

Биоинформатика в 
что ДНК говорит нам о
жизни и о ее истории

М.Гельфанд

50-лет ИППИ

29.12.1961 - 7.12.2011

2001: Лаборатория № 6 (В.А.Любецкий)

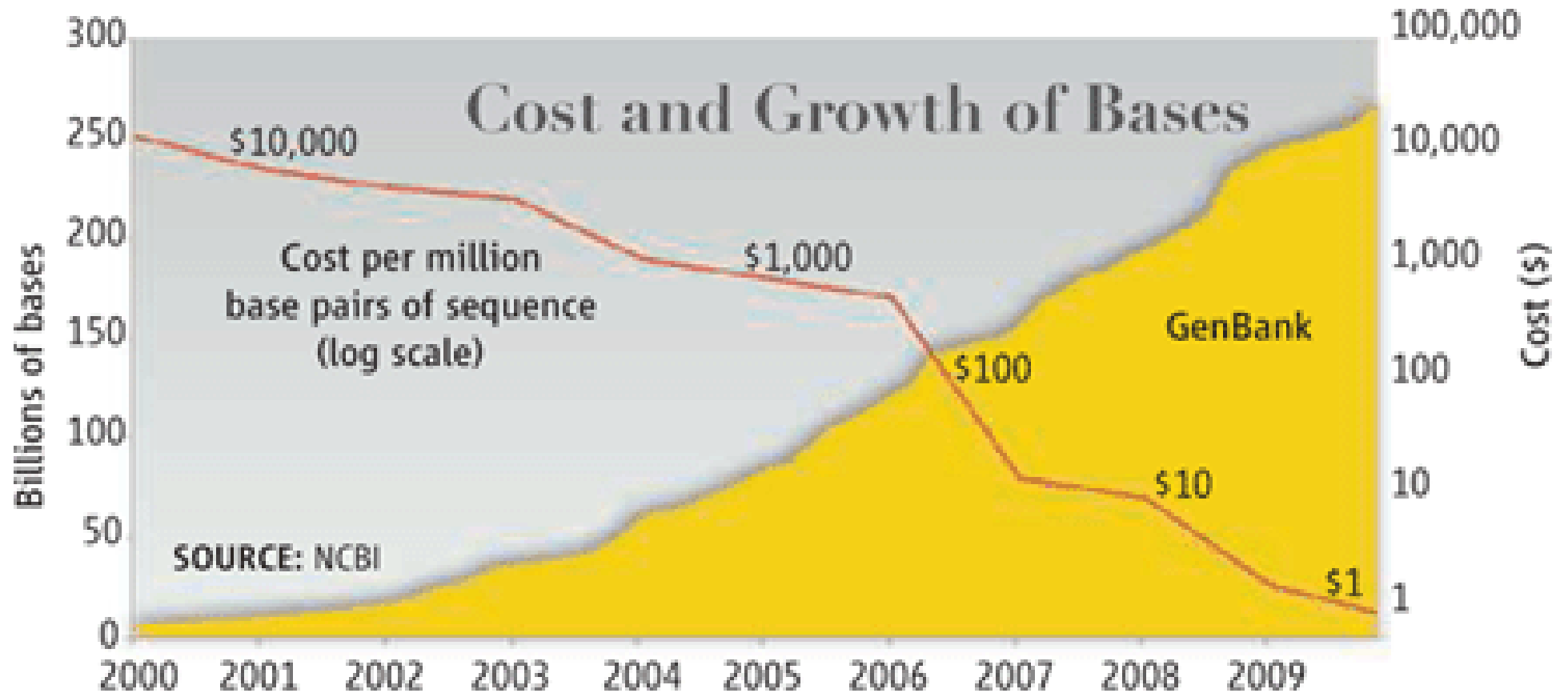
2003



2008: Сектор № 4 (Г.А.Базыкин)

2010: Сектор № 6 (Д.А.Родионов)

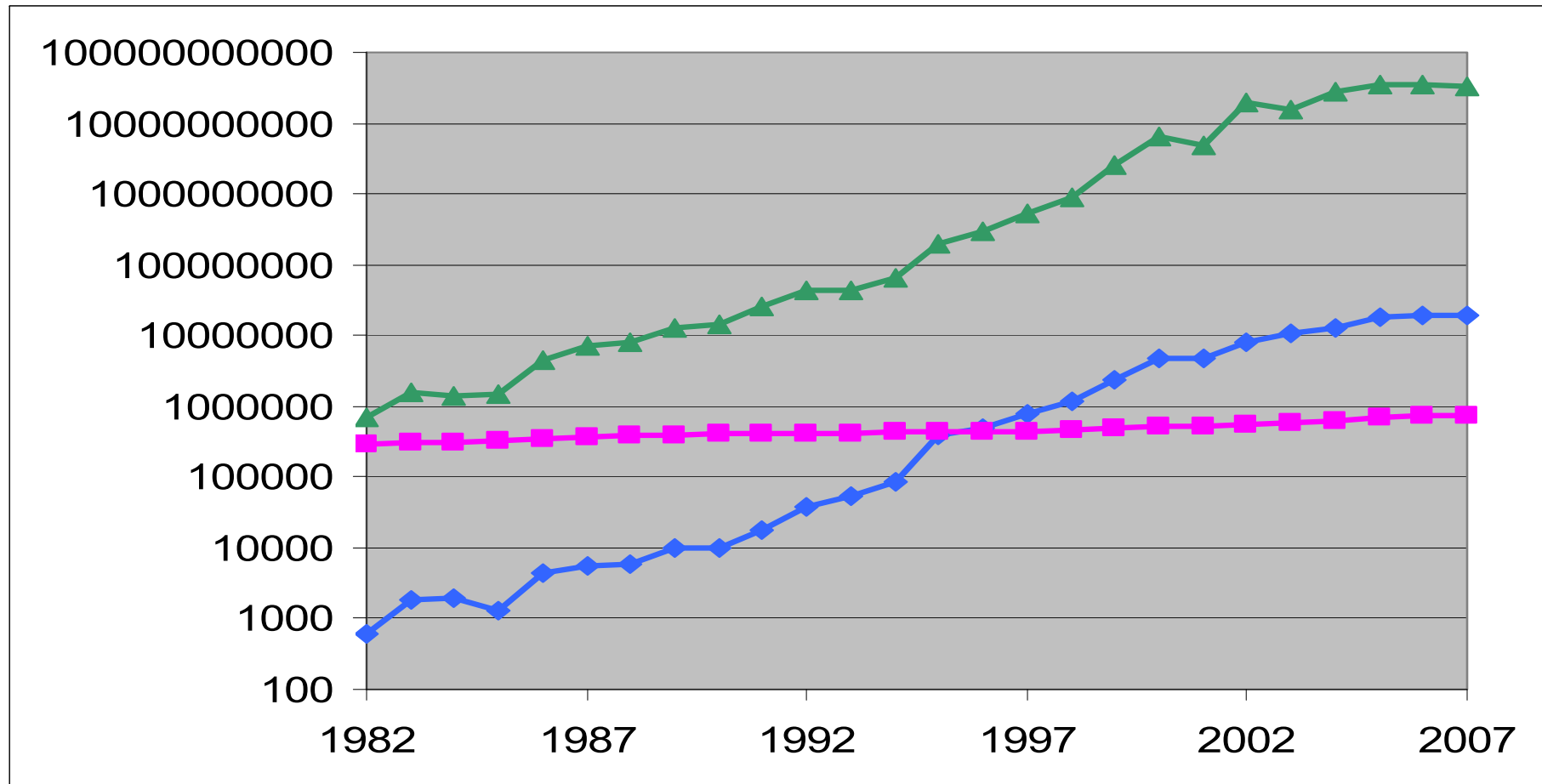
So much for so little. The decline in sequencing costs (red line) has led to a surge in stored DNA data.



Will Computers Crash Genomics?

Science 11 February 2011:
vol. 331 no. 6018 666-668

Рост объема нуклеотидных последовательностей



красный - статьи (PubMed)
голубой - фрагменты ДНК (GenBank)
зеленый - нуклеотиды (GenBank)

из 18 млн. статей в PubMed, ~675 тыс. имеют ключевое слово "bioinformat* OR comput*"

Полные бактериальные геномы

30

количество геномов, полная последовательность которых была опубликована в течение года

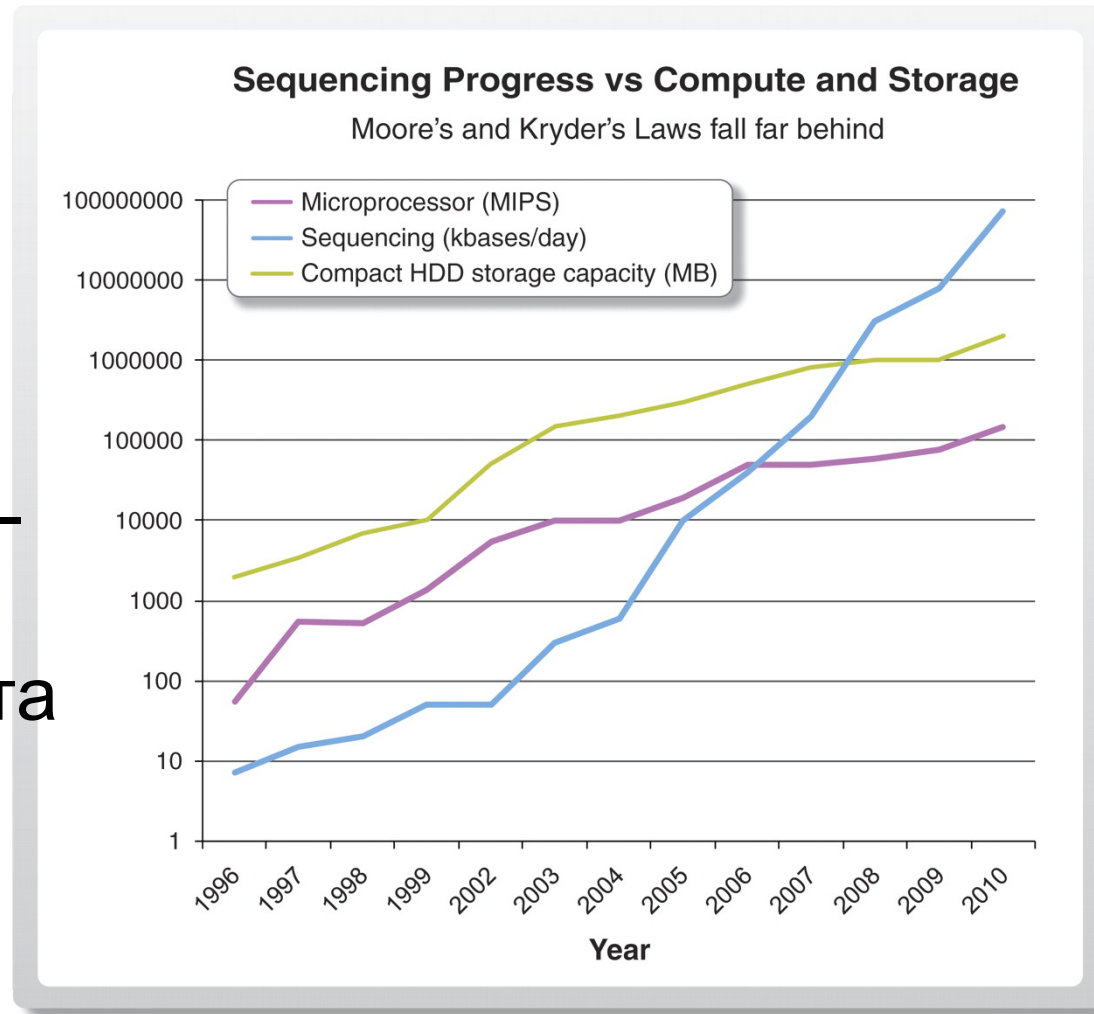
85

общее количество геномов с известной полной последовательностью



Fig. 1 A doubling of sequencing output every 9 months has outpaced and overtaken performance improvements within the disk storage and high-performance computation fields.

на самом деле, все-таки экспонента



S D Kahn Science 2011;331:728-729



0.1% генома *E. coli*

aacgggc aatat gtcct tgtgt ggatt aaaaa aagagt gttct gatag cagct tctga actgg ttacc tgcgg tgggt aaatt aaaa tttta ttgac ttagg tcaact aaata cttta accaa tatagg ca
tagcgcacagac agata aaaaat tacag agtac acaac atcca tgaac cgcac tagcaccacc attac caccaccatc accat tacc acagg taacg tgcgg ggcgt acgcg tacag gaaac acaga aaa
aagcccg cacct gacagt gtcgg gcttt ttttt cgacc aaagg taacg aggt acaaac catgc gagtgt tgaagt tgcgg ggt acat cagtgg caaaa tgcaga acgt tttct gcggt ttgcc gatat tct
ggaaagc aatgc caggc agggc caggt ggcga ccgtc ctctc tgcgc ccgcc aaaaat tacca accat ctggt agcga tgatt gaaaa aacc attag cggcc aagat gcttt accca atatc agcga tgc
cgaacgt atttt tgcga acttctgac gggac tgcgc ccgc ccagc cggga tttcc gctgg cacaat gaa aactt tctgc gacc aggaa tttgc ccaaa taaaacatgt cctgc atggc attag ttt
gttggggc cagtg cccgg atagc atcaa cgcct ggcgt gattt gcccgt ggcgag aaaaat gtcga tgcgc attat ggccggc gctgt ttaga agcgc gctgg tcaaac gtt accgt tatc gatcc ggtc gaaa
actgcttgcagt ggggc attac ctgca atctacc gtt gatat tgc t g agtcc acccg ccgta ttgcccga agcgc ca tccggc t gacc ac atggg tctga tggc t ggttt cactg ccggt aatga aaa
aggcagc tgggt ggttc tggga cgc aa cggtt ccgac tactc cgc tgcgg tgc tggc ggcct gttta cgcgc c gatt gttgc gaga tctgg acgga t gttg acggt gttta tacct cgcgat ccgcg tca
ggtgcc gatgc gaggt t gttg aagtc gatgt cctat cagga agcga tggag ctttc t t act t cggc gctaa agttc ttcac cccc gcacc atcac cccc tgcgc cagtt t caga tccct tgcct gat
taaaaa accg gaaatc ctc aagctcc aggtac gctc at tgg tgc a gccgt gatga agacg aatta ccggt caagg gcat t tcca atct gaataa catgg caatg ttcag cgttt ccggc ccggg gat
gaaaggat ggt tggca tggc ggcgc gctct ttgca gcgat gtcac ggcgc gctat ttcgg tgggt gctgat tacgc aatca tctt ccgaa tacag tatc agtttc tgcgt tccgc aaagc gactg tgt
gcgagc tgaac ggcga atgc ag gaa ga gttct accty gaact gaaag aaggc ttact gggagc cgttggc ggt gac gga accgct tggcc attatctc ggtgg taggt gatgg tatgc gcacc ttacg tgg
gatctcggc gaaattct ttgcc ggcct gggcc ggcgc aatat caaca ttgtc gccat tgcctc aggga tcttc tgaac gctca atctctgtc gtggc caata acgat gatgc gaccactggc gtcg cgt
tactcat cagat gctgt tcaat accga tcagg ttatc gaagt gttt g gatt ggctc ggtg gcttggc ggtgcgc tgc tggagc aactg aagcgc t cagc aaagc tgggt gaaga aaaa catat cga
cttacgt gctgc ggtg ttgct aactc gaagg cactg ctcac caatg tacat ggct taatc tggaa aac tggc agga aagaa c tggc gca aagc caa agagc cgtt aatct ccggc gctta attcg cct
cgtgaaagaata tcaatc tgc tga accc ggtca ttgtt gactg tactccagc caggc agtgg cggat caata tgcgc acttc ctgc gcgaa ggtttccag t gtt acgcc gaaca aaaag gccaac
ctcgtc gatgg a t t act accat cagtt gctt atg cggc gga aaaaat cgcggc gta aattcc tctat gacac caacg ttggg gctggatta ccggt tatc gagaac ctgca aaaa tgcct aatgc tgg
tgatgaa ttgat gaagt tctcc ggcac tcttt caggt tgcct tctt atatc ttcgg caagt tagac gaagg catga gtttc tccg aggcg accac actggc ggcg gaaat gggtt atacc gaacc gga
cccgagatg atcttct ggtatgga t gttggc gct aagct attga tctc gctcg tgaac cggga cgtga actgg agct tgcgg atatt gaaat tgaac ctgtg ctgcc cgcag agttt aacgc cga
gggtgat gtcgc cgttt ttagc gca tctgt cacag ctgcac gac tcttt gccgc gctgt tggcga aggc ccgtgat gaa gaaa agtt ttgcg ctatg ttggc aatat tgatg aagat ggcgt ctg
ccgctg aagat tgcga agtg gatgg taatgatcc gctgtt caaag tga aaaaatgg cga aacgcc ctggc cttct atagc cact attat cagcc gctgc cgtt ggtact gcgc g gatat ggtgc ggg
caatgac gttac agc tgcgg t gctt t gctgatct gctacg tacc tctca tggaa gttag gagtc tgaca tgggt aaagt ttat gccc ggttccag t gcaa tatga gctc ggggt tgatg tgc
tcggggc ggcgg tgaca cctgt tgatg tgc a ttgct cggag atgtagtcac ggtt gaggc g cagag acat t cag tctca caac ctggc agcgt ttgcc gataa gctgc cgtc agagcc acggg aaa
atctcgt ttatc agtgc tggga gctt tttgc caggag cttgg caa gcaaat tccag tggc atgac tctgg aaaa gaatat gccgat ccggt tccgg gctta ggtc cagc cctgt t cag t ggtc ggcg
cgtgat ggcga tgaat gaac actg cggc caagc cgt taatgac actc gttt gctggc tttg atggg c g agt tggaa gggc g tate tccggc agca ttcac tacgaca acg tggca ccggt ttttc ttg
gtggtat gca g t g at g att g a g a a a c g a c a t c a t c a g t c a g c a a g t g c c a g g g t t g a t g a g t g c t g t g g g t g c t g g c g t a t c c g g g a t t a a a g t c t g a c g g c a g a a c c a g g g c t a t t t t a c
cggcgc agtate cgcgc cagga ttgca ttgcccac ggcgc ac tctg gcagg cttca ttcac gctg ctatt cccgt cagcc tgagct tgc cgcga agct gatgaa agatg ttatc gctga accct acc
gtgaa cggttac tgcaggc cttccggc aggcgc ggcga ggcgg ttgccc gaaat cggcgc ggtga cgcagc ggtat tctcc ggcct cggc ccgac ttgt tgcct ctgtg tgaca agccg gatc cgcgc agc
gcgttgc c gact ggttgggtaa aaact acc tgc aaaa t cagga aggt tttgt tcaata tttgc cggct ggata cggcggg cgc acgag tact ggaaa actaa atgaa actct acaat ctgaa agatc aca
atgagc aggtc agcttt ggcga agcgc taacc cagggt tgggcaaaa aatca ggggc t gttt tccc gcacg acc tgcgg aatc agcct gact gaaatt gatga gatgc tgaag ctgga tttgt tca
ccgcag tgcga agatc ctctc ggcgt tttt ggtga t gaaa tccc g cagga aatcc tggaa gagc gctac cgcgc gctt tgcct tccc ggcct cggtc gcaa t gttgaaagc gatgt cgtt gct
tgg aatt gttcc acggg ccaac cctggc attt aaaa tttcggc ggt cgtt tatgg caca aat act gccc atatt gccc gctaa agcc agtga actt ctgac cgcga catcc ggtgat tctgag
cggc agt ggtc atgct t t t c a c g g c g a t t t c g a t g c
tcgacggc gatt t c g a t g c
actact t g a a g c t g t t g c
gttttat t g c t g c g a c c a a
gtgtgga agagt t g t t c g
ctgc cgt agctt atcgt g c g c t g c g t g a c c a g t t g a a t c c a g g c g a a t a t g g c t t g t c t c t c g g c a c c g c a t c c g g c g a a a t t t a a a g a g a g c g t g g a a g c g a t t c t c g g t g a a a c g t t g g a t c t g c
caaaag agctggc agaac g t g c t g a t t a c c c t t g c t t c g c a t a a a c c t g c c c g c a t t t t g c t g c g t t g c g t a a a t t g a t g a t a c a t c a g t a a c a t c t a t t a t t a t c t c a a t c a g g c c g g g t t
tgctttt atgca gccc ggttt tttat gaaga aaaa tggag aaaaa c g a c a g g g a a a a g g a g a a a t t c t c a a t a a t g c g g t a a a t t a g a g a t t a g g a t t g c g g a g a a t a c a a c t g c c g t t c a t

Геном бактерии: несколько миллионов нуклеотидов

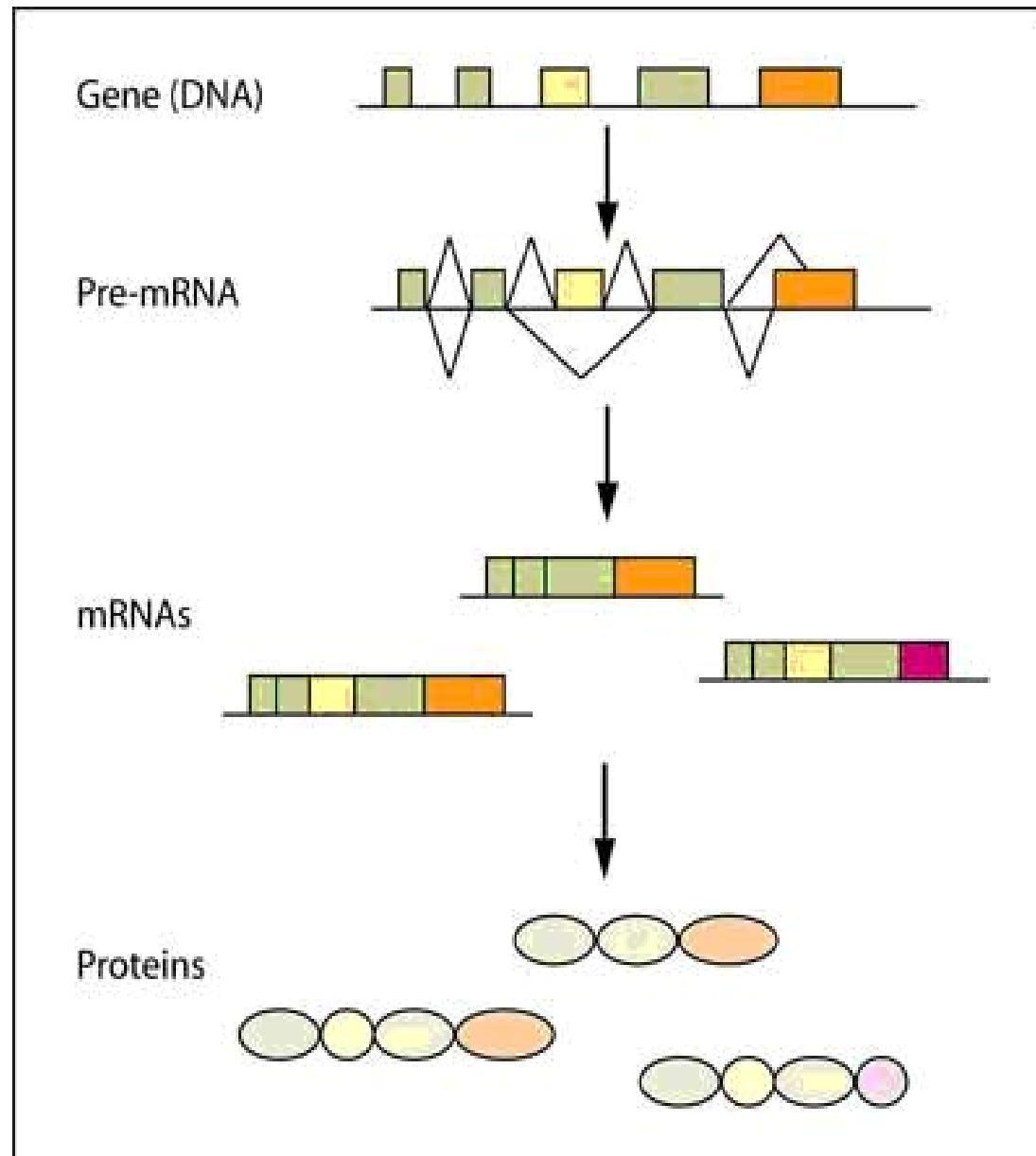
От 600 до 9 тысяч генов (примерно 90% генома кодирует белки)

0.0001% генома человека

cgtgcac ttctgaaggacttcaggtac cggcgtgccc cgcgtcctac tgtcc gcctgctcgc gtcctgggtgccgcc tctgagtagggcgggagagg
CAGCCAA GCGGGAGCTGATGGCTGCGC CGAGG GCGGG GCGGG GTGCA GGCTG GAGCC TTCGG GCATG GCGGG CTTTGGGGG CATT CGCTG GGGGAGGAG
ACCCCGTTTGAC CCCTGACCTC CGGGC CCTGC TGACGTCAGGA AACTTCTGAC CCCC GGGCC GAGTG ACTTATGGGAC CCCCAGTC TCTGGGCCCGTTG
TCTGTTGGGTC ACTGAACCCC GAGCATGCCTGACGTCTGGGACCCC GGGTC CCCGGGCACA ACTGACTGCGGTGAC CCCAGATAC CAGGACCCGGGAGG
CCTCAGAGAACTCTGGAACCCGTTCCG GCGCGTGGCTGGCGGTGGCGCTGGGCGCTGGGGGGCAGT GCTGTTGTTGTTGTGGGC GGGGGTCGGGTCC
TCCGGCC GTCCTCGCCG CCGTC CCTAG CCCGC CGCCC GCTTC TCCCC GGAGTCAGTACA AACTTCATC GCAGATGTGGTGGAGAAGACAGCA CCTGC CGTG
GTCTATATCGAGATCCTGGACC Ggtaa tgggtggggtagacc gggag gcaact gaagc cacag gctgg agggc gggcgggtaggagg ggtcagagcc tcc
cttatctgtgctttccc tccatttcagGCACC CTTTC TTGGGCCGCGAGGTC CCTAT CTCGAACGGC TCAGGATTCGTGGTGGCTGCCGATGGGCTCATT
GTCACCAACGCC CATGTGGTGGCTGATCGGCGCAGAGTCCGTGTGAGACTGCTAAGC GGGCA CACGTATGAGGCCGTGGTCA CAGCTGTGGATCCC GTGG
CAGACATCGCAA CCGCTGAGGATTCAGACTAAG gtgggggctg gggtagggcaggctctgggtgagctgcttattgctcgcac tcttcagatgacaggtct
cttttac ccatttccc ttaggagcctctccc cacgc tgcctctgggacgctcagctgatgtccggc aaggggagtttgtgtgtgc catgggaagtccct
ttgcaactgcagaacacgatcac atccggcattgttagctctgtcagcgtcc agccagagac ctgggactcc cccaa accaa tgtggaatacattc aac
tgatgcagctattgatgtgctcctgtaggagagaaatgac aatgatgggggaggggggagggc tgtgtggtac aagcaaccaactgatatatggtg
atgagcc tatatagagc ttaggctgcaaaaatgtggc cacttattca tgggc tgaga aagaagagaa tttggagaaagtacc taca tccgtggtatgcccc
cagacttagaatccccagatctctttc atgttttctc cttgtcctac agTTTGGAAA CTCTG GAGGT CCCCTGGTTAACCTG gtgagtgagacatcctc
cttccaagaatc cctgc cccaggtcagtggtgggaagggtaggtttcc cctaa ttcaaggatgtttggtcaagtttctgagcagttc tttgtggctatct
ctcaata tccaaccagatctcc ccaac acttgctggtacttt tgttc gggtgcccc atccc ctactatttgtttaggctagggaactgggggctgtatc
cctgcagGATGGGGAGGTGATTGGAGTGAACA CCATGAAGGTCACAGCTGGAATCTC CTTTGCCATC CCTTC TGATC GTCTTCGAGAGTTTCTGCATCGT
GGGGAAAAGAAG Agtga gcctgcctta tggggaaacgggttc cttaatgtggtgga aataggggaaagggca ttcagtgggacttc ctggagggtggtct
actgggagaagagggcaggyaaggaagatgt agctgggtggggctc atttgcctctgtc acagATTCCTCCTCC GGAATCAGTGGGTC CCAGC GGGC
CTACATTGGGGT GATGATGCTGACCCTGAGTC CCAGgtatgagctttagggacagtgacatg taatgtgacc agtgtaatcagaggggggc acctc tatt
gagctttgttctcatttctgtc tttatctaagatgaaactgtgtcacacttga aataa tcaca agagc tgtctccctcatca tcttgactttctta tccc
actccac tttgtacacc tgtcaccagattgat ttcactctgt tactgctttgatttc aagcc ttcaatccat taactggca ttttaagggc cattttcca
tctgtctgtaaa tcaac tttct tgt
tcttgtttatata tgggtaatat Геном человека: 3 000 000 000 нуклеотидов gct
attgattaacactggttgta cat
tgtattttgtagtacctagccc ttt
ggctaataagggtgatctgtgta aag
tcactcctgggctcccctgcacacgggtgagggagagggctgcagtgatgatggggatgggca aggtgtgcatgtgtc cttgaaactaggcctttgtac tcc
tcctttc tctctgtccattttctctatagGGCTGGTCTGCGGCCTGGTGATGTGATTTTGGCCATTGGGGAGCAGATGGTACAAAATGCTGAAGATGTT
TATGAAGCTGTT CGAAC CCAATCCCAGTTGGC AGTGCAGATC CGGCGGGGAC GAGAAACACTGACCTTATATGTGAC CCCTGAGGT CACAGAATGAATAG
ATCACCAAGAGTATGAGGCTCTGCTCTGATTTCTC CTTC CTTTC TGGCTGAGGTTCTGAGGGCA CCGAGACAGAGGGTTAAATGAACCAGTGGGGC
AGGTCCCTCCAAACCACCAGCACTGACTCCTGGGCTCTGAAgAATCACAGAAA CACTTTTTATATAAAATAAAATTATACCTAGCaacatatatagtaaa

Геном человека: 3 000 000 000 нуклеотидов

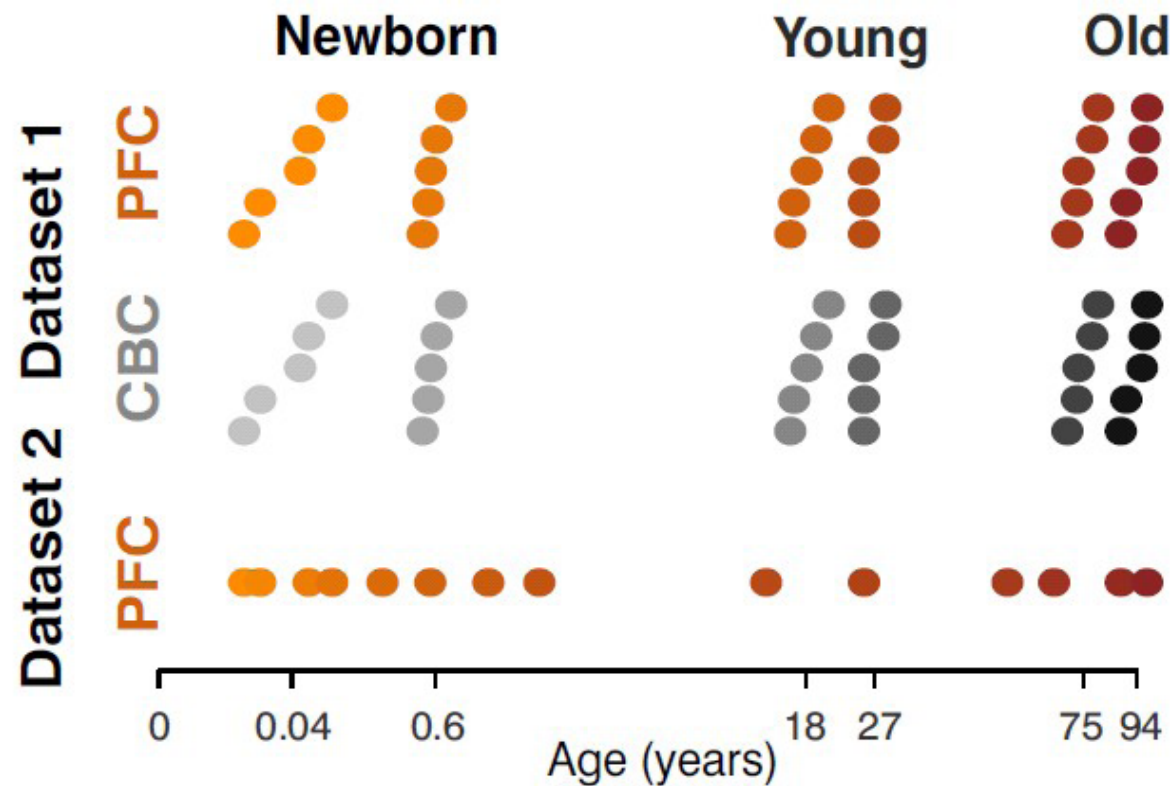
Примерно 20 тысяч генов, < 5% генома кодирует белки



Экзон-интронная структура и сплайсинг

- картирование генов, кодирующих белки
- альтернативный сплайсинг
- эволюция альтернативного сплайсинга
 - 2003: а.с. эволюционирует быстрее, чем конститутивный
 - 2008: положительный отбор в альт. экзонах
- количественный анализ

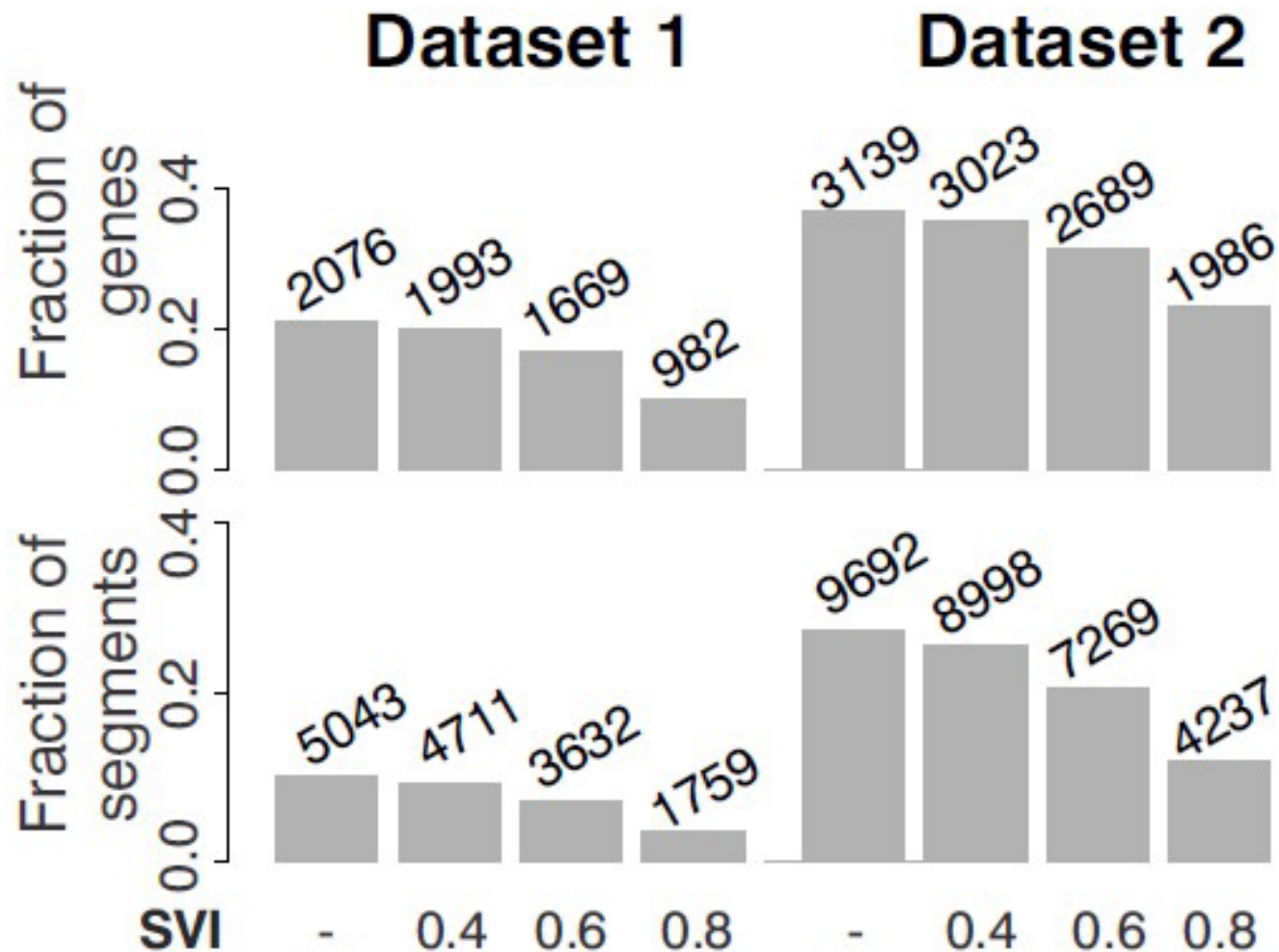
Возрастные изменения сплайсинга в мозге человека



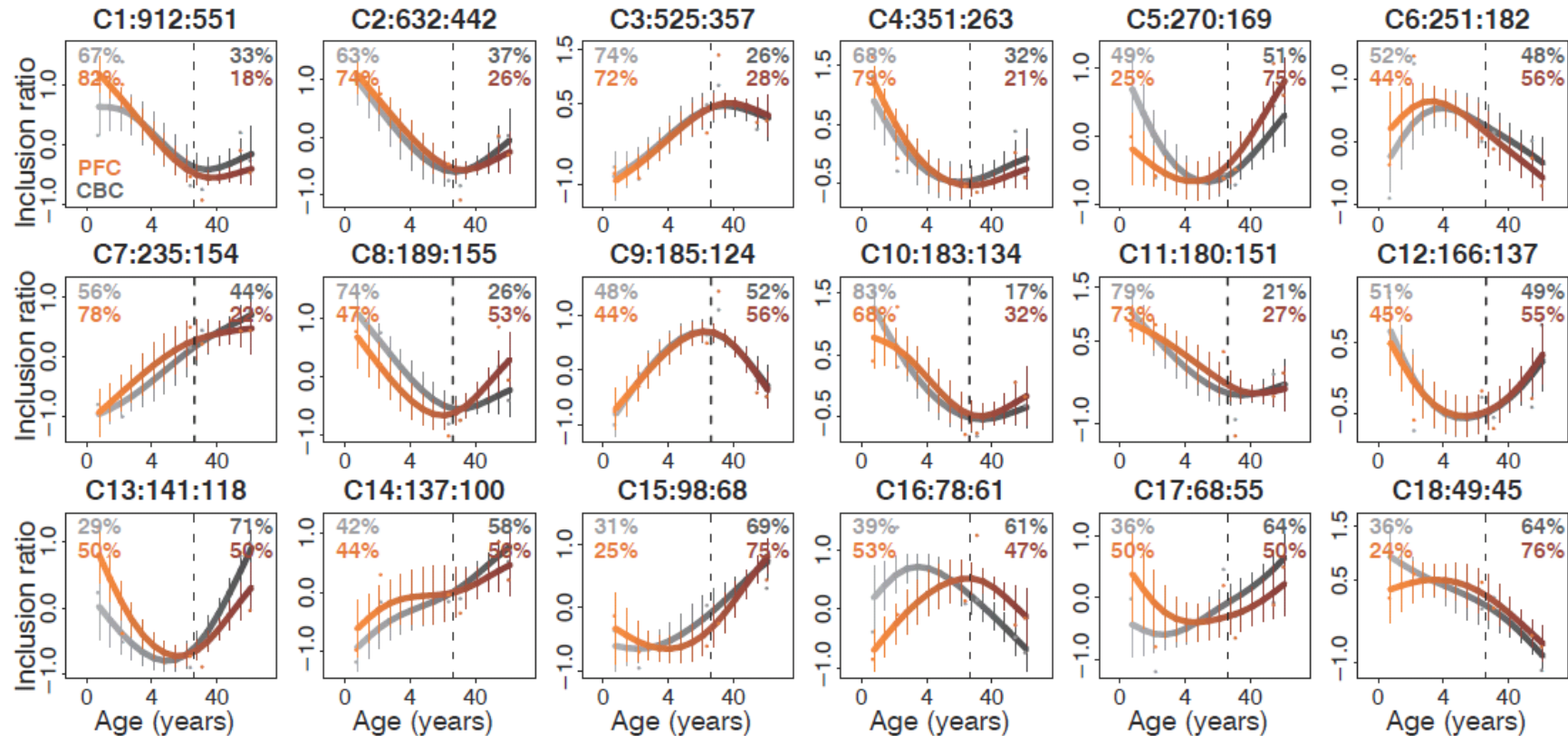
по 6 смешанных образцов (по 5 индивидуальных в каждом)

13 индивидуальных образцов

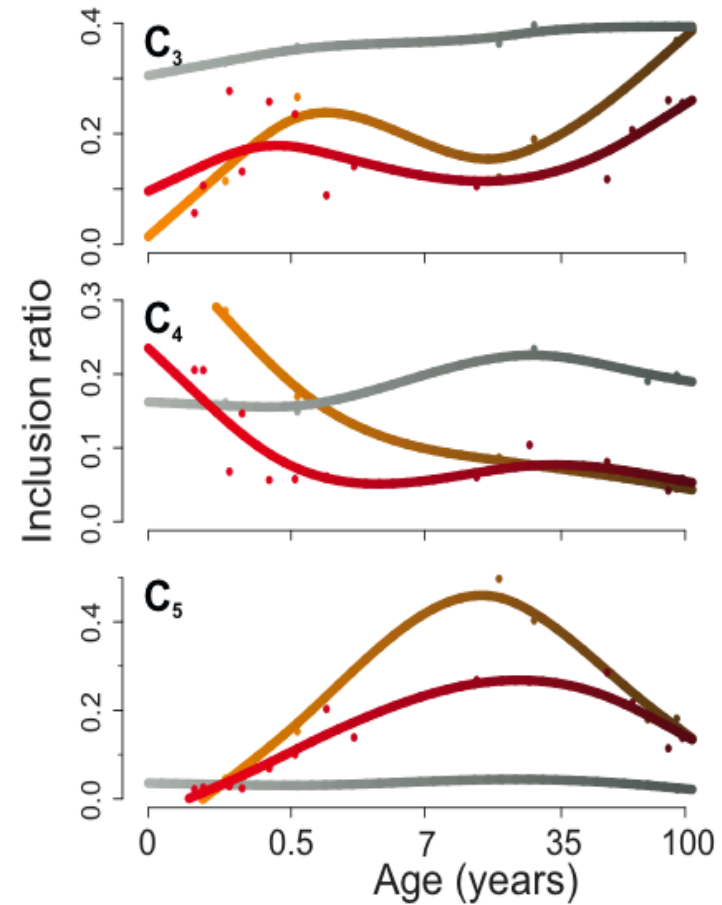
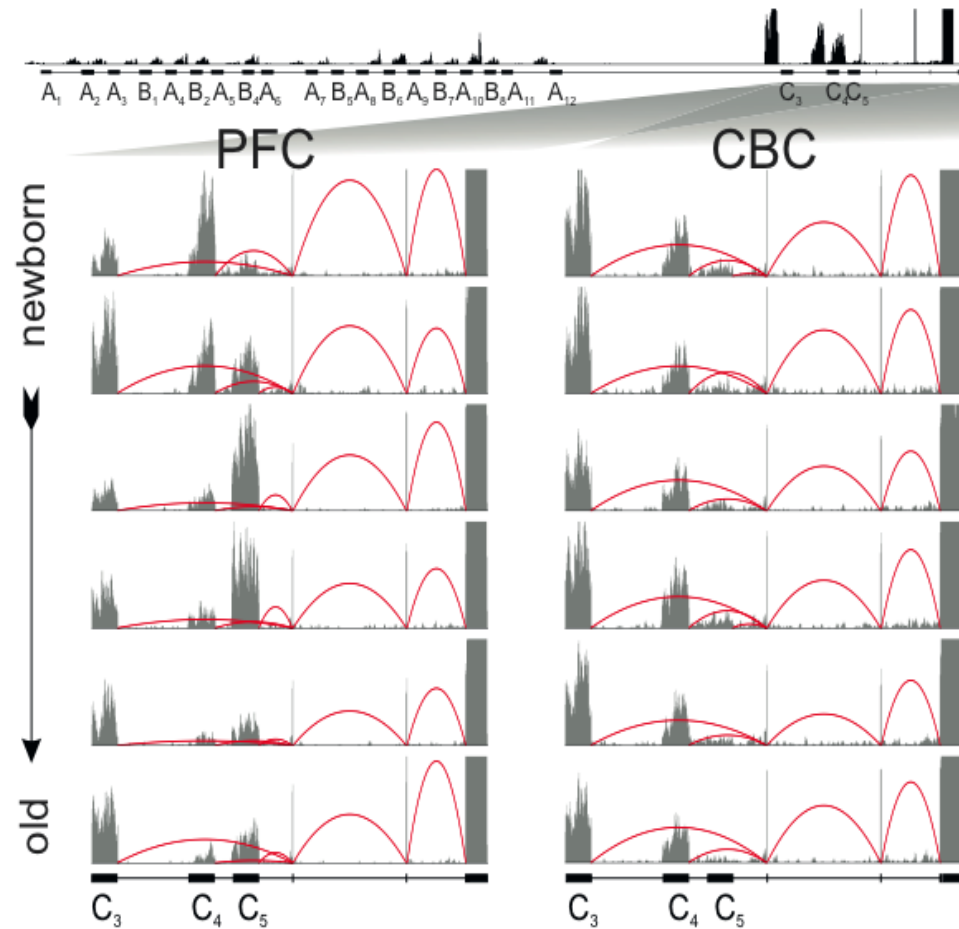
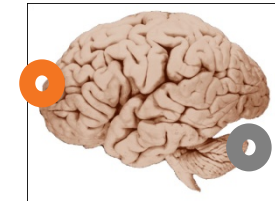
до 40% генов меняют сплайсинг с возрастом



Паттерны изменения IR с возрастом

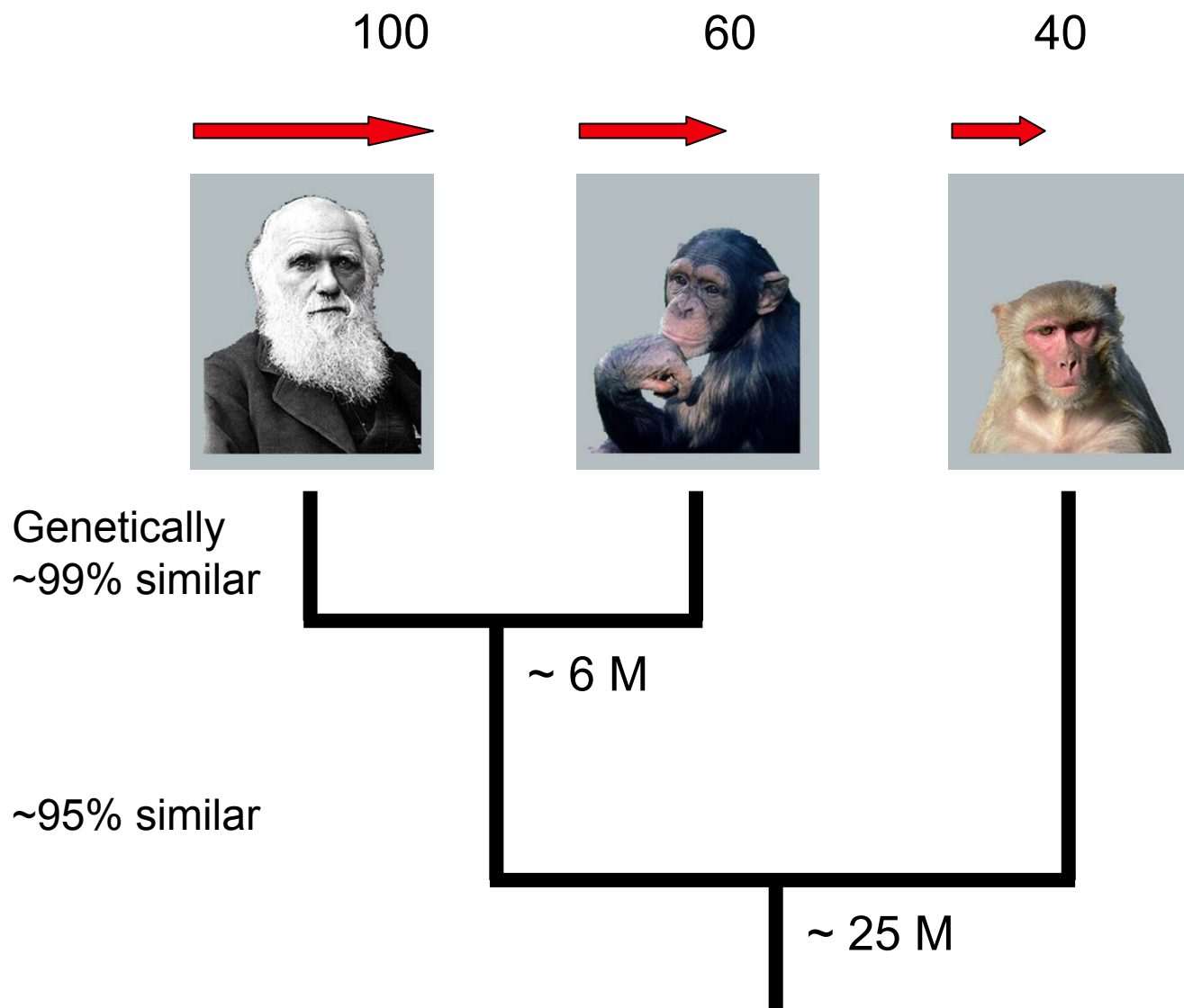


20% генов, меняющих сплайсинг с возрастом, имеют различные паттерны изменений в разных областях мозга

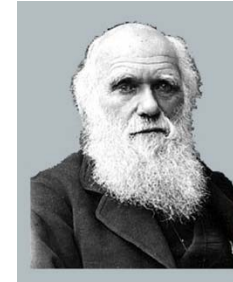
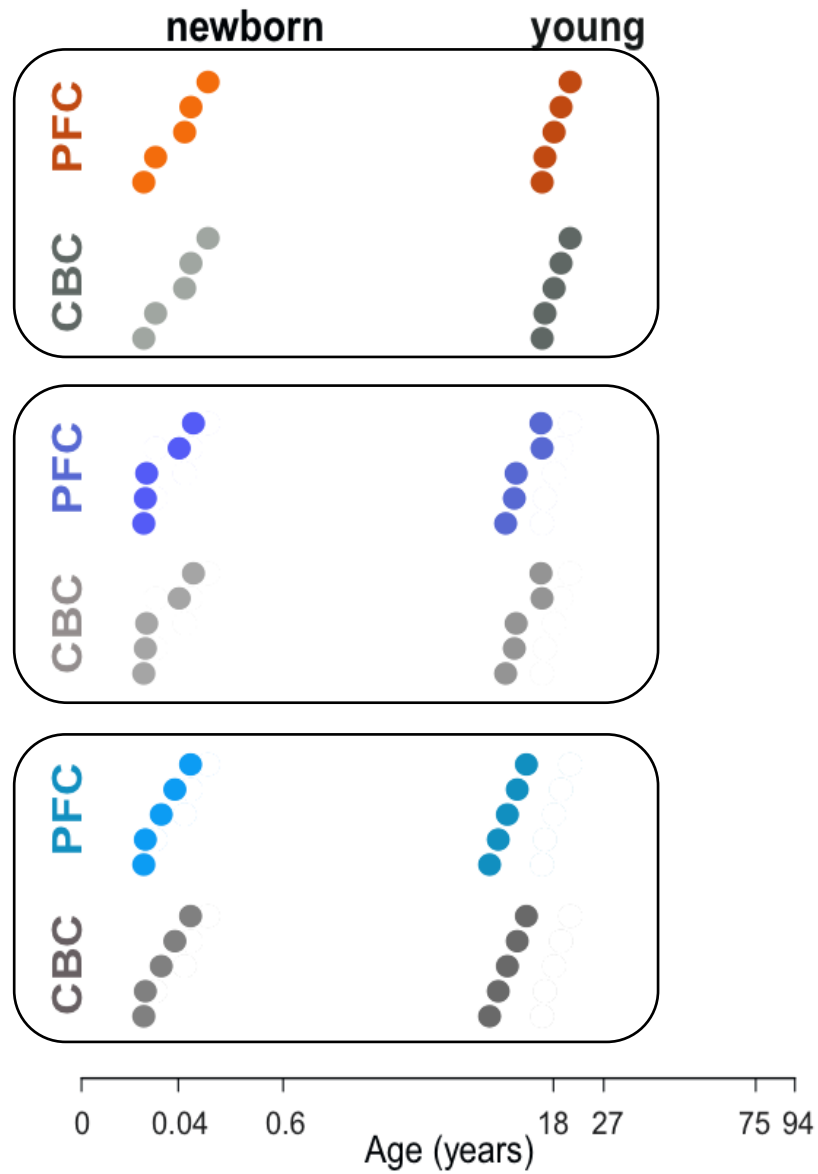


Protocadherin gamma cluster

Человек, шимпанзе, макака-резус



Выборка 1



Выборки 1 и 2

Dataset 1

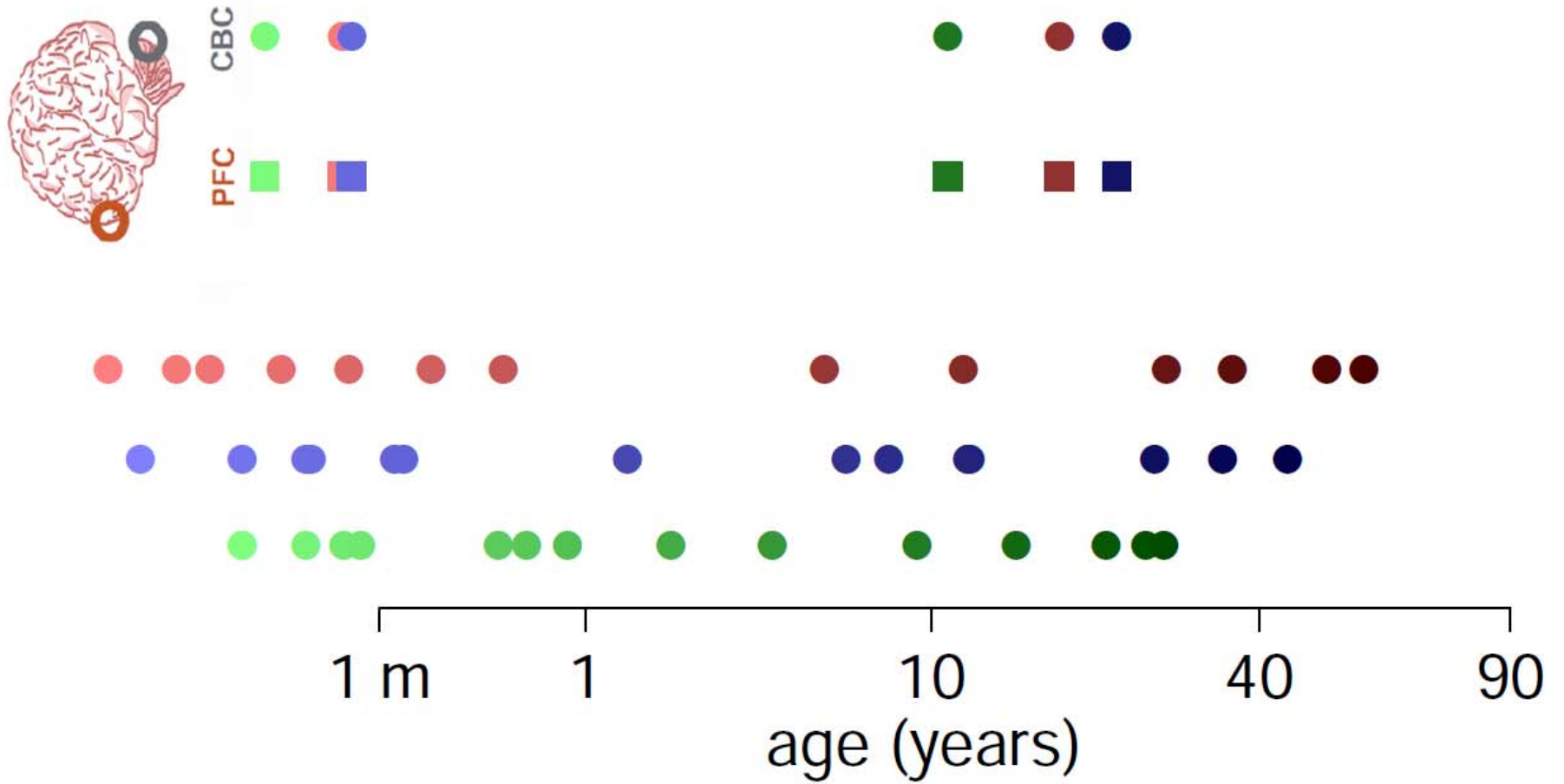


CBC
PFC

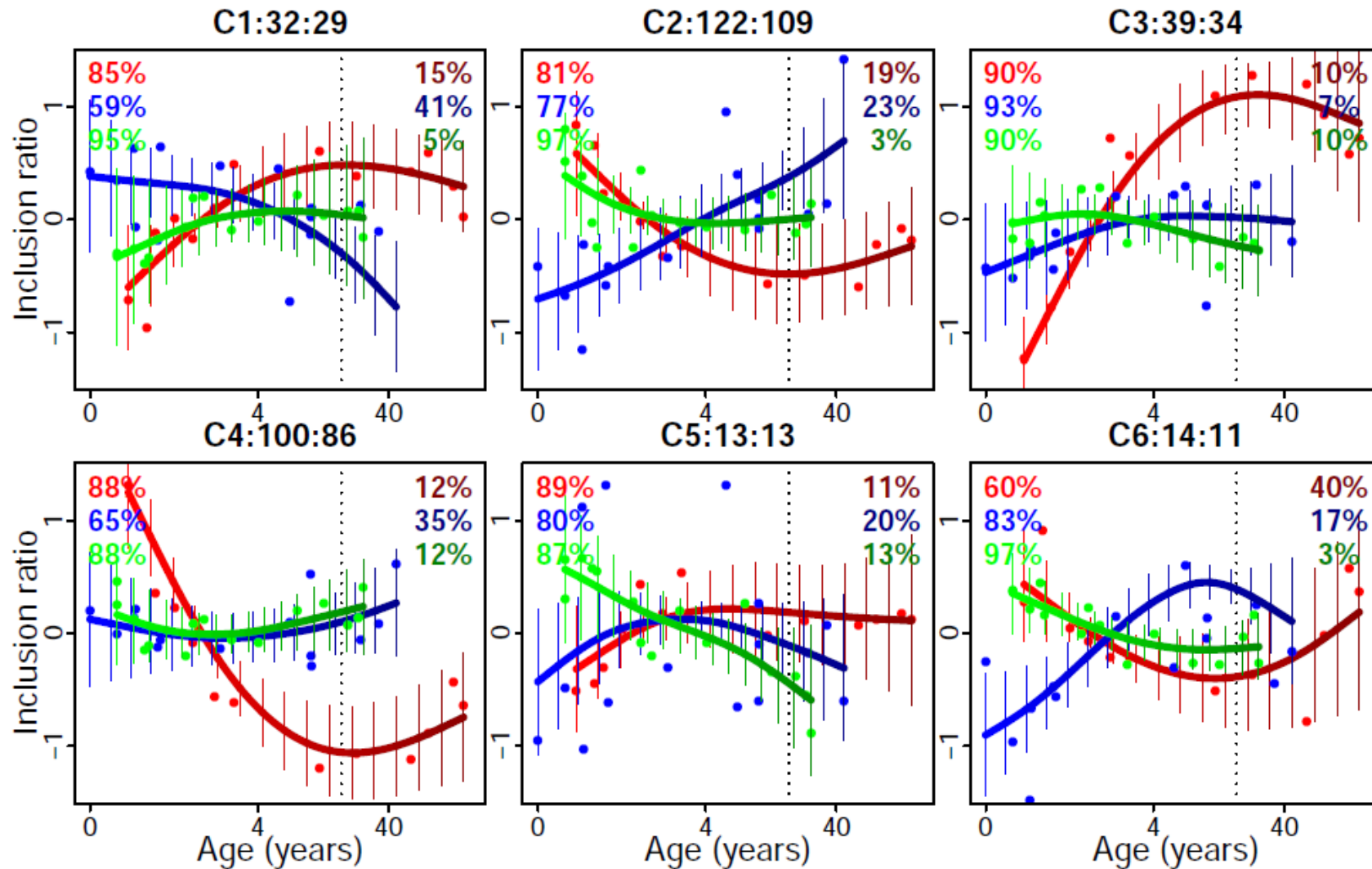
Dataset 2

1 m 1 10 40 90

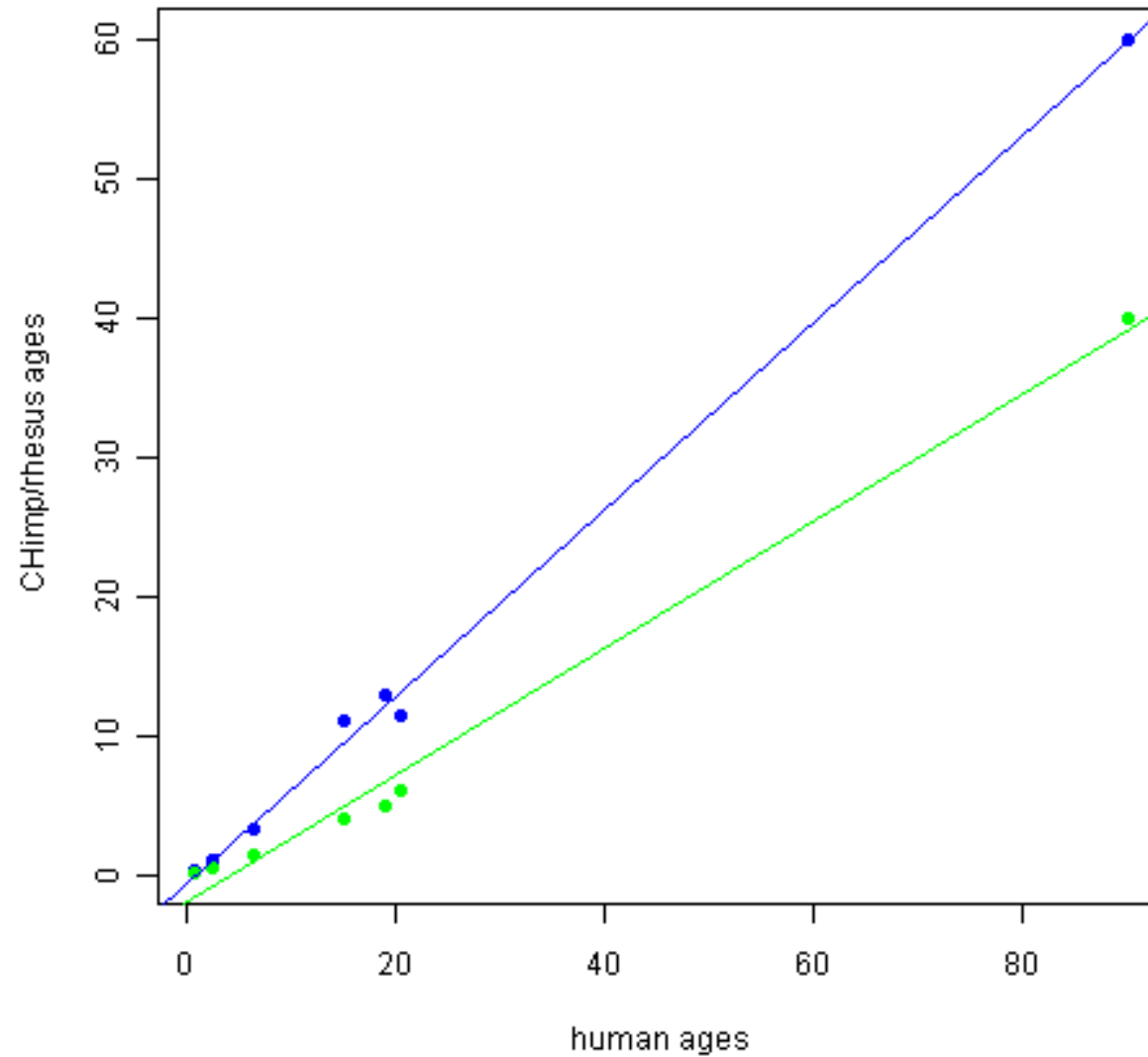
age (years)



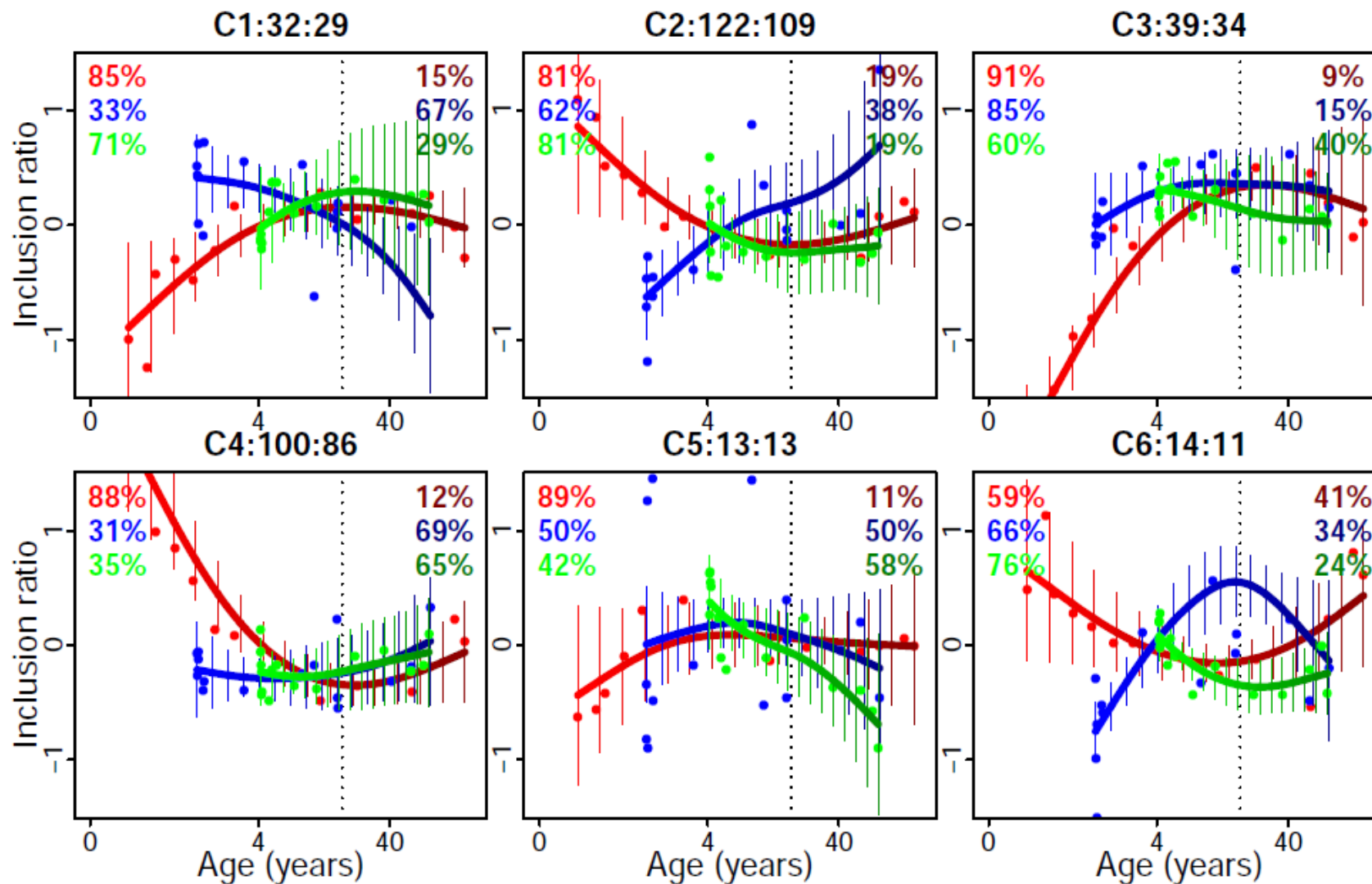
Слабое согласование профилей включения



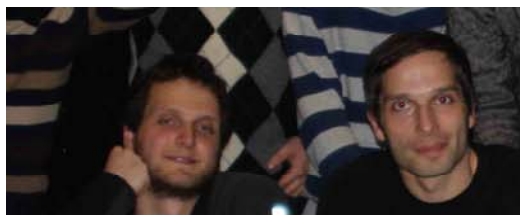
Поправка на физиологический возраст



Теперь лучше



- Павел Мазин
- Филипп Хайтович (Шанхай)

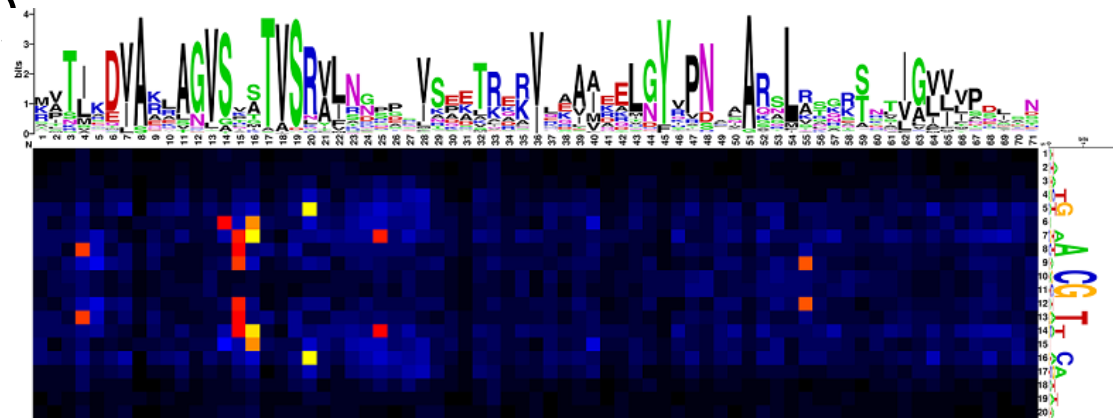


Другие проекты

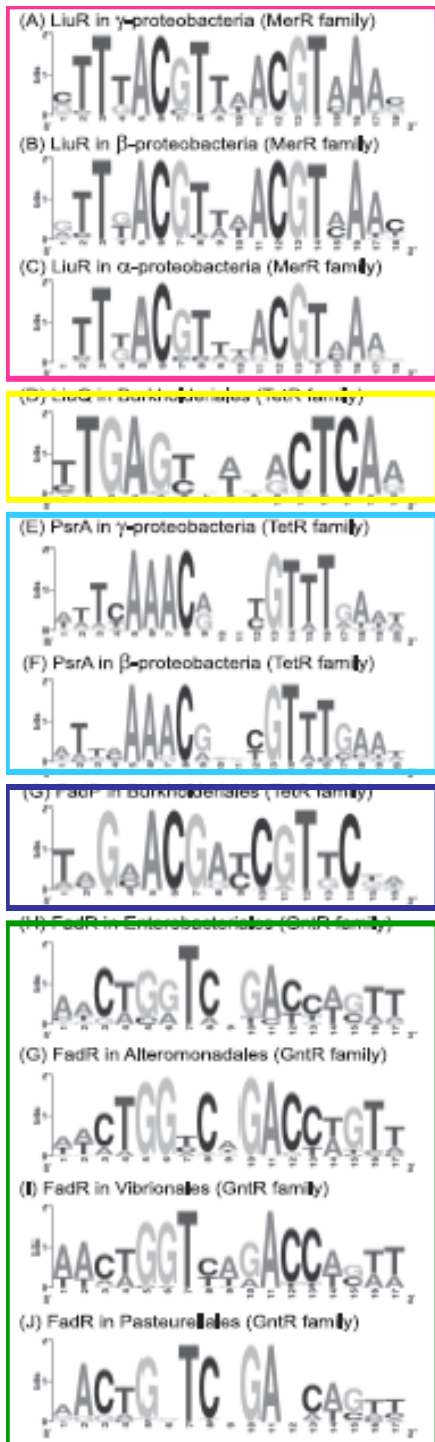
- Эволюция сплайсинга у дрозофил (Сергей Нуждин, USC + Екатерина Ермакова)
- РНК-связывающие белки (Альфред Биндерейф, Университет Гиссена + Екатерина Храмеева)
- Структура и состояние хроматина
- Старты транскрипции у бактерий (Гиссен + Институт Пастера, Елена Яловая)

Регуляторные системы

- Разработка алгоритмов (УНЦБ, Лаб. 6)
- Интернет-серверы (Инна Дубчак, LANL, Беркли)
- Сравнительная геномика и эволюция регуляторных систем
 - по функциональным системам (УНЦБ, Лаб. 6)
 - по таксонам (Д.А.Родионов, сектор 6)
 - по семействам факторов транскрипции (А.Б.Рахманинова, Юрий Коростелев)



- Моделирование регуляторных систем (В.А.Любецкий, В.И.Рубанов)



MerR

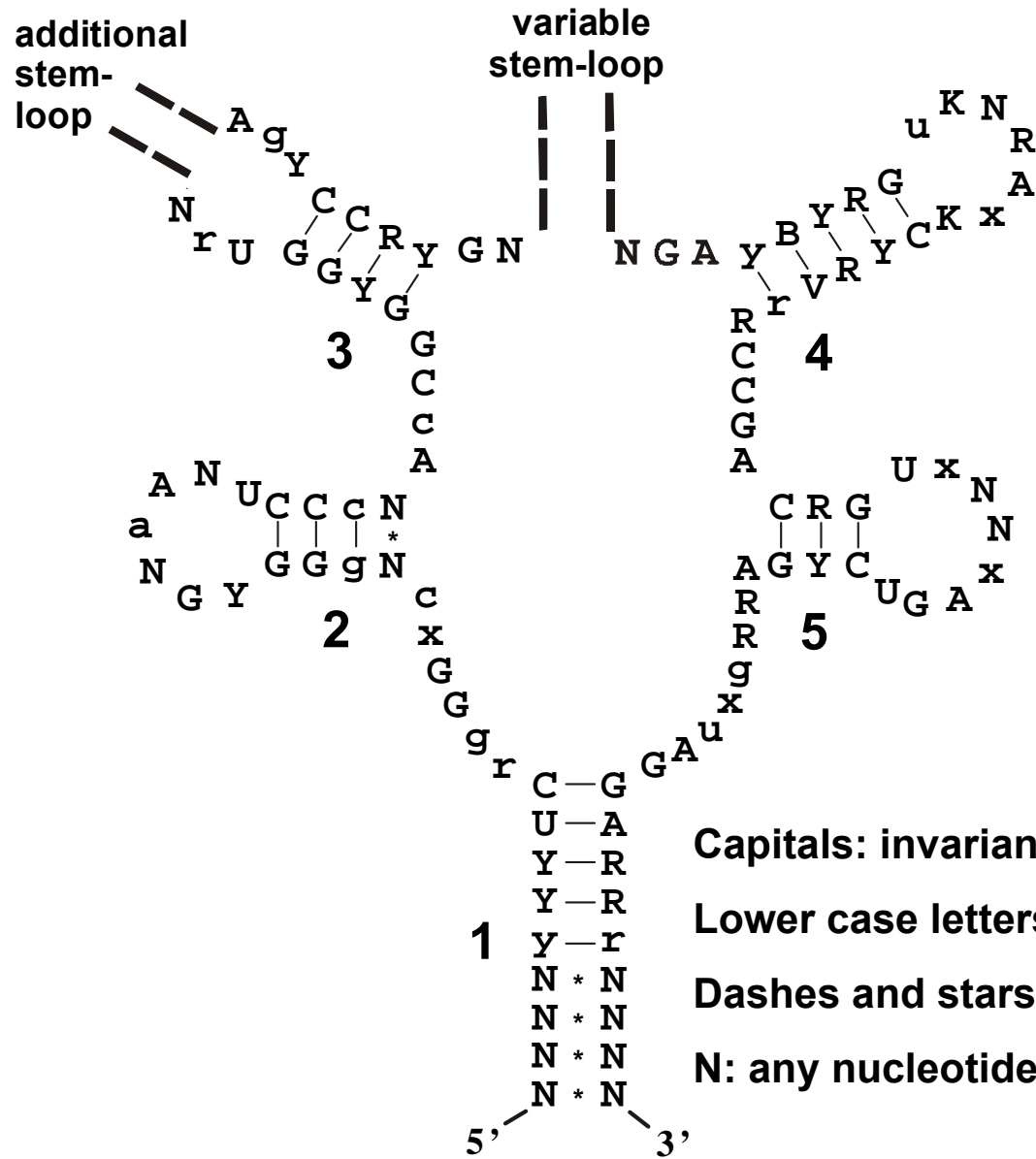
TetR

GntR

Genome of γ -, or β -proteobacteria	ILV degradation									ETF		FA degradation								
	<i>liuA</i>	<i>liuBCDE</i>	<i>liuFG</i>	<i>aacS</i>	<i>ivdA</i>	<i>ivdC</i>	<i>ivdG</i>	<i>ivdIBDEF</i>	<i>bkd</i>	<i>ldh</i>	<i>etfBA</i>	<i>etfD</i>	<i>fadBA</i>	<i>fadH</i>	<i>fadD</i>	<i>fadJ</i>	<i>fadE</i>	<i>fadL</i>	<i>acdAB</i>	<i>acdH</i>
Enterobacteriales (5 species)																				
Pasteurellales (7 species)																				
<i>Shewanella</i> spp.																				
<i>Idiomarina loihiensis</i>																				
<i>Colwellia psychrerythraea</i>																				
<i>Pseudoalteromonas haloplanktis</i>																				
<i>Pseudoalteromonas atlantica</i>																				
<i>Saccharophagus degradans</i>																				
<i>Vibrio cholerae</i>																				
<i>Vibrio fischeri</i>																				
<i>Vibrio parahaemolyticus</i>																				
<i>Vibrio vulnificus</i>																				
<i>Photobacterium profundum</i>																				
<i>Pseudomonas aeruginosa</i>																				
<i>Pseudomonas putida</i>																				
<i>Pseudomonas fluorescens</i>																				
<i>Pseudomonas syringae</i>																				
<i>Pseudomonas entomophila</i>																				
Xanthomonadales (3 species)																				
<i>Hahella chejuensis</i>																				
<i>Alcanivorax borkumensis</i>																				
<i>Chromohalobacter salexigens</i>																				
<i>Chromobacterium violaceum</i>																				
<i>Dechloromonas aromatica</i>																				
<i>Azoarcus</i> sp.																				
<i>Bordetella</i> (3 species)																				
<i>Ralstonia solanacearum</i>																				
<i>Ralstonia eutropha</i>																				
<i>Ralstonia metallidurans</i>																				
<i>Burkholderia xenovorans</i>																				
<i>Burkholderia</i> (4 species)																				
<i>Methylobium petroleiphilum</i>																				
<i>Polaromonas</i> sp.																				
<i>Rhodoferrax ferrireducens</i>																				

■ LiuR
 ■ LiuQ
 ■ FadR
 ■ PsrA
 ■ FadP
 ■ unknown regulation

RFN - первый РНК-переключатель



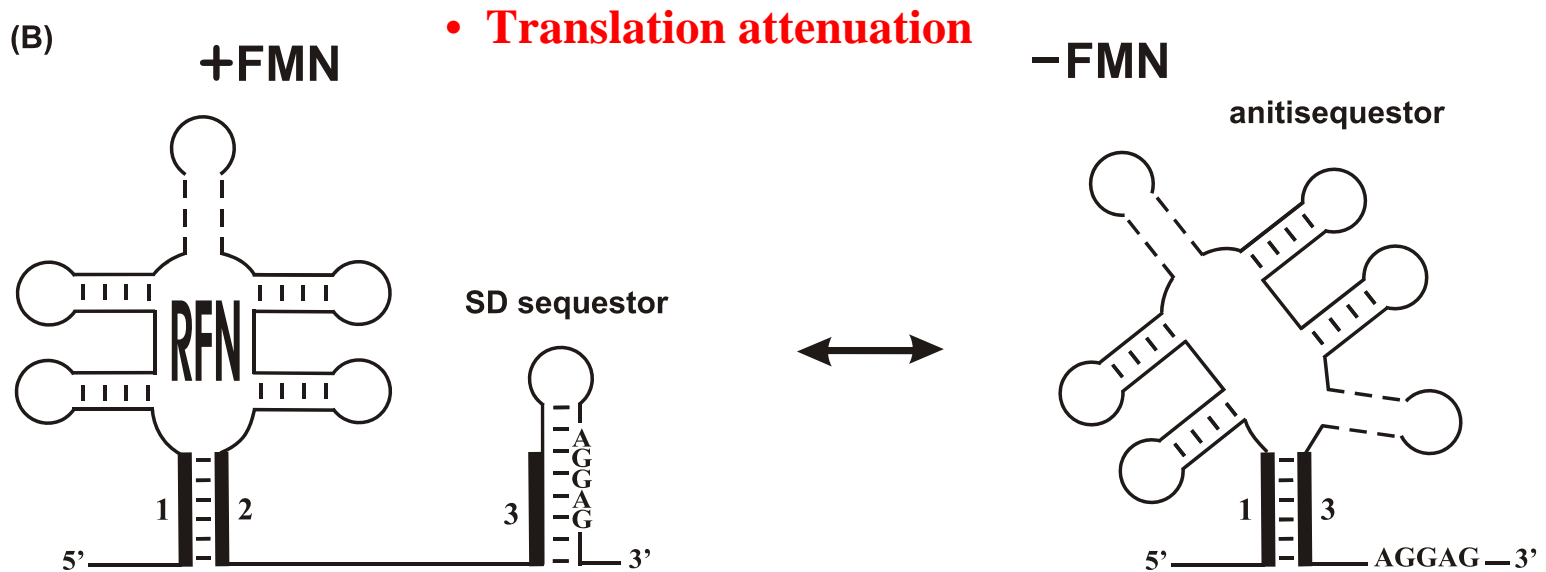
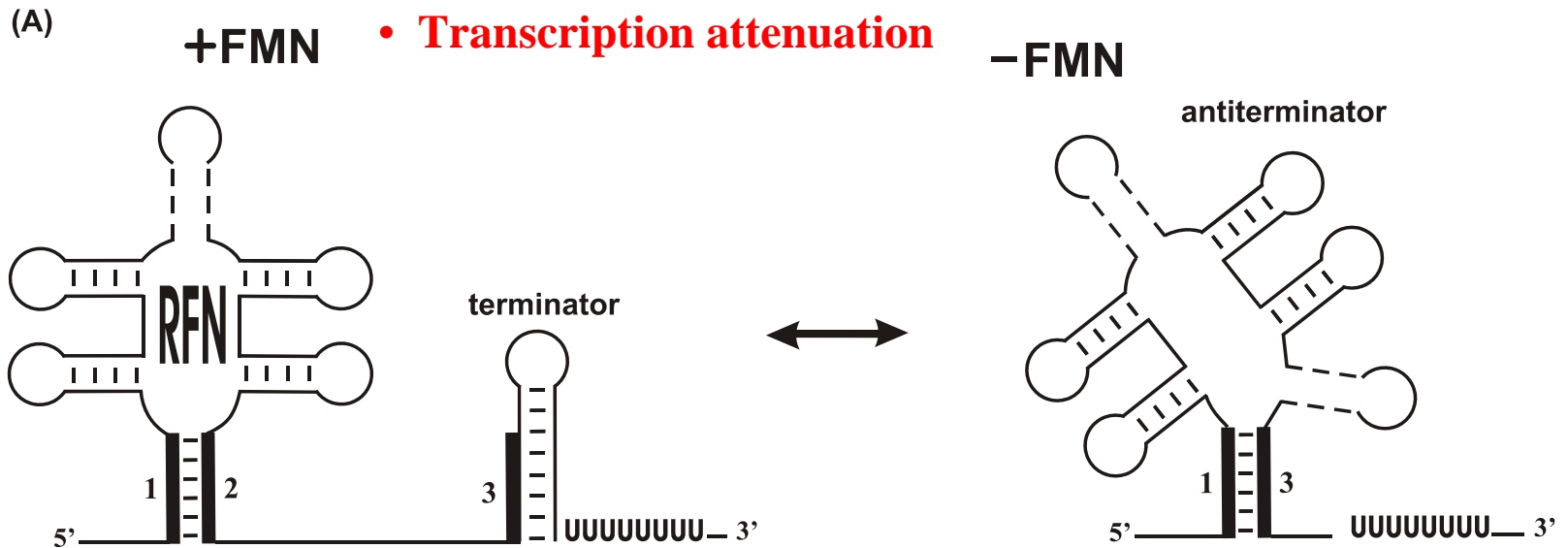
Capitals: invariant (absolutely conserved) positions.

Lower case letters: strongly conserved positions.

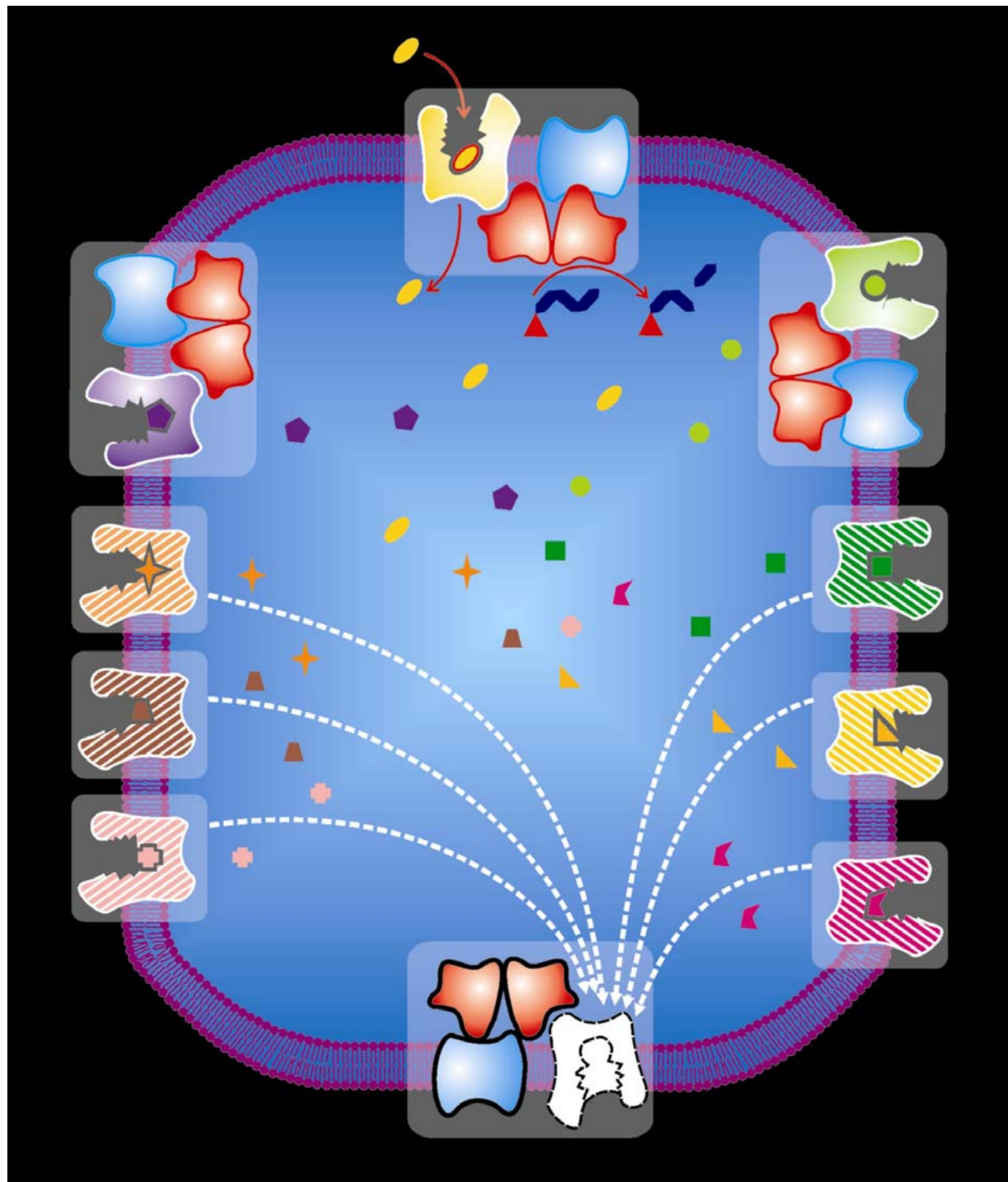
Dashes and stars: obligatory and facultative base pairs

N: any nucleotide. X: any nucleotide or deletion

RFN: механизм регуляции



Универ-
сальный
«энергети-
ческий
комплекс» +
компоненты,
определя-
ющие
специфич-
ность



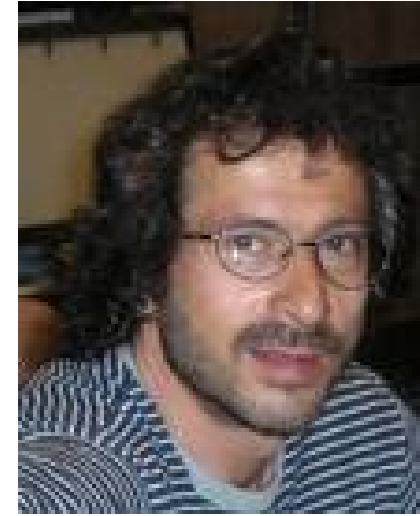
Дмитрий
Родионов



Thomas
Eitingger

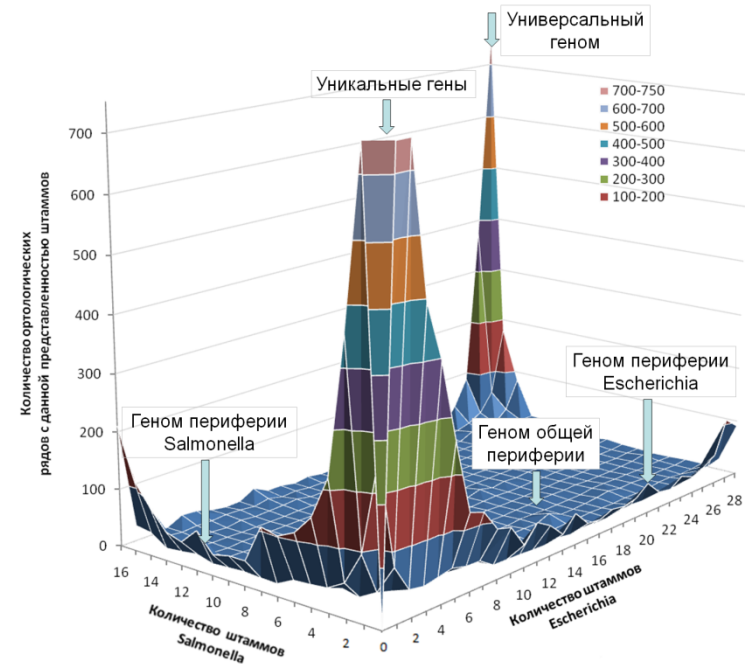
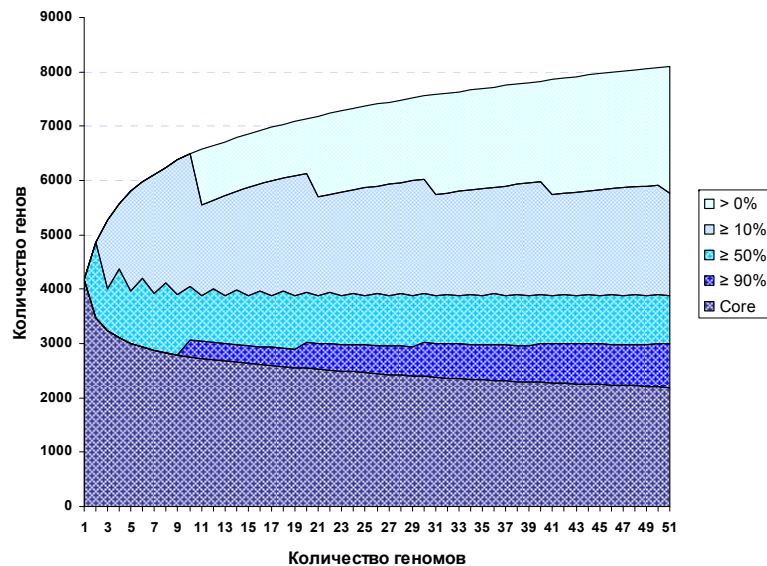


Андрей
Остерман



Эволюция бактерий

- Согласование деревьев, идентификация горизонтальных переносов (Лаб. 6)
- Микроэволюция штаммов / видов
 - на уровне наборов генов
 - геномных перестроек
 - межштаммовой рекомбинации



Лаборатория 6:

- 1) механизмы **регуляции инициации трансляции** у бактерий и в пластомах: изолейцил-тРНК синтетаза посредством Т-боксов, 2-изопропилмалатсинтаза посредством LEU-элемента у многих актиномицетов, включая возбудителей туберкулеза и дифтерии;
- 2) механизмы **регуляции транскрипции** генов метаболизма аминокислот у бактерий, метаболизма азота у цианобактерий и в хлоропластах водорослей;
- 3) механизм **процессинга мРНК** бруцелл, нарушение которого предотвращает незавершенный фагоцитоз; **регуляции транскрипции и трансляции** в пластидах споровиков – возбудителей протозойных инфекций, включая токсоплазмоз, эймериоз и малярию; предсказано воздействие антибиотиков на споровиков.
- 4) механизм **ответа изолированных хлоропластов на тепловой и холодовой шоки**; и количественные предсказания результатов **воздействия точечных мутаций и эпигенетических нарушений** в митохондриях животных на их фенотип. Объяснить влияние на фенотип генетических изменений в митохондриальной ДНК человека (**MELAS-болезни**) и нарушений метилирования ДНК при недостатке гормонов;
- 5) механизмы **задержки транскрипции до завершения сплайсинга**;
- 6) совместная **эволюция промоторов и сигма-субъединиц РНК-полимераз** пластид растений, водорослей и споровиков; **сайтов и факторов** у бактерий.





template

- text