FROM HUMAN GENOME PROJECT TO HUMAN PROTEOME PROJECT /HPP/: RUSSIAN PARTICIPATION

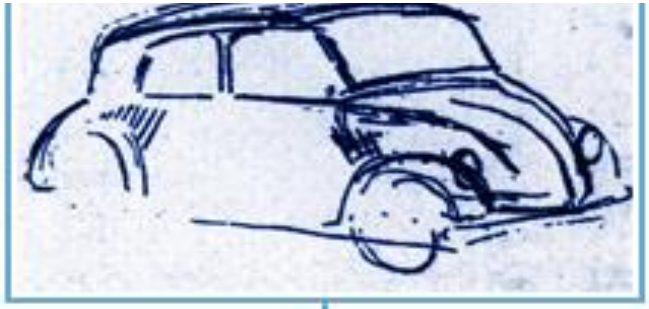
ARCHAKOV A.I.

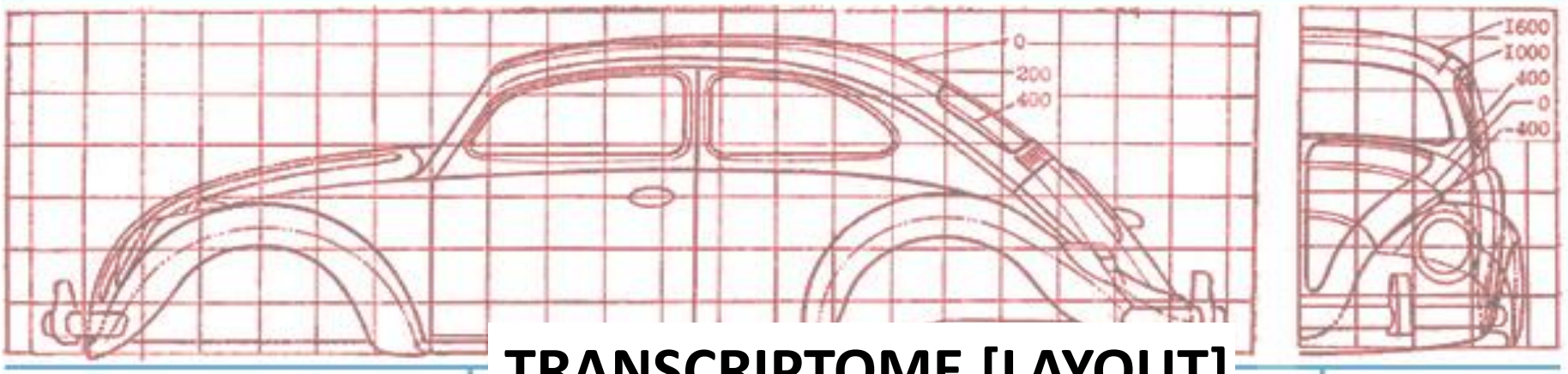
INSTITUTE OF BIOMEDICAL CHEMISTRY OF RAMS, MOSCOW, RUSSIA

www.ibmc.msk.ru

[HTTP://WWW.PROTEOME.RU/EN/ROADMAP]

GENOME [OUTLINE]





TRANSCRIPTOME [LAYOUT]

PROTEOME [PRODUCT]

THE MAIN DIFFERENCE BETWEEN GENOMICS AND PROTEOMICS:

THE SAME GENOME

CATERPILLAR

BUTTERFLY

THE DIFFERENT PROTEOME

STEPS OF HPP HUPO FORMATION:

BARBADOS CONFERENCE 2007 05/11-01-2007

H. PEARSEN. BIOLOGISTS INITIATE PLAN TO MAP HUMAN PROTEOME.
NATURE, 452, 24, 920-921, 2008.

HUMAN PROTEOME PROJECT HUPO 2008, 7th WORLD CONGRESS, 16/20-08-08,Amsterdam, NL

HPP MOSCOW WORKSHOP, RUSSIA 20/21-03,2009

SEOUL, KHUPO, 26/27-03-2009
CANADIAN HUPO, TORONTO, 26/30-09-2009

THE EXISTING PROTEOMIC TECHNOLOGIES ALLOW TO LAUNCH HPP

AFTER COMPLETION OF HPP PROJECT GENOME-BASED MEDICINE SHOULD BECOME PROTEOME-BASED, WHICH WILL BE THE BASIS FOR PERSONALIZED MEDICINE.

HOW MANY PROTEINS EXIST IN HUMAN BODY?

HUMAN GENOME

⇩

20 000 GENES[1]

⇩

~ 50 PROTEIN FORMS PER GENE[2]

⇩

~ 1[2] - 2[3] MLN.PROTEINS

22 PAIRS SOMATIC CHR
1 PAIR SEX CHR

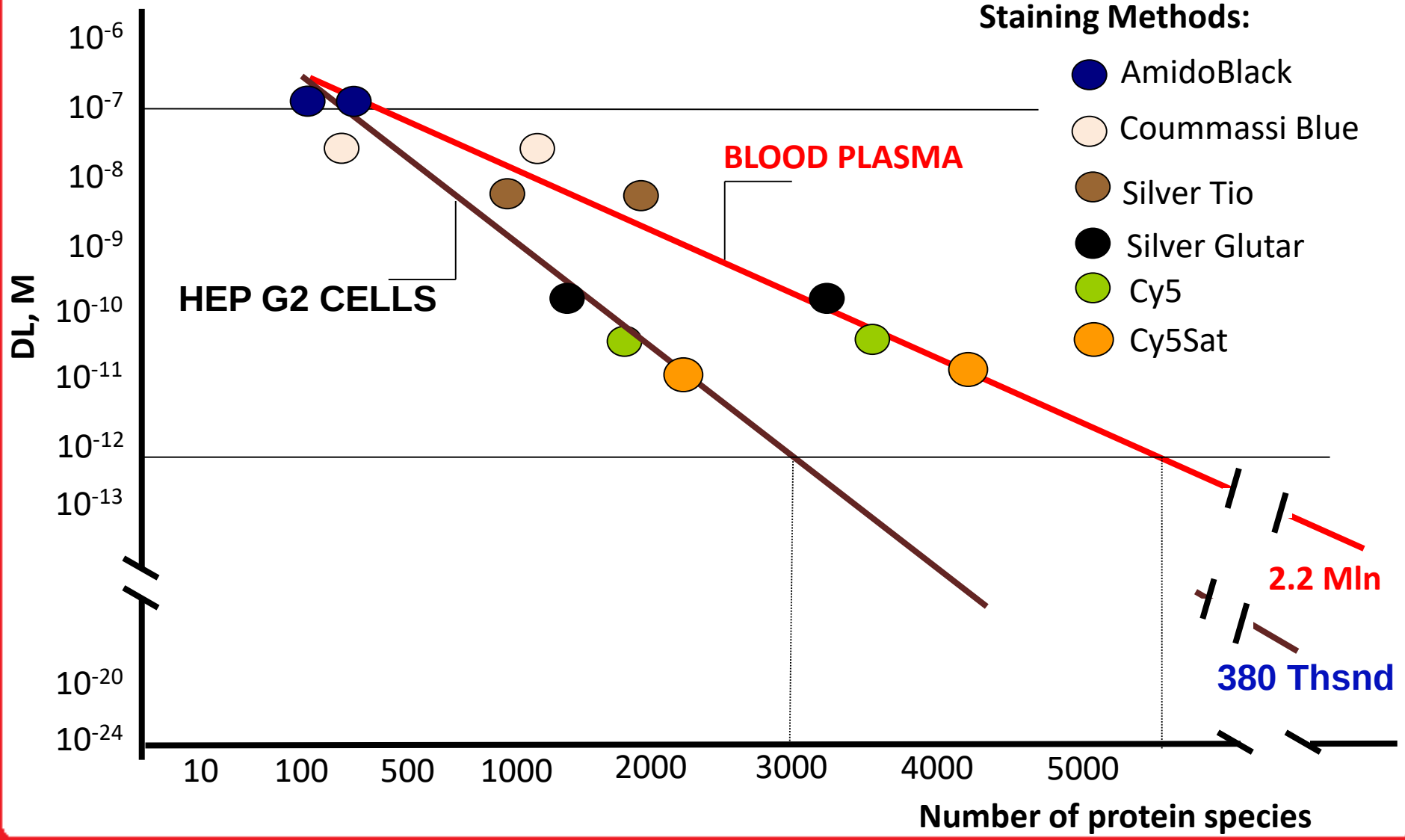
---

[1] *The sequence of the human genome. Venter et al., Science. 2001*

[2] *Extent of modifications in Human Proteome Samples…, Nielsen et al., MCP, 2009*

[3] *Biospecific irreversible fishing… Archakov et al., Proteomics, 2009*

HOW MANY PROTEINS IN HUMAN PROTEOME : dependence between detection limit of the staining method and number of proteins spots on 2DE. (Archakov et al., PROTEOMICS (2009), *9, 1326-1343).*



HUPO2010

Human Proteome World Congress Sydney 2010
Launch of the Human Proteome Project
19-23 September 2010 • Sydney Convention and Exhibition Centre, NSW, Australia

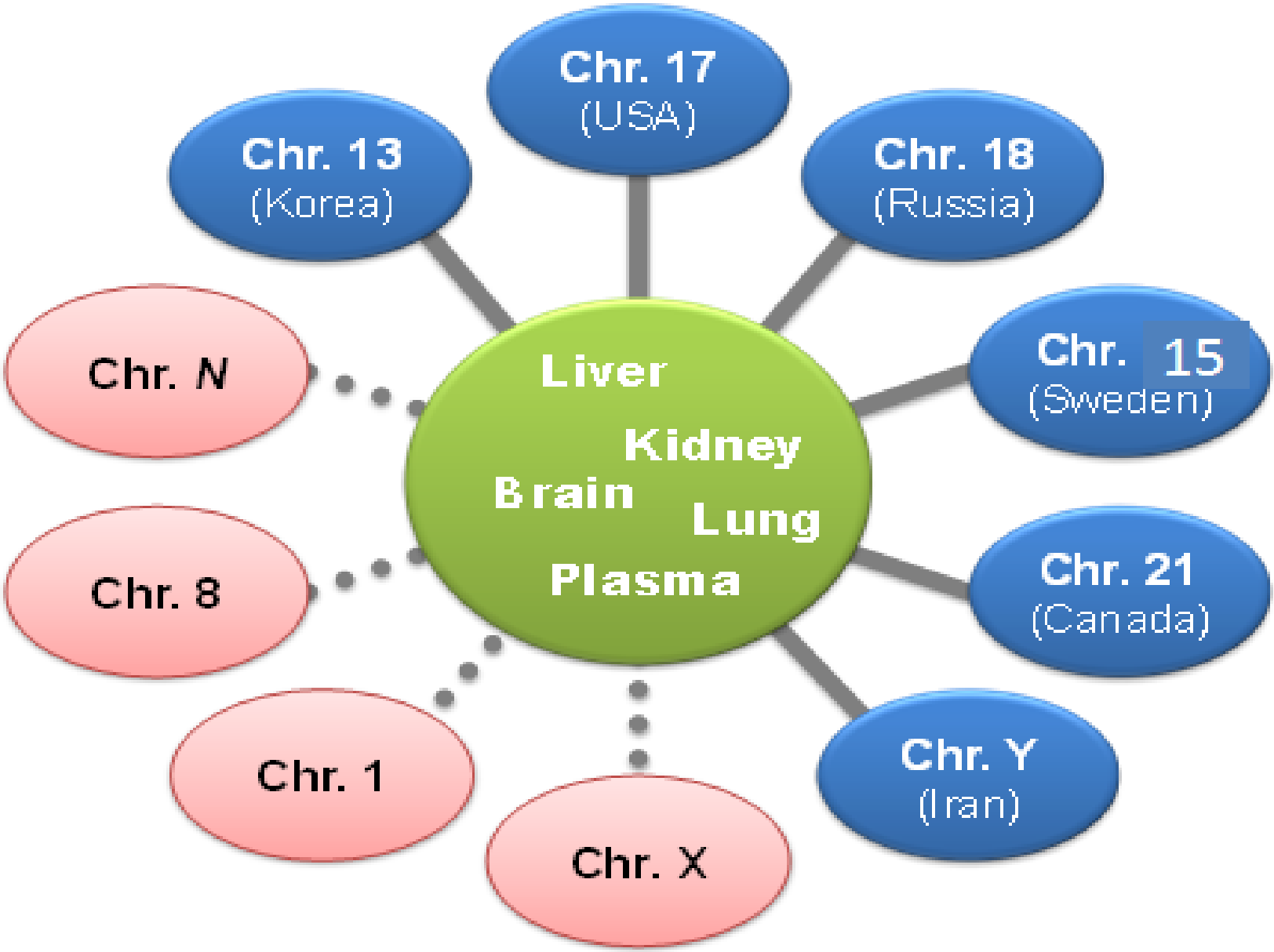
Program Book

SCIENTIFIC PROGRAM

Day 4: Thursday 23 September 2010

0830 – 1030 **Human Proteome Project Launch** *Bayside Auditorium B*
Chair: Gil Omenn and panel

0830 – 0845 **Strategic Overview of the Launch of the HUPO Human Proteome Project (HPP), on behalf of the HUPO HPP Working Group**
Pierre Legrain and Gil Omenn


Chr. 17
(USA)
Chr. 13
(Korea)
Chr. 18
(Russia)
Chr. N
Liver
Kidney
Brain
Lung
Plasma
Chr. 15
(Sweden)
Chr. 8
Chr. 21
(Canada)
Chr. 1
Chr. Y
(Iran)
Chr. X

WHY GENE CENTRIC PROTEOME PROJECT IS REALISTIC?

- **RECENT PROTEOMICS IS GENOME-BASED SCIENCE. GENE CENTRIC PROJECT IS APPLICATION OF EXISTING TECHNOLOGIES TO SIMPLIFYED TASKS**

- **CREATED MRM/IRREVERSIBLE BINDING TECHNOLOGY ALLOWS TO ANALYSE BLOOD PLASMA PROTEINS WITH SENSITIVITY 10^{-18}M (1 copy/1µL) OR 1 PROTEIN COPY PER 10^3 LIVER OR HEP2G CELLS .**
-
- **FOCUSED MRM ANALYSIS OF SINGLE CHROMOSOME WITH KNOWN GENE NUMBER ALLOWS TO GENERATE "GOLD STANDARD" FOR GENE CENTRIC PROTEOME PROJECT.**
- Molecular recognize mapping (*MRM*)

CRITERIA FOR CHROMOSOME SELECTION:

- TOTAL NUMBER OF PROTEIN-CODING GENES
- CLINICAL RELEVANCE
- NUMBER OF ALREADY IDENTIFIED PROTEINS
- ABSENCE OF IMMUNOGLOBULINS

CHROMOSOME STATISTICS

LENGTH (BPS):	78,077,248
GENES	513
KNOWN PROTEIN-CODING GENES:	285
PSEUDOGENE GENES:	64
miRNA Genes:	32
rRNA Genes:	13
snRNA Genes:	51
snoRNA Genes:	36
Misc RNA Genes:	25

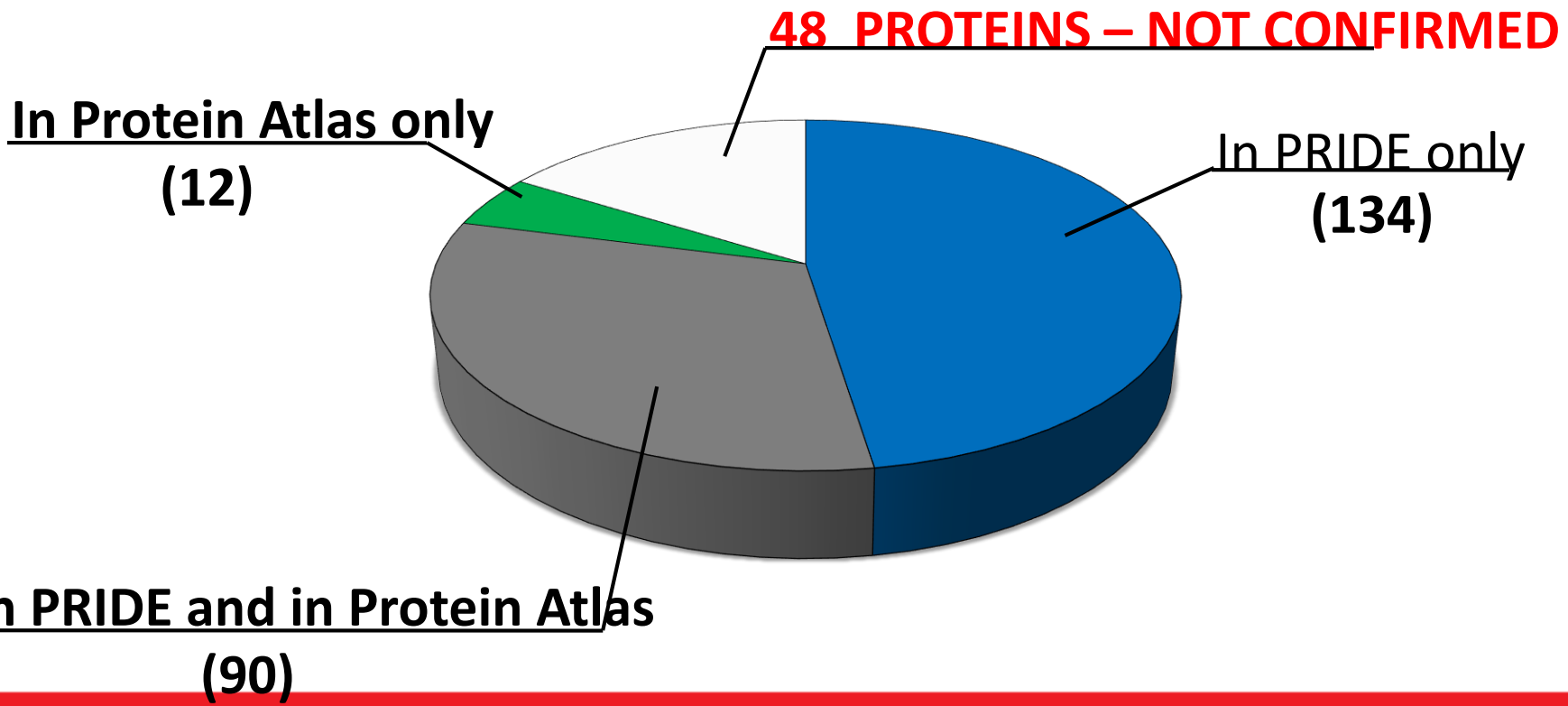
Ensembl release 60 - Nov 2010
http://www.ensembl.org

DATA MINING STATISTICS FOR 18th CHROMOSOME: 286 MASTER PROTEIN CODING GENES

PRIDE = 224 identified proteins

Protein Atlas = 102 proteins

*PRIDE (*PRoteomics IDEntifications database)



How many proteins can we expect for 18th chromosome?

285
+
230
||
516
x
50
||
25800

Trancriptome analisys of 18th chromosome

Over 80% of the genes undergo alternative splicing
(Kampa at. al, Genome Res., 2004, 14, 331-342)

On average, each of them can have 50 PTMs
(Nielsen M, Savitski M, Zubarev R, MCP, 2006, 5, 2384-)

Protein species can be expected as expressed in HepG2 cell line

HepG2 cell lines ซึ่งเป็นต้นกำเนิดจากมะเร็งชนิด hepatocellular adenocarcinoma

ROADMAP

PROTEOME OF THE 18-TH HUMAN CHROMOSOME: GENE CENTRIC IDENTIFICATION OF TRANSCRIPTS, PROTEINS AND PEPTIDES

http://www.proteome.ru/en/roadmap/

http://www.hupo.org/research/hpp/soc/

RusHPP CONSISTS OF TWO PHASES:

GOAL OF THE PILOT PHASE IS IDENTIFICATION OF ALL MASTER PROTEINS PRODUCED BY 18thCHR IN LIVER AND HEP2G CELLS AND IDENTIFICATION OF THEM IN PLASMA WITH SENSITIVITY 10^{-18} M (1 PROTEIN COPY/ML) AND 1 PROTEIN COPY PER 10^3 LIVER OR HEP2G CELLS.

MAIN PHASE OF RusHPP

THE GOAL IS THE IDENTIFICATION OF ALL MODIFIED PROTEINS(ABOUT 30000) EXPRESSED BY 18CHR AT SENSITIVITY 10^{-18} M.

BIOINFORMATIC AND EXPERIMENTAL CREATION OF 18THCHR PROTEIN INTERACTOME BY COMBINING OF OPTICAL BIOSENSOR WITH MS (BUNEEVA O. ET AL. PROTEOMICS 2010, 10, 23-57.)

CREATION OF 18THCHR PROTEIN KNOWLEDGE BASE

PRINCIPLE DIFFERENCE BETWEEN GENOMIC AND PROTEOMICS

PRINCIPLE DIFFERENCE BETWEEN GENOMIC AND PROTEOMICS IS EXISTENCE OF POLYMERASE CHAIN REACTION (PCR) IN GENOMICS ALLOWING TO MULTIPLY NUCLEIC ACID MOLECULES AND ABSENCE OF PCR-LIKE REACTION IN PROTEOMICS. DUE TO PCR GENOMICS DOES NOT HAVE DETECTION LIMIT (DL) OBSTACLE. PROTEOMICS HAS IT.

THREE BOTTLENECKS OF HPP

- **LOW SENSITIVITY OF RECENT PROTEOMIC TECHNOLOGIES. THE BEST ONE -- MRM-MS REACHES SENSITIVITY UP TO 10^{-14} M*, CORRESPONDING TO 10 000 PROTEIN COPIES IN 1ML OF PLASMA.**
- **THE ABSENCE OF "GOLD STANDARD" FOR SAMPLES AND SAMPLING.**
- **PROTEOMICS IS SITUATIONAL SCIENCE. WHAT IS BOARD LINE BETWEEN DIFFERENT SITUATIONS ??**

WHAT IS REVERSE AVOGADRO's NUMBER ?

$N_A = 6.022 \times 10^{23}$ MOLECULES/MOLE

1 MOLE - 1L - 1M

$1/N_A \approx 10^{-24}$ MOLE/MOLECULE

$\approx$1MOLECULE/L = 10^{-24} M

[Archakov A.I. et. al. PROTEOMICS 2007, 7, 4–9]

IT WOULD BE POSSIBLE TO INCREASE THE SENSITIVITY UP 10^{-18}M [ARCHAKOV ET AL., 2009], CORRESPONDING TO 1 PROTEIN COPY IN 1ML.

IT BECAME ACHIEVABLE DUE TO COMBINING AFM OR MRM TECHNOLOGIES WITH IRREVERSIBLE BINDING OF PROTEINS TO *BRCN- SEPHAROSE*.

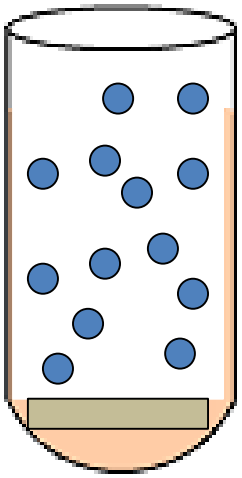
** *atomic force microscopy* (*AFM*) and magnetic resonance microscopy (*MRM*)

THE SENSITIVITY AT THE RANGE OF 10^{-18}M WILL BE QUITE ENOUGH FOR STARTING THE HPP.

COMBINING OF IRREVERSIBLE FISHING TECHNOLOGY WITH AFM

CONCENTRATION OF PROTEINS FROM SOLUTION ON THE ACTIVATED AFM –CHIP USING IRREVERSIBLE FISHING

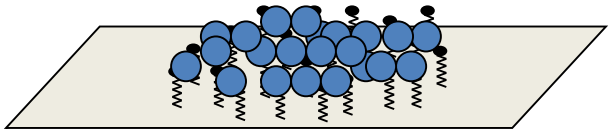
SOLUTION



AFM SURFACE

$$C_{vs} = \frac{C_v * V}{S * h}$$

FISHING



AFM SURFACE

C_v - CONCENTRATION OF PROTEINS IN THE SOLUTION – 10^{-11} M

V - VOLUME – 1mL

C_{vs} SURFACED CONCENTRATION OF FISHED PROTEINS – 10^{-3} M

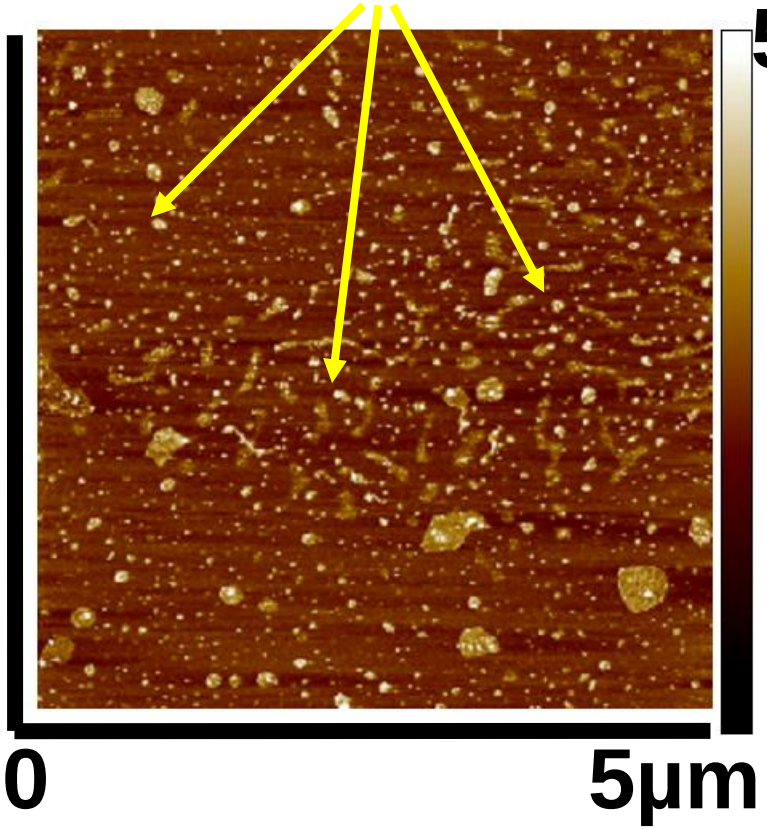
S - AFM SURFACE – $1mm^2$

h - 5 nm HEIGHT OF THE MOLECULE

AFM IRREVERSIBLE FISHING INCREASES SURFACE CONCENTRATION BY FACTOR OF 10^8

AFM IRREVERSIBLE CHEMICAL FISHING

AVIDIN ON AFM-SUPPORT



EXPERIMENT

SUCCINIMIDE MODIFIED MICA

$C_{AVIDIN}=10^{-13}M$;

V=1 ml; T=37^{0}C; t_{INCUB}= 60 min

$S_{ACTIVATION}=0,4\ mm^2$;

$S_{scan}=16FRAMES*25\mu m^2=4*10^{-4}$ mm

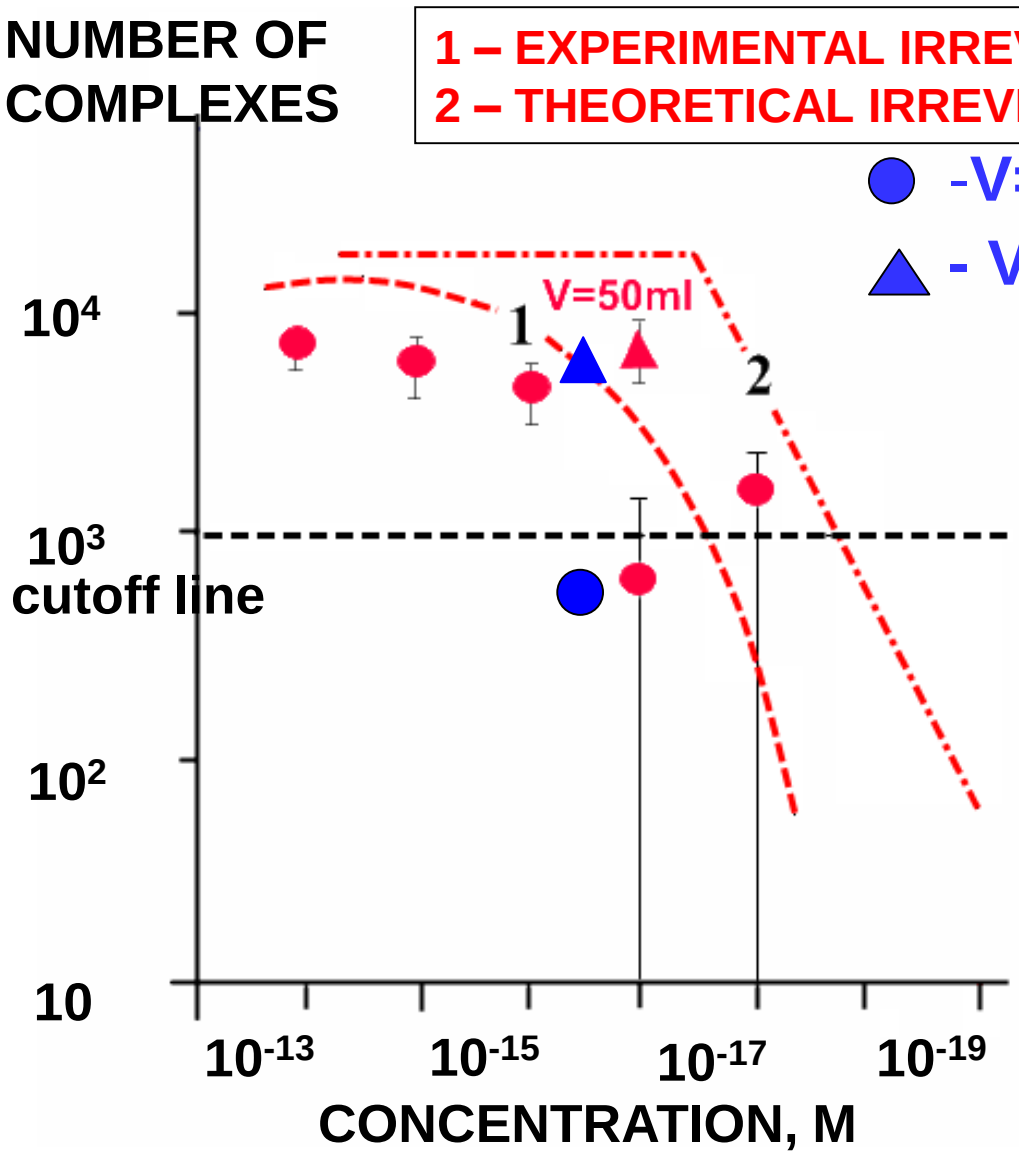
$t_{SCAN}=240min$

$N_{MOLECULES}=5122\pm500$ molecules /400 μm^2

$C_{VS}=5*10^{-6}M$ (FROM EXPERIMENT)

$C_{VS}=10^{-5}M$ (FROM THEORY)

THE DEPENDENCE OF ANTI-HCVcore$_{im}$ /HCVcoreAg PROTEIN COMPLEX NUMBER ON HCVcoreAg CONCENTRATION IN SOLUTION (IRREVERSIBLE BINDING) (ARCHAKOV et al 2009, 9, 1363-1343)



EXPERIMENTAL CONDITIONS:

anti-HCVcoreAg$_{imm}$ (PHOTO-CROSS LINKER MODIFIED)

SIMS=0,4 mm^2;

Scan=16FRAMES*25μm^2=4*10^{-4} mm^2

t_{SCAN}=240min

T = 37^{0}C; t = 60 min;

COMBINING OF IRREVERSIBLE FISHING WITH MRM MS

Experimental workflow for low and ultra low copied protein detection

Step 1

- **BM3 (CYP102)** Bifunctional P450/NADPH-P450reductase
- **BSA (P02769)** Bovin serum albumin

Step 2

Proteotypic peptides

charge state, retention time, fragment ions type

BM3P1 – (K)MHGAFSTNVVASK(E)

BM3P2 - (R)LWPTAPAFSLYAK(E)

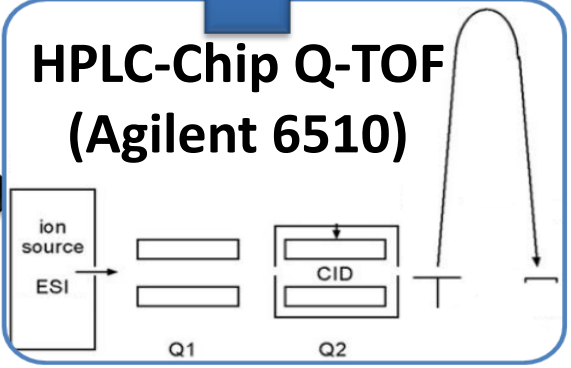
BM3P3 - (R)DFAGDGLFTSWTHEK(N)

BSA P1 - (R)HPEYAVSVLLR(L)

BSA P2 - (K)DAFLGSFLYEYSR(R)

BSA P3 - (K)TVMENFVAFVDK(C)

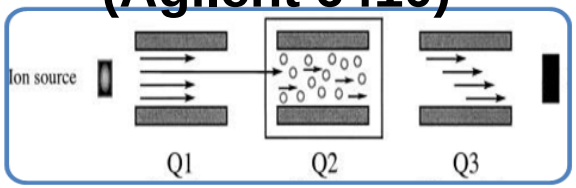
HPLC-Chip Q-TOF (Agilent 6510)



Step 3

QqQ dynamic MRM analysis of individual proteins and in Human Plasma

HPLC-Chip QqQ (Agilent 6410)



Transitions:

BM3P1 – 450.22 (2+)->617.35, 544.22, 404.24

BM3P2 – 732.88 (2+)->967.52, 896.49, 484.26

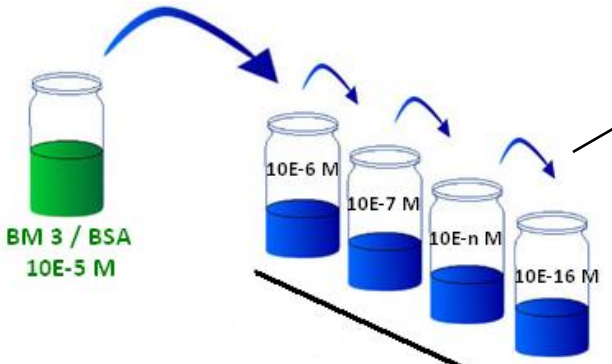
BM3P3 – 855.89 (2+)->724.86, 689.31, 563.19

BSA P1 – 428.59 (2+)->587.38, 346.14, 401.28

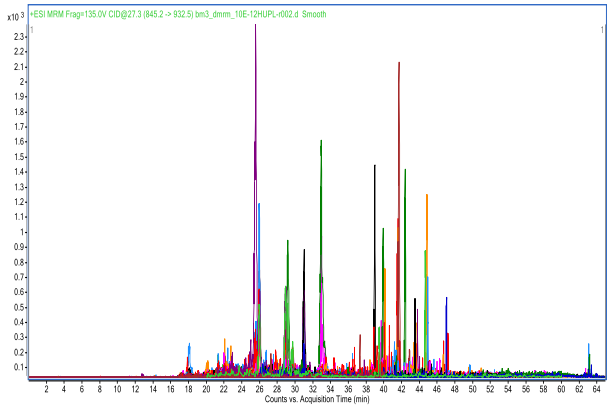
BSA P2 – 784.35 (2+)->1121.49, 717.29, 1064.47

BSA P3 – 700.34 (2+)-> 1199.55, 579.3, 508.26

MRM STRATEGY FOR LOW AND ULTRA-LOW COPIED PROTEIN DETECTION (BSA and CYP102/BM3/ as an example)

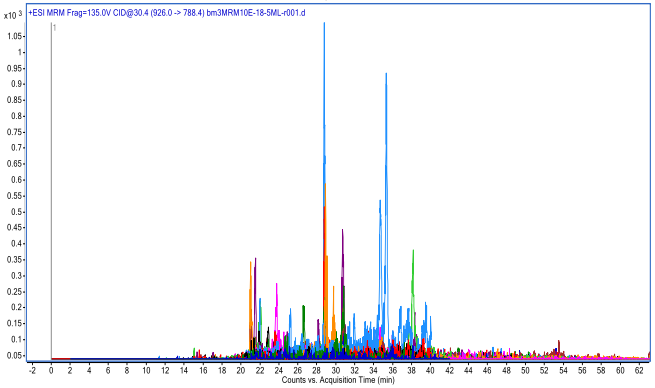


+ trypsin



Detection limit for purified proteins is 10^{-16} M, 600 copies/1 μl (s/n>7)

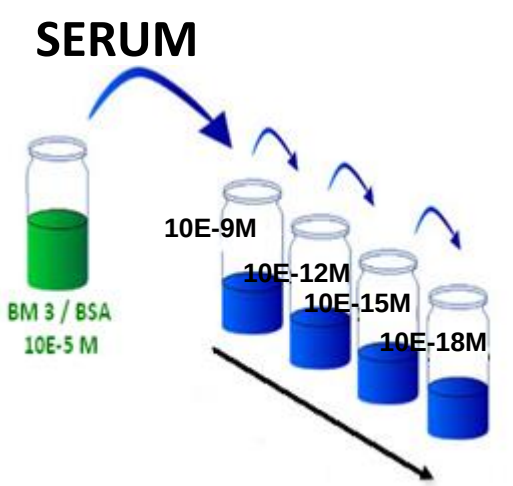
+serum
+ trypsin



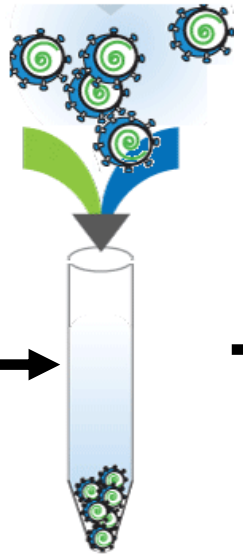
Detection limit for CYP102 in the presence of human serum is 10^{-14} M 60 000 copies/1 μl (s/n>7)

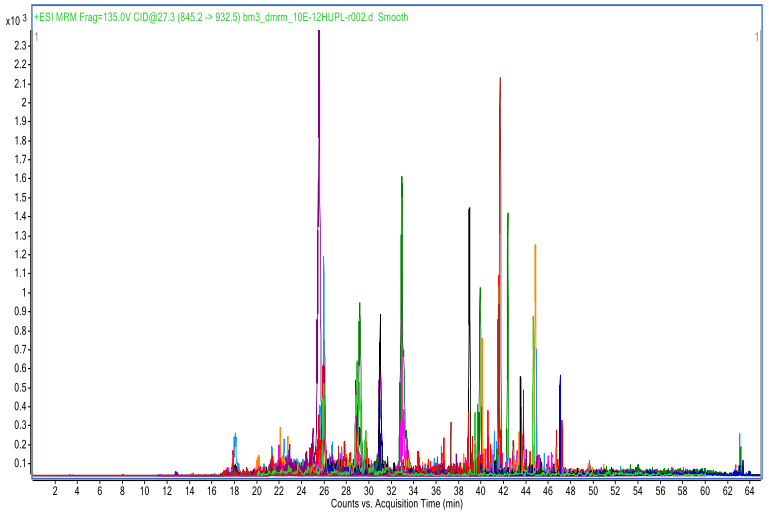
IRREVERSIBLE BINDING OF PROTEINS ON BrCN-SEPHAROSE BEADS

PURIFIED CYP102/BSA IN THE PRESENCE OF HUMAN SERUM



0.5-50 ml





IRREVERSIBLE BINDING OF PROTEINS ON BrCN-SEPHAROSE BEADS FOLLOWING BY DIGESTION WITH TRYPSIN

DETECTION LIMIT FOR CYP102/BSA FOR PURIFIED PROTEIN AND PROTEINS IN THE PRESENCE OF HUMAN SERUM IS 10^{-18} M, 1 COPY/1ML

NEW TECHNOLOGIES FOR Rus-HPP

1. ANALYTICAL COMPLEX BASED ON THE COMBINATION OF ATOMIC FORCE MICROSCOPY AND MASS-SPECTROMETRY

2. ANALYTICAL INSTRUMENTS BASED ON NANOWIRES

3. INFORMATIONAL CLOUD COMPUTING SYSTEM BASED ON THE PERSONAL SUPERCOMPUTER PLATFORM

WHAT IT MEANS PROTEOME-BASED MEDICINE?

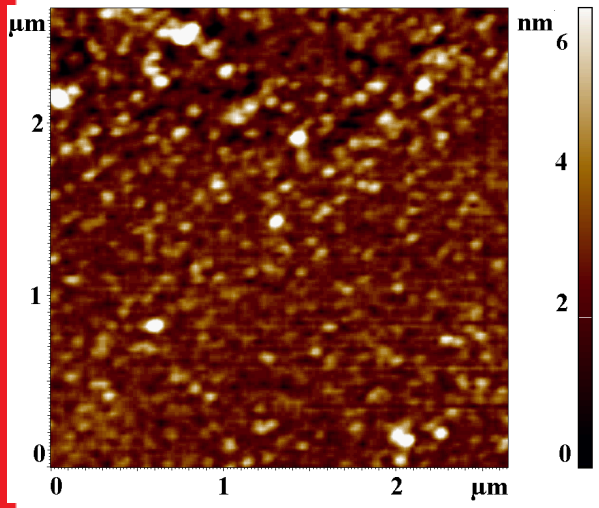
1. NEW DIAGNOSTIC TESTS BASED ON HIGH SENSITIVITY TECHNOLOGY SOLUTIONS (1A) AND PROTEOTYPING (1B) WILL BE CREATED.

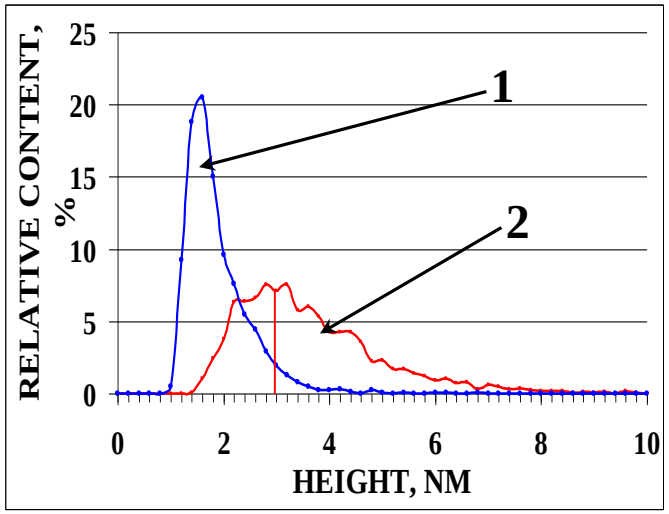
1A. UP-TO-DATE METHODS HAVE SENSITIVITY ABOUT 10^{-12} M. IT MEANS, THAT 1 BILLION OF PROTEIN COPIES COULD BE DETECTED IN 10μL OF BIOLOGICAL MATERIAL. IF SENSITIVITY INCREASE UP 10^{-18} M WE COULD DETECT 1 PROTEIN COPY PER CELL.

1B. AS THERE EXIST SNP, SAP, AS AND PTM, THE TOTAL NUMBER OF PROTEINS ENCODED IN THE GENOME (~20,000 GENES) COULD INCREASE UP 2 MLN. PRESUMABLY, THESE UNKNOWN PROTEINS CAN BE USED IN DIAGNOSTICS.

DETECTION OF HCVcoreAg AT TWO SPOTS AFM CHIP FROM HCVcoreAg SOLUTION (C=10^{-9} M)

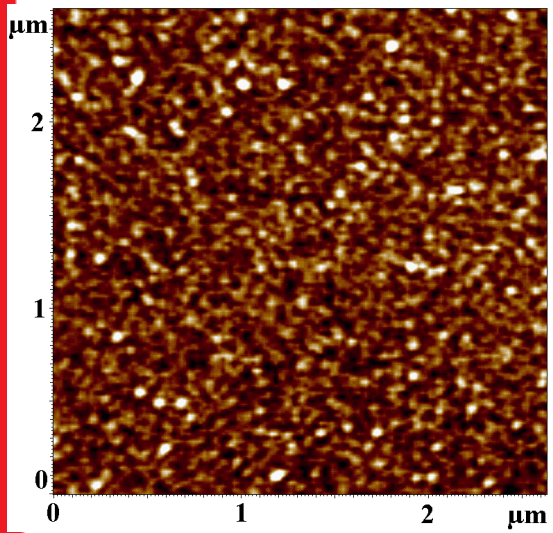
ANTI-HCVcoreAg$_{imm}$ SPOT

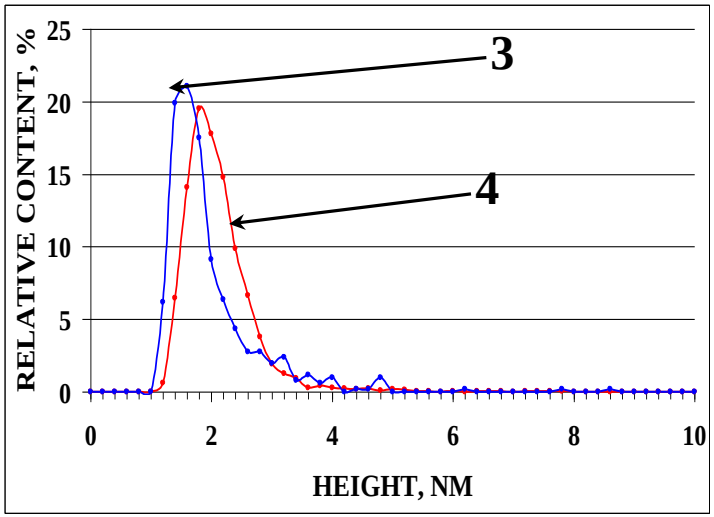




1 – ANTI HCVcoreAg
2 – anti-HCVcore$_{imm}$/HCVcoreAg
h= 3-7 nm

ANTI-HBsAg$_{imm}$ (CONTROL SPOT)





3 – ANTI-HBsAg$_{imm}$
4 – AFM CHIP AFTER INCUBATION IN HCVcore Ag (C=10^{-9}M) SOLUTION

SURFACE TOPOGRAPHY HAS NOT CHANGED

COMPARISON OF AFM AND OTHER METHODS (ELISA AND PCR) FOR DETECTION OF HBsAg AND HCVcoreAg

HCVcoreAg

hepatitis C virus (*HCV*)

	PCR (RNA HCV)	
AFM	**+**	**-**
+	**24**	**2**
-	**8**	**7**

COINCIDENCE – 76%

HBsAg

(Hepatitis B Surface Antigen : *HBsAg*)

	ELISA	
AFM	**+**	**-**
+	**25**	**3**
-	**1**	**6**

COINCIDENCE – 89%

POSSIBLE REASON OF DISAGREEMENT:

132 aa of 191 (70%) in HCVcore PROTEIN SEQUENCE ARE INVARIANT
[Bukh et.al., PNAS, 1994, 91, 8239-8234]

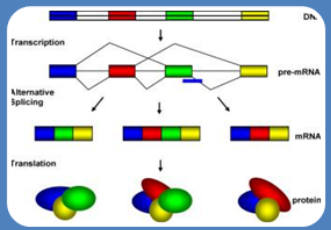
183 aa of 226 (80%) in HBsAg PROTEIN SEQUENCE ARE INVARIANT
[Norder et.al., J. Gener. Virol.,, 1992, 73, 1201-1208]

NEW DIAGNOSTIC TESTS BASED ON PROTEOTYPING

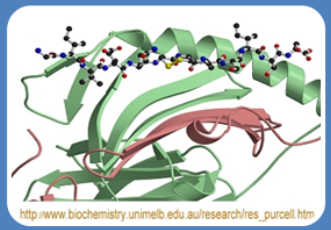
PROTEOTYPE IS A RESULT OF:

SINGLE AMINO ACID POLYMORPHISMS

- DELETIONS/INSERTIONS
- SUBSTITUTIONS



ALTERNATIVE SPLICING



POST-TRANSLATIONAL MODIFICATIONS

WHAT IT MEANS
PROTEOME-BASED MEDICINE/continuation/?

2. NEW DRUG TARGETS

THERE ARE ABOUT 500 DRUG TARGETS NOW IN USE IN PHARMACOLOGY. AT THE END OF HPP AROUND 5 000 - 10 000 NEW DRUG TARGETS COULD BE FOUND.

3. NEW MOLECULAR MECHANISMS OF DISEASES DEVELOPMENT WOULD BE DISCOVERED

IDENTIFICATION OF THOUSANDS OF NEW PROTEINS WOULD HELP TO DECIPHER THE COMPLEX METABOLIC PATHWAYS AND REVEAL UNCOMMON INDIRECT PROTEIN-PROTEIN RELATIONSHIPS.

ACKNOWLEDGEMENTS

- A. Lisitsa
- V. Zgoda
- Yu. Ivanov
- E. Ponomarenko

IBMCH RAMS, MOSCOW