university of groningen

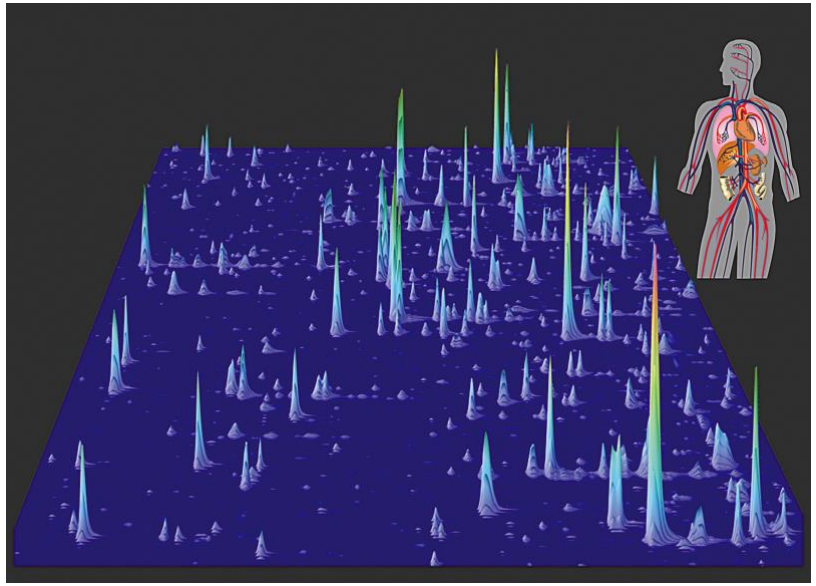
X-omics.nl

Department of Pharmacy
Analytical Biochemistry

The role of proteogenomics in understanding molecular mechanisms of COPD

Yanick Hagemeijer, Rainer Bischoff, Victor Guryev, Peter Horvatovich

p.l.horvatovich@rug.nl



X-Omics Festival, April 12, 2021

university of groningen

Population and intra-individual genomics variability

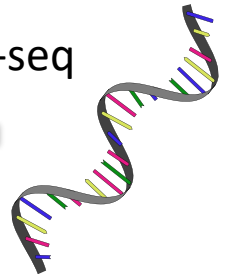
Genome/exome seq

DNA variants

Data integration

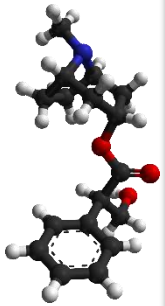
Transcripts

RNA-seq

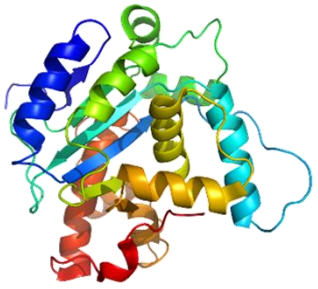


Proteins, peptides

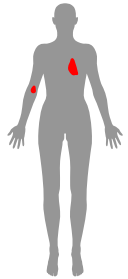
LC-MS/MS



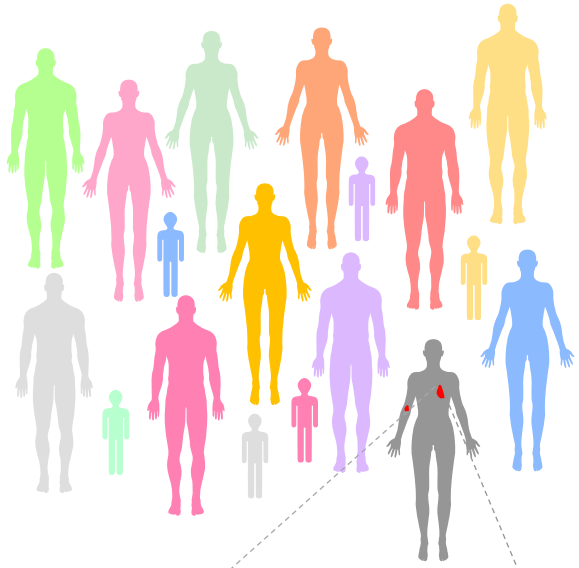
Metabolites



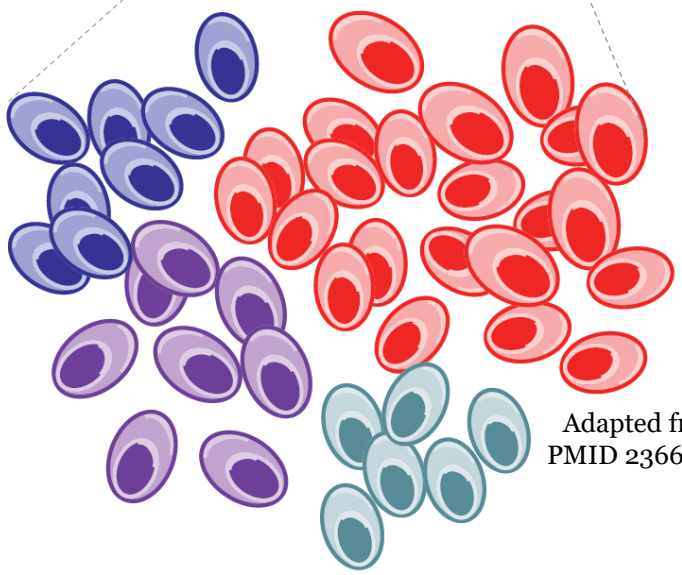
Disease



Omics layers:
Phenome
Microbiome
Metabolome
Lipidome
Glycome
Glycoproteome
Kinome
Phosphoproteome
Proteome
Translatome
Transcriptome
Epi-genome
Genome
Metabolome
Lipidome
Acetylome
…



Population genetic variability



Adapted from PMID 23664091

Tumor heterogeneity

university of groningen

Era of personal genomics

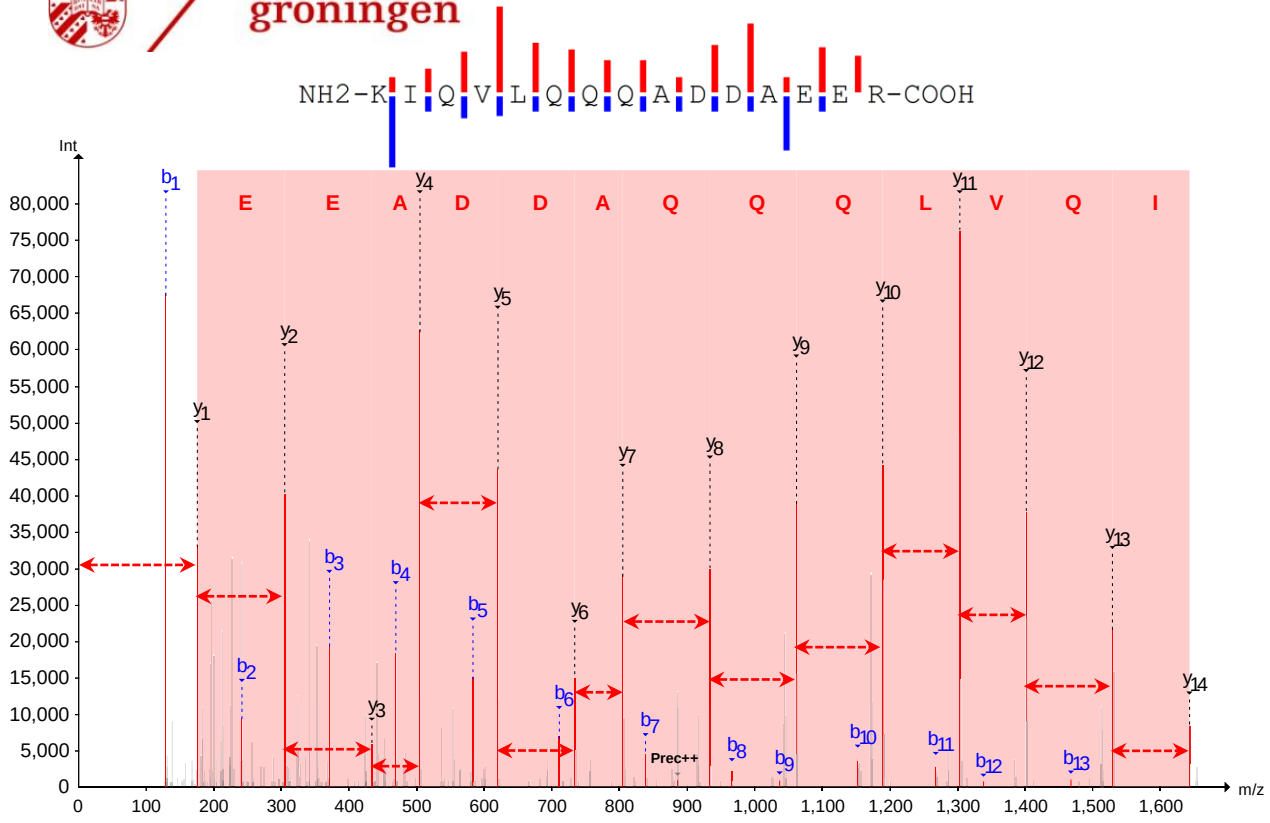
nature genetics
LETTERS
https://doi.org/10.1038/s41588-018-0273-y
OPEN

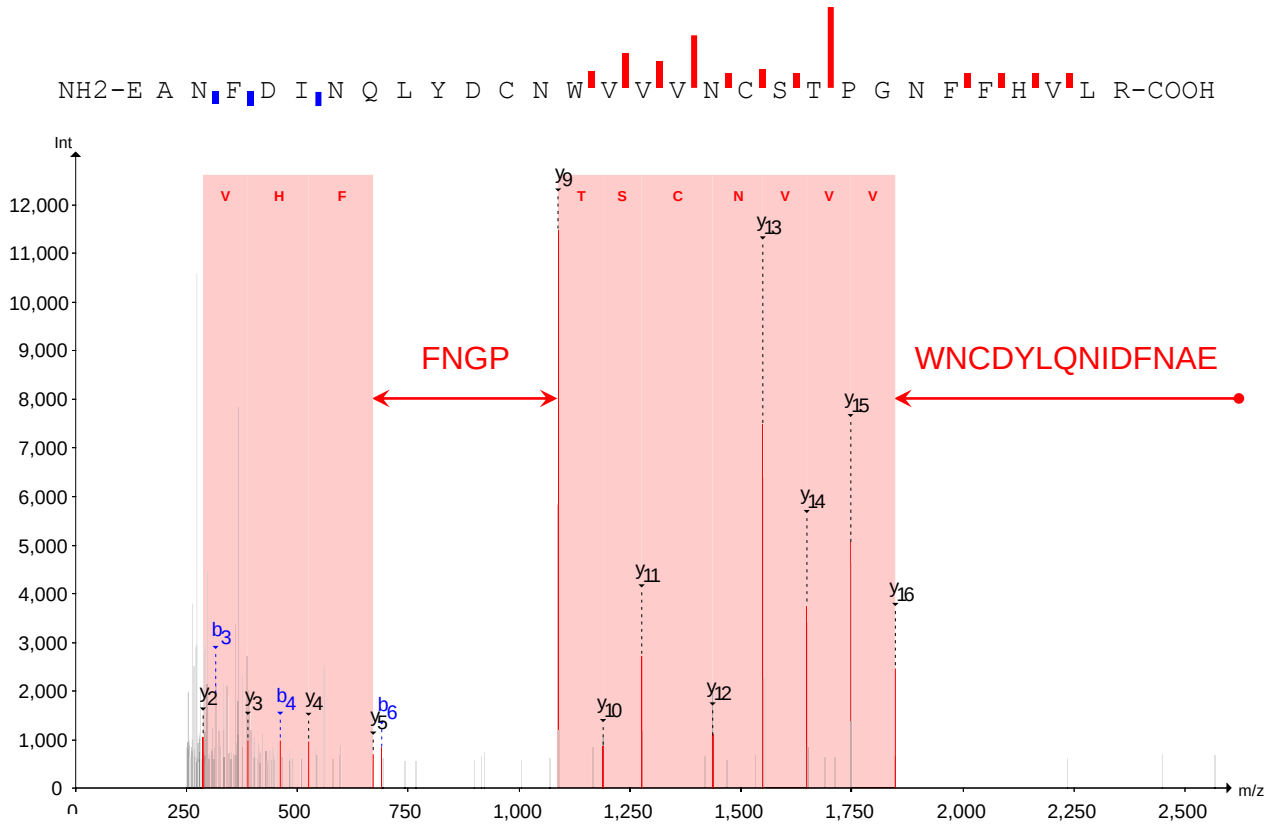
Assembly of a pan-genome from deep sequencing of 910 humans of African descent

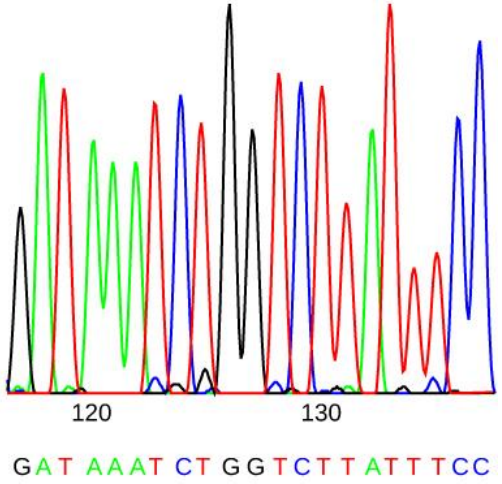
We used a deeply sequenced dataset of 910 individuals, all of African descent, to construct a set of DNA sequences that is present in these individuals but missing from the reference human genome. We aligned 1.19 trillion reads from the 910 individuals to the reference genome (GRCh38), collected all reads that failed to align, and assembled these reads into contiguous sequences (contigs). We then compared all contigs to one another to identify a set of unique sequences representing regions of the African pan-genome missing from the reference genome. Our analysis revealed 296,485,284 bp in 125,715 distinct contigs present in the populations of African descent, demonstrating that the African pan-genome contains ~10% more DNA than the current human reference genome. Although the functional significance of nearly all of this sequence is unknown, 387 of the novel contigs fall within 315 distinct protein-coding genes, and the rest appear to be intergenic.

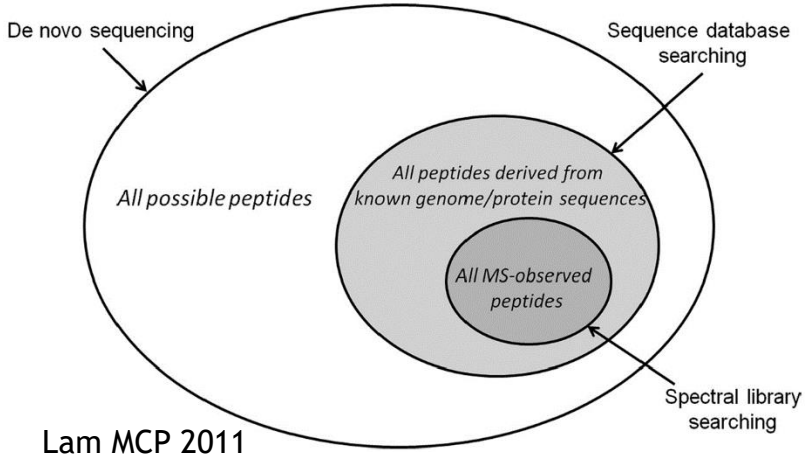
university of groningen

Difficulty in identifying variants de novo from MS/MS spectra







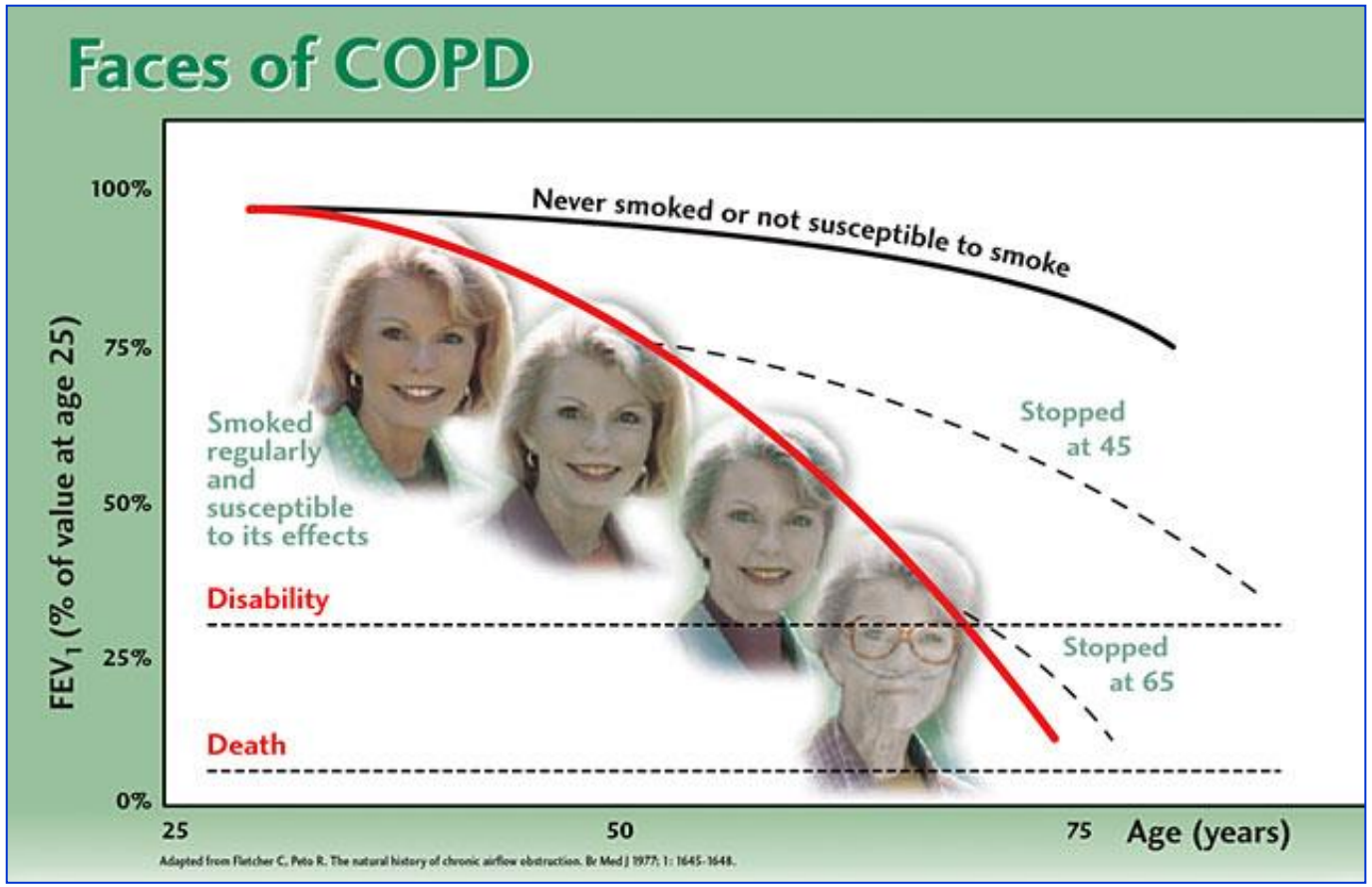


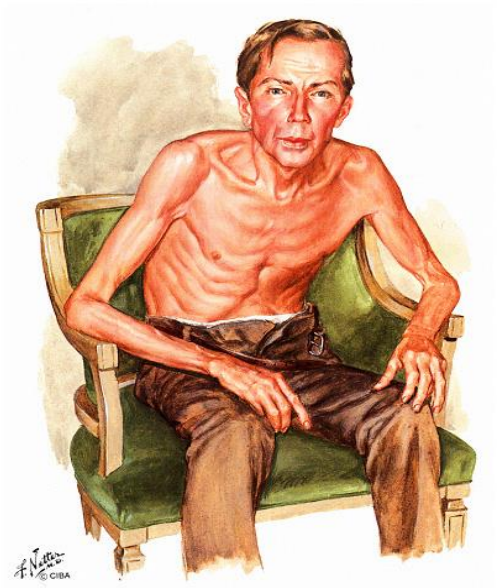
Lam MCP 2011

- Best identification method: database search (DBS)
- DBS requires list of protein sequences expected to be present in the sample
- Canonical sequences Swissprot and Uniprot (Ensembl) is used in common proteomics workflows
- In public databases low number of variants (20, 80 and 30 k proteins) are present.

university of groningen

Chronic Obstructive Pulmonary disease





- 20% of the smokers develop COPD, more than 200 millions persons have COPD
- Progressive loss of lung function with a large impact on the quality of life
- Insufficient insight in the molecular mechanisms of COPD
- Limited therapeutic options

university of groningen

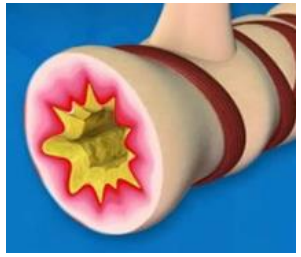
COPD phenotypes and treatment

chronic bronchitis

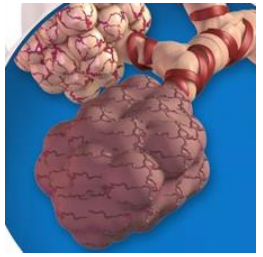
emphysema

Impaired lung function
Symptoms

airways



alveoli



Today treatment “**One size fits all**”

university of groningen

Study design

20/18 lung tissues

10 COPD stage IV (8 female/2 male)
10/8 ex-smokers control (6/4 female/4 male)

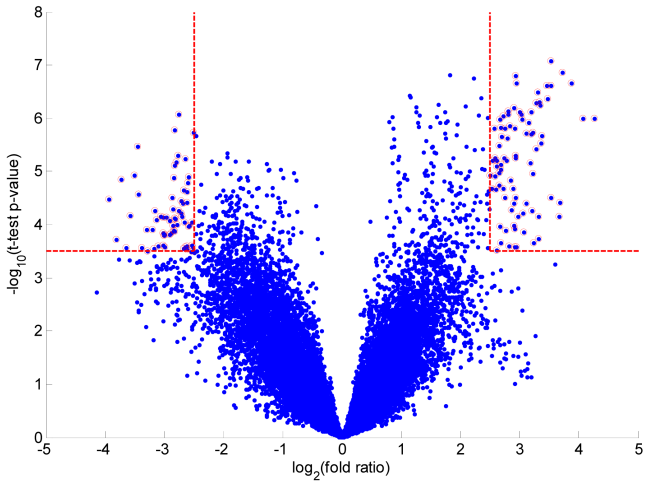
COPD
Control

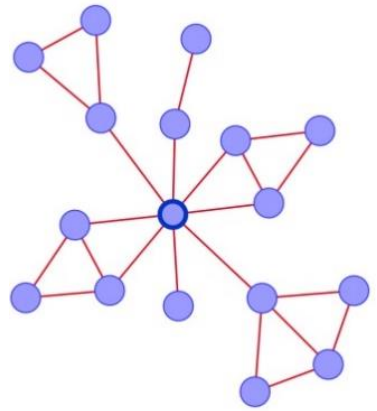
proteins

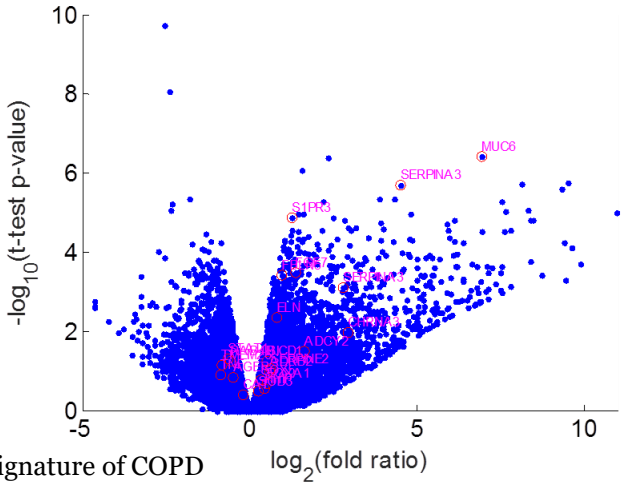
mRNA sequences

Orbitrap Q-Executive+
1D-LC-MS/MS
2D-LC-MS/MS

Illumina
20 million sequencing depth
Polyadenylated mRNA fraction







Brandsma CA, Guryev V, Timens W, ..., van den Berge M, Horvatovich P., Integrated proteogenomic approach identifying a protein signature of COPD and a new splice variant of SORBS1, Thorax, 2020, 75(2):180-183. PMID: 31937552

Proteogenomics data integration workflow

university of groningen

X-omics.nl

IBM

mRNA raw sequence
Illumina

FastQC
quality control

Trimmomatic
sequence trimming

- Annotated proteins
- SAAVs
- Splice isoforms
- indels
- New gene models
- New transcripts

STAR
alignment and annotation

TopHat2
alignment and annotation

Cufflinks
mRNA assembly and quantification

StringTie
mRNA assembly and quantification

TransDecoder
Translated protein sequence

mRNA quantitative table
Fasta format/FPKM

LC^2-MS/MS data
Orbitrap QE+

Grid
smoothing, resampling

Centroid
peak detection and quantification

Warp2D
time alignment

MetaMatch
peak matching

SearchGUI/PEAKS
peptide/protein identification

PeptideShaker/PEAKS
validation of peptide/protein identification

Peptide/protein quantitative table
Quantitative table (PSM, IC)

mRNA-Seq
mRNA-Seq pre-processing

Proteomics
MS1 quantification

Proteomics
peptide/protein identification

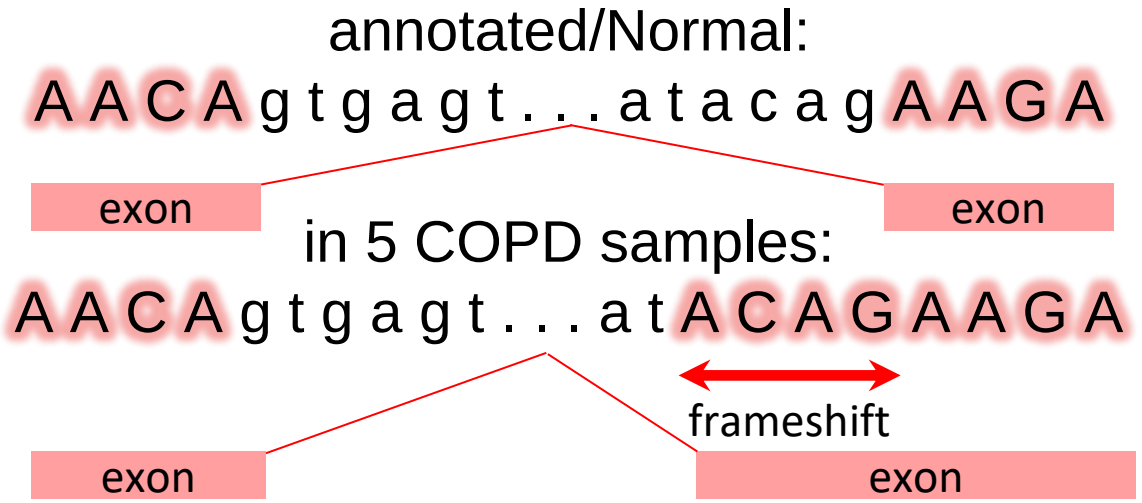
Input files
mRNA-Seq and LC-MS/MS

Output files
Protein/mRNA and quantities

nextflow nf-core

Suits F, Hoekman B, Rosenling T, Bischoff R, Horvatovich P., Threshold Avoiding Proteomics Pipeline, *Anal Chem.*, 2011, 83(20):7786-94.

university of groningen



One reason why transcript is present, and protein is absent

translated protein sequence of oncostatin M receptor gene without mutation

5'3' Frame 1

Met A L F A V F Q T T F F L T L L S L R T Y Q S E V L A E R L P L T P V S L K V S T N S T R Q S L H L Q W T V H N L P Y H Q E L K Met V F Q I Q I S R I E T S N V I W V G N Y S
T T V K W N Q V L H W S W E S E L P L E C A T H F V R I K S L V D D A K F P E P N F W S N W S S W E E V S V Q D S T G Q D I L F V F P K D K L V E E G T N V T I C Y V S R N
I Q N N V S C Y L E G K Q I H G E Q L D P H V T A F N L N S V P F I R N K G T N I Y C E A S Q G N V S E G Met K G I V L F V S K V L E E P K D F S C E T E D F K T L H C T W D
P G T D T A L G W S K Q P S Q S Y T L F E S F S G E K K L C T H K N W C N W Q I T Q D S Q E T Y N F T L I A E N Y L R K R S V N I L F N L T H R V Y L Met N P F S V N F E N V
N A T N A I Met T W K V H S I R N N F T Y L C Q I E L H G E G K Met Met Q Y N V S I K V N G E Y F L S E L E P A T E Y Met A R V R C A D A S H F W K W S E W S G Q N F T T L
E A A P S E A P D V W R I V S L E P G N H T V T L F W K P L S K L H A N G K I L F Y N V V V E N L D K P S S S E L H S I P A P A N S T K L I L D R C S Y Q I C V I A N N S V G A
S P A S V I V I S A D P E N K E V E E E R I A G T E G G F S L S W K P Q P G D V I G Y V V D W C D H T Q D V L G D F Q W K N V G P N T T S T V I S T D A F R P G V R Y D F R I
Y G L S T K R I A C L L E K K T G Y S Q E L A P S D N P H V L V D T L T S H S F T L S W K D Y S T E S Q P G F I Q G Y H V Y L K S K A R Q C H P R F E K A V L S D G S E C C K
Y K I D N P E E K A L I V D N L K P E S F Y E F F I T P F T S A G E G P S A T F T K V T T P D E H S S Met L I H I L L P Met V F C V L L I Met V Met C Y L K S Q W I K E T C Y P D I
P D P Y K S S I L S L I K F K E N P H L I I Met N V S D C I P D A I E V V S K P E G T K I Q F L G T R K S L T E T E L T K P N Y L Y L L P T E K N H S G P G P C I C F E N L T Y N Q
A A S D S G S C G H V P V S P K A P S Met L G L Met T S P E N V L K A L E K N Y Met N S L G E I P A G E T S L N Y V S Q L A S P Met F G D K D S L P T N P V E A P H C S E Y
K Met Q Met A V S L R L A L P P P T E N S S L S S I T L L D P G E H Y C Stop

with mutation

5'3' Frame 1

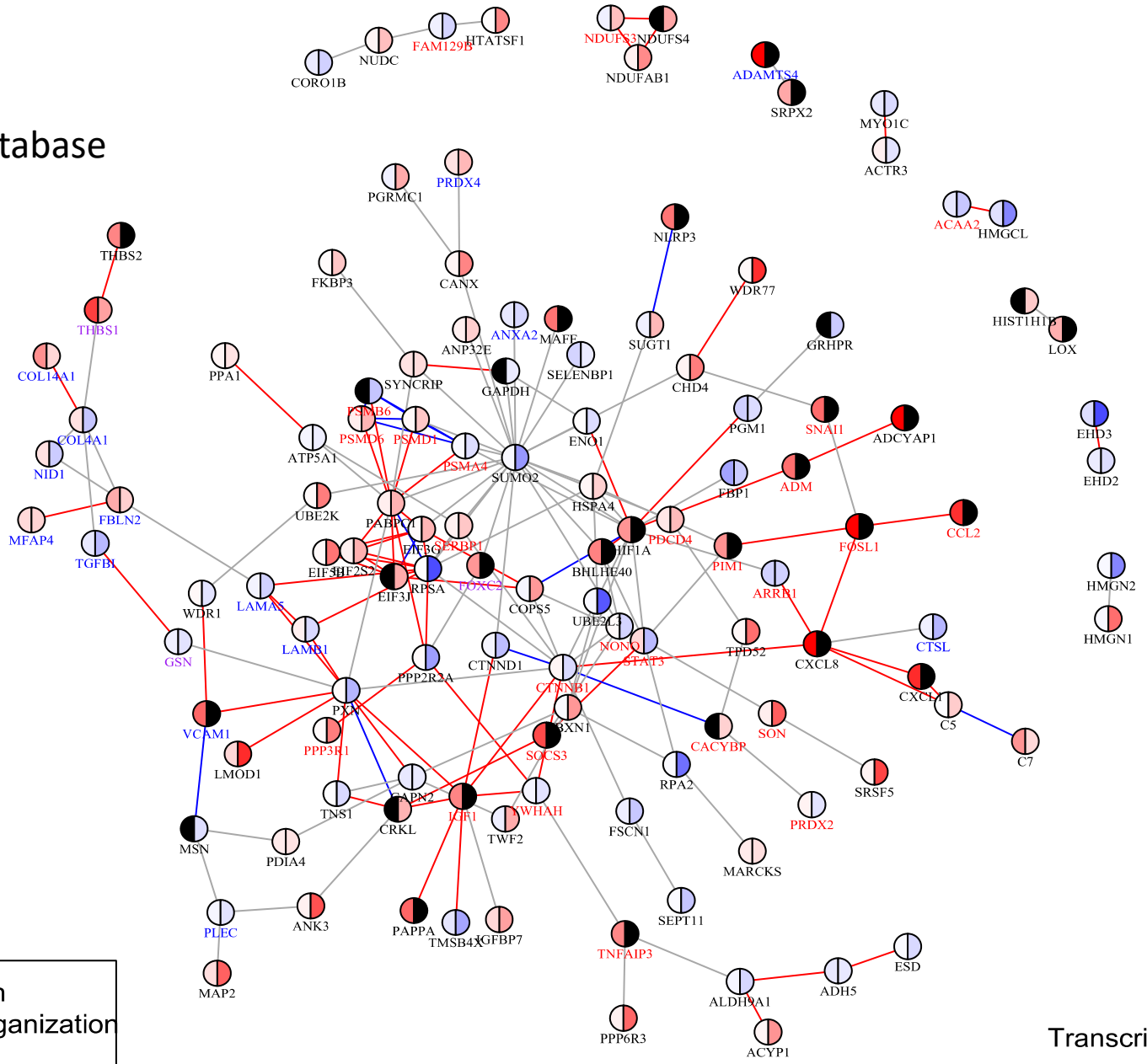
Met A L F A V F Q T T F F L T L L S L R T Y Q S E V L A E R L P L T P V S L K V S T N S T R Q S L H L Q W T V H N L P Y H Q E L K Met V F Q I Q I S R I E T S N V I W V G N Y S
T T V K W N Q V L H W S W E S E L P L E C A T H F V R I K S L V D D A K F P E P N F W S N W S S W E E V S V Q D S T G Q D I L F V F P K D K L V E E G T N V T I C Y V S R N
I Q N N V S C Y L E G K Q I H G E Q L D P H V T A F N L N S V P F I R N K G T N I Y C E A S Q G N V S E G Met K G I V L F V S K V L E E P K D F S C E T E D F K T L H C T W D
P G T D T A L G W S K Q P S Q S Y T L F E S F S G E K K L C T H K N W C N W Q I T Q D S Q E T Y N F T L I A E N Y L R K R S V N I L F N L T H R V Y L Met N P F S V N F E N V
N A T N A I Met T W K V H S I R N N F T Y L C Q I E L H G E G K Met Met Q Y N V S I K V N G E Y F L S E L E P A T E Y Met A R V R C A D A S H F W K W S E W S G Q N F T T L
E A A P S E A P D V W R I V S L E P G N H T V T L F W K P L S K L H A N G K I L F Y N V V V E N L D K P S S S E L H S I P A P A N S T K L I L D R C S Y Q I C V I A N N S V G A
S P A S V I V I S A D P E N N E R G Stop G R K N C R H R G W I L S V L E T P T W R C Y R L C C G L V Stop P Y P G C A R Stop F P V E E C R S Q Y H K H S H Stop H R C F
Stop A R S S I Stop L Q N L W V I Y K K D C L F I R E K N R I L S G T C S F R Q P S R A G G Y I D I P L L H S E L E R L L Y Stop I S T W F Y T R V P C L S E I Q G E A V P P T I
Stop K G S S F R W F R Met L Q I Q N Stop Q P G R K G I D C G Q P K A R I L L Stop V F H H S I H Stop C W Stop R P Q C Y V H E G H D S G Stop T L L D A D S Y P T A H G F
L R L A H H G H V L L E K S V D Q G D L L S Stop H P Stop P L Q E Q H P V I N K I Q G E P S P N N N E C Q Stop L Y P R C Y Stop S C K Q A R R D K D T V P R H Stop E V T H
R N R V D Stop A Stop L P L S P S N R K E S L W P W P L H L F Stop E L D L Stop P G S F Stop L W L L W P C S S I P K S P K Y A G T N D L T Stop K C T K G T R K K L H E L P
G R N P S W R N K F E L C V P V G F T H V W R Q G Q S P N K P S R G T T L F R V Stop N A N G S L P A S C L A S P D R E Stop Q P L L N Y P F R S R Stop T L L L

university of groningen

STRING protein-protein interaction network

STRING edges:
experiments and database
- experiments
- database
- both

Proteomics +
transcriptomics
FDR 0.01
Fixed layout



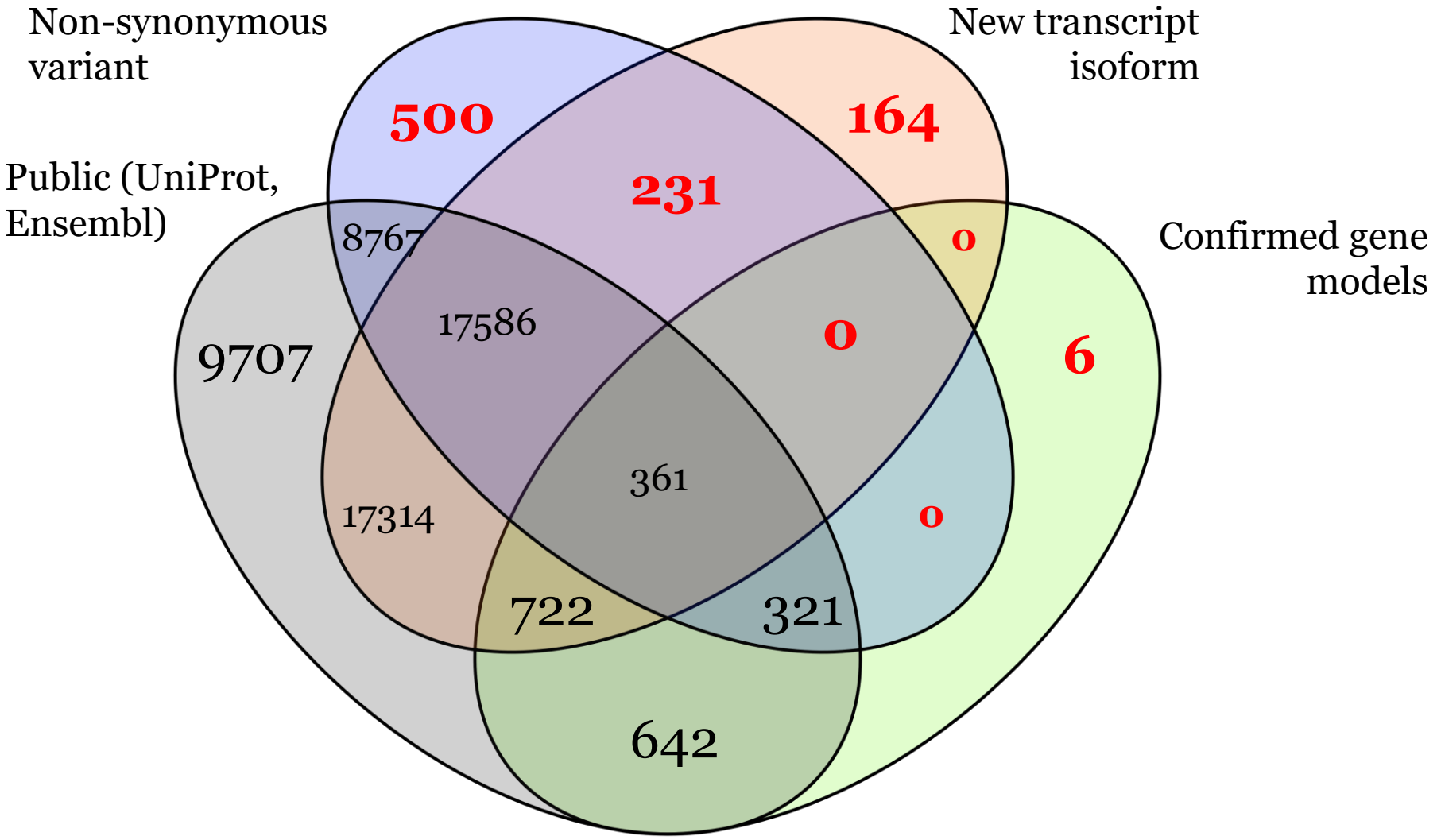
Fold Change
Up[4+]
Zero
Down[4-]

- regulation of cell death
- extracellular matrix organization
- common

Transcripts
Proteins

university of groningen

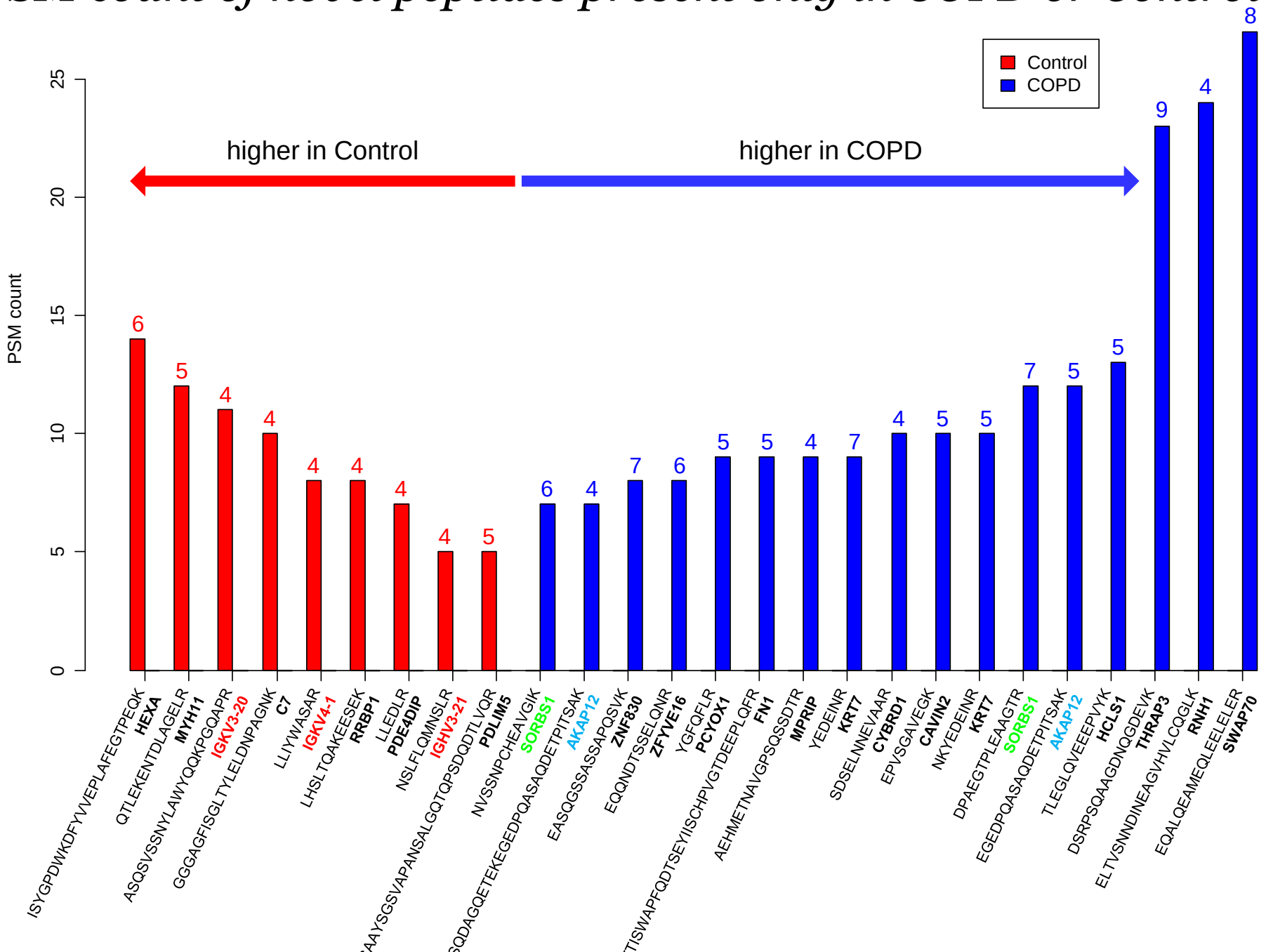
Peptide identification (combined dataset)



901 peptides are not in curated public databases

university of groningen

PSM count of novel peptides present only in COPD or Control

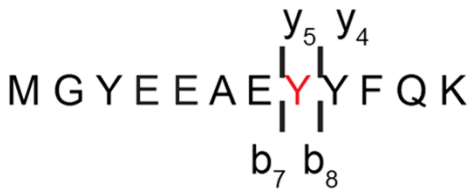


university of groningen

Group only "Novel peptides" identification PEAKS score and SpetcrumAI test

Peptides only present in COPD

Peptide sequence	Quality score	Gene	Effect	Ion support	Flanking ion support
ADSQDAGQETEKEGEDPQASAQDETPITSAK	134.83	AKAP12	E1600D amino acid substitution	×	×
DSRPSQAAGDNQGDEVK	132.86	THRAP3	A201V amino acid substitution	×	×
TGQEALSQTTISWAPFQDTSEYIISCHPVGTDEEPLQFR	131.44	FN1	Native	NA	NA
AEHMETNAVGPSQSSDTR	129.43	MPRIP	P327Q amino acid substitution	×	×
EGEDPQASAQDETPITSAK	129.34	AKAP12	E1600D amino acid substitution	×	×
DPAEGTPLEAAGTR	118.25	SORBS1	Splice variant, exon extension	NA	NA
ELTVSNNDINEAGVHVLCQGLK	115.80	RNH1	R188H amino acid substitution	×	×
TLEGLQVEEEPVYK	106.37	HCLS1	E361K amino acid substitution	×	×
EASQGSSASSAPQSVK	105.60	ZNF830	H99Q amino acid substitution	×	×
SDSELNNEVAAR	103.81	CYBRD1	S266N amino acid substitution	×	×
EQALQEAMEQLEELELER	100.83	SWAP70	Q505E amino acid substitution	×	×
EQQNDTSSELQNR	91.55	ZFYVE16	I192T amino acid substitution	×	×
YGFQFLR	81.46	PCYOX1	S149F amino acid substitution	×	×
NKYEDEINR	70.58	KRT7	H186R amino acid substitution	×	×
EPVSGAVEGK	62.17	CAVIN2	Q174E amino acid substitution	×	×
NVSSNPCHEAVGIK	56.53	SORBS1	Splice variant, exon extension	NA	NA
YEDEINR	49.80	KRT7	H186R amino acid substitution	×	×



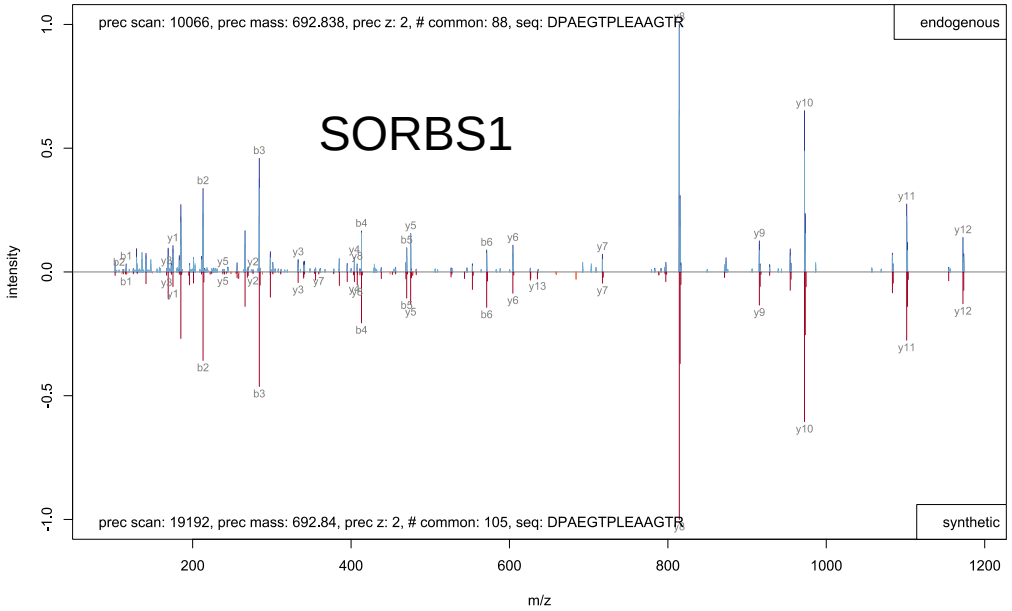
Peptides only present in Control

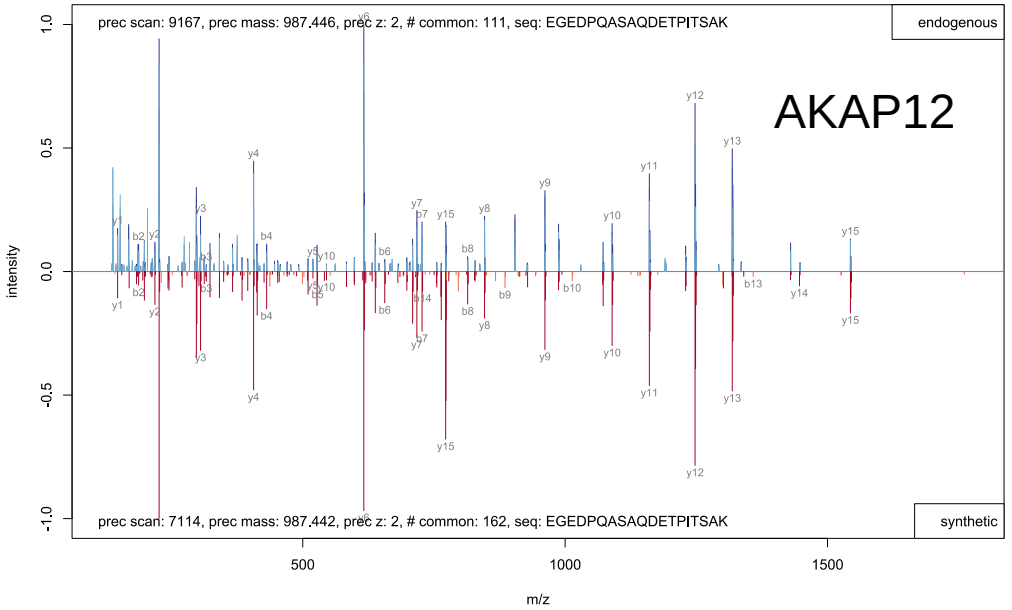
Peptide sequence	Quality score	Gene	Effect	Ion support	Flanking ion support
ISYGPDWKDFYVVEPLAFEGTPEQK	120.51	HEXA	I447V amino acid substitution	×	×
ISNSAAYSGSVAPANSALGQTQPSDQDTLVQR	110.84	PDLIM5	T410A amino acid substitution	×	×
LHSLTQAKEESEK	110.58	RRBP1	L1043H amino acid substitution	×	-
GGGAGFISGLTYLELDNPAGNK	108.34	C7	S389T amino acid substitution	×	×
ASQSVSSNYLAWYQQKPGQAPR	105.88	IGKV3-20	S52N amino acid substitution	×	×
QTLEKENTDLAGELR	77.52	MYH11	A1241T amino acid substitution	×	×
NSLFLQMNSLR	76.68	IGHV3-43	Y99F amino acid substitution	×	×
LLIYWASAR	68.41	IGKV4-1	T79A amino acid substitution	×	×
LLEDLR	33.62	OTOA/PDE4DIP	Native	NA	NA

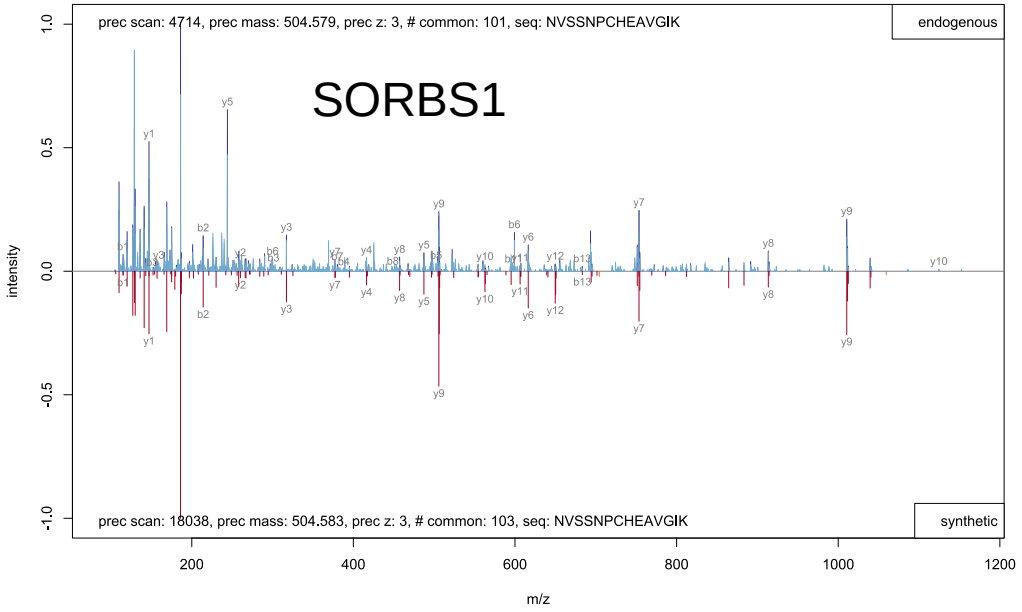
Zhu Y, Orre LM, Johansson HJ, Huss M, Boekel J, Vesterlund M, Fernandez-Woodbridge A, Branca RMM, Lehtiö J. Discovery of coding regions in the human genome by integrated proteogenomics analysis workflow. Nature Commun. 2018 Mar 2;9(1):903.

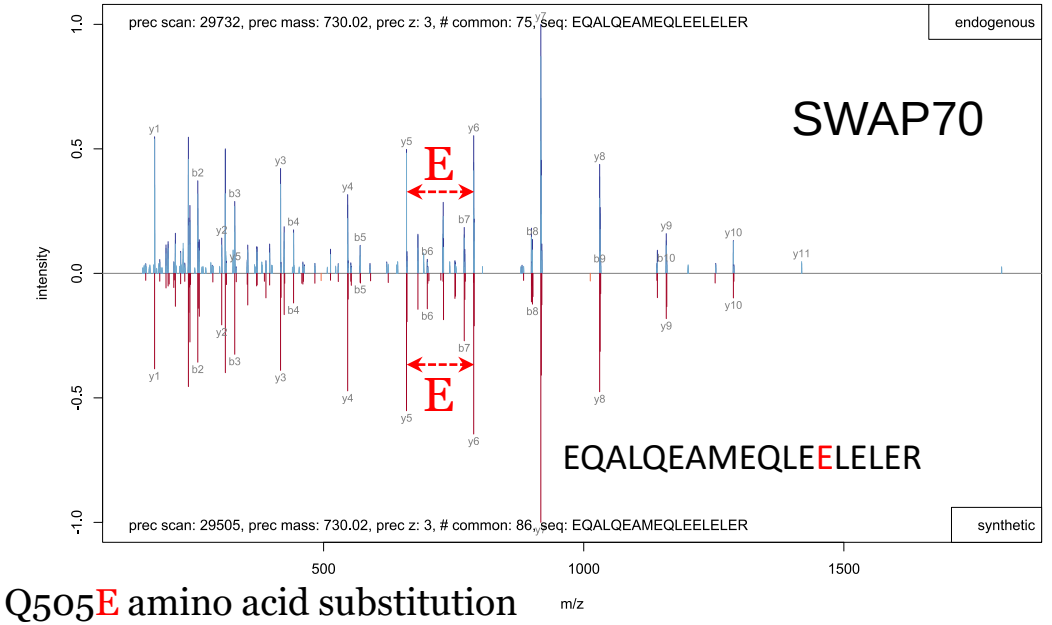
university of groningen

Validation of non-reference peptides with synthetic peptides MS/MS spectra







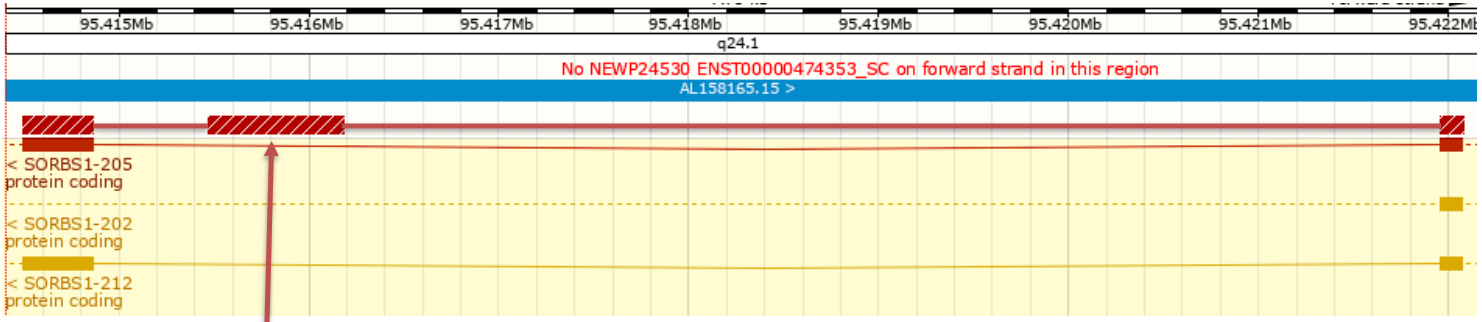


Q505E amino acid substitution

university of groningen

Two peptides uniquely mapping to a novel exon of SORBS1 gene

SORBS1 is encoded on minus strand of chr10 (band 10q24.1), between 95.31 and 95.56 Mbp (gene length: 249.64 kb)



Exon 8 — New exon — Exon 7

Adds 239 AAs between residues 270 and 271

2 peptides were found by MS/MS analysis

>NEWP24530

MSSECDGGSKAVMNGLAPGSNGQDKATADPLRARSISAVKIIPVKTVKNASGLVLPTDMD
LTKICTGKGAVTLRASSSYRETPSSSPASPQETRQHESKPGLEPEPSSADEWRLSSSADA
NGNAQPSSLAAKGYRSVHPNLPSDKSQDATSSSAAQPEVIVVPLYLVNTDRGQEGTARPP
TPLGPLGCVPTIPATASAASPLTFPTLDDFIPPHLQRWPHHSQPARASGSFAPISQTPPS
FSPPPPLVPPAPEDLRRVSEPDLTGAVSSTVLSPPRPPLPQKDRFAWQSPTIHNTYKDSL
YLSSPKPYVPLGTPRQQNPSQPQPISVLLAAGSAPKGVVCPGSLLPDSTFPSASSQPQQR
YAATRTVYHKNVSSNPCHEAVGIKKVSSLYVPCLSNNICLAASENSSRVARDPAEGTPLE
AAGTRAPAPGLVSRTAGTGKPPPAPPPDPPKLFFDIRKDAVNRGESPSLGTQASFPDVRP
PVLGPRVTSDPENRKSKESYLLQPSYPAK**DSSPLLNEVSSSLIGTDSQAFPSVSKPSSAY**
PSTTIVNPTIVLLQHNREQQKRLSSLSDPVSERRVGEQDSAPTQEKPTSPGKAIEKRAK...

Exons 1-7

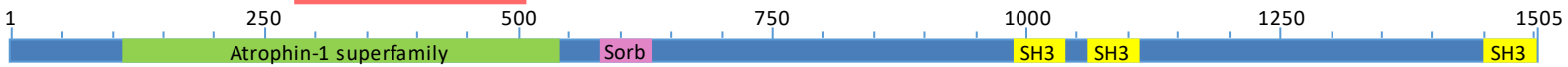
New exon

Exons 8 and further

Protein domains of SORBS1 splice variants

native SORBS1

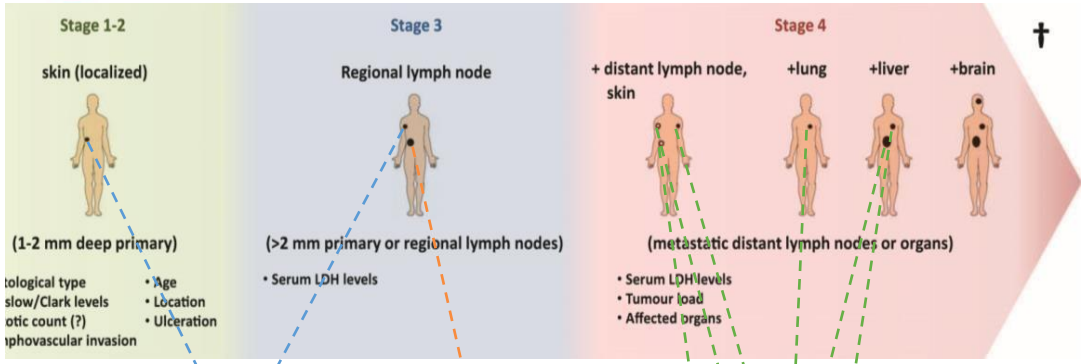
novel SORBS1 splice variant



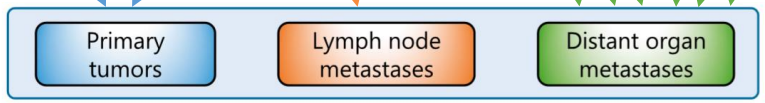
Cancer moonshot for melanoma

university of groningen

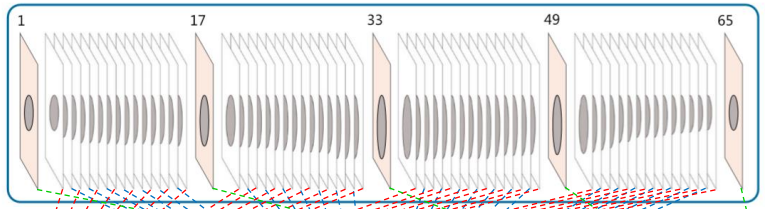
Clinically well characterized patients



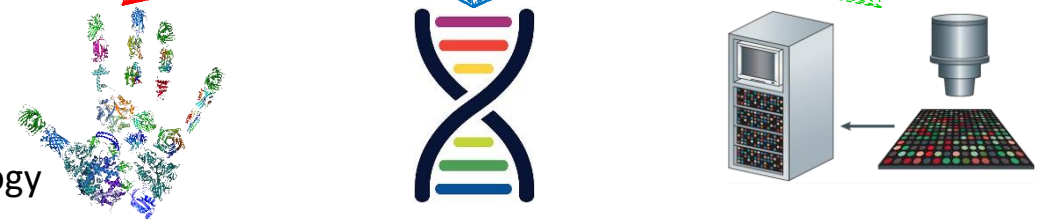
Biobanks



Tumor sections

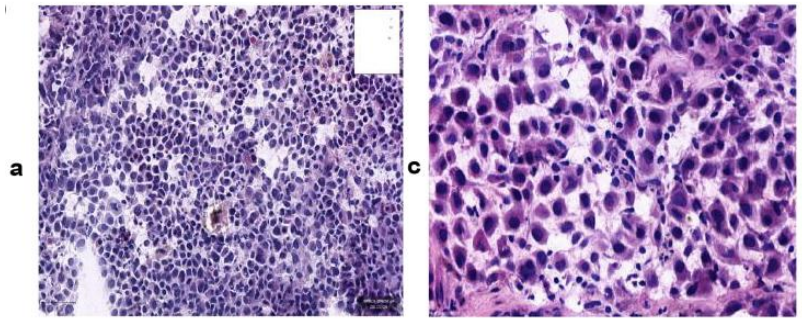


Multi-omics profiles & digital pathology

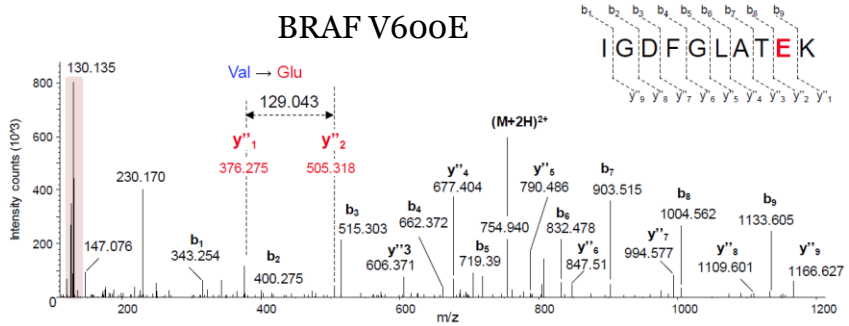


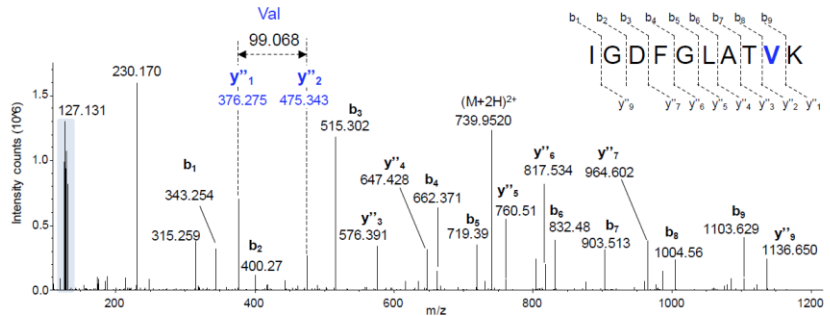
Proteomics | Genomics | Digital pathology

PMIDs: 31599373, 30914758, 30900145

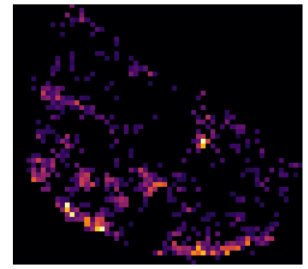


BRAF V600E





Vemurafenib melanoma lymph node metastasis



491.085 Da

Fundings and Collaborations

university of groningen

ERIBA

Yanick Hagemeijer
Ekaterina Ovchinnikova
Victor Guryev

university of groningen

Alejandro Sanchez Brotons
Karel Gerbrands
Ana Ciconelle
Hjalmar Permentier
Rainer Bischoff

umcg GRIAC Groningen

Corry-Anke Brandsma
Maarten van den Berge
Wim Timens
Karin Wolters
Dirkje Postma
Gyorgy Halmos
Renee Vehoeven

Maria Yakovleva
Melinda Rezeli
Jonatan Eriksson
Thomas Fehniger
György Markó-Varga

IBM

Frank Suits

X-omics.nl

ERIBA

NWO