# A Needle in a Haystack
# or
# Systematic Search for lncRNA Targets

Dmitri D. Pervouchine

Roderic Guigó (Center for Genomic Regulation, Spain)
Andrei Mironov & Mikhail Gelfand (Moscow State University, Russia)

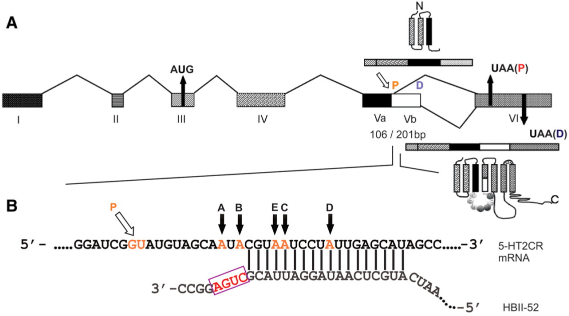
# Long non-coding RNAs

- $\simeq$ 40% of the entire human genome is transcribed
- $\simeq$ 18% of intergenic space is transcribed, generally at lower levels
- Some lncRNAs are involved in epigenetic silencing and imprinting
- Function most long non-coding transcripts yet unknown
- Diverse class of molecules with distinct functions

---

- Are there specific motifs in lncRNAs that are responsible for targeting to specific genomic loci?
- Do lncRNAs directly interact with DNA to form lncRNA:DNA hybrids or triplexes?
- **If the specificity of lncRNAs is achieved by sequence complementarity, do they directly interact with other RNAs**?
- **Are there any lncRNAs implicated in the regulation of alternative splicing?**

# snoRNA HBII-52 regulates splicing of 5-$\mathrm{HT_{2C}R}$ exon V


A
N
AUG
P
D
UAA(P)
I
II
III
IV
Va
Vb
VI
106 / 201bp
UAA(D)
C
B
P
A
B
E
C
D
5′ – .....GGAUCGGUAUGUAGCAAUACGUAAUCCUAUUGAGCAUAGCC.....–3′
5-HT2CR mRNA
3′ –CCGGAGUCGCAUUAGGAUAACUCGUACUAA.....–5′
HBII-52

- Exon V has two donor sites, proximal (P) and distal (D)
- A truncated protein is produced when P is used
- HBII-52, a brain specific C/D box snoRNA, serves as a patch base-pairing to a sequence downstream of P
- HBII-52 and 5-$\mathrm{HT_{2C}R}$ are on different chromosomes
- HBII-52 also affects splicing of at least five other genes[1]

[1] S. Kishore and S. Stamm, Science 311 no. 5758 pp. 230-232, 2006.

# Can we discover HBII-52 targets bioinformatically?

- BLAST or better GUUGle[2] (suffix trees + GU bps)
- Blasting SNORD115 against all human genes gives $\simeq 2,500$ hits
- After filtering out snoRNA paralogs, $\simeq 500$ hits left
- Other HBII-52 targets are imperfect, need internal loops

```
5' G        C        G 3'      5' A          A    G   3'      5' C     A   C          3'
    GUGAUUCU UUG GUAU              GU GUCCUGU UGAG AU              CGUGG CCU AUUG GCA
    CAUUAGGA AAC CGUA              CA UAGGAUA ACUC UA              GCAUU GGA UAAC CGU
3' G        U   U    U 5'      3' G  U            G   5'      3'       A        U   A 5'
```

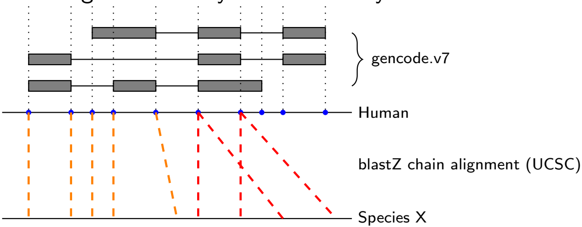
- Have internal loops? Sorry, no BLAST or GUUGle (but RNAplex[3]).
- **There is an emerging need for a computational method that would allow efficient detection of RNA-RNA interaction sites on transcriptome-wide scale**
- **Conservation is a powerful and restrictive filter to narrow down the search to phylogenetically conserved interactions.**

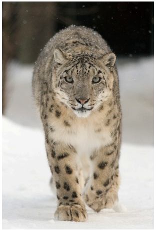
[2] Gerlach & Giegerich, Bioinformatics, 22(6):762-764, 2006

[3] Tafer *et al*, Bioinformatics 27(14):1934-40, 2011

# Methods

1. Gene segmentation by exon boundary:

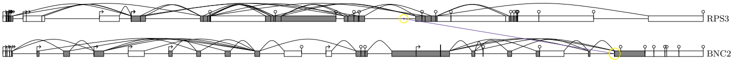


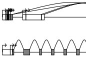


**IRBIS**

2. Sequence weights from phylogenetic tree (16 mammals)
3. IRBIS
   - Set $A$ = segments of non-coding genes (e.g., snoRNA, lncRNAs etc)
   - Set $B$ = non-coding segments of protein-coding genes
   - $\mathcal{R} = A \times B$ (all-to-all)
   - Pattern 4-2-4, at most 1 GT and at least 2 GC per seed
   - Low-comlexity regions excluded
   - Present in 75% of species
   - Length at least 12 after extention

# Gallery: intERmolecular structures





```
H.Sap   ctcagacagg gtgagcagct...311...-tatttattgatcgatttagaGGCATTTGTCTGAGAAGGG------tccagaccca-ggggtgcttg-----...74...tgtcctctag gtgtgctggg
R.Mac   ctcagacagg gtgagcagct...308...-tattcattgattgatttagaGGCATTTGTCTGAGAAGGGA------ccagaccca-ggagtgcttga----...73...tgtcctctag gtgtgctggg
P.Abe   ctcagacagg gtgagcagct...310...-tatttattgatcgatttagaGGCATTTGTCTGAGAAGGG------tccagaccca-ggggtgcttg-----...74...ggtcctctag gtgtgctggg
C.Jac   ctcagacagg gtaagcagct...313...-tattcattgattgatttagaGGCATTTGTCTGAGAAGGGA------ccagaccca--ggttgcttgag---...73...tgtcttctag gtgtgctggg
M.Mus   ctcagacagg gtaagtagct...265...-tgcgcactgattgatttagaGGCATTTGTCTGAGAAGGGA-----acagcgctca--gtgcttttgg----...60...cgtcttctag gtgtgctggg
R.Nor   ctcagacagg gtaagtagct...265...-tgcatactgattgatttagaGGCATTTGTCTGAGAAGGGA-----acagagctca--gtgcttttga----...67...tgtcttttag gcgtgctggg
C.Por   ctcagacagg gtaagcctct...306...-tgcttattgattgatttagaGGCATTTGTCTGAGAAGGgaTGAGAtgtagtagcactagtta--taa----...67...tgtcttctag gtgtgctggg
O.Cun   ctcagacagg gtgagccact...253...-tgttcattgattgatttagaGGCATTTGTCTGAGAAGGGA------ccagatcag--aggttcttggg---...54...ggtttcctag gtgtgctggg
C.Fam   ctcagacagg gtgagcagct...308...ctgtttattgattga-ttagaGGCATTTGTCTGAGAAGGGA------ctggatcca-tgggtgcttga----...74...catcttccag gtgtgctggg
F.Cat   ctcagacagg gtaagcagct...306...ctgttcattgattga-ttagaGGCATTTGTCTGAGAAGGGA------ccagaccca-ggggtgcttga----...72...catcttctag gtgtgctggg
B.Tau   ctcagacagg gtgagtagcc...270...ctgttcattgattga-ttagaGGCATTTGTCTGAGAAGGGG------ccagacac--aggatgcttgag---...80...catcttctag gtgtgctggg
E.Cab   ctcagacagg gtaagcgcct...299...ctgttcattgattga-ttagaGGCATTTGTCTGAGAAGGGA------ccagaccca-agggtgcttga----...81...tgtcctctag gtgtgctggg
S.Sus   ctcagacagg gtgagcagcc...297...ccgttcattggttga-ttagaGGCATTTGTCTGAGAAGGGG------ccagacccagaggatgcttg-----...78...catcttctag gtgtgctggg
O.Avi   ctcagacagg gtgagtagcc...272...ctgttcattgattga-ttagaGGCATTTGTCTGAGAAGGGG------ccagacac--aggatgcttgag---...80...catcttctag gtgtgctggg
L.Afr   ctcagacagg gtgagcagat...314...-tgttcattggttgatttagaGGCATTTGTCTGAGAAGGGA------ccaa---ca--gggtgcttgagtaa...65...catcctctag gtgtgctggg
M.Dom   ctcagacagg gtgagccact...224...-tggacattggttgatctagaGGCATTTGTCTGAGAAGGGG------ctaaa-----aagctgttctgtact...72...tgccattcag gtgtgttggg
        ********** ** **                  * ** * **  *********************                                             ** * *** ****
```

segm.id=61486 gene=ENSG00000149273 name=RPS3 segment=chr11_75115252_75115715_1 type=IN

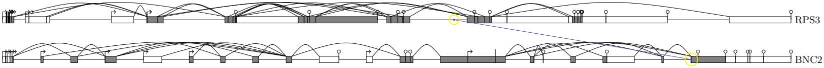
```
H.Sap   ttgtcacaag gtaagctcag...1568...-tccacgta--ccccccttccc-TCTTgtTTTCctAG ACAcAGTGCc
R.Mac   ttgtcacaag gtaagctcag...1491...-tccaccta--ccccccttccc-TCTTgTTTTCctAG ACAcAGTGCc
P.Abe   ttgtcacaag gtaagctcag...1800...-tccaccta--ccccccttccc-TCTTgtTTTCctAG ACAcAGTGCc
C.Jac   ttgtcacaag gtaagctcag...1521...-tccaccta--ccccccttccc-TCTTaTTTTCctAG ACAcAGTGCc
M.Mus   ttgtcacaag gtaagctcag...1414...-tgtgcgcc--tccccctcccc-TCTTgTTTTCctAG ACAcAGTGCc
R.Nor   ttgtcacaag gtaagctcag...1392...-cccccgtg--ccccccttccc-TCTTgTTTTCctAG ACAcAGTGCc
C.Por   ttatcacaag gtaagctcag...1455...ctgcaccta---cccccttccc-TCTTGTTTTCcTAG ACAcAGTGCc
O.Cun   ttatcacaag gtaagctcag...1520...-tgcaccca--ccccccttccc-TCTTgTTTTCctAG ACAcAGTGCc
C.Fam   ttatcacaag gtaagctcag...1502...-atacacct--ccccccttccc-TCTTgTTTTCctAG ACAcAGTGCc
F.Cat   ttatcacaag gtaagctcag...1532...---cacctatcccccccttccc-TCTTgTTTTCctAG ACAcAGTGCc
B.Tau   ttatcacaag gtaagctcag...1564...acctacccg--cccccacttcc-CTCTtgTTTTCctAG ACAcAGTGCc
E.Cab   ttatcacaag gtaagctcag...1541...-tcacacct--ccccccttccc-TCTTgTTTTCctAG ACAcAGTGCc
S.Sus   ttatcacaag gtaagctcag...1522...-ccacctcc--cccaccccttccCTCTtgTTTTCctAG ACAcAGTGCc
O.Avi   ttatcacaag gtaagctcag...1568...acctacccg--ccccccttcc-CTCTtgTTTTCctAG ACAcAGTGCc
L.Afr   ttatcacaag gtaagctcag...1533...-cctccacc--tccccccttccc-TCTTgTTTTCctAG ACAcAGTGCc
M.Dom   ttatcacaag gtaagctcag...1527...-cccacctg--cccaccccttccCTCTtaTTTTCctAG ACAcAGTGCc
        ** ******* **********                          ** * *   *  **** ********* **********
```

segm.id=324610 gene=ENSG00000173068 name=BNC2 segment=chr9_16419648_16421259_-1 type=IN

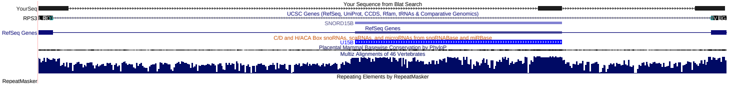
Figure 24: CS=21.87

```
H
P
C
M
R
C
O
C
F
B
E
S
O
L
M
```

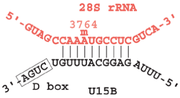
# Conserved box in RPS3 intron is a snoRNA



```
H.Sap   ctcagacagg gtgagcagct...311...-tatttattgatcgatttagaGGCATTTGTCTGAGAAGGG------tccagaccca-ggggtgcttg-----...74...tgtcctctag gtgtgctggg
R.Mac   ctcagacagg gtgagcagct...308...-tattcattgattgatttagaGGCATTTGTCTGAGAAGGGA------ccagaccca-ggagtgcttga----...73...tgtcctctag gtgtgctggg
P.Abe   ctcagacagg gtgagcagct...310...-tatttattgatcgatttagaGGCATTTGTCTGAGAAGGG------tccagaccca-ggggtgcttg-----...74...ggtcctctag gtgtgctggg
C.Jac   ctcagacagg gtaagcagct...313...-tattcattgattgatttagaGGCATTTGTCTGAGAAGGGA------ccagaccca--ggttgcttgag---...73...tgtcttctag gtgtgctggg
M.Mus   ctcagacagg gtaagtagct...265...-tgcgcactgattgatttagaGGCATTTGTCTGAGAAGGGA-----acagcgctca--gtgcttttgg----...60...cgtcttctag gtgtgctggg
R.Nor   ctcagacagg gtaagtagct...265...-tgcatactgattgatttagaGGCATTTGTCTGAGAAGGGA-----acagagctca--gtgcttttga----...67...tgtcttttag gcgtgctggg
C.Por   ctcagacagg gtaagcctct...306...-tgcttattgattgatttagaGGCATTTGTCTGAGAAGGgaTGAGAtgtagtagcactagtta--taa----...67...tgtcttctag gtgtgctggg
O.Cun   ctcagacagg gtgagccact...253...-tgttcattgattgatttagaGGCATTTGTCTGAGAAGGGA------ccagatcag--aggttcttggg---...54...ggtttcctag gtgtgctggg
C.Fam   ctcagacagg gtgagcagct...308...ctgtttattgattga-ttagaGGCATTTGTCTGAGAAGGGA------ctggatcca-tgggtgcttga----...74...catcttccag gtgtgctggg
F.Cat   ctcagacagg gtaagcagct...306...ctgttcattgattga-ttagaGGCATTTGTCTGAGAAGGGA------ccagaccca-ggggtgcttga----...72...catcttctag gtgtgctggg
B.Tau   ctcagacagg gtgagtagcc...270...ctgttcattgattga-ttagaGGCATTTGTCTGAGAAGGGG------ccagacac--aggatgcttgag---...80...catcttctag gtgtgctggg
E.Cab   ctcagacagg gtaagcgcct...299...ctgttcattgattga-ttagaGGCATTTGTCTGAGAAGGGA------ccagaccca-agggtgcttga----...81...tgtcctctag gtgtgctggg
S.Sus   ctcagacagg gtgagcagcc...297...ccgttcattggttga-ttagaGGCATTTGTCTGAGAAGGGG------ccagacccagaggatgcttg-----...78...catcttctag gtgtgctggg
O.Avi   ctcagacagg gtgagtagcc...272...ctgttcattgattga-ttagaGGCATTTGTCTGAGAAGGGG------ccagacac--aggatgcttgag---...80...catcttctag gtgtgctggg
L.Afr   ctcagacagg gtgagcagat...314...-tgttcattggttgatttagaGGCATTTGTCTGAGAAGGGA------ccaa---ca--gggtgcttgagtaa...65...catcctctag gtgtgctggg
M.Dom   ctcagacagg gtgagccact...224...-tggacattggttgatctagaGGCATTTGTCTGAGAAGGGG------ctaaa-----aagctgttctgtact...72...tgccattcag gtgtgttggg
        ********** ** **                    * ** * **  **********************                                                  ** * *** ****
```

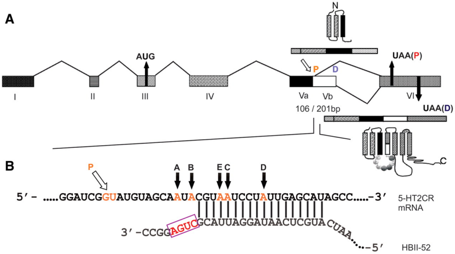

YourSeq
Your Sequence from Blat Search
RPS3
UCSC Genes (RefSeq, UniProt, CCDS, Rfam, tRNAs & Comparative Genomics)
SNORD15B
RefSeq Genes
RefSeq Genes
C/D and H/ACA Box snoRNAs, scaRNAs, and microRNAs from snoRNABase and miRBase
U15B
Placental Mammal Basewise Conservation by PhyloP
Multiz Alignments of 46 Vertebrates
Repeating Elements by RepeatMasker
RepeatMasker

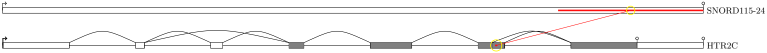
- U15B is predicted to guide the 2'O-ribose methylation of 28S rRNA



- Why conservation extends beyond D-box?
- U15B is complementary to 11 other targets

# HBII-52 and splicing of 5-$\mathrm{HT}_{2C}\mathrm{R}$ exon V


A
N
AUG
UAA(P)
P
D
I
II
III
IV
Va
Vb
VI
106 / 201bp
UAA(D)
C
B
P
A
B
E
C
D
5′ – .....GGAUCGGUAUGUAGCAAUACGUAAUCCUAUUGAGCAUAGCC.....–3′
5-HT2CR mRNA
3′ –CCGGAGUCGCAUUAGGAUAACUCGUACUAA.....–5′
HBII-52



```
H.Sap   gccttgggtt cagtcagtgtcgagaaccttatattgttctgaagagaggtggtgacttaaaaatcATGCTCAATAGGATTACGctgaggccc agcctagttg
R.Mac   gtcctaggtt gggtcaatgatgagaaccttataatgttctgaagagaggtgatgacttaaaaatcATGCTCAATAGGATTACGctgaggccc agcctaggtg
P.Tro   gccttgggtt cagtcagtgtcgaggaccttatattgttctgaagagaggtggtgacttaaaaatcATGCcCAATAGGAcTACGctgaggccc agcctagttg
P.Abe   gccttgggtt tggtcaatgttgagaaccttatattgttctgaagagaggtggtgacttaaaaatcATGCTCaATAGGATTACGctgaggcat agcctagttg
M.Mus   gtctaggcct gggtcaataatggcaatc---caatgtcatgaagaaaggtgatgacataaaattcATGCTCAATAGGATTACtctgaggccc aaccagggcg
R.Nor   gcctgggtct tggacaatgatgacaaca---aaaagtcatgaagaaagatgatgacataacaatcATGCTCAGTAGGATTACGctgaggcct aaccatggag
O.Cun   tcctgtggct gggtcgatgatgag-acctcatactgtcttgaagagagatgatgacttaaaaatcATGCTCAATAGGATTACGctgaggccc agcctagtga
          *   *  *   * *  *   *   *      *  **  ****** ** ** **** *** * ****** ** ***** *** *******   * **   *
```

segm.id=54259 gene=ENSG00000200398 name=SNORD115-24 segment=chr15_25458795_25458876_1 type=EN

```
H.Sap   cgctggatcg gtatgtagcaataCGTAATCCTATTGAGCATagccgtttcaattcgcggac...34...atttctatag gtaaataaaa
R.Mac   cgctggatcg gtatgtagcaataCGTAATCCTATTGAGCATagccgtttcaattcgcggac...34...atttctatag gtaaataaaa
P.Tro   cgctggatcg gtatgtagcaataCGTAaTCCTATTGaGCATagccgtttcaattcgcggac...34...atttctatag gtaaataaaa
P.Abe   ccctggatcg gtatgtagcaataCGTAATCCTA-TGAGCATagccgtttcaattcgcggac...34...atttctatag gtaaatgaaa
M.Mus   cgctggaccg gtatgtagcaatacGTAATCCTATTGAGCATagccggttcaattcgcggac...34...atatcaatag gtaattatac
R.Nor   cgctggaccg gtatgtagcaataCGTAATCCTATTGAGCATagccggttcaattcgcggac...34...atatcaatag gtaaatatac
O.Cun   cgctggatcg gtatgtagcaataCGTAATCCTATTGAGCATagccgtttcaattcgcggac...34...atttctatag gtaactaaac
        * ***** ** *********************** ************ **************        ** ** **** **** *  *
```

segm.id=592300 gene=ENSG00000147246 name=HTR2C segment=chrX_114082672_114082766_1 type=AC

Figure 1: CS=

# Confounding factors

- There are reasons for a pair of motifs to be complementary and conserved **other than RNA secondary structure**
- Conserved bi-directional cis-elements on the DNA will always be found as such
- We can't distinguish them from conserved RNA-RNA interaction sites in principle
- (Sense-antisense pairs have to be excluded forever)

# Control 1: Search the opposite strand

- $A$ = segments of lncRNAs
- $B$ = (intronic) segments of protein coding genes
- Search $A$ vs. $B'$, sequences on the opposite strand to ones in $B$
- Conservation rate and dinucleotide content don't change

# Control 1: Search the opposite strand

In the tables: $\#\text{hits}[A,B]/\ \#\text{hits}[A,B']$ (% enrichment)

**16 placental mammals**

| | snoRNA | snRNA | lncRNA | introns |
|---|---|---|---|---|
| snoRNA | | 3/0 (NA) | 277/241 (+14%) | 1439/1099 (+30%) |
| snRNA | | | 15/2 (NA) | 120/92 (+30%) |
| lncRNA | | | | 7974/6329 (+25%) |
| introns | | | | |

**12 drosophilids**

| | snoRNA | snRNA | ncRNA | introns |
|---|---|---|---|---|
| snoRNA | | 71/95 (-25%) | 34/39 (-12%) | 1158/1122 (+3%) |
| snRNA | | | 60/175 (-65%) | 3432/2695 (+27%) |
| ncRNA | | | | 963/921 (+4%) |
| introns | | | | |

**6 nematodes**

| | snoRNA | snRNA | ncRNA | introns |
|---|---|---|---|---|
| snoRNA | | 107/69 (+55%) | 514/512 (0%) | 362/355 (+1%) |
| snRNA | | | 2273/2584 (-12%) | 1088/1117 (-2%) |
| ncRNA | | | | 5635/4950 (+13%) |
| introns | | | | |

**Non-coding RNAs have higher potential to basepair introns of protein-coding genes at sense strand compared to antisense strand**

# Control 2: Random sampling

- $A =$ segments of lncRNAs
- Search set $A_1$ vs. set $B$, where $A$ is sampled randomly from "non-lncRNAs"
- $A_1 =$ segments of protein coding genes (equivalent random sample)
- $B_1 = B \setminus A_1$
- Search $A$ against $B_1$ against $A_1$ against $B_1$

- Conservation rate and GC content of the random sample are confounding
- How enrichment in $A$ against $B_1$ vs. $A$ against $B_1'$ relates to the enrichment $A_1$ against $B_1$ vs. $A$ against $B_1'$

# Control 2: Random sampling

- $A$ = segments of lncRNAs
- $B$ = intronic segments of protein coding genes
- $A_1$ = random sample of segments of protein coding genes
- $B_1 = B \setminus A_1$
- $B_1'$ = reverse complements to sequences in $B_1$

| $i$ | $\#[A, B_1]$ | $\#[A, B_1']$ | % | $\#[A_1, B_1]$ | $\#[A_1, B_1']$ | % | % − % |
|---|---|---|---|---|---|---|---|
| 1 | 6653 | 5126 | 29.79 | 8482 | 7790 | 8.88 | 20.91 |
| 2 | 6756 | 5329 | 26.78 | 8166 | 7673 | 6.43 | 20.35 |
| 3 | 6661 | 5354 | 24.41 | 7922 | 7753 | 2.18 | 22.23 |
| 4 | 6581 | 5268 | 24.92 | 8864 | 8370 | 5.90 | 19.02 |
|  | ... | ... | ... | ... | ... | ... | ... |
| 20 | 6737 | 5218 | 29.11 | 8252 | 7566 | 9.07 | 20.04 |

- **Long non-coding RNAs have higher potential to basepair introns of protein-coding genes than do protein-coding genes themselves**
- True in mammals (+20%), drosophilids (+7%), and nematodes (+15%)

# Summary

- Non-coding RNAs have higher potential to basepair introns of protein-coding genes at sense strand as compared to antisense strand
- lncRNAs have higher potential to basepair introns of protein-coding genes than do protein-coding genes themselves
- lncRNAs predicted to be complementary to introns of protein-coding genes are, on average, more correlated (by absolute value) with the respective splicing events than do mock target pairs
- In spite of statistical evidence, we still don't know which pairs are functional

# Acknowledgments

Centre de Regulació Genòmica

Roderic Guigó
Alessandra Breschi
Rory Johnson
Angelika Merkel
Andrea Tanzer
Sarah Djebali
Maik Röder
Julien Lagarde
Cedric Notredame
Giovanni Bussotti
Veronica Raker
Juan Valcárcel

Moscow State University

Katya Khrameeva
Marina Pichugina
Ilya Kurochkin
Anya Gerasimova
Petr Rubtsov
Andrei Mironov
Mikhail Gelfand

Oleksii Nikolaienko
Inessa Skripkina
Alla Ryndich

**Postdocs wanted**

**in R. Guigó's lab**

**and in C. Notredame's lab**