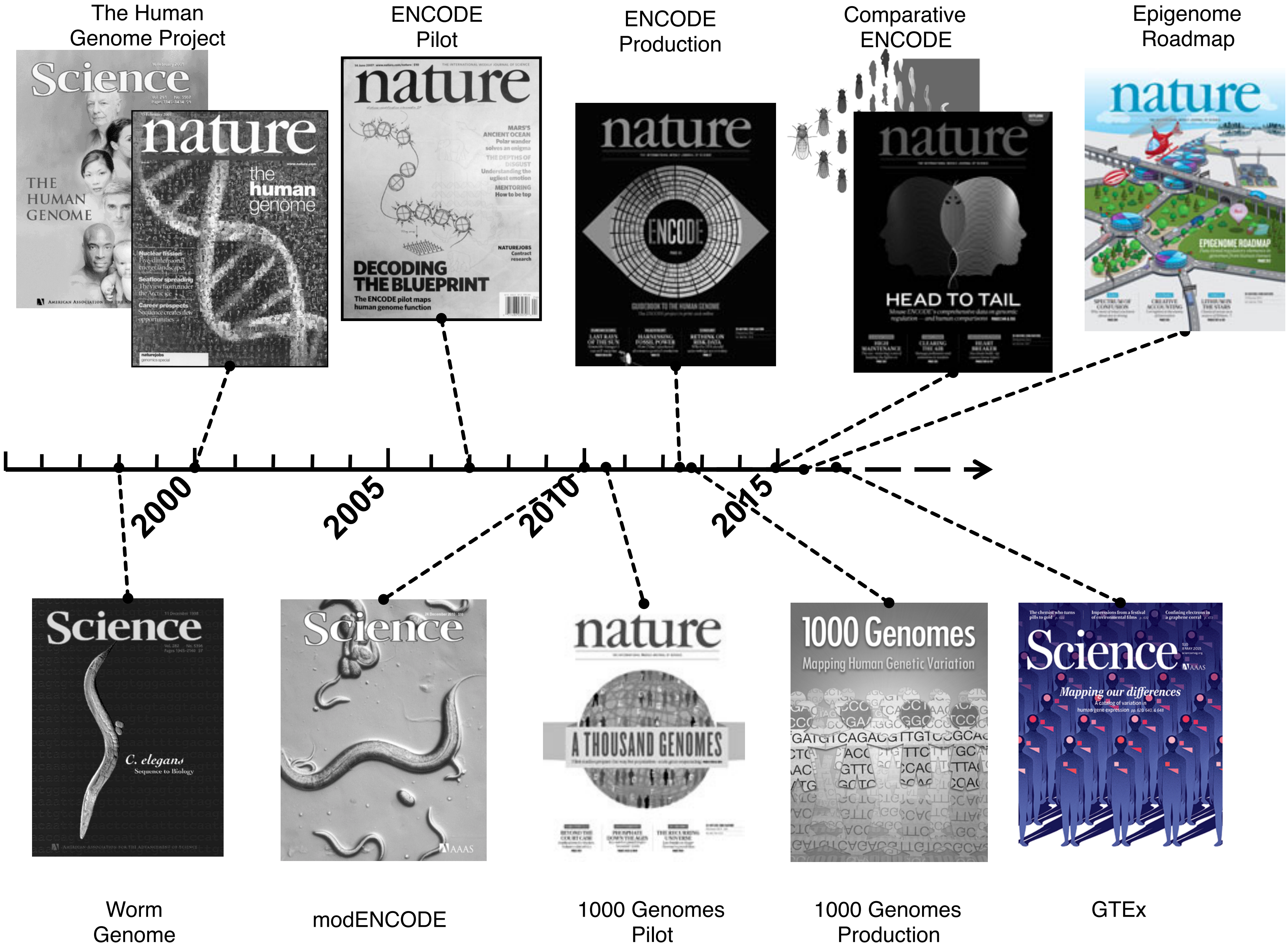

The Human
Genome Project
Science
THE HUMAN
GENOME
nature
the
human
genome
Nuclear fission
Five-dimensional energy landscapes
Seafloor spreading
The view from under the Arctic ice
Career prospects
Sequence creates new opportunities
2000
2005
2010
2015
Science
C. elegans
Sequence to Biology
Worm
Genome



The Human Genome Project
ENCODE Pilot
ENCODE Production
Comparative ENCODE
2000
2005
2010
2015
Worm Genome
modENCODE

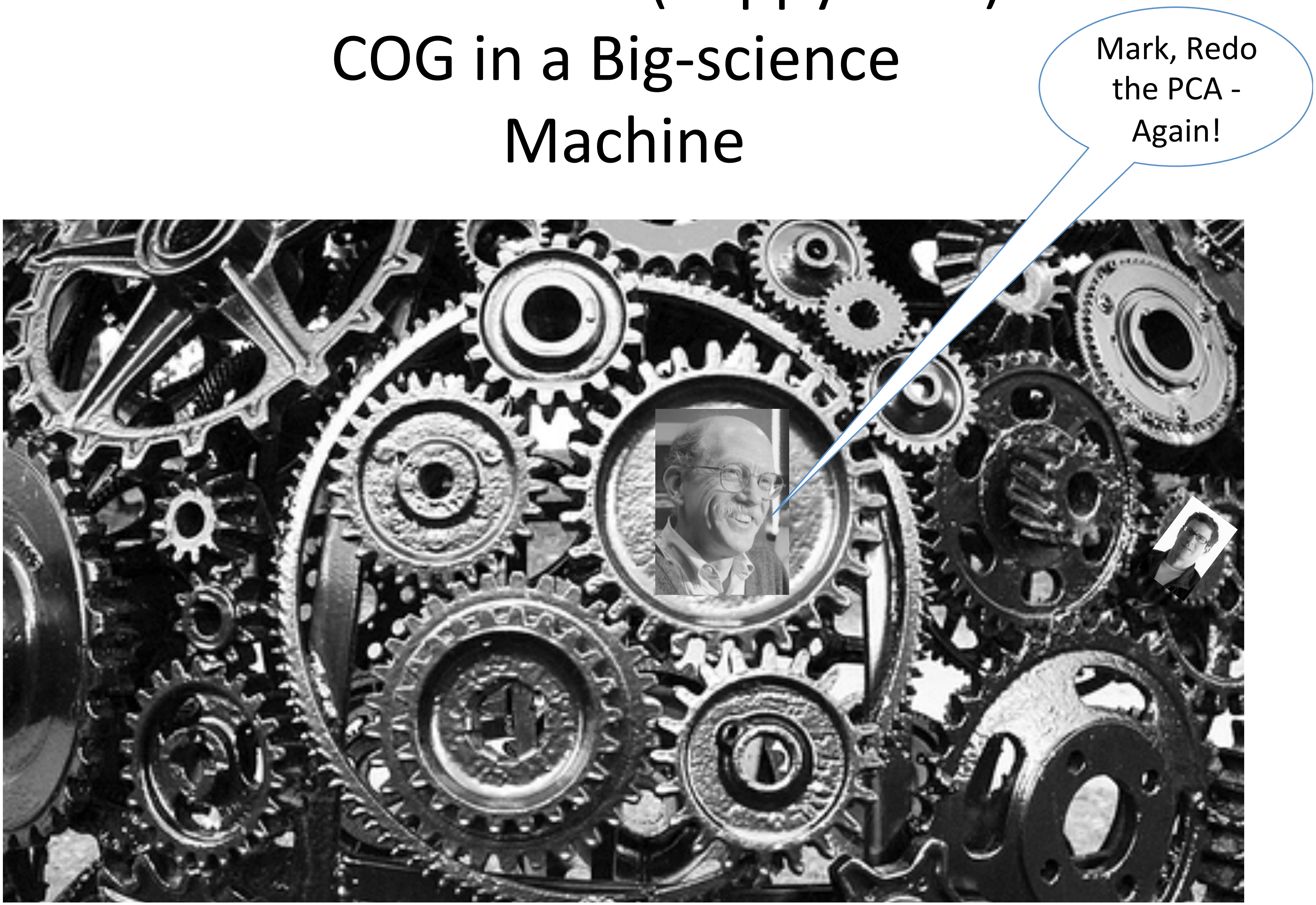


The Human Genome Project
ENCODE Pilot
ENCODE Production
Comparative ENCODE
2000
2005
2010
2015
Worm Genome
modENCODE
1000 Genomes Pilot
1000 Genomes Production



The Human Genome Project
ENCODE Pilot
ENCODE Production
Comparative ENCODE
Epigenome Roadmap
2000
2005
2010
2015
Worm Genome
modENCODE
1000 Genomes Pilot
1000 Genomes Production
GTEx

MBG: ‘03-’15: (happy little) COG in a Big-science Machine



Encode Production People

Dunham I, Kundaje A, Aldred SF, Collins PJ, Davis CA, Doyle F, Epstein CB, Frietze S, Harrow J, Kaul R, Khatun J, Lajoie BR, Landt SG, Lee BK, Pauli F, Rosenbloom KR, Sabo P, Safi A, Sanyal A, Shoresh N, Simon JM, Song L, Trinklein ND, Altshuler RC, Birney E, Brown JB, Cheng C, Djebali S, Dong X, Dunham I, Ernst J, Furey TS, Gerstein M, Giardine B, Greven M, Hardison RC, Harris RS, Herrero J, Hoffman MM, Iyer S, Kellis M, Khatun J, Kheradpour P, Kundaje A, Lassmann T, Li Q, Lin X, Marinov GK, Merkel A, Mortazavi A, Parker SC, Reddy TE, Rozowsky J, Schlesinger F, Thurman RE, Wang J, Ward LD, Whitfield TW, Wilder SP, Wu W, Xi HS, Yip KY, Zhuang J, Pazin MJ, Lowdon RF, Dillon LA, Adams LB, Kelly CJ, Zhang J, Wexler JR, Green ED, Good PJ, Feingold EA, Bernstein BE, Birney E, Crawford GE, Dekker J, Elnitski L, Farnham PJ, Gerstein M, Giddings MC, Gingeras TR, Green ED, Guigó R, Hardison RC, Hubbard TJ, Kellis M, Kent W, Lieb JD, Margulies EH, Myers RM, Snyder M, Stamatoyannopoulos JA, Tenenbaum SA, Weng Z, White KP, Wold B, Khatun J, Yu Y, Wrobel J, Risk BA, Gunawardena HP, Kuiper HC, Maier CW, Xie L, Chen X, Giddings MC, Bernstein BE, Epstein CB, Shoresh N, Ernst J, Kheradpour P, Mikkelsen TS, Gillespie S, Goren A, Ram O, Zhang X, Wang L, Issner R, Coyne MJ, Durham T, Ku M, Truong T, Ward LD, Altshuler RC, Eaton ML, Kellis M, Djebali S, Davis CA, Merkel A, Dobin A, Lassmann T, Mortazavi A, Tanzer A, Lagarde J, Lin W, Schlesinger F, Xue C, Marinov GK, Khatun J, Williams BA, Zaleski C, Rozowsky J, Röder M, Kokocinski F, Abdelhamid RF, Alioto T, Antoshechkin I, Baer MT, Batut P, Bell I, Bell K, Chakrabortty S, Chen X, Chrast J, Curado J, Derrien T, Drenkow J, Dumais E, Dumais J, Duttagupta R, Fastuca M, Fejes-Toth K, Ferreira P, Foissac S, Fullwood MJ, Gao H, Gonzalez D, Gordon A, Gunawardena HP, Howald C, Jha S, Johnson R, Kapranov P, King B, Kingswood C, Li G, Luo OJ, Park E, Preall JB, Presaud K, Ribeca P, Risk BA, Robyr D, Ruan X, Sammeth M, Sandhu KS, Schaeffer L, See LH, Shahab A, Skancke J, Suzuki AM, Takahashi H, Tilgner H, Trout D, Walters N, Wang H, Wrobel J, Yu Y, Hayashizaki Y, Harrow J, Gerstein M, Hubbard TJ, Reymond A, Antonarakis SE, Hannon GJ, Giddings MC, Ruan Y, Wold B, Carninci P, Guigó R, Gingeras TR, Rosenbloom KR, Sloan CA, Learned K, Malladi VS, Wong MC, Barber GP, Cline MS, Dreszer TR, Heitner SG, Karolchik D, Kent W, Kirkup VM, Meyer LR, Long JC, Maddren M, Raney BJ, Furey TS, Song L, Grasfeder LL, Giresi PG, Lee BK, Battenhouse A, Sheffield NC, Simon JM, Showers KA, Safi A, London D, Bhinge AA, Shestak C, Schaner MR, Kim SK, Zhang ZZ, Mieczkowski PA, Mieczkowska JO, Liu Z, McDaniell RM, Ni Y, Rashid NU, Kim MJ, Adar S, Zhang Z, Wang T, Winter D, Keefe D, Birney E, Iyer VR, Lieb JD, Crawford GE, Li G, Sandhu KS, Zheng M, Wang P, Luo OJ, Shahab A, Fullwood MJ, Ruan X, Ruan Y, Myers RM, Pauli F, Williams BA, Gertz J, Marinov GK, Reddy TE, Vielmetter J, Partridge E, Trout D, Varley KE, Gasper C, Bansal A, Pepke S, Jain P, Amrhein H, Bowling KM, Anaya M, Cross MK, King B, Muratet MA, Antoshechkin I, Newberry KM, McCue K, Nesmith AS, Fisher-Aylor KI, Pusey B, DeSalvo G, Parker SL, Balasubramanian S, Davis NS, Meadows SK, Eggleston T, Gunter C, Newberry J, Levy SE, Absher DM, Mortazavi A, Wong WH, Wold B, Blow MJ, Visel A, Pennachio LA, Elnitski L, Margulies EH, Parker SC, Petrykowska HM, Abyzov A, Aken B, Barrell D, Barson G, Berry A, Bignell A, Boychenko V, Bussotti G, Chrast J, Davidson C, Derrien T, Despacio-Reyes G, Diekhans M, Ezkurdia I, Frankish A, Gilbert J, Gonzalez JM, Griffiths E, Harte R, Hendrix DA, Howald C, Hunt T, Jungreis I, Kay M, Khurana E, Kokocinski F, Leng J, Lin MF, Loveland J, Lu Z, Manthravadi D, Mariotti M, Mudge J, Mukherjee G, Notredame C, Pei B, Rodriguez JM, Saunders G, Sboner A, Searle S, Sisu C, Snow C, Steward C, Tanzer A, Tapanari E, Tress ML, van Baren MJ, Walters N, Washietl S, Wilming L, Zadissa A, Zhang Z, Brent M, Haussler D, Kellis M, Valencia A, Gerstein M, Reymond A, Guigó R, Harrow J, Hubbard TJ, Landt SG, Frietze S, Abyzov A, Addleman N, Alexander RP, Auerbach RK, Balasubramanian S, Bettinger K, Bhardwaj N, Boyle AP, Cao AR, Cayting P, Charos A, Cheng Y, Cheng C, Eastman C, Euskirchen G, Fleming JD, Grubert F, Habegger L, Hariharan M, Harmanci A, Iyengar S, Jin VX, Karczewski KJ, Kasowski M, Lacroute P, Lam H, Lamarre-Vincent N, Leng J, Lian J, Lindahl-Allen M, Min R, Miotto B, Monahan H, Moqtaderi Z, Mu XJ, O'Geen H, Ouyang Z, Patacsil D, Pei B, Raha D, Ramirez L, Reed B, Rozowsky J, Sboner A, Shi M, Sisu C, Slifer T, Witt H, Wu L, Xu X, Yan KK, Yang X, Yip KY, Zhang Z, Struhl K, Weissman SM, Gerstein M, Farnham PJ, Snyder M, Tenenbaum SA, Penalva LO, Doyle F, Karmakar S, Landt SG, Bhanvadia RR, Choudhury A, Domanus M, Ma L, Moran J, Patacsil D, Slifer T, Victorsen A, Yang X, Snyder M, Auer T, Centanin L, Eichenlaub M, Gruhl F, Heermann S, Hoeckendorf B, Inoue D, Kellner T, Kirchmaier S, Mueller C, Reinhardt R, Schertel L, Schneider S, Sinn R, Wittbrodt B, Wittbrodt J, Weng Z, Whitfield TW, Wang J, Collins PJ, Aldred SF, Trinklein ND, Partridge EC, Myers RM, Dekker J, Jain G, Lajoie BR, Sanyal A, Balasundaram G, Bates DL, Byron R, Canfield TK, Diegel MJ, Dunn D, Ebersol AK, Frum T, Garg K, Gist E, Hansen R, Boatman L, Haugen E, Humbert R, Jain G, Johnson AK, Johnson EM, Kutyavin TV, Lajoie BR, Lee K, Lotakis D, Maurano MT, Neph SJ, Neri FV, Nguyen ED, Qu H, Reynolds AP, Roach V, Rynes E, Sabo P, Sanchez ME, Sandstrom RS, Sanyal A, Shafer AO, Stergachis AB, Thomas S, Thurman RE, Vernot B, Vierstra J, Vong S, Wang H, Weaver MA, Yan Y, Zhang M, Akey JM, Bender M, Dorschner MO, Groudine M, MacCoss MJ, Navas P, Stamatoyannopoulos G, Kaul R, Dekker J, Stamatoyannopoulos JA, Dunham I, Beal K, Brazma A, Flicek P, Herrero J, Johnson N, Keefe D, Lukk M, Luscombe NM, Sobral D, Vaquerizas JM, Wilder SP, Batzoglou S, Sidow A, Hussami N, Kyriazopoulou-Panagiotopoulou S, Libbrecht MW, Schaub MA, Kundaje A, Hardison RC, Miller W, Giardine B, Harris RS, Wu W, Bickel PJ, Banfai B, Boley NP, Brown JB, Huang H, Li Q, Li JJ, Noble WS, Bilmes JA, Buske OJ, Hoffman MM, Sahu AD, Kharchenko PV, Park PJ, Baker D, Taylor J, Weng Z, Iyer S, Dong X, Greven M, Lin X, Wang J, Xi HS, Zhuang J, Gerstein M, Alexander RP, Balasubramanian S, Cheng C, Harmanci A, Lochovsky L, Min R, Mu XJ, Rozowsky J, Yan KK, Yip KY, Birney E.

Pubmed record

Worm modENCODE

Mark Gerstein, Zhi John Lu[1,2,*], Eric L. Van Nostrand[4,*], Chao Cheng[1,2,*], Bradley I. Arshinoff[5,6,*], Tao Liu[7,8,*], Kevin Y. Yip[1,2,*], Rebecca Robilotto[1,*], Andreas Rechtsteiner[9,*], Kohta Ikegami[10,*], Pedro Alves[1,*], Aurelien Chateigner[11,*], Marc Perry[5,*], Mitzi Morris[12,*], Raymond K. Auerbach[1,*], Xin Feng[5,22,*], Jing Leng[1,*], Anne Vielle[13,*], Wei Niu[14,15,*], Kahn Rhrissorrakrai[12,*], Ashish Agarwal[2,3], Roger P. Alexander[1,2], Galt Barber[16], Cathleen M. Brdlik[4], Jennifer Brennan[10], Jeremy Jean Brouillet[4], Adrian Carr[11], Ming-Sin Cheung[13], Hiram Clawson[16], Sergio Contrino[11], Luke O. Dannenberg[17], Abby F. Dernburg[18], Arshad Desai[19], Lindsay Dick[10], Andréa C. Dosé[18], Jiang Du[3], Thea Egelhofer[9], Sevinc Ercan[10], Ghia Euskirchen[14], Brent Ewing[20], Elise A. Feingold[21], Reto Gassman[19], Peter J. Good[21], Phil Green[20], Francois Gullier[11], Michelle Gutwein[12], Mark S. Guyer[21], Lukas Habegger[1], Ting Han[23], Jorja G. Henikoff[24], Stefan R. Henz[29], Angie Hinrichs[16], Heather Holster[17], Tony Hyman[26], A. Leo Iniguez[17], Judith Janette[15], Morten Jensen[10], Masaomi Kato[28], W. James Kent[16], Ellen Kephart[5], Vishal Khivansara[23], Ekta Khurana[1,2], John K. Kim[23], Paulina Kolasinska-Zwierz[13], Eric C. Lai[30], Isabel Latorre[13], Amber Leahey[20], Suzanna Lewis[31], Paul Lloyd[5], Lucas Lochovsky[1], Rebecca F. Lowdon[21], Yaniv Lubling[32], Rachel Lyne[11], Michael MacCoss[20], Sebastian D. Mackowiak[33], Marco Mangone[12], Sheldon McKay[34], Desirea Mecenas[12], Gennifer Merrihew[20], David M. Miller III[27], Andrew Muroyama[19], John I. Murray[20], Siew-Loon Ooi[24], Hoang Pham[18], Taryn Phippen[9], Elicia A. Preston[20], Nikolaus Rajewsky[33], Gunnar Rätsch[25], Heidi Rosenbaum[17], Joel Rozowsky[1,2], Kim Rutherford[11], Peter Ruzanov[5], Mihail Sarov[26], Rajkumar Sasidharan[2], Andrea Sboner[1,2], Paul Scheid[12], Eran Segal[32], Hyunjin Shin[7,8], Chong Shou[1], Frank J. Slack[28], Cindie Slightam[35], Richard Smith[11], William C. Spencer[27], E.O. Stinson[31], Scott Taing[7], Teruaki Takasaki[9], Dionne Vafeados[20], Ksenia Voronina[19], Guilin Wang[15], Nicole L. Washington[31], Christina Whittle[10], Beijing Wu[35], Koon-Kiu Yan[1,2], Georg Zeller[25,36], Zheng Zha[5], Mei Zhong[14], Xingliang Zhou[10], modENCODE Consortium, Julie Ahringer[13,#], Susan Strome[9,#], Kristin C. Gunsalus[12,37,#], Gos Micklem[11,#], X. Shirley Liu[7,8,#], Valerie Reinke[15,#], Stuart K. Kim[35,4,#], LaDeana W. Hillier[20,#], Steven Henikoff[24,#], Fabio Piano[12,37,#],

Michael Snyder[4,14,#], **Lincoln Stein[34,5,6,#], Jason D. Lieb[10,#], Robert H. Waterston**[20,#]


ENCODE and modENCODE Consortia
June 16-19, 2008
Rockville, Maryland, US

ENCODE

(11 Main Projects, ~50 labs, >700 substantial contributors + NHGRI)

Network Acknowledgements

Networks/Elements (~60 participants):

A Kundaje, M Hariharan, S Landt, K Yan, C Cheng, X Mu, E Khurana, J Rozowsky, R Alexander, R Min, P Alves, A Abyzov, N Addleman, N Bhardwaj, A Boyle, P Cayting, A Charos, D Chen, Y Cheng, **D Clarke**, C Eastman, G Euskirchen, S Frietze, **Y Fu**, J Gertz, F Grubert, **A Harmanci,** P Jain, M Kasowski, P Lacroute, **J Leng,** J Lian, H Monahan, H O'Geen, Z Ouyang, E Partridge, D Patacsil, F Pauli, D Raha, L Ramirez, T Reddy, B Reed, M Shi, T Slifer, J Wang, L Wu, X Yang, **K Yip,** G Zilberman-Schapira, S Batzoglou, A Sidow, P Farnham, **R Myers,** S Weissman, **M Snyder**

Also:
E Birney, B Wold, J Stam, T Gingeras, B Bernstein, R Hardison, Z Weng, P Bickel, B Noble, M Kellis, R Guigo, T Hubbard, J Harrow, E Feingold, B Graveley, M Pazin, J Kent, M Cherry

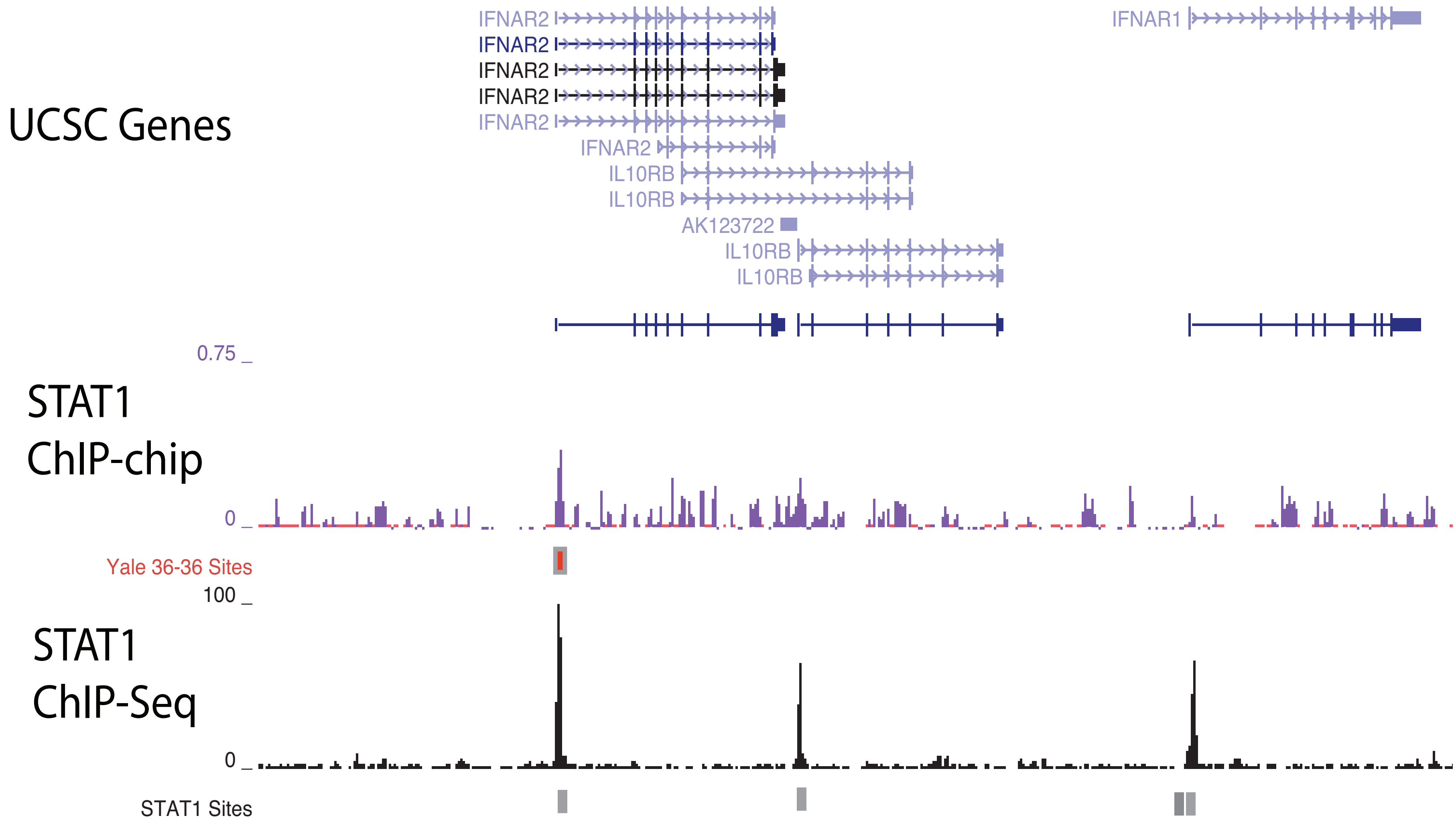
- **Introduction**
  - Being a happy cog in a 500+ person Big-science project
- **Evolution of Linear Annotation based on Functional Genomics**
  - Chip-Chip, Chip-Seq, Thresholding v Control, Segmentation, Multi-scale site calling
- **Its Relation to Conservation: An Enduring Puzzle from Pilot to Production**
  - Many unconstrained regulatory sites
  - But finding small number of sites particularly sensitive to mutations
- **Development of a 2nd Level Network Annotation**
  - Creating it from the linear annotation & connecting it to network science & hubs
  - More connectivity, more constraint
- **New Direction: Applying the Annotation to Prioritize Mutations**
  - Tools (eg FunSeq) for systematically weighting non-coding features
- **Postscript**
  - Culture Clash: Open Data in Genomics v Patient Privacy
  - Genomics Legacy: the discipline as a exemplar for Data Science

ENCODE: "Evolution" of Approaches to Annotate the Human Genome & Interpret its Variants

- **Introduction**
  - Being a happy cog in a 500+ person Big-science project
- **Evolution of Linear Annotation based on Functional Genomics**
  - Chip-Chip, Chip-Seq, Thresholding v Control, Segmentation, Multi-scale site calling
- **Its Relation to Conservation: An Enduring Puzzle from Pilot to Production**
  - Many unconstrained regulatory sites
  - But finding small number of sites particularly sensitive to mutations
- **Development of a 2nd Level Network Annotation**
  - Creating it from the linear annotation & connecting it to network science & hubs
  - More connectivity, more constraint
- **New Direction: Applying the Annotation to Prioritize Mutations**
  - Tools (eg FunSeq) for systematically weighting non-coding features
- **Postscript**
  - Culture Clash: Open Data in Genomics v Patient Privacy
  - Genomics Legacy: the discipline as a exemplar for Data Science

ENCODE: "Evolution" of Approaches to Annotate the Human Genome & Interpret its Variants

ChIP-seq vs ChIP-chip: Much cleaner signal from sequencing than arrays



[Rozowsky et al. Nat. Biotech ('09)]

Summarizing the Signal: "Traditional" ChipSeq Peak Calling

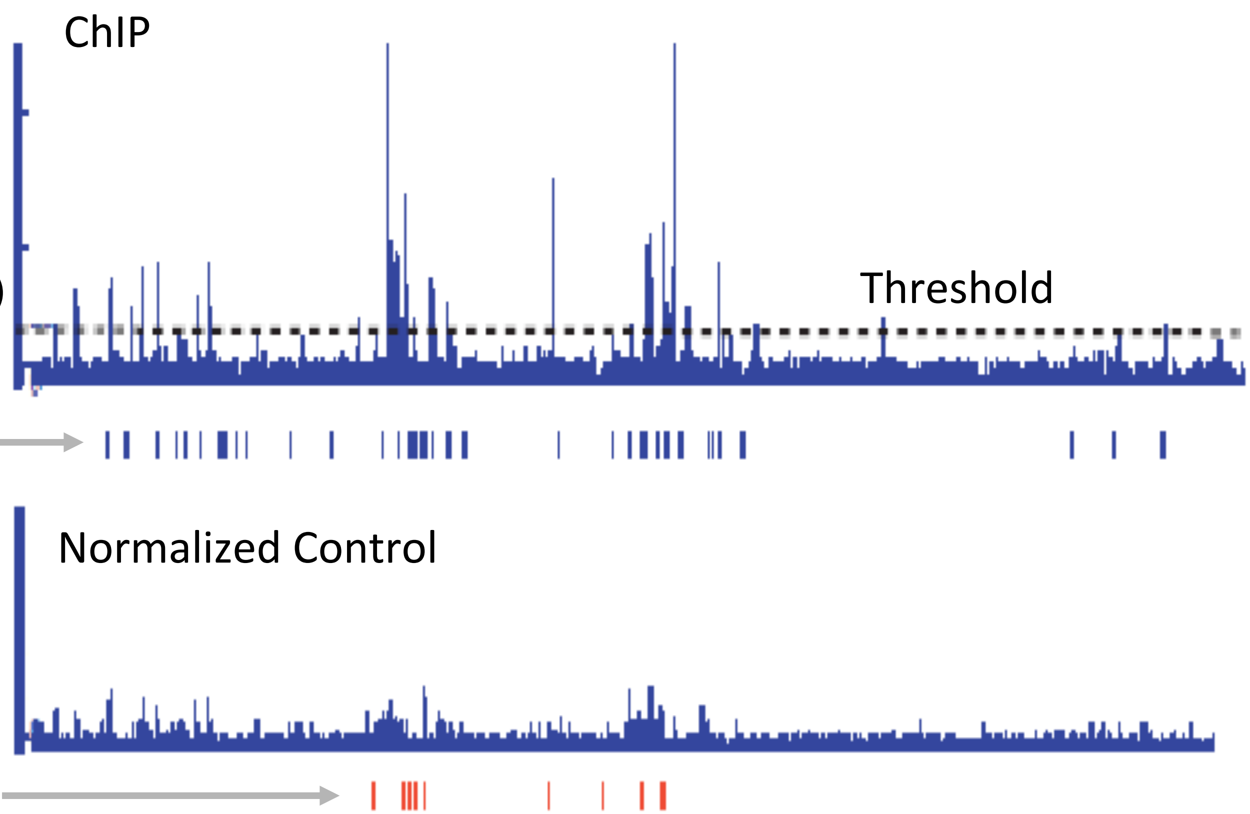
Generate & **threshold** the signal profile to identify candidate target regions

- Simulation (PeakSeq)
- Local window based Poisson (MACS)
- Fold change statistics (SPP)

Potential Targets

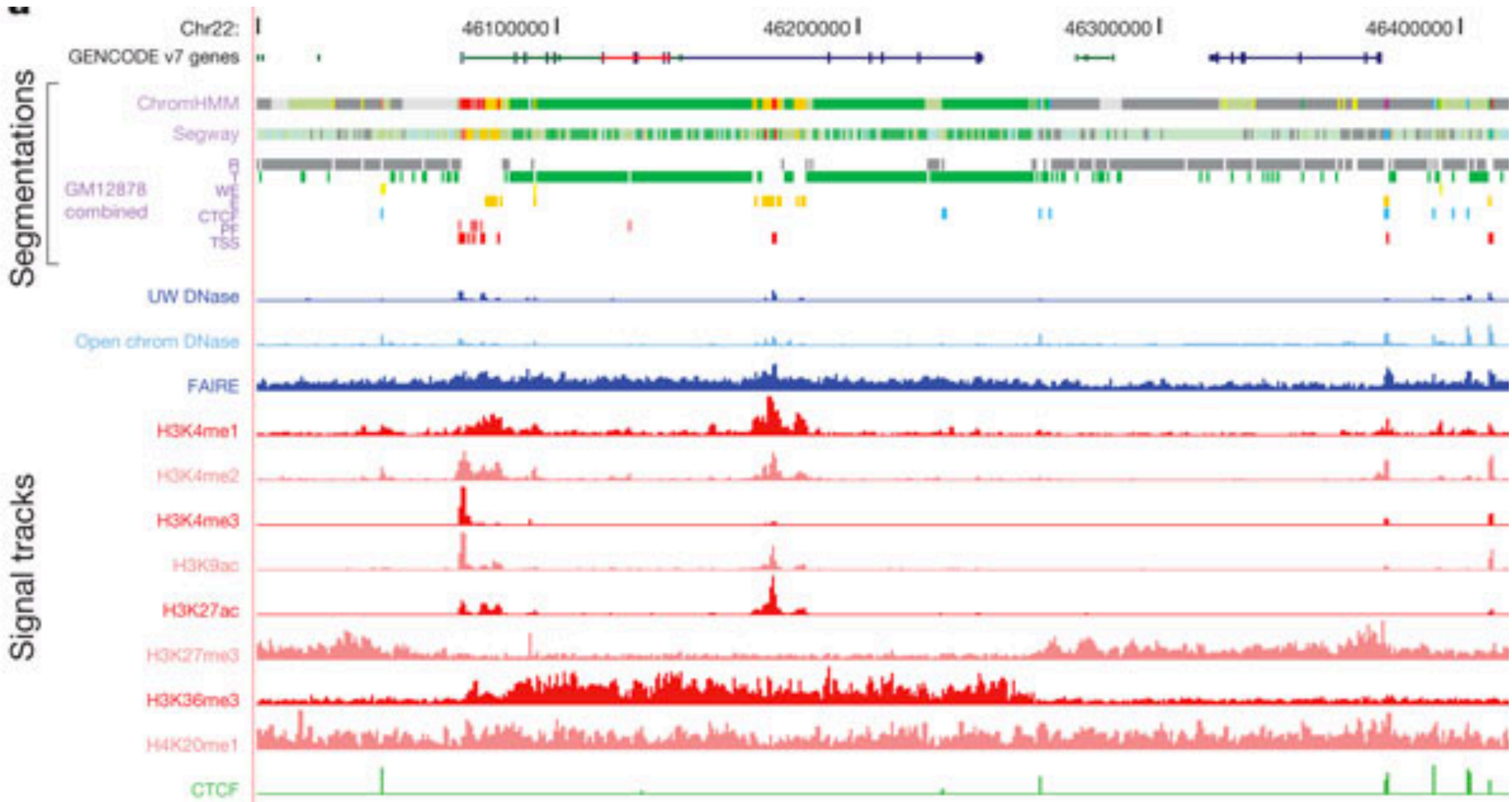
Score against the **control**

Significantly Enriched targets



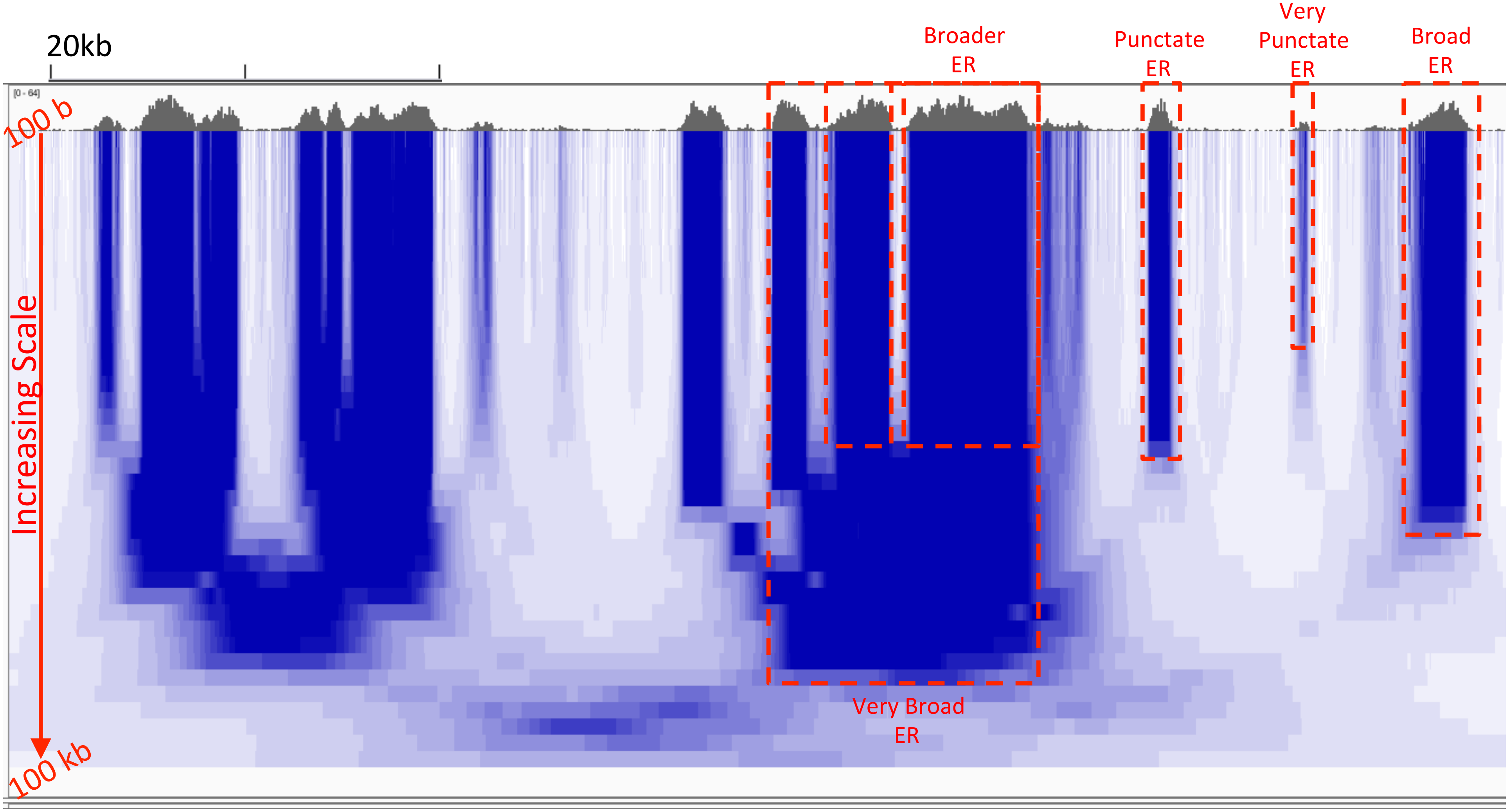
[Rozowsky et al. ('09) *Nat Biotech*]

Multi-track analysis: Segmentation



[Encode Consortium ('12), Nature; Ernst & Kellis, Hoffman & Noble]

Multiscale Decomposition

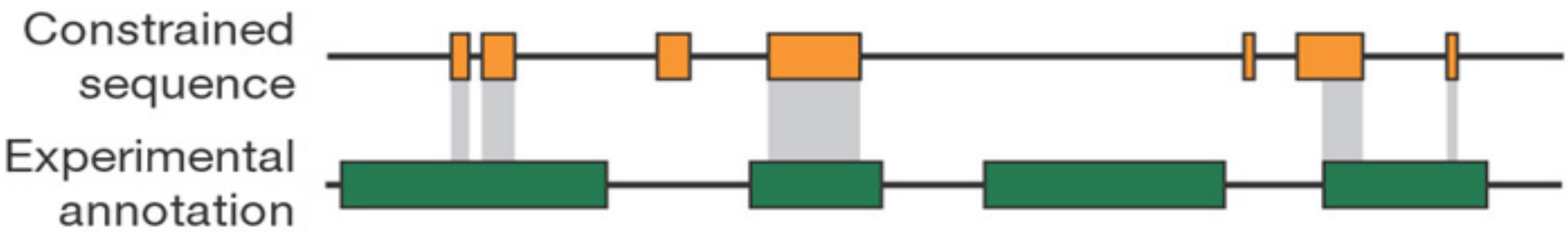


[Harmanci *et al, Genome Biol.* ('14)]

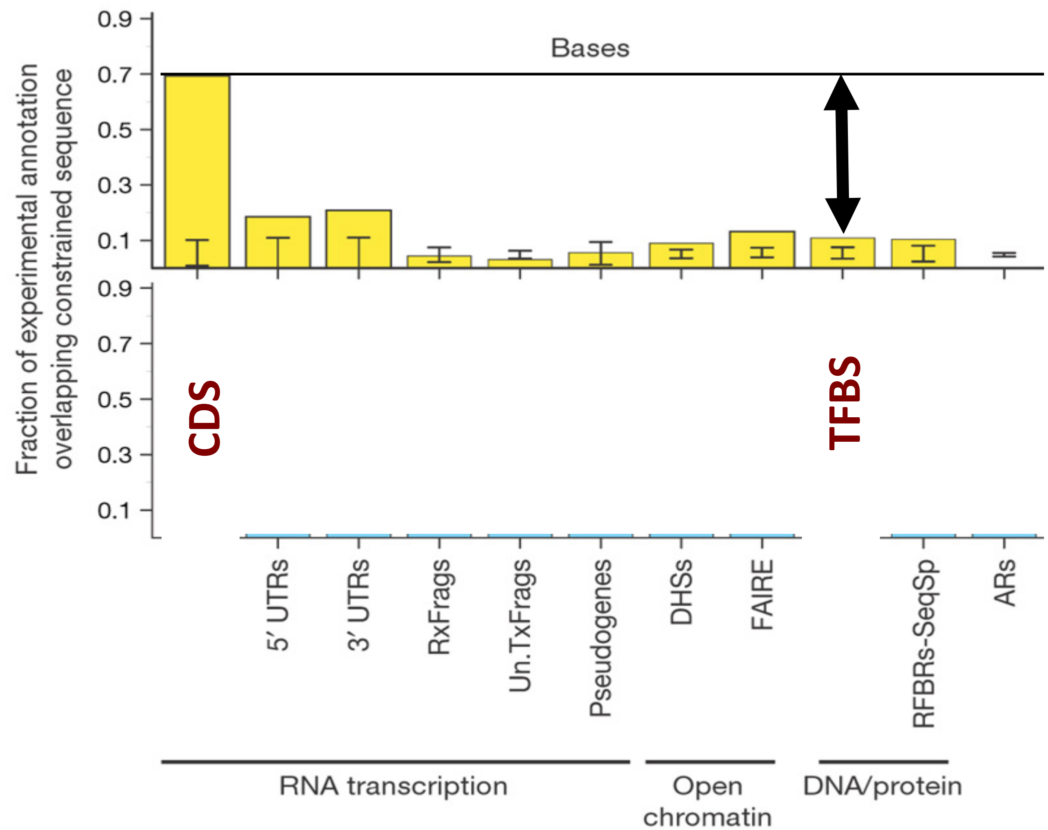
- **Introduction**
  - Being a happy cog in a 500+ person Big-science project
- **Evolution of Linear Annotation based on Functional Genomics**
  - Chip-Chip, Chip-Seq, Thresholding v Control, Segmentation, Multi-scale site calling
- **Its Relation to Conservation: An Enduring Puzzle from Pilot to Production**
  - Many unconstrained regulatory sites
  - But finding small number of sites particularly sensitive to mutations
- **Development of a 2nd Level Network Annotation**
  - Creating it from the linear annotation & connecting it to network science & hubs
  - More connectivity, more constraint
- **New Direction: Applying the Annotation to Prioritize Mutations**
  - Tools (eg FunSeq) for systematically weighting non-coding features
- **Postscript**
  - Culture Clash: Open Data in Genomics v Patient Privacy
  - Genomics Legacy: the discipline as a exemplar for Data Science

ENCODE: "Evolution" of Approaches to Annotate the Human Genome & Interpret its Variants

A Puzzle from the Pilot: Why so much biochemical activity w/o Sequence Constraints

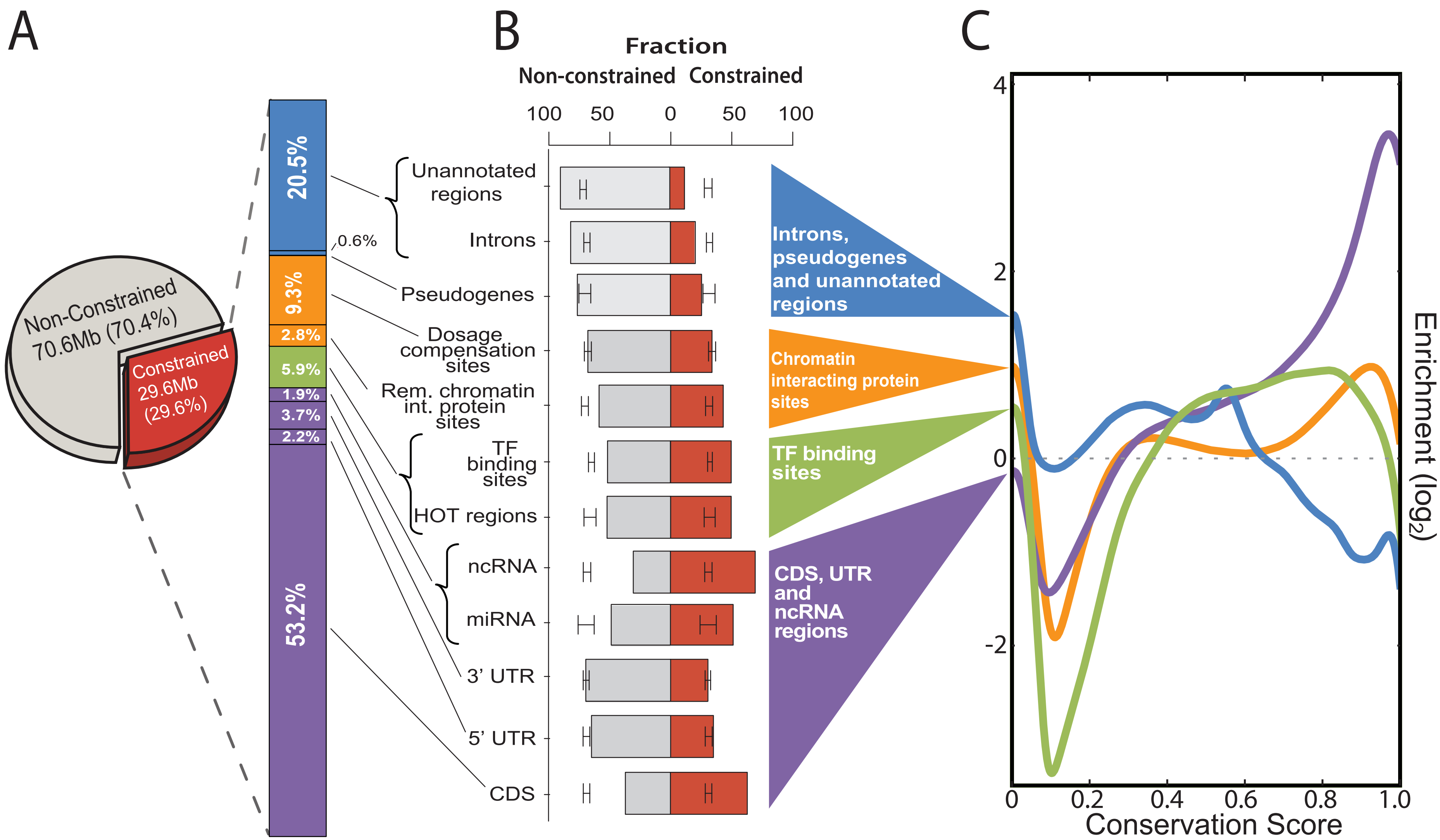


"At the outset of the ENCODE Project, many believed that the broad collection of experimental data would nicely dovetail with the detailed evolutionary information derived from comparing multiple mammalian sequences to provide a neat 'dictionary' of conserved genomic elements, each with a growing annotation about their biochemical function(s). In one sense, this was achieved; the majority of constrained bases in the ENCODE regions are now associated with at least some experimentally-derived information about function. However, we have also encountered a remarkable excess of unconstrained experimentally-identified functional elements, and these cannot be dismissed for technical reasons. This is perhaps the biggest surprise of the pilot phase of the ENCODE Project, and suggests that we take a more 'neutral' view of many of the functions conferred by the genome. "



[ENCODE Consortium, *Nature* 447, 2007]

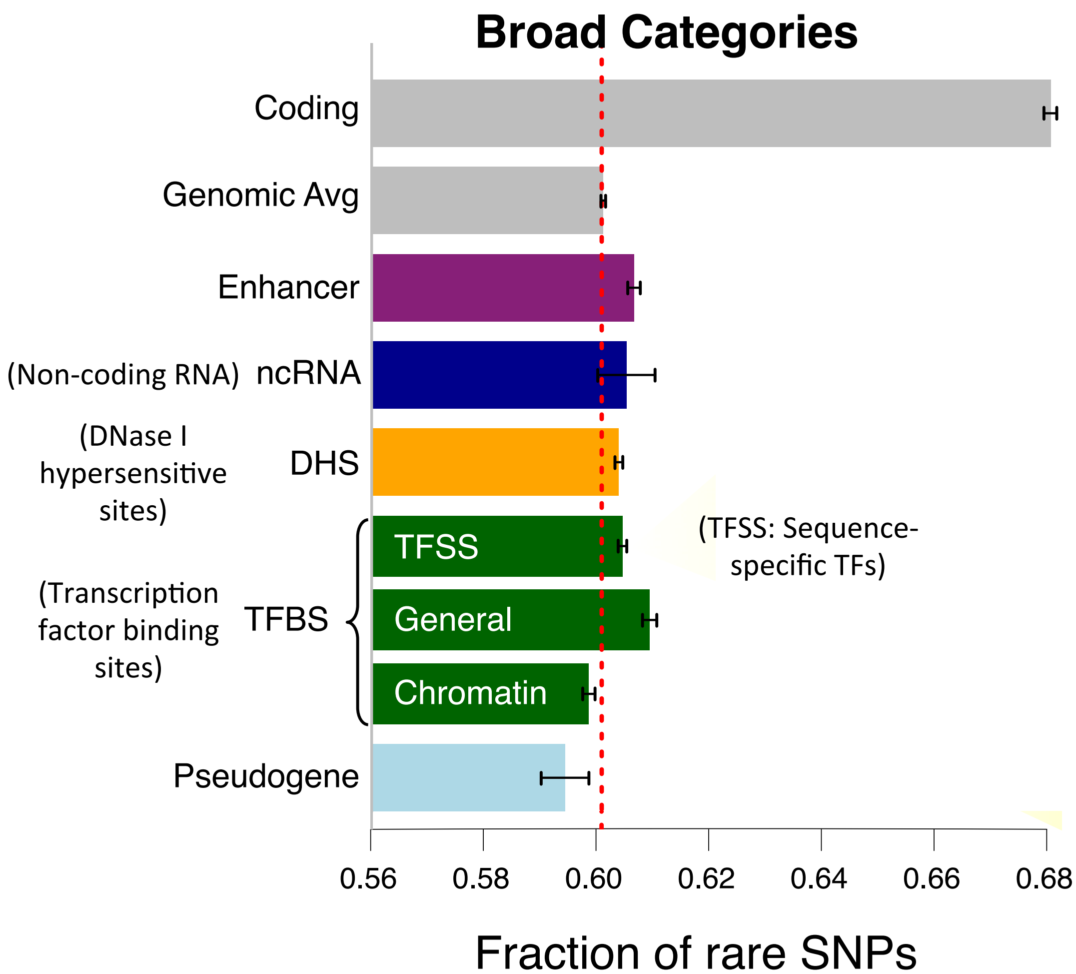
Many Regulatory Sites still unconstrained in Model Organism Analysis (Worm)



[*Science* 330:6012]

Finding "Conserved" Sites in the Human Population:

Negative selection in non-coding elements based on Production ENCODE & 1000G Phase 1



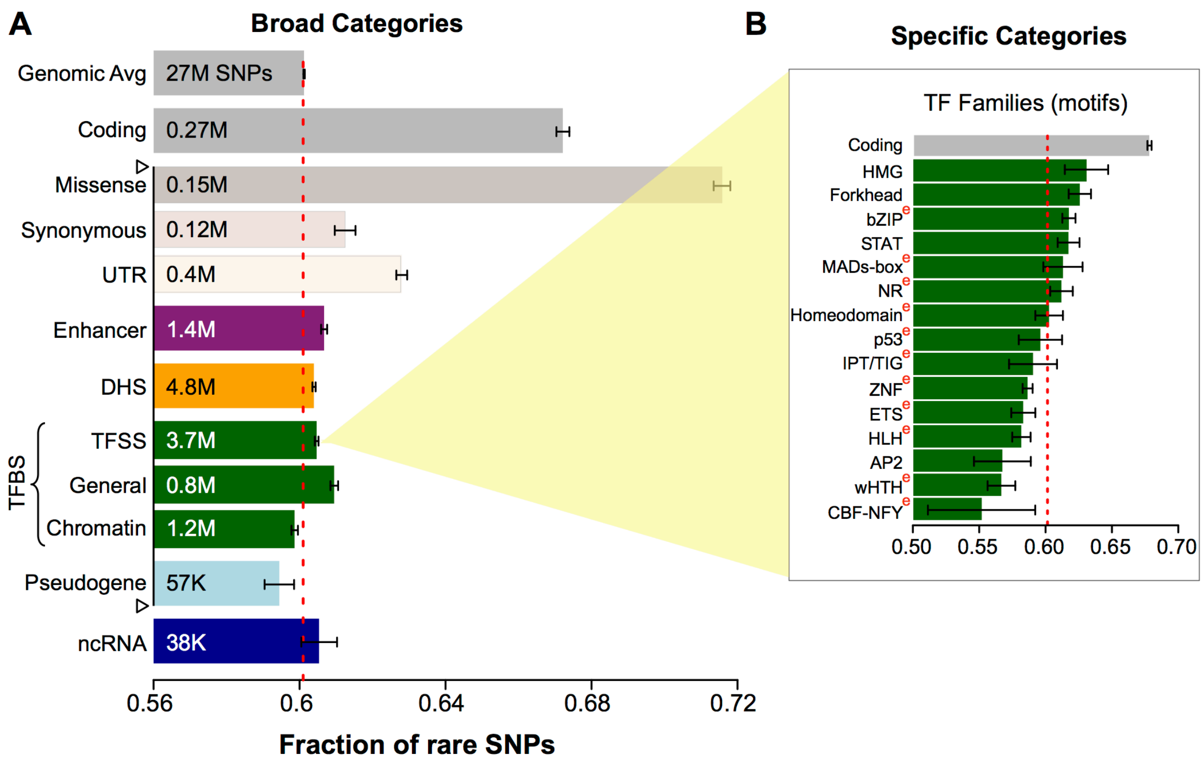
Depletion of Common Variants in the Human Population

- Broad categories of regulatory regions under negative selection
- Related to:

ENCODE, *Nature*, 2012
Ward & Kellis, *Science*, 2012
Mu et al, *NAR*, 2011

[Khurana et al., *Science* ('13)]

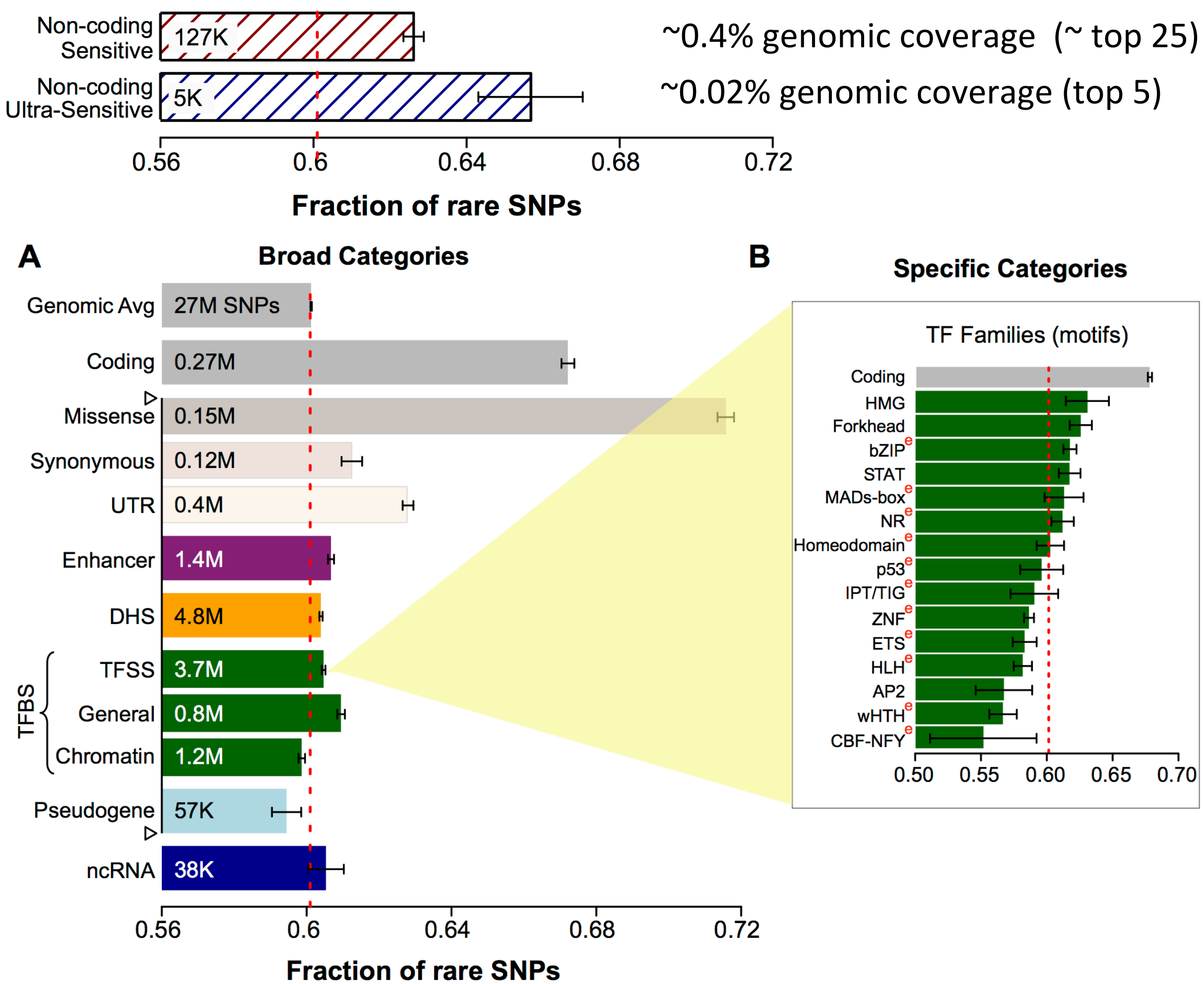
Differential selective constraints among specific sub-categories



Sub-categorization possible because of better statistics from 1000G phase 1 v pilot

[Khurana et al., *Science* ('13)]

Defining Sensitive non-coding Regions



Start **677** high-resolution non-coding categories; Rank & find those under strongest selection

Sub-categorization possible because of better statistics from 1000G phase 1 v pilot

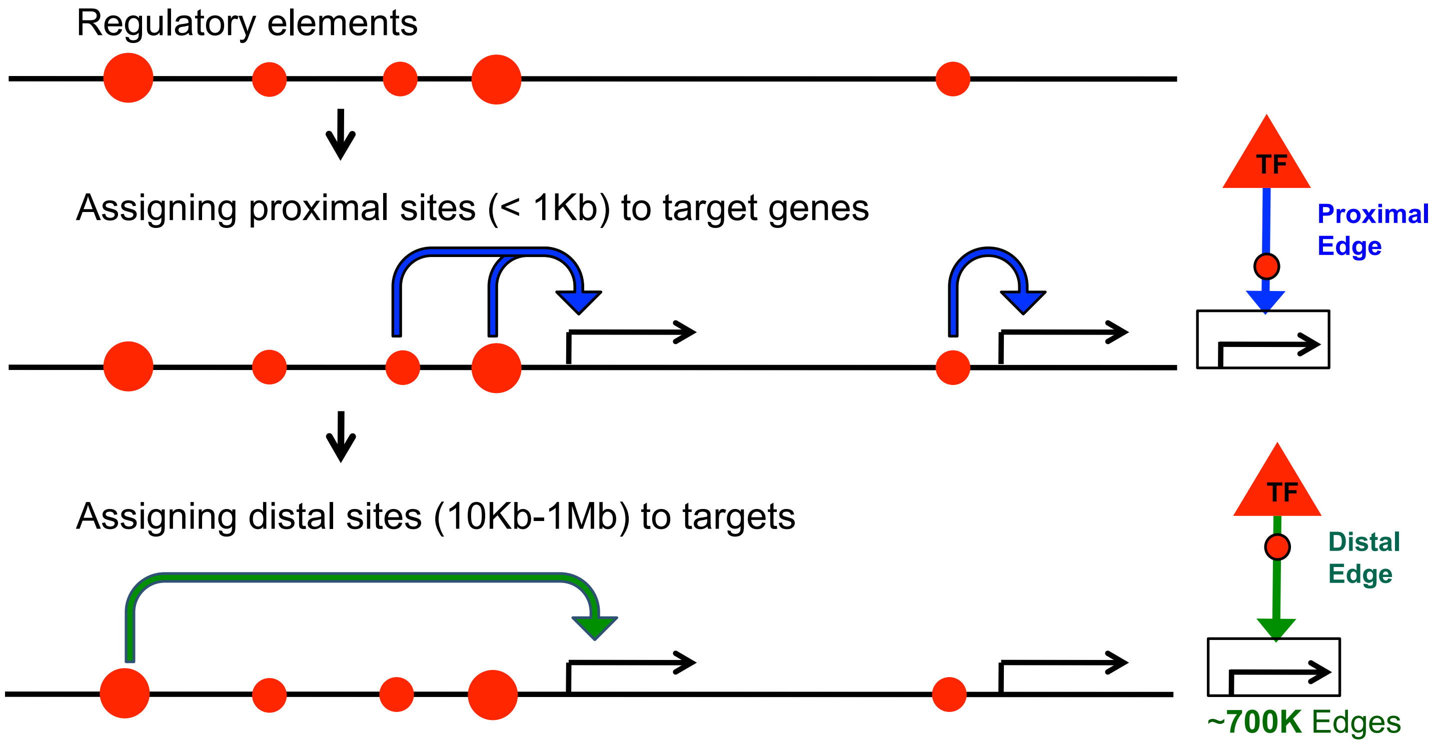
[Khurana et al., *Science* ('13)]

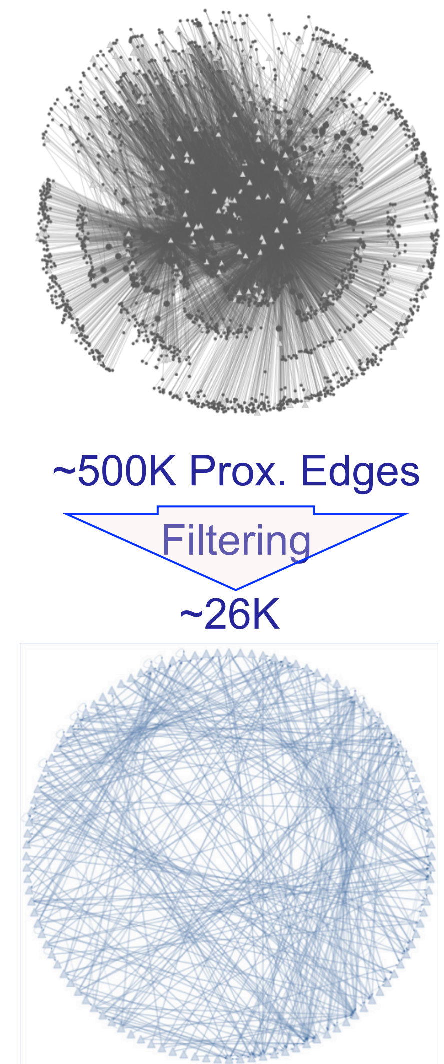
- **Introduction**
  - Being a happy cog in a 500+ person Big-science project
- **Evolution of Linear Annotation based on Functional Genomics**
  - Chip-Chip, Chip-Seq, Thresholding v Control, Segmentation, Multi-scale site calling
- **Its Relation to Conservation: An Enduring Puzzle from Pilot to Production**
  - Many unconstrained regulatory sites
  - But finding small number of sites particularly sensitive to mutations
- **Development of a 2nd Level Network Annotation**
  - Creating it from the linear annotation & connecting it to network science & hubs
  - More connectivity, more constraint
- **New Direction: Applying the Annotation to Prioritize Mutations**
  - Tools (eg FunSeq) for systematically weighting non-coding features
- **Postscript**
  - Culture Clash: Open Data in Genomics v Patient Privacy
  - Genomics Legacy: the discipline as a exemplar for Data Science

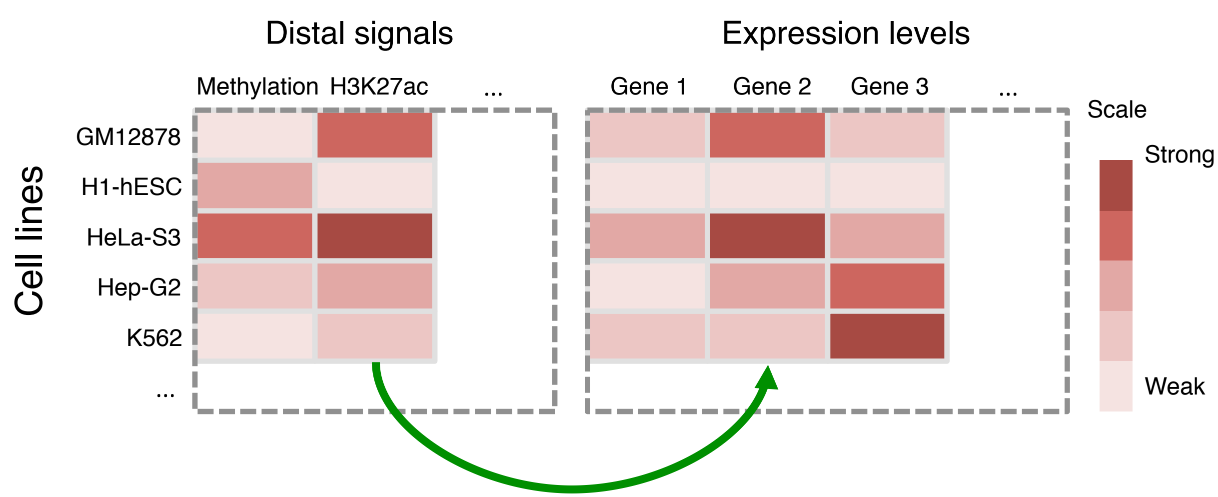
ENCODE: "Evolution" of Approaches to Annotate the Human Genome & Interpret its Variants

Relating Non-coding Annotation to Protein-coding Genes via Networks

[Cheng et al., *Bioinfo.* ('11), Gerstein et al., *Nature* ('12) , Yip et al., *GenomeBiology* ('12), Fu et al., *GenomeBiology*('14)]





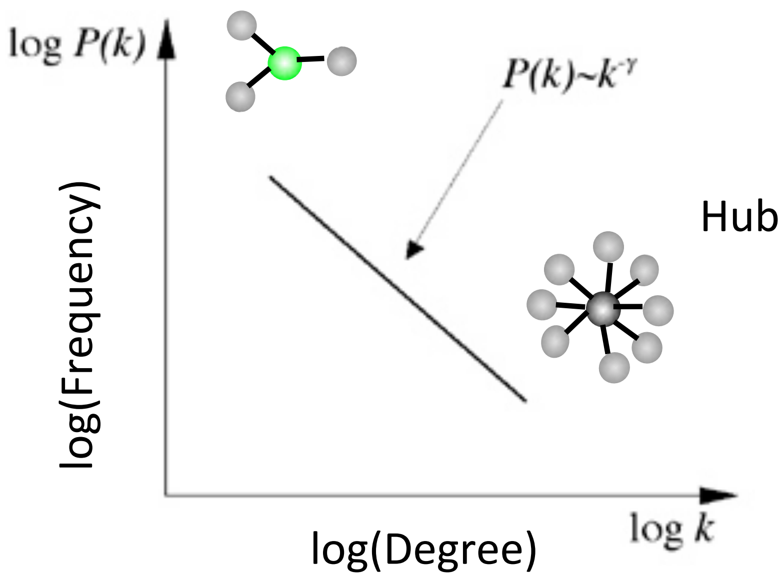


Connecting Distal Elements via **Activity Correlations**.

Other strategies to create linkage incl. eQTL and Hi-C. Much in recent Epigenomics Roadmap.

Hubs Under Constraint: A Finding from the Network Biology Community

Power-law distribution



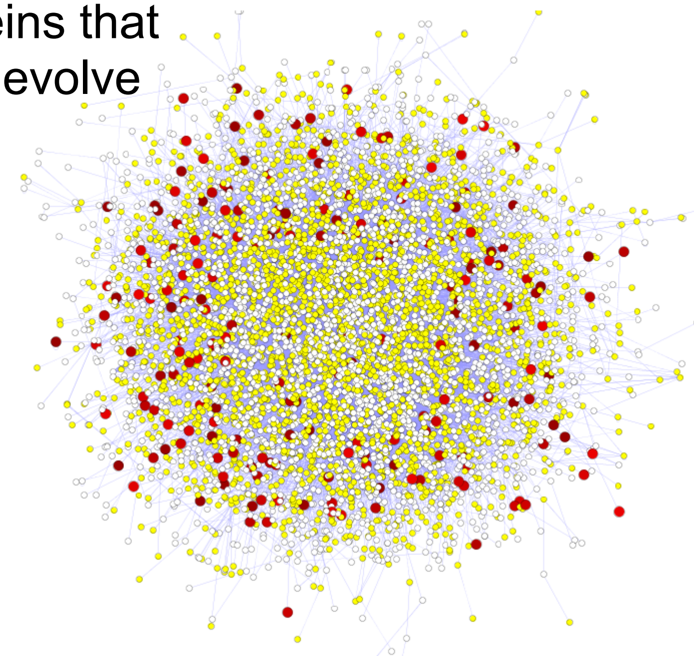
High likelihood of positive selection

Lower likelihood of positive selection

Not under positive selection

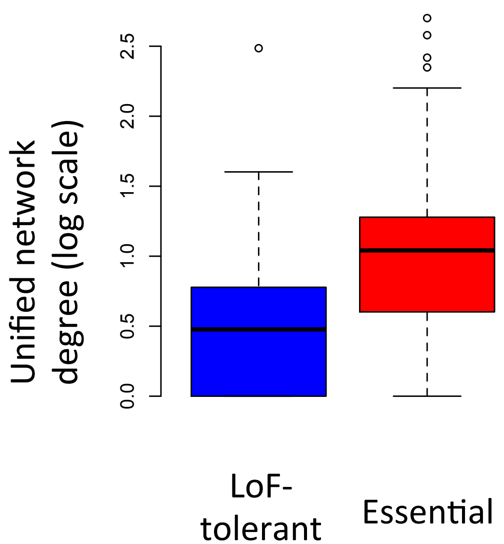
No data about positive selection

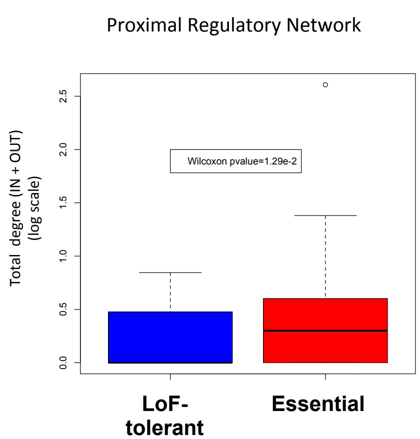
[Nielsen et al. *PLoS Biol.* (2005), HPRD, Kim et al. PNAS (2007)]

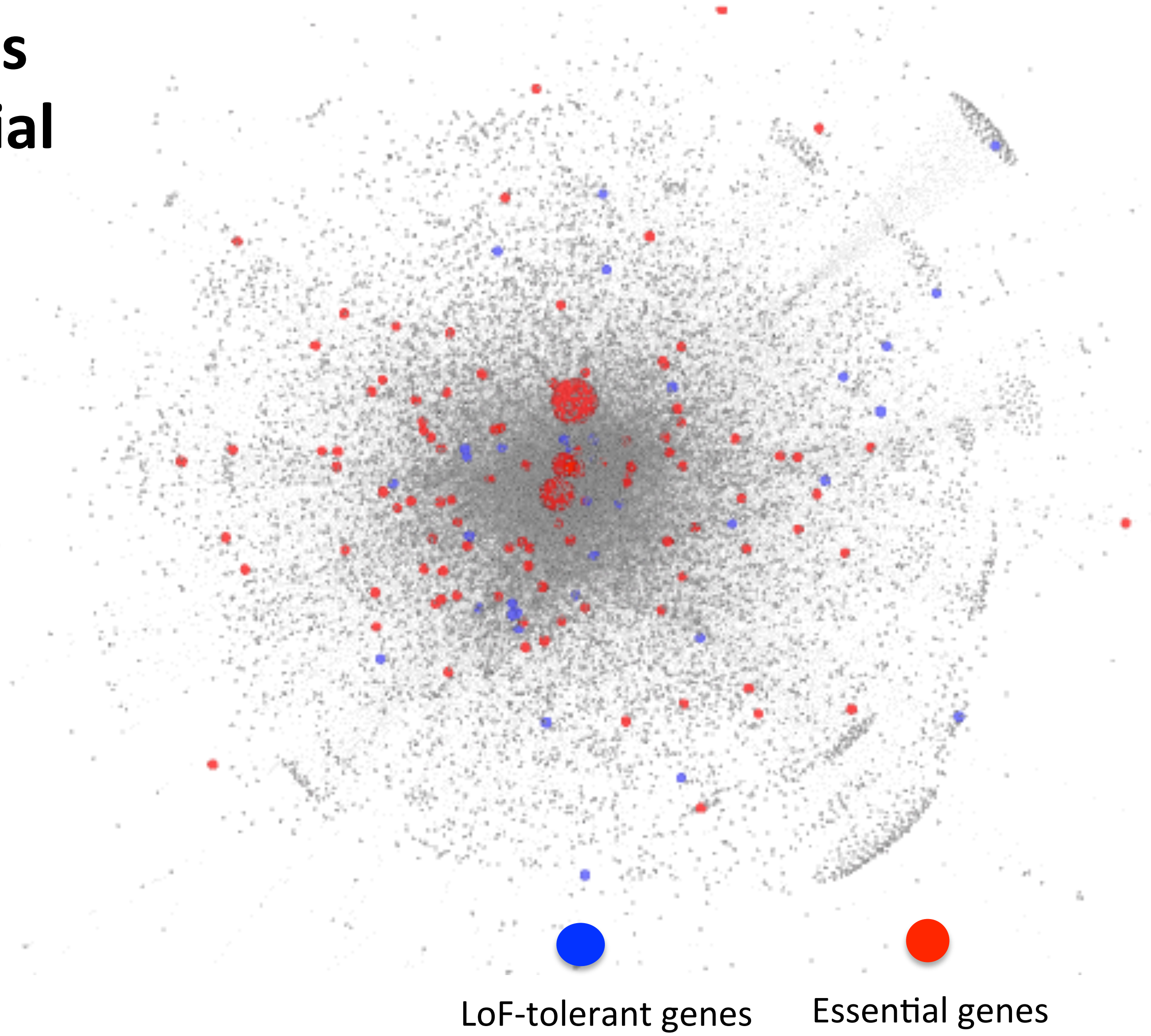


- <u>More Connectivity, More Constraint:</u> Genes & proteins that have a more central position in the network tend to evolve more slowly and are more likely to be essential.
- This phenomenon is observed in **many organisms & different kinds of networks**
  - **yeast PPI** - Fraser et al ('02) Science, ('03) BMC Evo. Bio.
  - **Ecoli PPI** - Butland et al ('04) Nature
  - **Worm/fly PPI** - Hahn et al ('05) MBE
  - **miRNA net** - Cheng et al ('09) BMC Genomics

Regulatory Hubs are more Essential







Size of nodes scaled by total degree

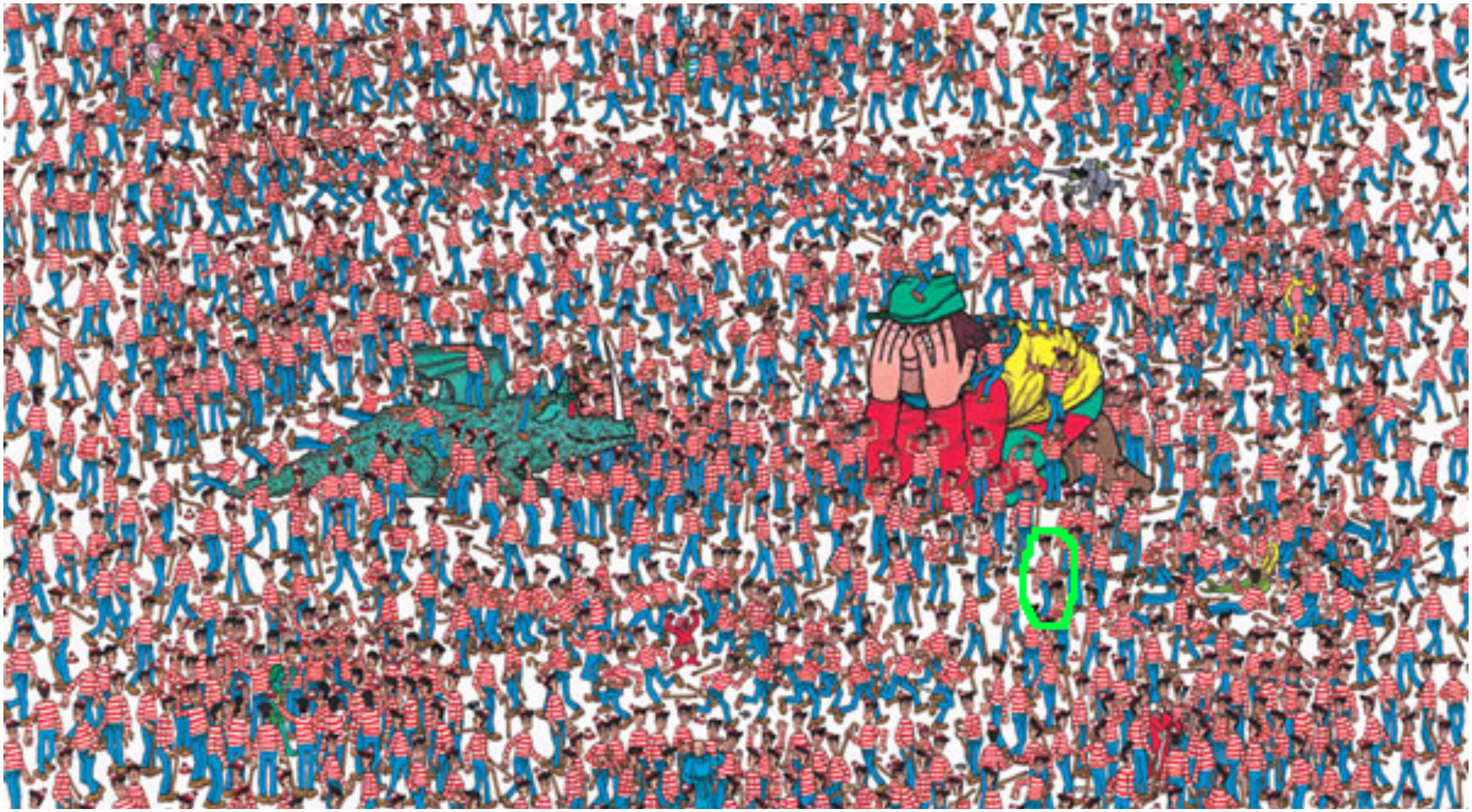
[Khurana et al., *PLOS Comp. Bio.* '13]

- **Introduction**
  - Being a happy cog in a 500+ person Big-science project
- **Evolution of Linear Annotation based on Functional Genomics**
  - Chip-Chip, Chip-Seq, Thresholding v Control, Segmentation, Multi-scale site calling
- **Its Relation to Conservation: An Enduring Puzzle from Pilot to Production**
  - Many unconstrained regulatory sites
  - But finding small number of sites particularly sensitive to mutations
- **Development of a 2nd Level Network Annotation**
  - Creating it from the linear annotation & connecting it to network science & hubs
  - More connectivity, more constraint
- **New Direction: Applying the Annotation to Prioritize Mutations**
  - Tools (eg FunSeq) for systematically weighting non-coding features
- **Postscript**
  - Culture Clash: Open Data in Genomics v Patient Privacy
  - Genomics Legacy: the discipline as a exemplar for Data Science

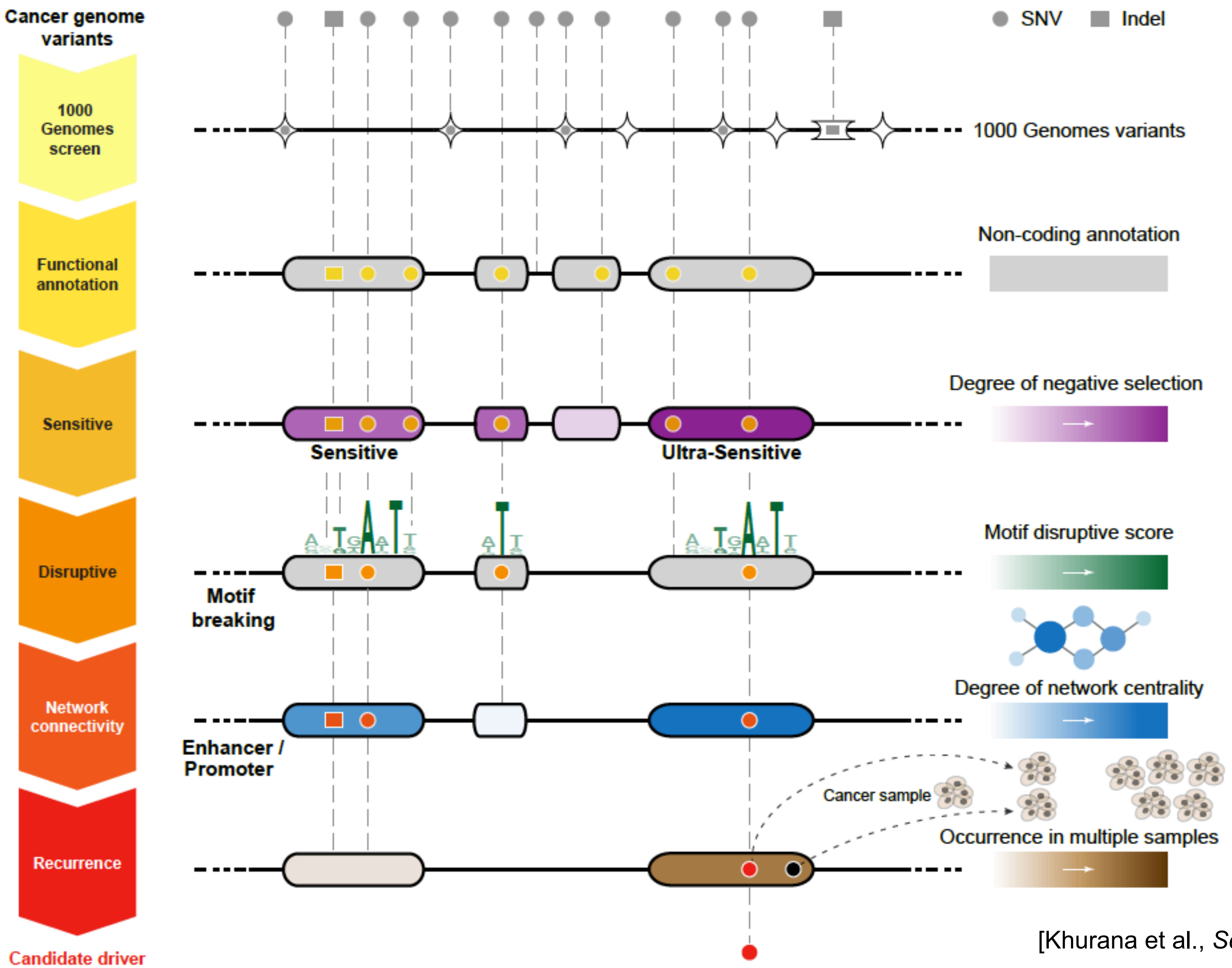
ENCODE: "Evolution" of Approaches to Annotate the Human Genome & Interpret its Variants

Where is Waldo?

(Finding the key mutations in ~4M Germline variants & ~5K Somatic Variants in a Tumor Sample)



Applying Linear & Network Annotation to Prioritize Somatic Mutations as Possible Drivers



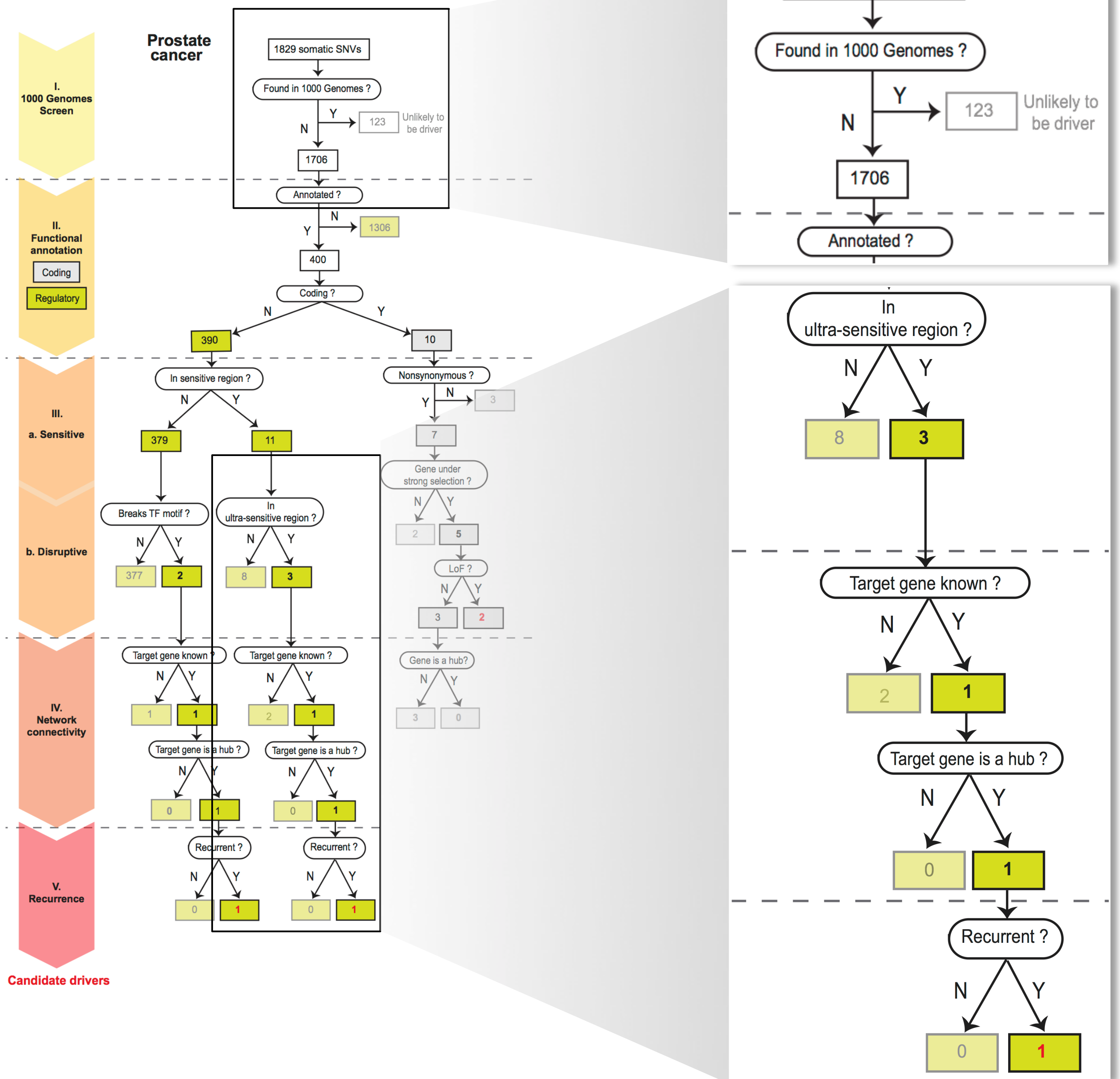
FunSeq

Other Tools for non-coding prioritization: HaploReg, VEP, CADD....

[Khurana et al., *Science* ('13)]

Flowchart for 1 Prostate Cancer Genome
(from Berger et al. '11)

Start: 1829
Somatic SNVs



End: 1 Somatic SNV in ultra-sensitive region & hub. Potential non-coding Driver

[Khurana et al., *Science* ('13)]

- **Introduction**
  - Being a happy cog in a 500+ person Big-science project
- **Evolution of Linear Annotation based on Functional Genomics**
  - Chip-Chip, Chip-Seq, Thresholding v Control, Segmentation, Multi-scale site calling
- **Its Relation to Conservation: An Enduring Puzzle from Pilot to Production**
  - Many unconstrained regulatory sites
  - But finding small number of sites particularly sensitive to mutations
- **Development of a 2nd Level Network Annotation**
  - Creating it from the linear annotation & connecting it to network science & hubs
  - More connectivity, more constraint
- **New Direction: Applying the Annotation to Prioritize Mutations**
  - Tools (eg FunSeq) for systematically weighting non-coding features
- **Postscript**
  - Culture Clash: Open Data in Genomics v Patient Privacy
  - Genomics Legacy: the discipline as a exemplar for Data Science

ENCODE: "Evolution" of Approaches to Annotate the Human Genome & Interpret its Variants

ENCODE: "Evolution" of Approaches to Annotate the Human Genome & Interpret its Variants

- **Introduction**
  - Being a happy cog in a 500+ person Big-science project
- **Evolution of Linear Annotation based on Functional Genomics**
  - Chip-Chip, Chip-Seq, Thresholding v Control, Segmentation, Multi-scale site calling
- **Its Relation to Conservation: An Enduring Puzzle from Pilot to Production**
  - Many unconstrained regulatory sites
  - But finding small number of sites particularly sensitive to mutations
- **Development of a 2nd Level Network Annotation**
  - Creating it from the linear annotation & connecting it to network science & hubs
  - More connectivity, more constraint
- **New Direction: Applying the Annotation to Prioritize Mutations**
  - Tools (eg FunSeq) for systematically weighting non-coding features
- **Postscript**
  - Culture Clash: Open Data in Genomics v Patient Privacy
  - Genomics Legacy: the discipline as a exemplar for Data Science

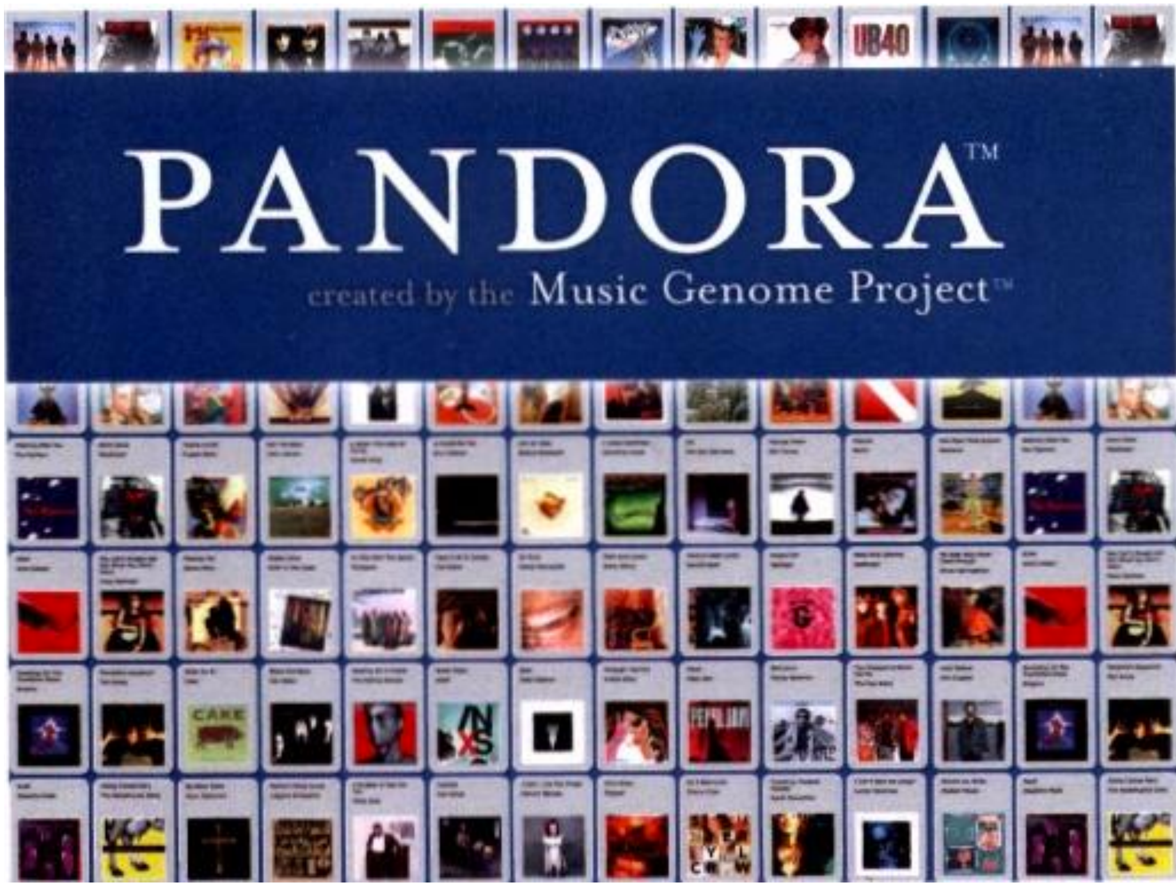
Culture Clash: Open Data in Genomics v Patient Privacy

- **Open Data, Open Source, &c is the culture of Genomics ("its meta-DNA")**
  - **Origins in worm project**

- **Strong Reasons for Genomic Privacy in the Future**
  - Personal Genomic info. essentially meaningless currently but will it be in 20 yrs? 50 yrs?
  - Genomic sequence very revealing about one's children
  - Once put on the web it can't be taken back

[D Greenbaum & M Gerstein ('08). Am J. Bioethics; D Greenbaum & M Gerstein, Hartford Courant, 10 Jul. '08 ; SF Chronicle, 2 Nov. '08; Greenbaum et al. *PLOS CB* ('11) ; Greenbaum & Gerstein ('13), The Scientist; Photo from NY Times]

Legacy of Human Genome Annotation? Is it an early exemplar for Data Science



Harvard Business Review

[Oct. '12 issue]

Data Scientist: The Sexiest Job of the 21st Century

by Thomas H. Davenport and D.J. Patil

Artwork: **Tamar Cohen, *Andrew J Buboltz,*** 2011, silk screen on a page from a high

When Jonathan Goldman arrived for work in June 2006 at LinkedIn, the business ne up. The company had just under 8 million accounts, and the number was growing qu friends and colleagues to join. But users weren't seeking out connections with the pe rate executives had expected. Something was apparently missing in the social expe

Info about content in this slide pack

- General PERMISSIONS
  - This Presentation is copyright Mark Gerstein, Yale University, 2015.
  - Please read permissions statement at http://www.gersteinlab.org/misc/permissions.html .
  - Feel free to use slides & images in the talk with PROPER acknowledgement (via citation to relevant papers or link to gersteinlab.org).
  - Paper references in the talk were mostly from Papers.GersteinLab.org.

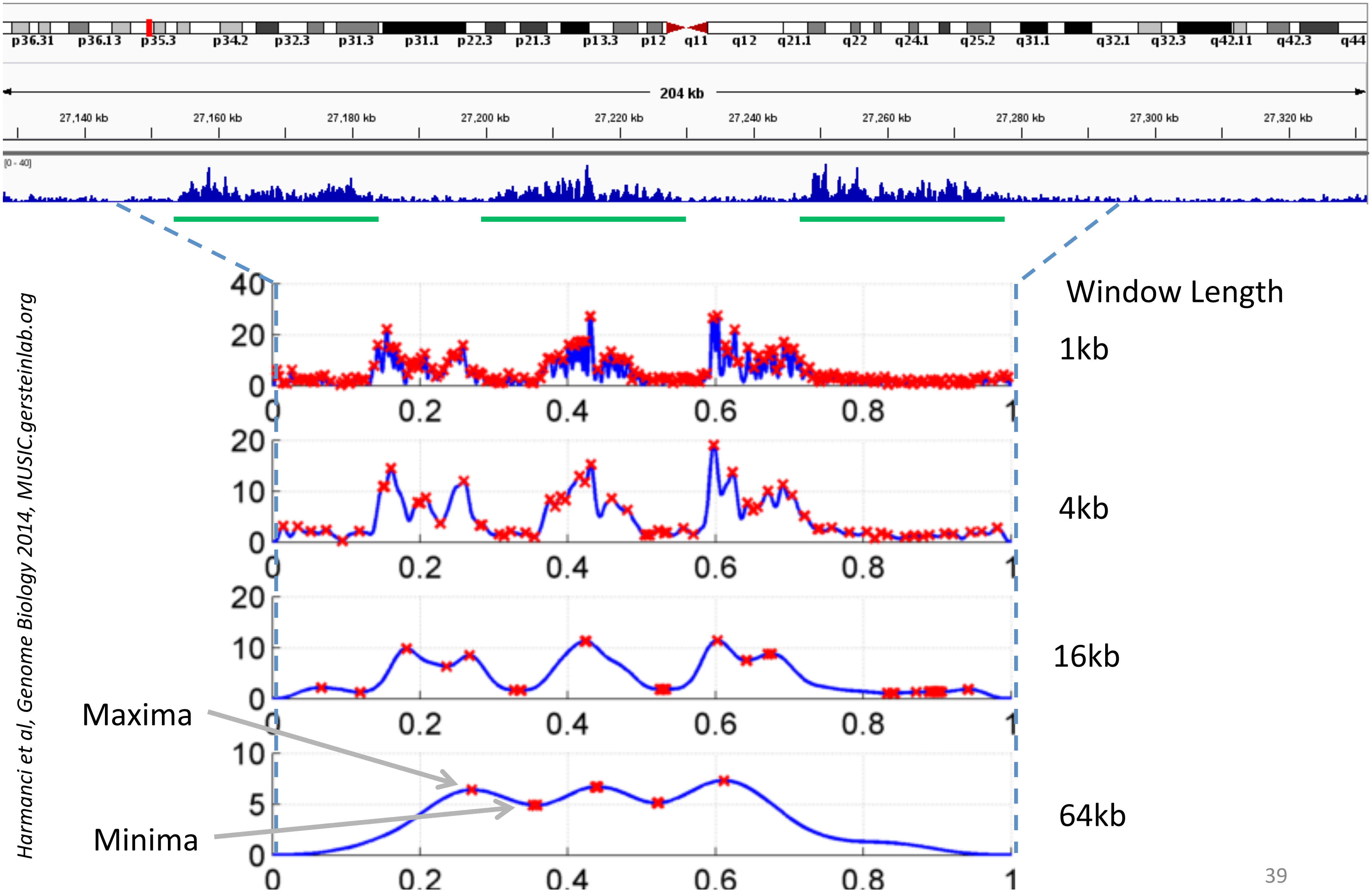
- For SeqUniverse slide, please contact Heidi Sofia, NHGRI

- PHOTOS & IMAGES. For thoughts on the source and permissions of many of the photos and clipped images in this presentation see http://streams.gerstein.info .
  - In particular, many of the images have particular EXIF tags, such as kwpotppt , that can be easily queried from flickr, viz: http://www.flickr.com/photos/mbgmbg/tags/kwpotppt

MUSIC makes music

- -get_multiscale_music: Generates a .wav file using the aggregate multiscale decomposition
- Listen to K562 H3K36me3 chromosome 1: http://archive.gersteinlab.org/proj/MUSIC/music/H3K36me3.mp3
  - Telomeres are vocal, centromeres (46:00-53:00) are silent
- Listen K562 H3K4me3 chromosome 1: http://archive.gersteinlab.org/proj/MUSIC/music/H3K4me3.mp3
  - More "clicky" than H3K36me3 with more punctate enriched regions

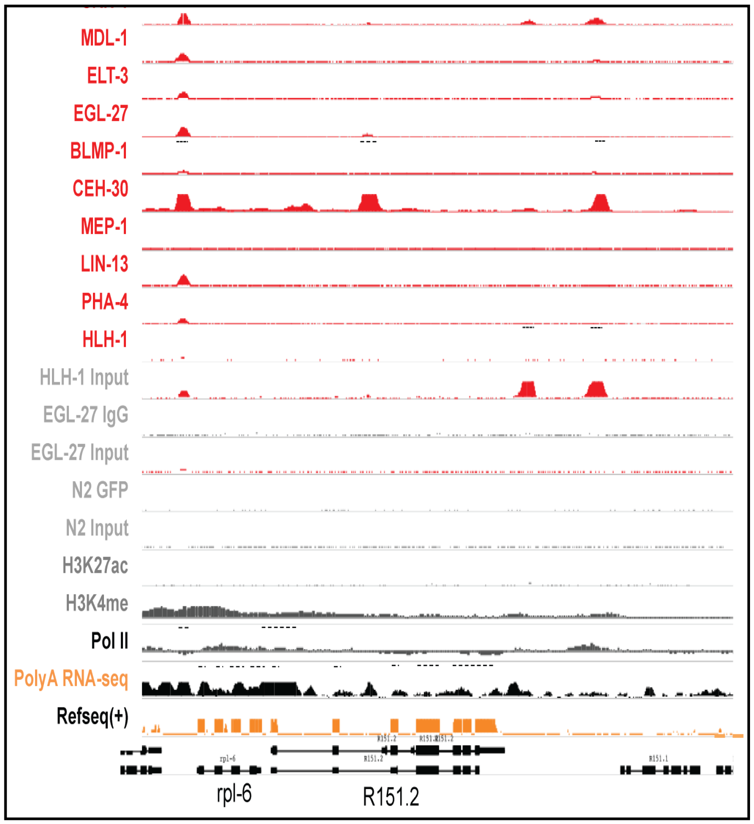
Multiscale Analysis, Minima/Maxima based Coarse Segmentation

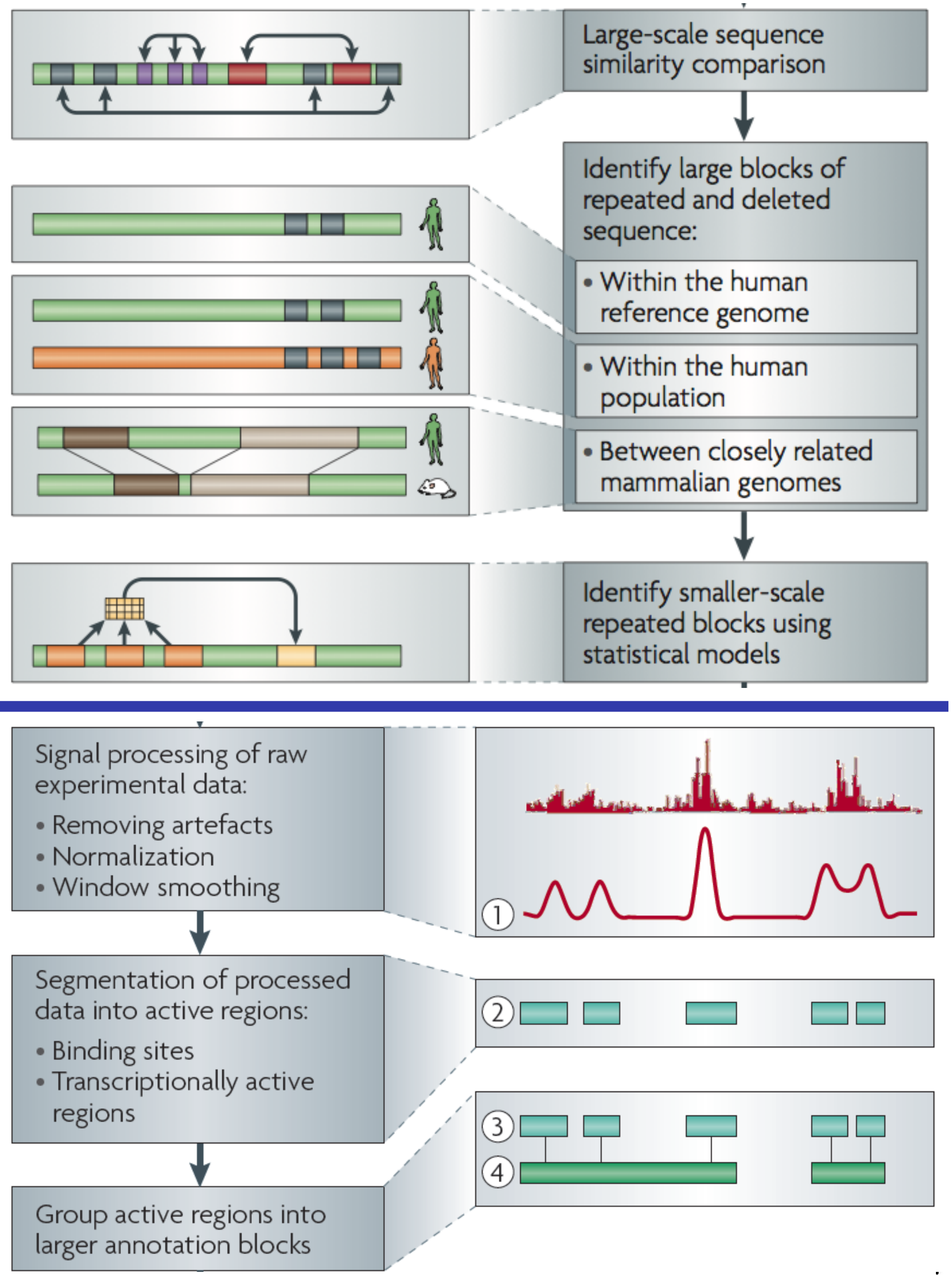


Harmanci et al, Genome Biology 2014, MUSIC.gersteinlab.org

Sources of Annotation: Comparative & Functional

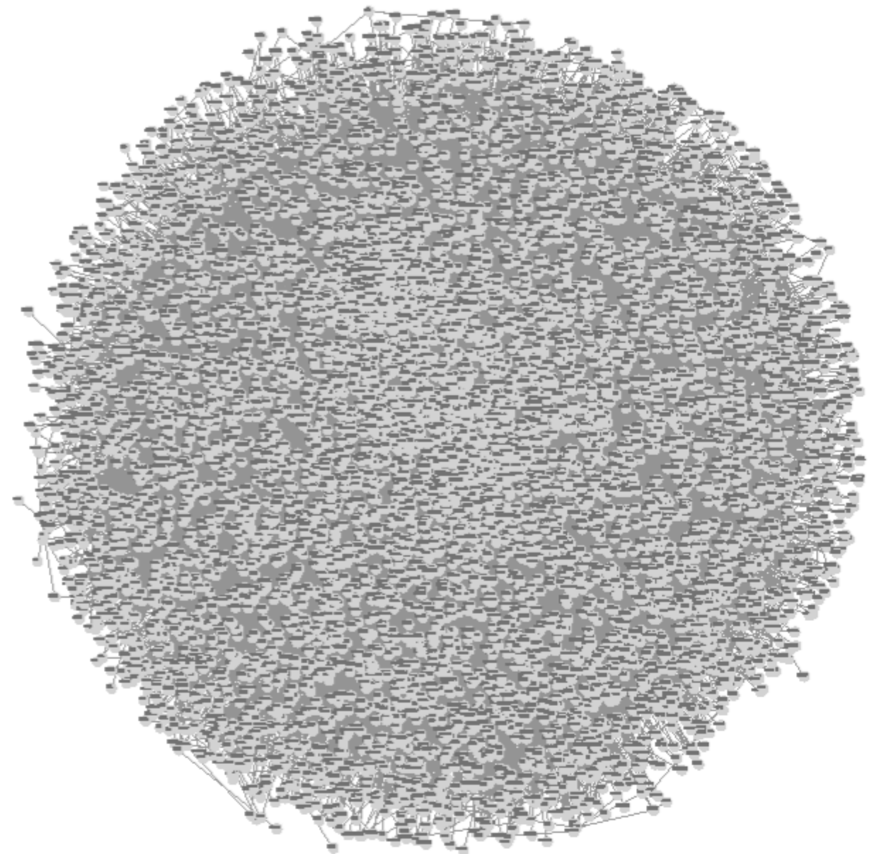
[Nat. Rev. Genet. (2010) 11: 559]
[Science 330:6012]





Multinet – the ultimate hairball!

Genes participate in many networks and no single network captures the global picture of gene interactions



Combine **regulatory** interactions with other networks : **physical protein-protein**, **signaling**, **metabolic**, **phosphorylation** and **genetic** to create a **unified network (Multinet)**

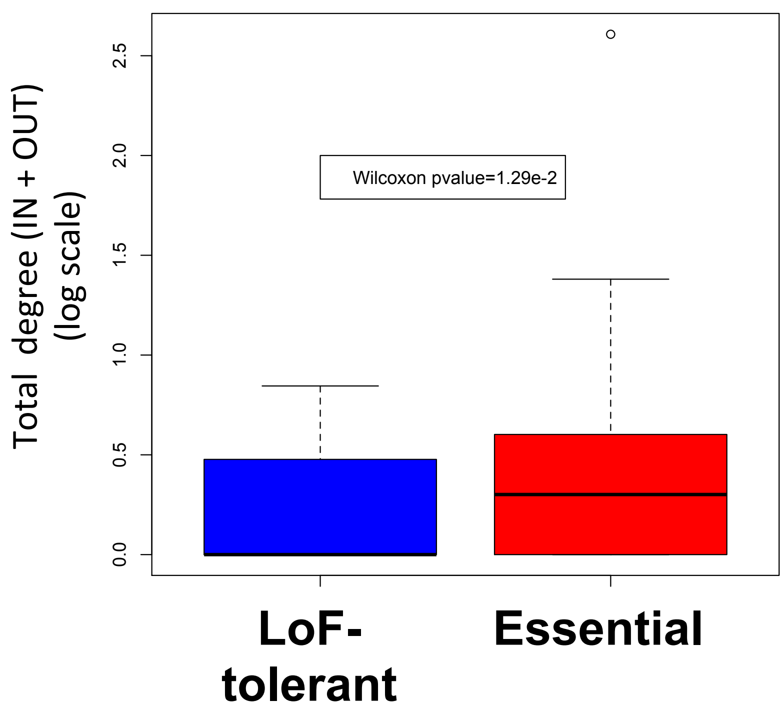
Nodes: ~15,000 genes
Edges: ~110,000 interactions

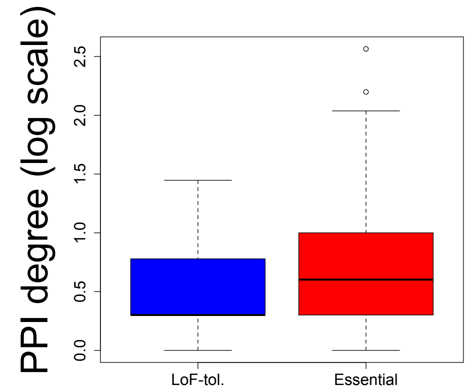
Edges shown in gray

[Khurana et al., *PLOS Comp. Bio.* '13]

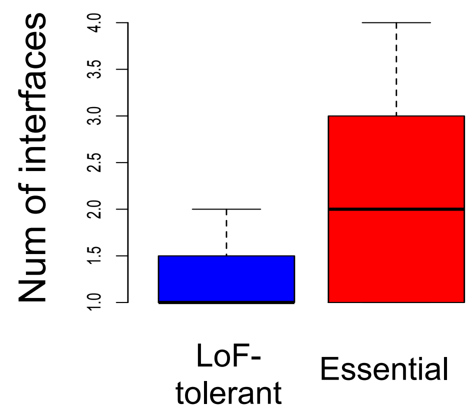
Centrality in Gene Networks Weakly Associated with Essentiality

Proximal Regulatory Network





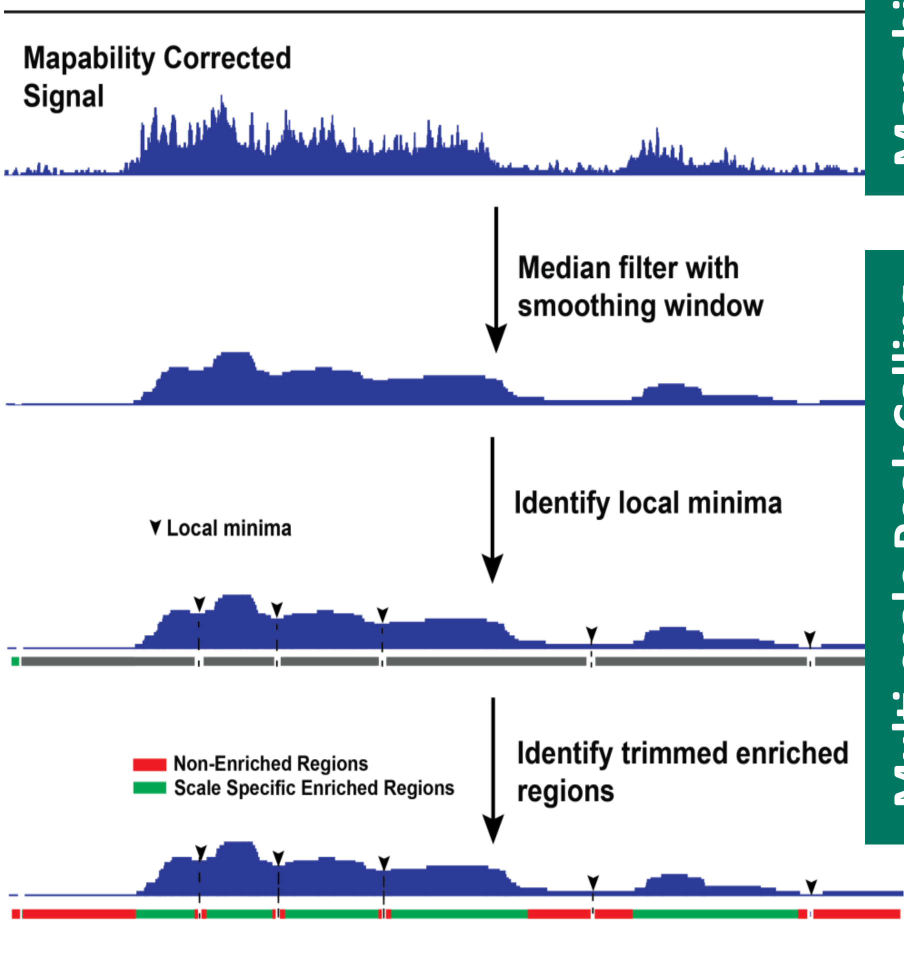
Higher Centrality In PPI

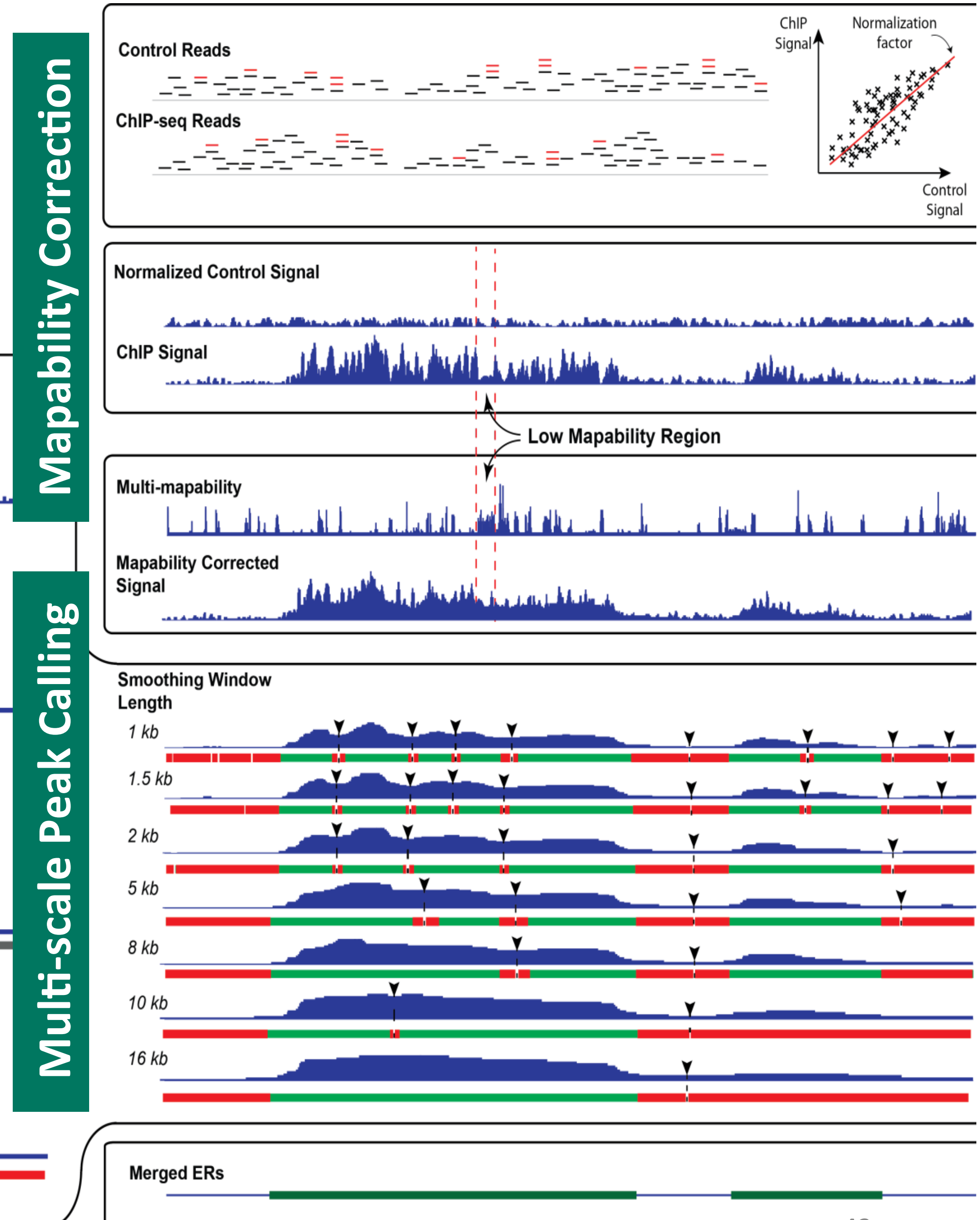


More interaction interfaces

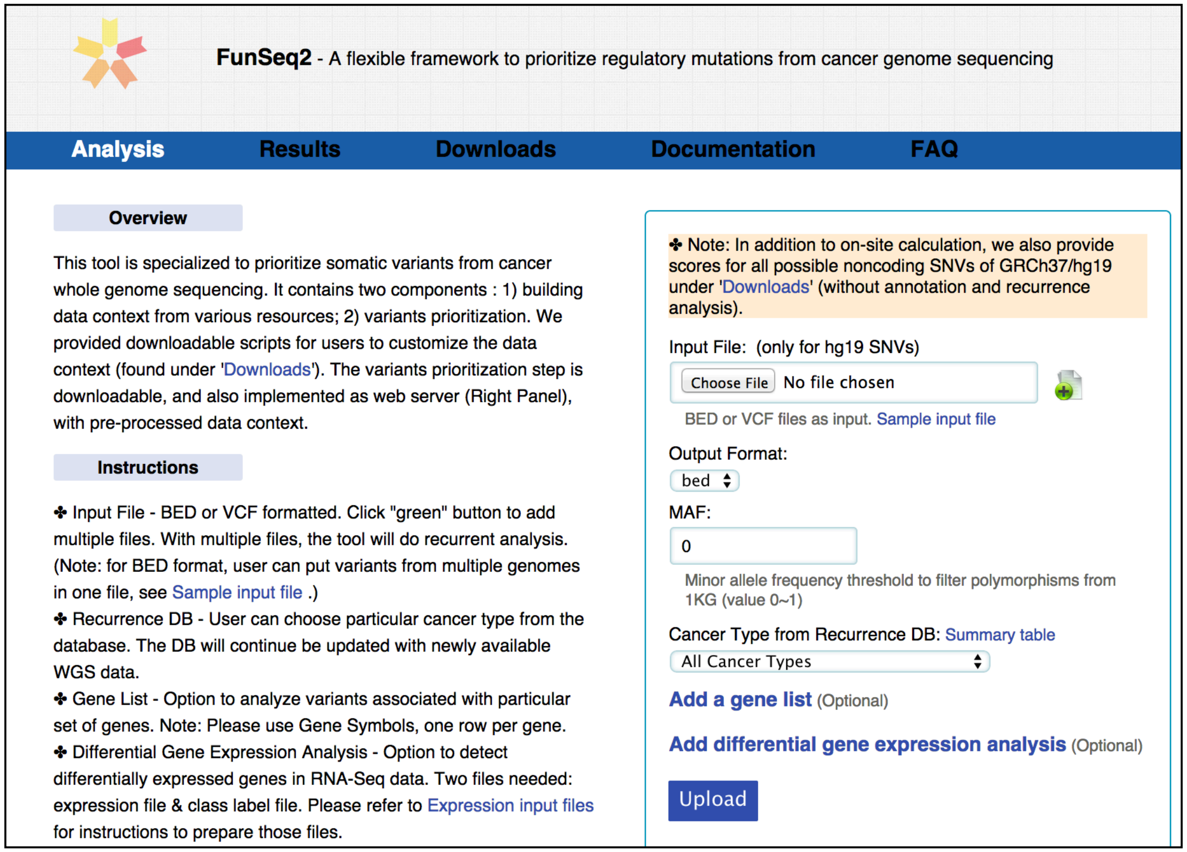
Khurana et al., *PLoS Comp. Bio.*, 2013

MUSIC.gersteinlab.org Algorithm

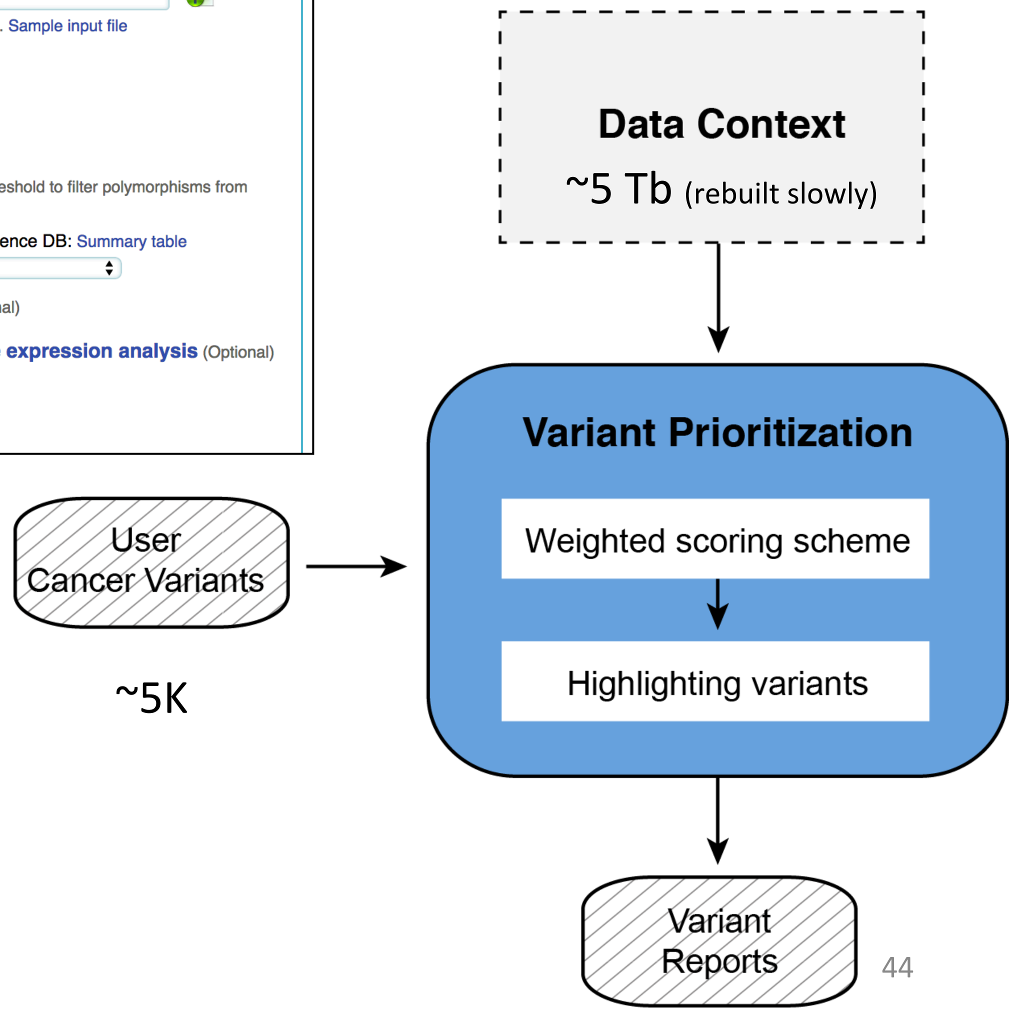




[Harmanci *et al, Genome Biol.* ('14)]



Site integrates user variants with large-scale context

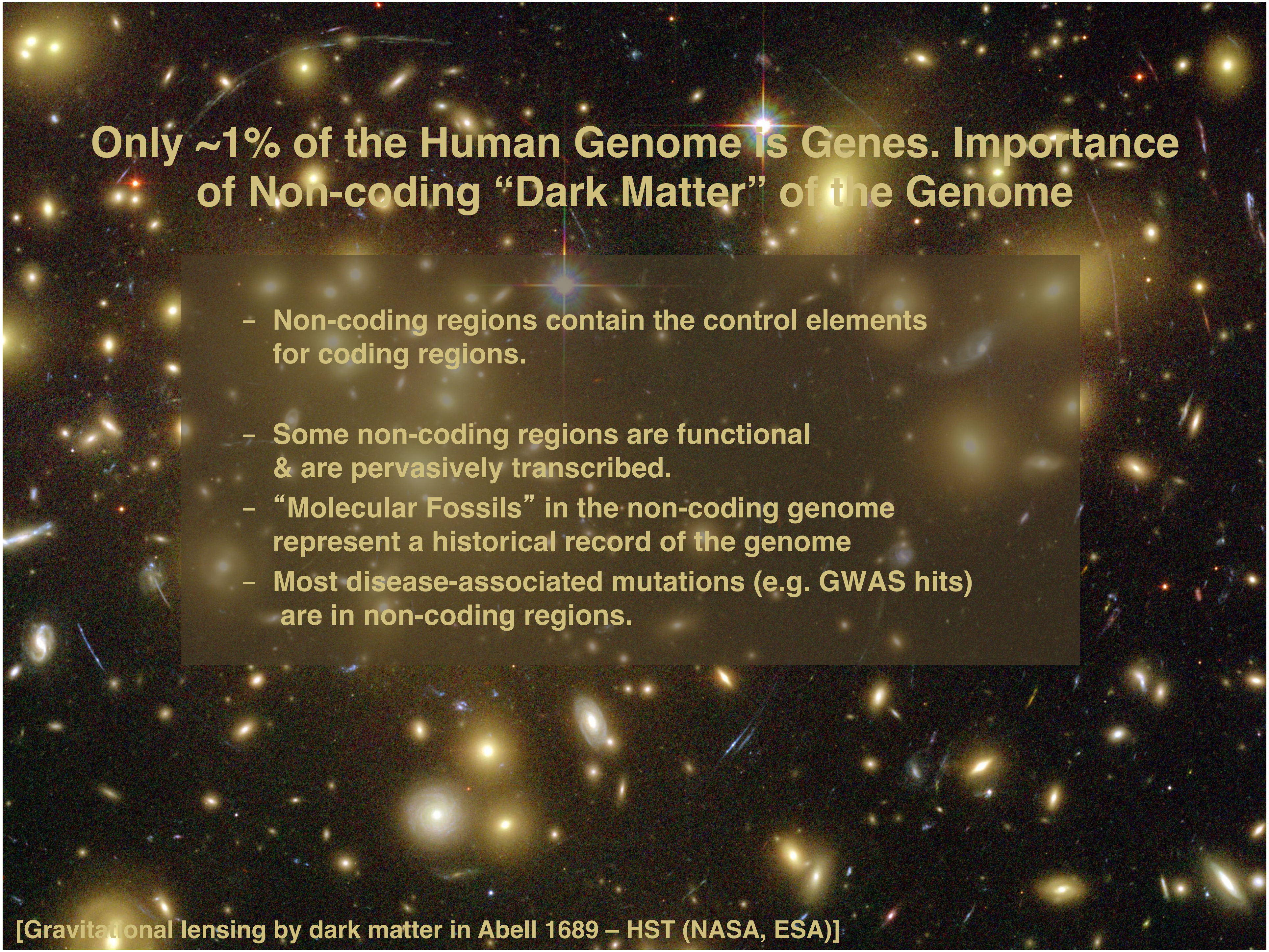


FunSeq.gersteinlab.org

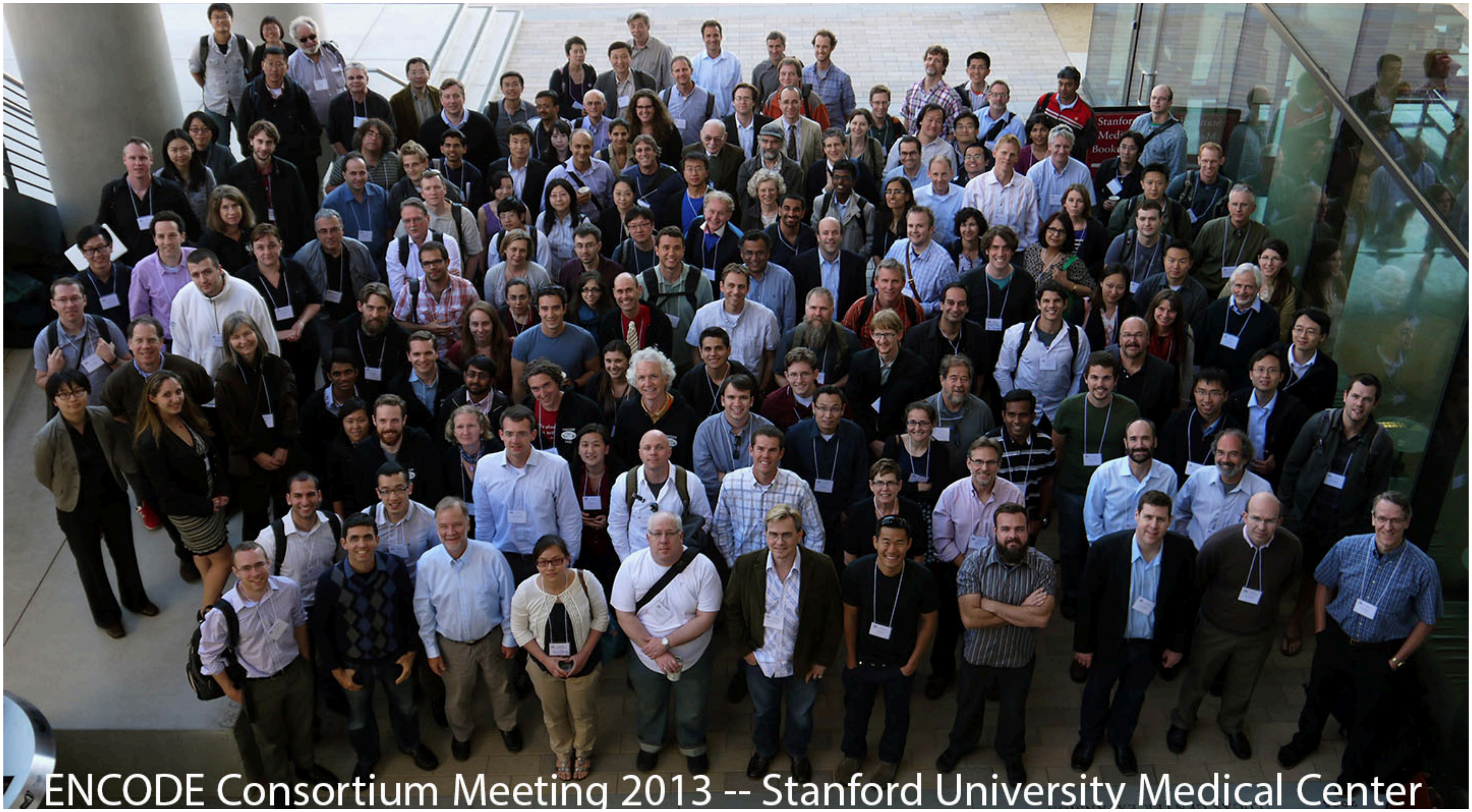
[Fu et al., GenomeBiology ('14)]

Only ~1% of the Human Genome is Genes. Importance of Non-coding "Dark Matter" of the Genome

- Non-coding regions contain the control elements for coding regions.
- Some non-coding regions are functional & are pervasively transcribed.
- "Molecular Fossils" in the non-coding genome represent a historical record of the genome
- Most disease-associated mutations (e.g. GWAS hits) are in non-coding regions.

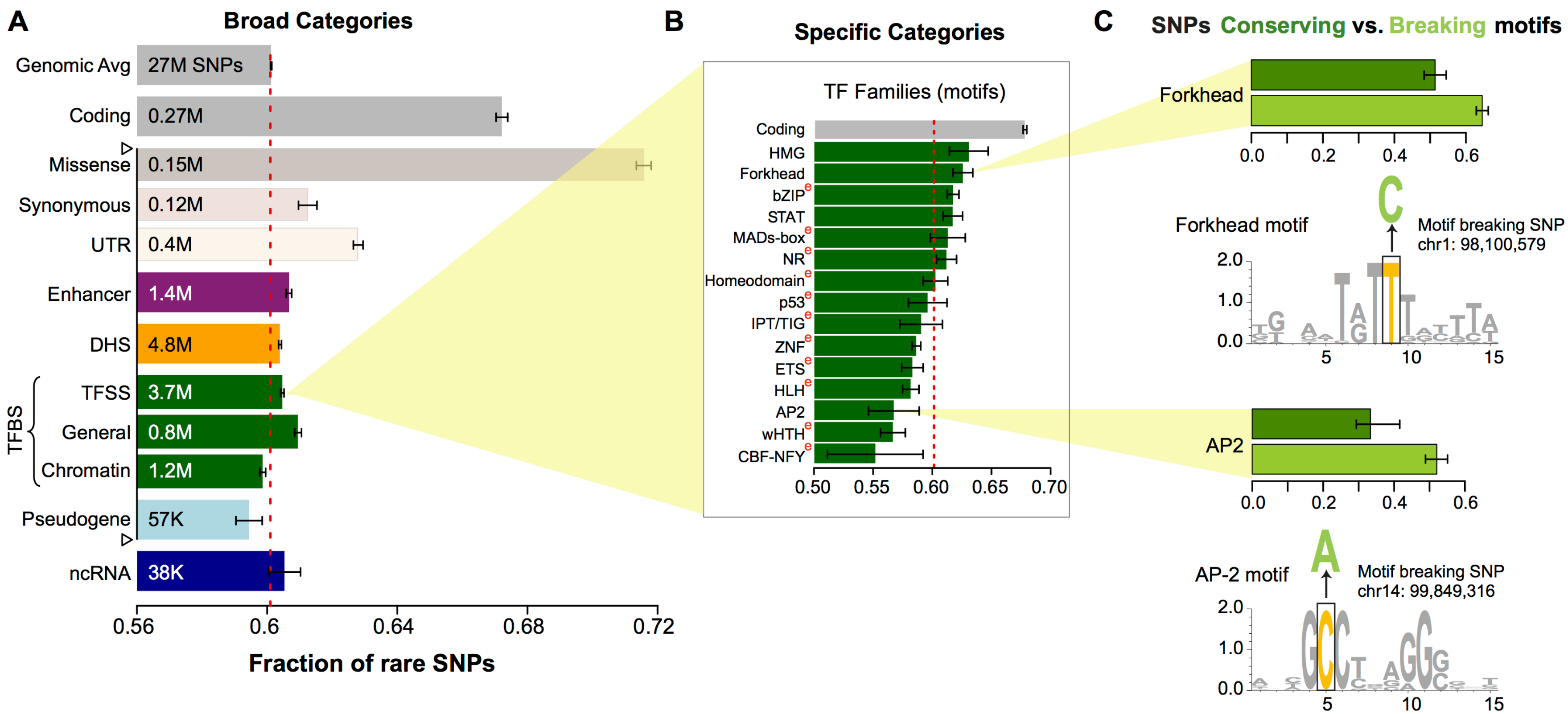


[Gravitational lensing by dark matter in Abell 1689 – HST (NASA, ESA)]



ENCODE Consortium Meeting 2013 -- Stanford University Medical Center

SNPs which break TF motifs are under particularly strong selection



[Khurana et al., *Science* ('13)]

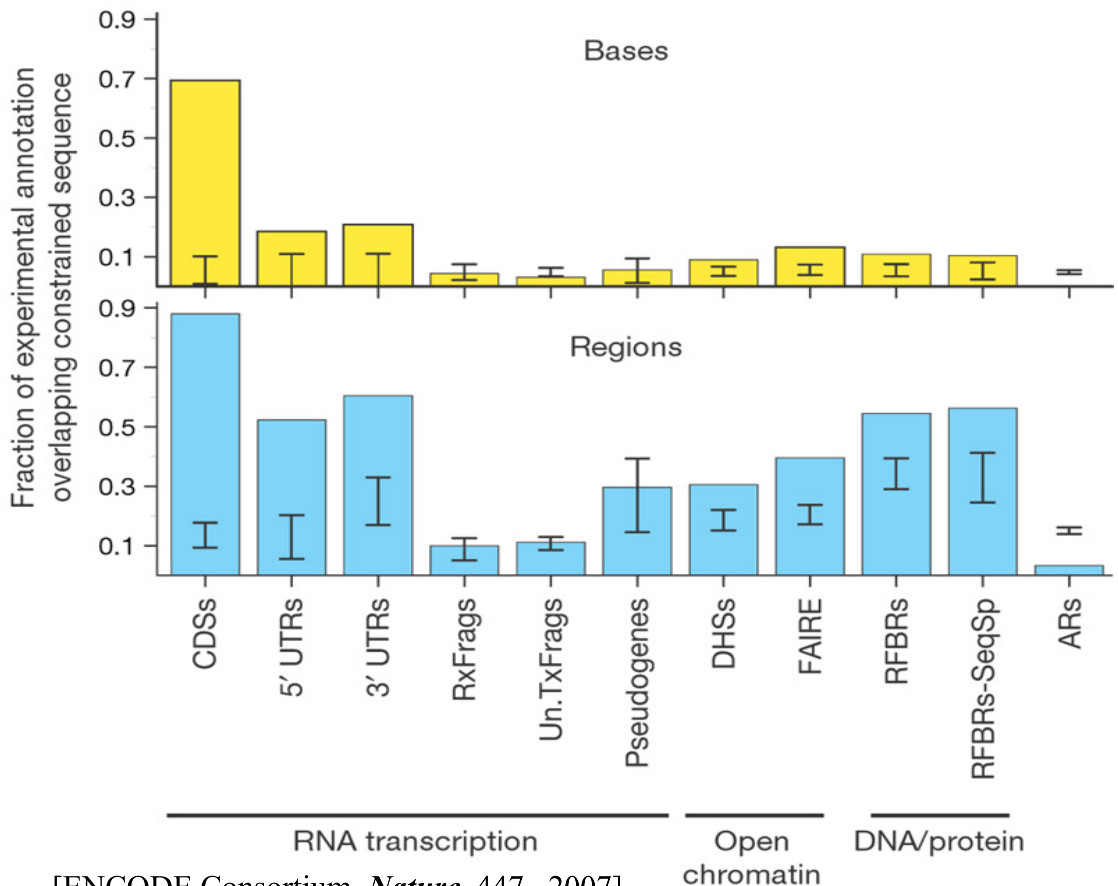
Hiring postdocs, see

GersteinLab.org/jobs

Acknowledgements

- **MUSIC**.gersteinlab.org
  - A **Harmanci**, J Rozowsky
- **FunSeq2**.gersteinlab.org
  - Y **Fu**, Z Liu, S Lou, J Bedford, X Mu, K Yip, E **Khurana**

Biochemically Active Regions Don't all Appear to be Under Constraint



[ENCODE Consortium, ***Nature*** 447, 2007]

- Integrating & averaging results over larger and larger sets
- Comparison of integrated quantities

Fusion Group

1000G FIG-Cancer

Acknowledgements

Yale

Ekta Khurana, Yao Fu, Jieming Chen,
Xinmeng Mu, Lucas Lochovsky, Arif Harmanci, Alexej Abyzov, Suganthi Balasubramanian, Cristina Sisu,
Declan Clarke, Mike Wilson

Sanger

Vincenza Colonna, Yali Xue, **Chris Tyler-Smith**

Cornell

Steven Lipkin, Jishnu Das, Robert Fragoza, Xiaomu Wei, **Haiyuan Yu**

Andrea Sboner, Dimple Chakravarty, Naoki Kitabayashi, Vaja Liluashvili, Zeynep H. Gümüş, **Mark A. Rubin**

~50 people ← ~1000 "authors"

Functional Interpretation Subgroup

A THOUSAND GENOMES

US, UK, Switzerland….

Hyun Min Kang, Tuuli Lappalainen, Kathryn Beal, Daniel Challis, Yuan Chen, Laura Clarke, Fiona Cunningham, Emmanouil T. Dermitzakis, Uday Evani, Paul Flicek, Erik Garrison, Javier Herrero, Yong Kong, Kasper Lage, Daniel G. MacArthur, Gabor Marth, Donna Muzny, Tune H. Pers, Graham R. S. Ritchie, Jeffrey A. Rosenfeld, Fuli Yu, Richard Gibbs