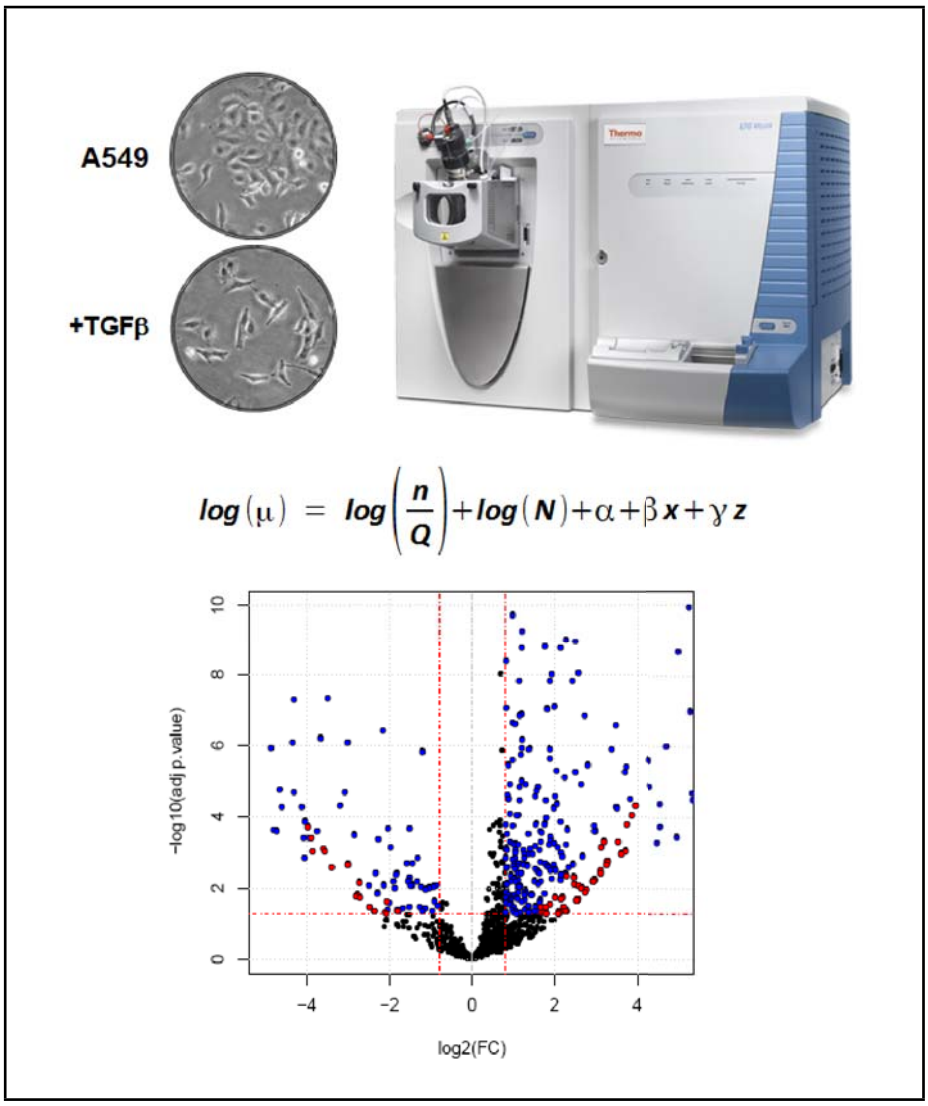
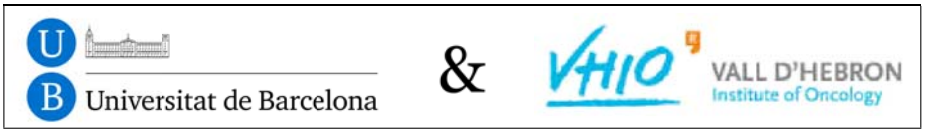

# Statistical Methods for the Modelling of Label-Free Shotgun Proteomic Data in Cell Line Biomarker Discovery

Josep Gregori i Font

# Statistical Methods for the Modelling of Label-Free Shotgun Proteomic Data in Cell Line Biomarker Discovery




**Josep Gregori i Font**




# Statistical Methods for the Modelling of Label-Free Shotgun Proteomic Data in Cell Line Biomarker Discovery

Mètodes estadístics pel modelat de dades de proteòmica sense marcatge, en descobriment de biomarcadors sobre línies cel·lulars

Memòria presentada per Josep Gregori i Font
per optar al grau de doctor per la Universitat de Barcelona

Tesi realitzada al laboratori de Biomarcadors Tumorals
del Institut d'Oncologia de la Vall d'Hebron

Signatures

EL DOCTORAND:

Josep Gregori i Font .....

ELS DIRECTORS DE LA TESI:

Alexandre Sánchez i Pla (UB) .....

Josep Villanueva i Cardús (VHIO) .....

EL TUTOR:

Jorge Ocaña i Rebull .....

Programa de doctorat en estadística
Departament d'Estadística, Facultat de Biologia
Universitat de Barcelona

A Montse, Marc i Laia, amb joia i agraïment.
Als petits Mar i Guillem, esperança de noves fites.

# Acknowledgments

*"No one can whistle a symphony; it takes a whole orchestra to play it"*
Halford E. Luccock
*"If people are not making mistakes, they are not trying new things."*
Walter C. Wright Jr.

It is my pleasure to thank all those people who contributed to this exciting second experience in the making of a doctoral thesis.

First and foremost I am heartly thankful to my supervisor, **Prof. Alex Sanchez**, who made possible for me this sort of rebirth to the top wave of the science after a long professional life in industry. Thanks for your instruction and advice at the UOC, thanks for your guidance during these last years at VHIR-VHIO. I am so thankful to my second supervisor and PI at the Biomarkers Lab, **Dr. Josep Villanueva**. Thank you for allowing me to pursue all those things I tried to pursue, thank you for all those MS/MS experiments I needed in this work, thank you for your enthusiasm and for believing in the project.

These thanks are extended to the team at the lab, **Dr. Laura Villareal**, the formula 1 driver of the brave Orbitrap, **Dr. Olga Mendez**, the Mycosomes Blaster, caring about all that is biological in our lab, **Dr. Theodora Katsila**, our cheerful breath of Greek fresh air, and **Candida Salvans** our indispensable technician.

This day could not have arrived without the complicity and generosity of my employers. Many thanks have to be given to **Roche Diagnostics**, and specially to **Dr. Artur Palet**, my boss. Many thakns also to **Dr. Josep Quer** and **Dr. Francisco Rodriguez** at the VHIR lab where I use to analyse and process data from NGS sequencing of merciless virus like the HCV and the HBV.

And last but certainly not least my thanks and love to **Montse**, my wife, who patiently supported plenty of holidays, week-ends, and even vacations devoted to this work.

Thanks to all those who I don't mention and who contributed in some manner to this modest achievement.

# Abstract

A data analysis pipeline for the discovery of biomarkers on cancer cell lines by label-free shotgun proteomics has been developed and implemented in this thesis. Specifically the solution has been optimized for the analysis of secretomes of cancer cell lines measured in spectral counts by LC-MS/MS. Along the development it has been shown the incidence and relevance of batch effects in the comparative analysis of label-free proteomics by LC-MS/MS. Also the features providing reproducibility to potential biomarkers have been identified. The model has been developed on empirical data obtained from a series of spiked experiments, and with the help of simulations, to evaluate its performance. The pipeline comprises an exploratory data analysis (EDA) R/Bioconductor package based on multidimensional analysis tools, and a R/Bioconductor inference package based on generalized linear models (GLM) with Poisson or negative binomial distributions, or the quasi-likelihood GLM extension. Two graphical interfaces have also been developed to ease the use of the provided solution in a MS lab by non experts. The designed model is devised to discover differentially expressed proteins in cancer cell line secretomes, using the cell as the unit of interest. The model allows blocking factors as a mean for batch effects correction. The normalization to cell units is embedded in the model through the use of offsets, and no previous data treatment is required. The two packages developed are called `msmsEDA` and `msmsTests`, and allow for:

- Dataset quality assessment.
- The identification of outliers
- The identification of confounding factors or batch effects.
- The discovery of potential biomarkers by using the distribution best fitting the available data.
- Improving the degree of reproducibility by a post test filter based of effect size and signal levels.

Different papers have been published in proteomics journals both, developing each data treatment step, and demonstrating its use and value in biological experiments carried out in our lab at VHIO.

# Compendi

En la tesi s'ha desenvolupat, dissenyat i implementat una solució per l'anàlisi de dades de proteòmica en descobriment de biomarcadors. Específicament la solució s'ha optimitzat per l'anàlisi de secretomes de línies cel·lulars de càncer. Durant el desenvolupament de la metodologia s'ha demostrat la incidència i rellevància dels efectes batch en l'anàlisi comparatiu de pèptis sense marcar per LC-MS/MS. Així com les característiques que identifiquen un potencial biomarcador com a reproductible. Els models s'han desenvolupat amb l'ajut de dades empíriques obteses de mostres amb mescles controlades de proteïnes, i de simulacions. La solució informàtica que implementa el model desenvolupat consta de dos paquets R/Bioconductor, amb les respectives interfícies gràfiques que faciliten el seu ús a no experts. El primer paquet, `msmsEDA`, consta de funcions útils en l'anàlisi exploratòria de dades, i permet avaluar la qualitat del conjunt de dades d'un experiment de LC-MS/MS basat en comptatge d'espectres, així com explorar l'eventual presència de valors extrems, factors de confusió, o d'efectes batch. El segon paquet, `msmsTests`, encapsula funcions per la inferència en el descobriment de biomarcadors. Els tests ajusten un model GLM que permet la inclusió de factors per blocs com a mecanisme de correcció d'efectes batch, i que incorpora una normalització generalitzada mitjançant la tècnica dels offsets que permet la comparació de secretoma al nivell d'una cel·lula. Les distribucions implementades són la de Poisson i la binomial negativa, així com l'extensió de la quasiversemblança. En conjut, el model desenvolupat i la implementació informàtica que se'n ha fet permet:

- Avaluar la qualitat d'un conjunt de dades de LC-MS/MS basades en SpC.
- Identificar valors extrems.
- Identificar la presència de factors de confusió o d'efectes batch.
- El descobriment de biomarcadors emprant la distribució que millor s'ajusti a les dades.
- Asegurar un bon nivell de reproductibilitat mercès a un filtre post-test que té en compte la intensitat del senyal i la mida de l'efecte.

Els paquets, llur documentació i tutorials, estan disponibles a bioconductor.org, i les interfícies gràfiques amb tutorials i manuals d'usuari a github.com.

S'han publicat articles en revistes internacionals de proteòmica desenvolupant cada pas en el tractament de dades, i demostrant el seu ús i aplicabilitat en experiments biològics de rellevància clínica.

# Contents

# List of Figures

# List of Tables

# Acronyms and symbols

**ASCO** American Society of Clinical Oncology

**BD** Biomarker discovery

**DEP** Differentially expressed protein

**DGE** Digital gene expression

**DSP** Differentially secreted protein

**EDA** Exploratory data analysis

**EGF** Epithelial growth factor

**EGFR** Epithelial growth factor receptor

**ELISA** Enzyme-Linked ImmunoSorbent Assay

**EMT** Epithelial to mesenchymal transition

**ESI** Electrospray ionization

**FBS** Fetal bovine serum

**FC** Fold change

**FDA** Food and Drugs Administration

**FDR** False discovery rate

**FFPE** Formalin-fixed paraffin-embedded

**FISH** Fluorescence in situ hybridation

**GLM** Generalized linear model

**GLMM** Generalized linear mixed effects model

**HC** Hierarchical clustering

**HPLC** High performance liquid chromatography

**HUPO** Human Proteome Organization

**ICAT** Isotope-coded affinity tags

**ITRAQ** Isobaric tags for relative and absolute quantitation

**IVDMIA** In-vitro diagnostic multivariate index assay

**LC** Liquid chromatography

**LC-MS** Liquid chromatography-mass spectrometry

**LC-MS/MS** liquid chromatography-tandem mass spectrometry

**logFC** Base 2 log fold change

**MA** Microarray

**MALDI** Matrix-assisted laser desorption/ionization

**MAQC** Microarray Quality Control Consortium

**MS** Mass spectrometry

**m/z** Mass to charge ratio

**NGS** Next generation sequencing

**NIH** National Institutes of Health

**NSAF** Normalized spectral abundance factor

**OTU** Operational taxonomic unit

**PAF** Protein abundance factor

**PAGE** Polyacrylamide gel electrophoresis

**PAI** Protein abundance index

**PLGEM** Power Law Global Error Model

**PCA** Principal Components Analysis

**PSA** Prostate-Specific Antigen

**QL** Quasi-likelihood extension of GLM

**QLLL** Quasi-likelihood with log link

**RNA-seq** Differential gene expression by NGS

**RP** Reverse phase chromatography

**SAGE** Serial analysis of gene expression

**SCX** Strong cation exchange chromatography

**SDS-PAGE** Sodium Dodecyl Sulfate Polyacrylamide gel electrophoresis

**SEC** Size exclusion chromatography

**SILAC** Stable isotope labeling by amino acids in cell culture

**SpC** Spectral count

**TGFb** Transforming growth factor beta

**TOF** Time of flight

# Capítol 1

# RESUM EN CATALÀ

Aquest resum del treball presentat consisteix en una introducció, que situa i contextualitza el treball desenvolupat, un apartat on es fixen els objectius, una secció de resultats on es presenten els treballs publicats i el software desenvolupat, una discussió on s'expressen mancances i limitacions, i s'assenyalen possibles noves vies de recerca, i una final de conclusions on es recullen les aportacions al camp estudiat.

## 1.1 Introducció

En el que portem de segle s'ha viscut l'emergència d'un nombre de tecnologies d'alt rendiment en el camp de les *òmiques*, que han despertat altes expectatives com a eines en el descobriment de biomarcadors útils en diagnòstic i pronòstic clínics. Aquestes tecnologies han comportat l'emergència d'un nou paradigma estadístic, el de *moltes-variables-poques-rèpliques* que constitueix un problema de dimensionalitat [Bellman, 1961]. La primera d'aquestes tecnologies, els microarrays d'expressió (MA), emprats per mesurar d'un sol cop el transcriptoma complet d'una mostra biològica, s'ha enfrontat més que cap altra als reptes de no decebre en les seves capacitats, i de resoldre els corresponents reptes estadístics i d'anàlisi de dades.

Si bé des del principi es va reconèixer la necessitant d'un bon disseny experimental en els estudis amb microarrays, també es van anar evidenciant problemes de reproductibilitat en biomarcadors que havien estat publicats amb altes taxes de sensibilitat i especificitat [Kuo et al., 2002; Tan et al., 2003; Ransohoff, 2004; Marshall, 2004; Dupuy & Simon, 2007; Borst & Wessels, 2010; Baggerly & Coombes, 2009; Potti et al., 2011]. Com a resposta a l'aparent atzucac es va constituir el MicroArray Quality Control Consortium (MAQC) amb autoritats acadèmiques,

de l'administració (FDA), i de la indústria per avaluar els factors que podien intervenir en la reproductibilitat de resultats entre diferents laboratoris i diferents plataformes [Shi et al., 2006] (figura 2.1), i per construir i avaluar discriminadors basats en resultats de microarrays [Shi et al., 2010].

Aquest treball pretén explorar temes específics en proteòmica comparativa, i traslladar i implementar algunes de les lliçons primer apreses amb microarrays en la fase de descobriment de biomarcadors.

### 1.1.1 Biomarcadors

Un biomarcador es pot definir com una molècula o conjunt de molècules, quin nivell és indicatiu d'un estat biològic [Madu & Lu, 2010]. En clínica un biomarcador és rellevant com a predictor de l'estat d'una malaltia (diagnòstic) o de l'evolució de la malatia (pronòstic). Els biomarcadors de diagnòstic són mesures basals que proporcionen informació sobre quins pacients es poden beneficiar d'un tractament. Els biomarcadors de pronòstic són mesures de pre-tractament informatives sobre l'evolució de la malaltia i el seu resultat en el llarg termini. En aquest sentit es poden considerar com a mesures de risc que poden aconsellar tractaments més agressius [Madu & Lu, 2010].

El descobriment de biomarcadors (BD) constitueix només el primer pas en el seu desenvolupament. Un procés complex i llarg que inclou les següents etapes [Rifai et al., 2006]:

- **Descobriment**: Identificar candidats a biomarcadors.
- **Qualificació**: Confirmar l'expressió diferencial en el fluid biològic d'interès per una tècnica de més baix rendiment i més exacta.
- **Validació**: Determinar la sensibilitat i l'especificitat amb mostres independents.
- **Desenvolupament en assaig clínic**: Establir la sensibilitat i l'especificitat amb una cohort suficient, i independent, donades les regles d'eligibilitat. I optimització de l'assaig.

### 1.1.2 Proteòmica i línies cel·lulars

Mentre que el DNA conté informació completa sobre els plans de la cèl·lula, i el mRNA fa la funció de missatger, amb peces d'informació enviades a la maquinària

de la cèl·lula, les proteïnes són la part funcional i reflecteixen més acuradament el fenotip. Aquesta idea ve reforçada per l'escassa correlació observada entre els mRNA i les proteïnes de la cèl·lula [Gygi et al., 1999]. Ja que les proteïnes representen un nivell funcional més alt que els mRNA, s'espera que la proteòmica pugui portar al descobriment de dianes terapèutiques i biomarcadors d'una manera més efectiva que la transcriptòmica ho ha fet. Tanmateix aquest camp encara presenta un conjunt de reptes a resoldre.

La gran complexitat del proteoma d'un teixit, i l'enorme rang d'abundàncies que s'hi observa, desaconsellen una anàlisi directa atès que la fracció de proteïna que pot estar específicament relacionada amb una malaltia es condidera gaire bé negligible. L'alternativa del sèrum, que seria molt apreciada en permetre proves no invasives, presenta semblants dificultats (figura 2.4) ja que les 22 proteïnes més abundants representen el 99% del seu proteoma [Tirumalai et al., 2003]. Sabent que els biomarcadors específics d'una malaltia donada es produeixen localment, en el teixit afectat, s'espera que els fluids proximals estiguin enriquits en aquestes substàcies, i presentin una complexitat molt menor. Així el líquid intersticial representa una font molt interessant de potencials biomarcadors [Rifai et al., 2006].

Per altra banda la disponibilitat de models animals i de línies cel·lulars simplifica notablement el problema de l'obtenció de mostres, particularment en l'estadi de descobriment de biomarcadors. En concret la facilitat de manipulació i el control que es té sobre les línies cel·lulars facilita l'estudi de les respostes a tractaments, i a diferents condicions biològiques. Per extensió dels conceptes anteriors, el conjunt de proteïnes secretades, conegut com a secretoma, que és una aproximació al liquid intersticial d'un teixit tumoral, constitueix una font molt vàlida per al descobriment de biomarcadors tumorals (figura 2.5).

**El treball que aquí es presenta va dirigit a desenvolupar eines i tècniques estadístiques que facilitin el descobriment de biomarcadors tumorals en el secretoma de línies cel·lulars fent èmfasi en la seva reproductibilitat**.

### 1.1.3 Espectroscopia de masses

Les proteïnes mostren un ventall molt ampli de pesos moleculars i propietats físicoquímiques, i moltes només són solubles sota condicions molt específiques, si és que ho arriben a ser. Això fa que l'anàlisi directa sobre un proteoma revesteixi una immensa complexitat, i que sigui força més simple treballar sobre pèptids. Les

unitats constituents de les proteïnes. Aquests són de pesos moleculars força més reduits, i generalment solubles sota diverses condicions.

Les tècniques d'alt rendiment emprades en la identificació i quantificació de proteïnes, parteixen d'una mostra que és inicialment purificada per eliminar-ne els components aliens, i posteriorment digerida amb l'ajut d'enzims (usualment tripsina) que les degraden a una mescla complexa de pèptids. L'avantatge de la digestió amb tripsina està en que ataca uns enllaços específics (l'enllaç carboxi-terminal de l'arginina i la lisina) que posteriorment facilita la identificació dels peptids.

El digerit de pèptids es sotmet a fraccionament de manera que es poden obtenir diverses submostres de menor complexitat i de propietats físico-químiques semblants. El fraccionament pot ser en columnes cromatogràfiques d'exclusió molecular, que permeten classificar els pèptids per pes molecular; en columnes de bescanvi iònic de diversos tipus que permeten la separació en funció del seu caràcter iònic; o en columnes que separen pel seu grau d'hidrofobicitat. Al final es té un conjunt de submostres que es passen per una columna cromatogràfica de nano fluxe en fase reversa. Aquesta columna es conecta a una agulla de silica o acer inoxidable de poques micres de diametre intern, que es sotmet a alt voltatge, on per electro-esprai es produeix un núvol de nanogotes carregades que alimenta un espectròmetre de masses, on es quantifica i s'identifica cada pèptid de la mescla.

Així un experiment de proteòmica d'alt rendiment (figura 2.6) consisteix en els següents passos [Steen & Mann, 2004]:

1. Aïllament i purificació de la mostra de proteïnes.
2. Digestió a una complexa mescla de pèptids.
3. Separació dels pèptids per propietats físico-químiques.
4. Ionització en l'electro-spray a la sortida de la nano-columna.
5. Caracterització dels pèptids que coelueixen per espectrometria de masses (MS), mitjançant la seva relació massa a càrrega (m/z) i la intensitat de l'ió.
6. Aïllament i posterior fragmentació per MS/MS de cada ió en els aminoàcids constituents per identificar l'ió precursor.

Les variables que permeten quantificar tot el proteoma analitzat són, per cada ió: el temps de retenció a la columna, la seva relació massa/càrrega, i la seqüència

d'aminoàcids constituents. Cada fracció es sotmet a idèntica anàlisi, i un cop analitzades totes les fraccions s'integren els resultats oferint una quantificació de les diverses proteïnes constituents de la mostra original [Nesvizhskii et al., 2007].

L'expressió LC-MS/MS, indica la connexió d'una o diverses etapes cromatogràfiques, a un equip d'espectroscopia de masses en tandem. Una primera etapa per separar els ions que co-elueixen per la seva relació càrrega/massa (m/z), i una segona per determinar-ne la seqüència d'aminoàcids.

La quantificació pot portar-se a terme bàsicament per dos mètodes. Per la senyal d'intensitat dels ions en el primer MS, o pel nombre d'espectres de MS/MS (SpC) que s'associen a cada proteïna. **El mètode seguit en aquest treball és per SpC**.

### 1.1.4 Quantificació per nombre d'espectres

La quantificació per SpC ha estat avalada per nombroses publicacions [Old et al., 2005; Gao et al., 2005; Zybailov et al., 2005]. Una recent revisió experta [Lundgren et al., 2010] considera avantatges i inconvenients en l'ús d'aquest mètode, i explora normalitzacions proposades en la literatura que tenen en compte la longitud de la proteïna o el seu pes molecular en quantificació absoluta i relativa. El propòsit de les normalitzacions esmentades és el de tenir en compte que, en igualtat de condicions, proteïnes més grans i més pesades produiran un nombre d'espectres major que proteïnes més curtes o lleugeres.

El descobriment de biomarcadors es basa en la quantificació relativa d'una mateixa proteïna entre dos estats biològics determinats. En aquestes circumstàncies la normalització de SpC tenint en compte alguna mesura de complexitat protèica contribueix el mateix a les dues situacions comparades, i per la majoria de mètodes estadístics no hauria de tenir efecte, tal com ha estat demostrat [Lundgren et al., 2010]. **Els mètodes desenvolupats en aquesta tesi no fan ús de cap transformació que tingui en compte la complexitat de la proteïna.**

### 1.1.5 Expressió diferencial sense marcatge químic

En expressió diferencial un mètode tradicionalment emprat per minimitzar el biaix en la comparació, tant en transcriptòmica [Shalon et al., 1996] com en proteòmica [Patel et al., 2009], consisteix en marcar molecularment les mostres. Això permet mesclar-les el més aviat possible en el protocol de preparació i anàlisi, garantint que els factors no controlats incideixin de la mateixa forma en totes les mostres

a comparar. Les marques, o etiquetes, permeten la identificació posterior de les proteïnes de cada condició biològica. Malgrat els avantatges, aquesta metodologia presenta certes limitacions. Els punts més importants són: la incorporació química incompleta de les etiquetes, la necessitat de concentracions més elevades en les mostres, més baixa sensibilitat en mesurar menys quantitat de cada mostra, i procediments complexes de preparació. Per altra banda el mateix avantatge d'aquesta metodologia en constitueix una limitació. Els experiments estan dedicats a una única comparació, i les dades adquirides difícilment poden ser reutilitzades en altres comparacions.

L'anàlisi de mostres proteiques sense marcar [Zhu et al., 2010] ofereix una alternativa (figura 2.10) més flexible i eficient, sempre que es puguin evitar biaixos en la comparació. Això exigeix experiments dissenyats i planificats amb cura. El major risc és el de factors no controlats que afectin de manera diversa una condició respecte l'altre.

**L'objectiu d'aquesta tesi és l'anàlisi d'expressió diferencial en secretomes de línies cel·lulars per LC-MS/MS sense marcar**.

### 1.1.6 Inferència amb SpC

Les dades que s'obtenen d'un experiment de LC-MS/MS consisteixen en una matriu d'expressió on cada columna correspon a una mostra i cada fila a una proteïna identificada. Els valors en les cel·les són els SpC observats. En aquesta matriu d'expressió hi poden haver diverses rèpliques tècniques i/o biològiques de diverses condicions biomèdiques. Aquestes dades són la base del BD, que consisteix en trobar proteïnes diferencialment expressades en termes estadístics entre les condicions biomèdiques d'interès. El que segueix descriu els principals mètodes estadístics emprats en la comparació de comptatges, per acabar amb una revisió de l'estat de l'art en proteòmica comparativa per SpC.

#### Taules de contingència

En considerar una sola de les proteïnes identificades, la manera més simple de representar les dades és en forma d'una taula de contingència com en la taula 1.1. La primera fila ens dóna els SpC observats per aquesta proteïna en cada condició. La segona fila ens dóna els SpC observats per la resta de proteïnes identificades. Els SpC totals, a la tercera fila, donen mesura de la quantitat total de proteïna de cada condició.

Taula 1.1: Taula de contingència

| | $Cond_1$ | $Cond_2$ | ... | $Cond_c$ | Total |
|---|---|---|---|---|---|
| SpC proteïna d'interès | $x_{11}$ | $x_{12}$ | ... | $x_{1c}$ | $x_{1o}$ |
| SpC altres proteïnes | $x_{21}$ | $x_{22}$ | ... | $x_{2c}$ | $x_{2o}$ |
| SpC totals en la mostra | $x_{o1}$ | $x_{o2}$ | ... | $x_{oc}$ | $n$ |

Els mètodes estadístics de directa aplicació són el test exacte de Fisher, el test $\chi^2$ de Pearson, o el test de raó de versemblança $G^2$ [Agresti, 2002]. El test de Fisher s'aplica a taules 2x2, i s'anomena exacte perquè se'n coneix la distribució i no depèn de propietats assimptòtiques. El Test de Pearson avalua la hipòtesi nul·la d'independència, que per mostres multinomials independents correspon a la homogeneïtat de resposta entre les condicions comparades. El test $G^2$ avalua un estadístic de raó de versemblança comparant el model nul contra el saturat. Els dos estadístics $X^2$ i $G^2$ són assimptòticament equivalents.

### Transformació de SpC

Quan es disposi d'un nombre rèpliques per condició que permeti una suficient estimació de la variància, un mètode alternatiu és el d'emprar una transformació del nombre d'espectres que possibiliti usar tècniques basades en la distribució normal. En condicions ideals els comptatges poden descriure's per una distribució de Poisson, on la variància és igual a la mitjana. Amb aquesta mena de comptatges la variància s'estabilitza amb transformacions del tipus $x' = \sqrt{x}$ o $x' = \sqrt{x} + \sqrt{x+1}$ [Kutner et al., 2005]. Sobre aquests valors transformats pot emprar-se el test de t. L'aproximació és poc exacte per a valors molt baixos d'expressió.

### Models lineals generalitzats

Una solució més general, que a més permet la introducció de més d'un factor i de covariables, és la dels models lineals generalitzats (GLM) [Agresti, 2002]. Aquests són aplicables sobre dades que puguin modelar-se segons una distribució de la família exponencial. Aquestes família presenta una funció densitat de probabilitat factoritzable com en l'equació 1.1

$$f(y_i;\theta_i) = a(\theta_i)b(y_i)exp[y_iQ(\theta_i)] \tag{1.1}$$

Exemples en són la distribució de Poisson o la binomial negativa. Un GLM

s'especifica amb tres components: i) Una variable aleatòria de resposta $Y$, i la seva distribució de probabilitat. ii) Una component sistemàtica donada per una combinació lineal de variables predictores. I iii) Una funció d'enllaç que relaciona $E[Y]$ amb el predictor lineal. En el nostre cas la funció d'enllaç és el logaritme neperià i el model linial el de l'equació 1.2.

$$log\,\mu_i = \sum_j \beta_j\, x_{ij} \tag{1.2}$$

La distribució de Poisson queda definida per un sol paràmetre $\mu$ que equival a la mitjana i a la variància de la distribució:

$$\mu = E[X] = Var[X] \tag{1.3}$$

La distribució binomial negativa queda definida amb dos paràmetres $\mu$ i $\phi$. Com en la distribució de Poisson $\mu$ equival a la mitjana de la distribució, mentre que la variància és una funció d'ambdós paràmetres:

$$Var(X) = \mu + \phi\mu^2 \tag{1.4}$$

El model de Poisson és aplicable en aquells casos en que l'única font de variabilitat és la pròpia del mostreig. Mentre que el model basat en la binomial negativa pot incorporar a més la variabilitat típica entre individus o espècimens. De fet la binomial negativa equival a un model mixte de Poisson on $\mu$ es distribueix segons una gamma.

Una extensió dels GLM és la quasiversemblança, on l'ajust es fa segons un model de Poisson, però la inferència té en compte una funció de variància com en l'equació 1.5, on el coeficient de dispersió $\psi$ s'estima a partir de les dades.

$$Var(Y) = \psi\mu_i \tag{1.5}$$

Aquest model també pot explicar fonts addicionals de variància com la variabilitat biològica entre individus, cultius o espècimens.

### Estat de l'art en proteòmica comparativa per SpC

Tots els mètodes descrits més amunt han estat emprats en proteòmica diferencial. Els primers a explorar-se varen ser els basats en taules de contingència, en el test de t, o en varacions de mètodes desenvolupats específicament per a microarrays, SAGE o DGE [Zybailov et al., 2006; Zhang et al., 2006; Pavelka et al., 2008].

L'estat de l'art actual es basa en la implementació de GLMs. Bé de models d'efectes mixtes [Choi et al., 2008], de quasiversemblança [Li et al., 2010], o de la binomial negativa [Leitch et al., 2012]. L'objectiu de tots aquests models és el de tenir en compte la variabilitat biològica de les mostres, causant de la sobredispersi, s a dir d'una variancia major que la mitjana. Aquests mètodes han anat proposant-se paral·lelament al desenvolupament de la tesi, i com evidencia el darrer treball citat encara hi ha molt potencial de recerca en el camp.

**La solució adoptada en la tesi és la dels GLM.**

### 1.1.7 Lliçons en el descobriment de biomarcadors

La reproductibilitat en les llistes de biomarcadors és el punt més crucial en BD. En aquest sentit resulta molt útil considerar l'experiència adquirida en gaire bé vint anys de transcriptòmica, en benefici d'altres *òmiques* més joves, i en particular de la proteòmica. En aquest apartat es presenten algunes de les pricipals lliçons apreses en BD, tant en transcriptòmica com en proteòmica. Aquestes lliçons porten als objectius de la tesi.

#### Normalització

La normalització s'entén com un procediment per eliminar diferències sistemàtiques de naturalesa tècnica entre les mostres. Atès que de cada mostra es mesura la mateixa quantitat de substància, el procediment més estès consisteix a referir els SpC de cada proteïna al total de SpC observats en la mostra. En el context dels models GLM aquesta normalització s'incorpora al model mitjançant un terme d'*offset* [Agresti, 2002; Choi et al., 2008; Li et al., 2010; Leitch et al., 2012]. Així el model esdevé:

$$log\left(\frac{\mu}{size}\right) = \alpha + \beta x$$

$$log(\mu) = log(size) + \alpha + \beta x \tag{1.6}$$

on $\mu$ és l'expressió esperada d'una proteïna donada, $size$ és el factor de normalització, i $\alpha$ i $\beta$ són els paràmetres del model, amb $x$ igual a 0 per la condició de control, o a 1 per la condició de tractament. El terme $log(size)$ és l'*offset* en aquest cas. Aquest model admet la inclusió d'altres factors o covariables. La formulació del model és independent de la distribució subjacent.

En termes biològics aquesta normalització es basa a més en la suposició que les cèl·lules produeixen la mateixa quantitat global de proteïna en les dues condicions que es comparen. **Aquesta suposició pot acceptar-se en general quan s'analitzen mostres de llisats cel·lulars, però resulta més qüestionable quan s'analitzen secretomes de línies cel·lulars. Aquest tema crucial també s'investiga i es resol en la tesi, proposant un esquema específic de normalització basat en *offsets*.**

### P-valors i mida de l'efecte

En els inicis dels microarrays el BD es basava simplement en els FC observats. Paulatinament es van a anar introduint mètodes estadístics de creixent sofisticació, i ajustaments de multitest en els p-valors, amb control del FDR [Allison et al., 2006]. Els resultats d'un estudi de microarrays es donaven en una llista de gens diferencialment expressats en ordre creixent de p-valor. Els gens del capdemunt de la llista se solen considerar els més significatius.

En aquest context, amb diferents plataformes de microarrays comercialment disponibles, i amb les grans expectatives que la introducció dels microarrays van despertar, aviat es va evidenciar que no es podien reproduir els resultats d'alguns biomarcadors que s'havien publicat amb altes taxes de sensibilitat i especificitat.

D'entre els projectes que es van endegar per estudiar els factors que afectaven la reproductibilitat dels resultats per microarrays destaca el MicroArray Quality Control Consortium [Shi et al., 2006], format per membres de les autoritats reguladores (FDA), autoritats acadèmiques, i institucions comercials. El número de setembre de 2006 de Nature Biotechnology va estar completament dedicat als resultats de l'estudi MAQC-I. Les seves recomanacions, per tal d'obtenir llistes reproductibles de gens diferencialment expressats, més enllà d'emprar acurats dissenys experimentals, i de recòrrer a transformacions apropiades de dades, es basaven en limitar el nombre de transcrits identificats com a diferencialment expressats, i amb ordenar-los atenent al FC amb un llindar de p-valor no massa astringent. Aquests resultats varen ser qüestionats [Chen et al., 2007] i posteriorment confirmats comparant diferents mètodes de selecció de gens i amb simulacions [Shi et al., 2008].

Resumint, els gens del capdemunt de la lista han de ser no els de major significació estadística, si no els que mostren major efecte amb un p-valor ajustat raonable.

**Aquesta lliçó s'implementa en el treball en forma d'un filtre post-test**

**que marca les proteïnes amb poca senyal i baix efecte com d'escassa reproductibilitat malgrat el seu baix p-valor. La llista de proteïnes diferencialment expressades s'ordena per p-valor amb el corresponent indicador**.

### Efectes batch

Les conclusions de l'estudi MAQC-I estan directament relacionades amb el problema de *molts-gens-poques-rèpliques*. Això suggereix que augmentant el nombre de rèpliques augmentarà la potència i millorarà la reproductibilitat. Malhauradament s'ha observat que augmentant el nombre de rèpliques augmenta també la possibilitat d'introduir biaix (Speed T. en [Scherer, 2009]). Quan els experiments es recullen durant un periode llarg de temps el biaix pot resultar inevitable. Això està relacionat amb els anomenats efectes *batch*. Malgrat que un bon disseny experimental i l'ús de randomitzacions i blocs en cada pas de tot el procés, des de la recollida de mostres fins l'anàlisi final, puguin reduir molt els efectes de variables no controlades, els efectes batch es mostren ubics i inevitables en les *òmiques* [Ransohoff, 2005a; Scherer, 2009; Leek et al., 2010; Auer & Doerge, 2010; Schloss et al., 2011; Valsesia et al., 2013].

Els efectes batch es defineixen com a sistemàtics en contraposició al soroll tècnic o experimental, que és de natura aleatòria. Degut a la seva natura sistemàtica la pitjor manifestació dels efectes batch s'observa quan les mostres tractament es preparen i analitzen separadament de les mostres control, en lots diferents o en temps diferents. Les diferències que s'observin estaran confoses entre l'efecte del tractament i dels factors no controlats que puguin influir. Aquesta ha estat la causa més freqüent i estrepitosa de fracassos en BD [Baggerly et al., 2004, 2005; Ransohoff, 2005b; Baggerly et al., 2008; Baggerly & Coombes, 2009]. Una manifestació menys dramàtica s'observa quan les mostres s'han analitzat de manera balancejada en les condicions a comparar, però en moments espaiats en el temps. Això causa un augment de la variabilitat intraclasse reduint la potència dels tests.

La presència d'efectes batch es pot visualitzar amb l'ús de tècniques multidimensionals com l'anàlisi en components principals (PCA), la descomposició en valors singulars (SVD), o el clustering jeràrquic (HC). Idealment les mostres han d'agrupar-se per condició biològica. La presència d'efectes batch es manifesta quan les mostres s'agrupen pel moment en que es van recollir o pel moment en que es van tractar i analitzar, en comptes de per la seva condició biològica (veure figura 2.11).

**Com a resposta a aquesta lliçó s'han implementat eines multidimensionals de visualització en un paquet R/Bioconductor, i s'ha dissenyat una interfície gràfica que facilita l'anàlisi exploratòria de dades de LC-MS/MS basades en SpC per evidenciar la presència de valors extrems o d'efectes batch. Per altre banda els models GLM proposats i implementats en un altre paquet R/Bioconductor permeten portar a terme la inferència tenint en compte la presència dels efectes batch eventualment detectats en l'anàlisi exploratòria prèvia.**

## 1.2 Objectius

L'establiment d'un nou laboratori de biomarcadors tumorals a l'Institut d'Oncologia de la Vall d'Hebron (VHIO) ha ofert l'oportunitat de desenvolupar eines d'anàlisi de dades en proteòmica comparativa, i contribuir a posar en valor l'experiència adquirida en el tractament de dades *òmiques* de microarrays durant més d'una dècada, traslladant part de les lliçons apreses al camp de la poteòmica comparativa. En aquest sentit, els objectius d'aquesta tesi són la implementació d'eines i mètodes específicament dirigits a:

- Anàlisi de dades d'experiments de proteòmica diferencial sense marcatge químic i basats en el nombre observat d'espectres de pèptids.
- Avaluació de la qualitat d'un conjunt de dades en termes de detecció de valors extrems, i de la influència de factors no controlats.
- Modelització i normalització de dades de secretomes de línies cel·lulars.
- Millora en la reproductibilitat de les llistes de proteïnes diferencialment expressades.
- Produir paquets de R amb les eines desenvolupades.
- Produir interfícies gràfiques que facilitin l'ús de les eines proposades.

## 1.3 Resultats

Totes les publicacions presentades en la tesi han estat sotmeses a revistes internacionals amb avaluació d'experts.

El sofware desenvolupat ha estat encapsulat en paquets R sotmesos a Bioconductor, i en dues interfícies gràfiques disponibles a GitHub.

Tot seguit es presenta una llista de les publicacions amb els abstracts corresponents, i una breu descripció dels paquets R i les respectives interfícies. A continuació s'exposa una discussió general, i es donen les conclusions de la tesi amb les contribucions al camp estudiat. Per acabar s'inclou una llista d'altres publicacions de l'autor en el camp de la bioestadística/bioinformàtica no relacionades amb la tesi però realitzades durant el seu desenvolupament.

## 1.4 Articles de la tesi

1. La correcció d'effectes batch millora la sensibilitat dels tests en proteòmica comparativa basada en comptatge d'espectres.
   *Batch effects correction improves the sensitivity of significance tests in spectral counting-based comparative discovery proteomics.*
   Gregori J, Villarreal L, Méndez O, Sánchez A, Baselga J, Villanueva J.
   J Proteomics. 2012 Jul 16; 75(13):3938-51. doi: 10.1016/j.jprot.2012.05.005. Epub 2012 May 12.
   Factor d'impacte: 4.1

   RESUM

   La proteòmica per shotgun ha esdevingut la tècnica estàndar per a la mesura a gran escala d'abundància de proteïnes en mostres biològiques. Malgrat que la proteòmica quantitativa ha emprat usualment tècniques de marcatge molecular, la quantificació sense marcadors ofereix avantatges considerables. Entre elles: i) Evitar els procediments de marcatge. ii) No presentar limitació en el nombre de mostres a comparar. I iii) Augment de la sensibilitat en la detecció de proteïnes. Tanmateix atès que les mostres són tractades i analitzades de forma separada, el disseny experimental esdevé crític. L'exploració de la quantificació per nombre d'espectres que es presenta en aquest treball recull evidència experimental de la influència dels effectes batch en proteòmica comparativa. Aquests effectes, demostrats amb experiments amb mescles controlades, interfereixen clarament amb el senyal biològic. Per tal

de minimitzar la interferència dels effectes batch es proposa i implementa una correcció estadística. Els resultats demostren que aquests efectes es poden atenuar emprant un disseny experimental adequat. La correcció implementada porta a un augment substancial de la sensibilitat dels tests. L'aplicabilitat de la correcció proposada es mostra sobre dos projectes de descobriment de biomarcadors amb secretomes de càncer. El mètode proposat permet millorar el disseny i l'execució de projectes de proteòmica comparativa, i contribueix a evitar falses conclusions en el procés de descobriment de biomarcadors en proteòmica.

2. Un filtre de mida de l'efecte millora la reproductibilitat en proteòmica comparativa basada en comptatge d'espectres.
   *An effect size filter improves the reproducibility in spectral counting-based comparative proteomics.*
   Gregori J, Villarreal L, Sánchez A, Baselga J, Villanueva J.
   J Proteomics. 2013 Dec 16; 95:55-65.
   doi: 10.1016/j.jprot.2013.05.030. Epub 2013 Jun 11
   Factor d'impacte: 4.1

   RESUM

   La comunitat en el camp dels microarrays ha demostrat que la baixa reproductibilitat observada en alguns estudis de descobriment de biomarcadors en expressió genòmica és parcialment deguda a basar les llistes de gens diferencialment expressats exclusivament en p-valors. Les seves conclusions recomanen complementar el llindar de p-valor amb l'ús de criteris d'efecte. La intenció d'aquest treball ha estat avaluar la influència d'un filtre per mida de l'efecte i intensitat del senyal en l'anàlisi de proteòmica comparativa basada en nombre d'espectres. Els resultats han provat que el filtre augmenta el nombre de positius certs i disminueix el nombre de falsos positius en el seu conjunt. Aquests resultats s'han confirmat amb conjunts de dades simulades, amb augment progressiu en la fracció de proteïnes diferencialment expressades. Els resultats suggereixen que relaxant el llindar de p-valor i emprant un filtre posterior als test, basat en llindars en el nivell del senyal i en la mida de l'efecte, pot augmentar la reproductibilitat dels resultats. En base als resultats d'aquest treball, es recomana com a pràctica general emprar un filtre exigint un senyal mínim entre 2 i 4 SpC en la condició més abundant, i un efecte no inferior a un LogFC en valor absolut de 0.8. La

implementació d'aquests filtres pot millorar els resultats en el descobriment de biomarcadors a l'assegurar una major reproductibilitat entre laboratoris independents i diferents plataformes de MS.

3. Millora en el valor biològic de la proteòmica de secretomes, vinculant la proliferaci cel·lular del tumor i la secreció de proteïnes.
   *Enhancing the Biological Relevance of Secretome-based Proteomics by Linking Tumor Cell Proliferation and Protein Secretion.*
   Gregori J., Méndez O., Katsila T., Pujals M., Salvans C., Villarreal L., Arribas J., Tabernero J., Sánchez A., Villanueva J.
   Sotmès a J. of Proteome Research, pendent de publicació.

   RESUM

   La determinació del perfil dels secretomes ha esdevingut una metodologia útil en el descobriment de biomarcadors tumorals secretats. Els secretomes són proteomes molt dinàmics que contenen proteïnes directament involucrades en diferents aspectes de la tumorogènesi. Degut a la seva naturalesa dinàmica es va formular la hipòtesi que algunes pertorbacions cel·lulars podien no només afectar la composició del secretoma si no també canviar la taxa de secreció cel·lular. De resultar certa, aquesta observació seria molt rellevant en el descobriment de biomarcadors, ja que la unitat biològica sobre la que es cerca la comparativa és la cèl·lula. En aquest treball s'ha desenvolupat i implementat un model que incorpora una normalització que permet referir els resultats a la quantitat de proteïna secretada per cèl·lula. El model desenvolupat correspon a l'equació 1.7:

$$log(\mu) = log\left(\frac{n}{Q}\right) + log(size) + \alpha + \beta x + \gamma z \tag{1.7}$$

   on $Q$ és la quantitat de proteïna secretada per $n$ cèl·lules, en una mostra amb $size$ SpC totals, $X$ és el factor tractament, i $Z$ un eventual factor per blocs.

   S'han detectat diferències substancials en la quantitat global de proteïna secretada entre cèl·lules sotmeses a diferents pertorbacions biològiques, i també entre línies cel·lulars en el seu estat basal. L'aplicació del model a dos escenaris biològics diferents amb cèl·lules tumorals ha mostrat un fort efecte sobre la llista de proteïnes diferencialment secretades. En aquest sentit s'ha vist que efectors de la trancisió epitelial a mesenquimal només resulten estadísticament significatius quan s'aplica el model descrit. L'estudi ha

permès també individualitzar altres proteïnes encara no descrites en l'esmentada trancisió que poden resultar d'interès com a biomarcadors. Finalment l'estudi suggereix que la taxa de secreció global de proteïnes en cèl·lules tumorals està relacionada amb el seu estat de proliferació cel·lular. El treball confirma la hipòtesi inicial i mostra que la naturalesa dinàmica dels secretomes pot esbiaixar els resultats en el descobriment de biomarcadors de no emprar un model adequat. Des del punt de vista oncològic el vincle entre secreció proteica i proliferació cel·lular suggereix que els tumors de creixement lent poden ser susceptibles de majors taxes de secreció, i en conseqüència contribuir en major grau a la senyalització paracrina.

4. La secreció no convencional és un contribuent major en els secretomes de línies cel·lulars de càncer.
   *Unconventional secretion is a major contributor of cancer cell line secretomes.*
   Villarreal L, Méndez O, Salvans C, Gregori J, Baselga J, Villanueva J.
   Mol Cell Proteomics. 2013 May; 12(5):1046-60.
   doi: 10.1074/mcp.M112.021618. Epub 2012 Dec 26.
   Factor d'impacte: 7.4

   RESUM

   Un repte per aconseguir una gestió òptima del càncer és el descobriment de biomarcadors secretats que representin l'estat de la malaltia i es puguin mesurar de forma no invasiva. Degut a la problemàtica que planteja l'anàlisi del proteoma del plasma, s'ha proposat el secretoma com a font alternativa de marcadors, ja que pot estar enriquit en proteïnes secretades rellevants de la malaltia. Tanmateix, l'anàlisi del secretoma planteja també els seus reptes. En particular distingir les proteïnes realment secretades. En aquest treball s'han estudiat dos dels principals reptes en l'anàlisi de secretomes en proteòmica comparativa. En primer lloc, s'ha portat a terme un estudi cinètic en el que s'ha analitzat el secretoma i el llisat cel·lular per monitoritzar la viabilitat cel·lular durant la producció de secretoma. S'ha determinat que un grup de proteïnes secretades es correlaciona bé amb l'apoptosi induïda en el periode d'inanició per sèrum, i que pot emprar-se com a indicador intern de viabilitat cel·lular. En segon lloc, s'han determinat les interferències causades pel necessari ús de sèrum en el cultiu cel·lular. L'anàlisi proteòmica comparativa entre línies cel·lulars marcades amb SILAC ha mostrat un cert

nombre de falsos positius que provenen del sèrum, i que diverses proteïnes es troben tant en el serum com en el secretoma de cèl·lules tumorals.

Per altra banda un estudi minuciòs de la metodologia d'obtenció de secretoma ha revelat que sota condicions experimentals òptimes hi ha una fracció substancial de proteïnes que són secretades per mecanismes no convencionals. Finalment s'ha mostrat que algunes proteïnes nuclears detectades en el secretoma canvien de localització cel·lular en tumors de mama, suggerint que les cèl·lules tumorals usen una secreció no convencional durant la tumorogènesi. La secreció no convencional de proteïnes en l'espai extracel·lular exposa un nou nivell de regulació genòmica post-translacional que pot constituir una font potencial de biomarcadors tumorals i de dianes terapèutiques.

| | |
|---|---|
| `pp.msms.data` | preprocessat de dades per convertir NAs en 0 i eliminar les files amb tot zeros. |
| `gene.table` | extreure els símbols de gen de la descripció de proteïna. |
| `count.stats` | estadístics de SpC i nombre de proteïnes per mostra. |
| `counts.pca` | anàlisi de components principals de la matriu de SpC. |
| `counts.hc` | dendrograma del clustering jeràrquic de les mostres. |
| `norm.counts` | normalització de la matriu de SpC. |
| `counts.heatmap` | heatmap de la matriu de SpC. |
| `disp.estimates` | anàlisi de dispersió residual. |
| `spc.barplots` | gràfic de barres dels divisors de normalització relatius. |
| `spc.boxplots` | gràfic de caixes mostrant la distribució de SpC per mostra. |
| `spc.densityplots` | gràfic de densitat mostrant la distribució de SpC per mostra. |
| `filter.flags` | marques llògiques per les proteïnes segons llindars de senyal i variabilitat. |
| `bacth.neutralize` | correcció d'efectes batch en la matriu de SpC. |

Taula 1.2: Funcions en el paquet `msmsEDA`.

## 1.5 Software

### 1.5.1 Paquets R/Bioconductor

El software ha estat desenvolupat en el llenguatge i entorn R [R Core Team, 2012]. S'han produit dos paquets R amb el codi desenvolupat durant els treballs que han portat a la publicació dels articles esmentats més amunt. Aquests paquets han estat adaptats a la infraestructura de Bioconductor [Gentleman et al., 2004], i adaptats específicament per treballar amb instàncies de la classe S4 `MSnSet` definida en el paquet `MSnbase` [Gatto & Lilley, 2012].

- `msmsEDA` recull les funcions emprades en l'anàlisi exploratòria de matrius d'expressió amb SpC.
- `msmsTests` ofereix funcions útils en inferència sobre matrius d'expressió amb SpC, basades en models GLM.

Les funcions per l'anàlisi exploratòria de dades (EDA) permeten la identificació de valors extrems, efectes batch, o factors de confusió. Qualsevol estudi de BD hauria de començar sistemàticament per un EDA en profunditat per validar les mostres i el model que s'usarà posteriorment en l'estudi. Les principals funcions del paquet `msmsEDA` es llisten en la taula 1.2.

El procés de BD es porta a terme per l'aplicació del mateix model i test sobre cada fila de la matriu d'expressió. El model general considerat en les funcions del paquet `msmsTests` és el de l'equació 1.7. Es disposa d'una funció per el GLM basat en la distribució de Poisson, d'una altra basada en la quasiversemblança, i d'una altra basada en la binomial negativa. Per la Poisson i la quasiversemblança el test és el de la raó de versemblances entre el model alternatiu i el null. Per la binomial negativa s'usa l'aproximació implementada en el paquet `edgeR` [Robinson et al., 2010]. Les principals funcions d'aquest paquet es llisten en la taula 1.3.

Ambdós paquets inclouen funcions que ajuden en la interpretació dels resultats.

#### Disponibilitat

Els paquets s'han integrat en el projecte Bioconductor [Gentleman et al., 2004] i usen la classe S4 `MSnSet` en el paquet `MSnbase` [Gatto & Lilley, 2012].

Els paquets estan disponibles a Bioconductor:
`http://www.bioconductor.org/packages/2.13/bioc/html/msmsEDA.html`
`http://www.bioconductor.org/packages/2.13/bioc/html/msmsTests.html`

| | |
|---|---|
| `msms.glm.pois` | Model GLM de Poisson |
| `msms.glm.qlll` | Model GLM de quasiversemblança |
| `msms.edgeR` | Model de la binomial negativa del paquet edgeR |
| `pval.by.fc` | Taula creuada de freqüència de proteïnes per p-valors en blocs de LogFC |
| `test.results` | Ajustament multitest de p-valors amb control de FDR, i filtre post-test per marcar els DEPs més reproductibles. |
| `res.volcanoplot` | Volcanplot dels resultats. |

Taula 1.3: Funcions en el paquet `msmsTests`.

**Documentació**

Els manuals dels paquets i tutorials en forma de vignettes estan disponibles on-line a `http://www.bioconductor.org`.

- Manual de `msmsEDA`
  `http://www.bioconductor.org/packages/2.13/bioc/manuals/msmsEDA/man/msmsEDA.pdf`

- Vignette de `msmsEDA`: *Analisi exploratory de dades de LC-MS/MS.*
  `http://www.bioconductor.org/packages/release/bioc/vignettes/msmsEDA/inst/doc/msmsData-Vignette.pdf`

- Manual de `msmsTests`
  `http://www.bioconductor.org/packages/2.13/bioc/manuals/msmsTests/man/msmsTests.pdf`

- Vignette de `msmsTests`: *Filtres post test per millorar la reproductibilitat.*
  `http://www.bioconductor.org/packages/2.13/bioc/vignettes/msmsTests/inst/doc/msmsTests-Vignette.pdf`

- Vignette de `msmsTests`: *Disseny per bocks per compensar efectes batch.*
  `http://www.bioconductor.org/packages/2.13/bioc/vignettes/msmsTests/inst/doc/msmsTests-Vignette2.pdf`

S'adjunten a la tesi un tutorial, els manuals, i les vignettes dels dos paquets.

### 1.5.2 Interfícies gràfiques

S'han desenvolupat dues interfícies gràfiques (GUI) per facilitar els càlculs rutinaris en un entorn de laboratori, i per acostar les solucions incorporades en els paquets descrits als investigadors en el camp de la proteòmica que no disposin d'habilitats de programació. Els GUI s'han desenvolupat sobre les funcions dels dos paquets descrits, i amb l'ajut de la infraestructura proporcionada pels paquets `gWidgets` i `RGtk2` [Verzani, 2012; Lawrence & Verzani, 2012; Lawrence & Temple Lang, 2010].

- `msmsEDA_GUI`

  Proporciona una anàlisi exploratòria completa d'un experiment LC-MS/MS donats dos fitxers. El fitxer de descripció de les mostres (targets), amb identificadors de mostra, etiquetes, i eventuals factors (divisors) de normalització. I un fitxer amb la matriu d'expressió en SpC, i descriptors de les proteïnes identificades.

  Mitjançant gràfics de caixes i de densitat de distribució de SpC de cada mostra, i gràfics de barres dels valors relatius dels factors de normalització per mostra, es porta a terme una avaluació de la qualitat del conjunt de dades de l'experiment. La matriu d'expressió per SpC s'analitza per les tècniques de PCA, HC i heatmaps. I aquesta anàlisi es fa sobre la matriu de SpC sense tractar, sobre la matriu de SpC normalitzada, i si s'escau sobre la matriu de SpC normalitzada i corregida d'effectes batch. La distribució de proteïnes informatives segons el factor principal es visualitza a cada pas del tractament de dades. Finalment s'explora la distribució de valors de coeficient de dispersió residual. El procés genera un conjunt de fitxers de text amb resultats i gràfics que permeten visualitzar i avaluar els diferents passos de l'EDA. També es genera un fitxer html que serveix com a índex de tots els fitxers generats, amb noms, descripció i vincles a cadascun.

- `msmsTests_GUI`

  Donats els fitxers de descripció de mostres i de matriu de SpC, com en l'altra interfície, proporciona una gran flexibilitat en BD amb controls en el GUI que permeten escollir el mètode de normalització, el test estadístic, el mètode de correcció de p-valors amb control de la FDR, el llindar de significància, i els llindars de senyal i mida d'efecte en el filtre post-test. El control de sortida en la interfície mostra el desenvolupament dels càlculs, i els principals resultats.

Es generen un conjunt de fitxers de text i pdf amb gràfics, amb resultats intermedis i la llista final.

### Disponibilitat i documentació

Les iterfícies i la seva documentació estan disponibles on-line a GitHub.com

| | |
|---|---|
| `msmsEDA_GUI` | `https://github.com/JosepGregori/msmsEDA_GUI_repos` |
| `msmsTests_GUI` | `https://github.com/JosepGregori/msmsTests_GUI_repos` |

S'adjunten a la tesi les guies d'usuari de les dues interfícies.

## 1.6 Discussió general

Aquest treball s'ha concentrat en tres aspectes rellevants del procés de descobriment de biomarcadors (BD): efectes batch, reproductibilitat, i el disseny d'un model per la comparació de secretomes en línies cel·lulars entre dues condicions biològiques cel·lula-a-cel·lula.

Hem estudiat la incidència dels efecte batch [Scherer, 2009] en proteòmica comparativa sense marcatge basada en SpC, i hem fixat la finestra de temps més estreta en la que els assajos de LC-MS/MS estan afectats de manera equivalent pels factors no controlats en un sol dia d'adquisició de dades [Gregori et al., 2012].

La presència d'efectes batch s'ha determinat amb tècniques multidimensionals com l'anàlisi de components principals (PCA) o el clustering jeràrquic (HC). Llavors hem estudiat mètodes de correcció d'aquests efectes [Chen et al., 2011] en experiments balancejats en les condicions a comparar, i hem trobat que el millor mètode és una correcció d'escala implementada en un GLM amb funció d'enllaç logarítmica que incorpora un factor per blocs [Quinn & Keough, 2002] agrupant cada lot de mostres.

Basant-nos en els aspectes conceptuals de les recomanacions de l'estudi MAQC-I [Shi et al., 2008], hem estudiat els avantatges d'emprar un filtre post-test per nivell de senyal i mida de l'efecte en la millora de la reproductibilitat de biomarcadors descoberts per LC-MS/MS lliure de marcatge amb SpC. Hem determinat que aquests filtres poden millorar el nombre de positius certs (TP), tot i restringir el nombre de falsos positius (FP), relaxant el llindar de significància. Hem vist també que aquest tipus de filtre ofereix l'avantatge addicional de millorar el solapament entre llistes de proteïnes declarades com diferencialment expressades per tests diferents. Ambdós factors contribuixen a millorar la reproductibilitat dels resultats [Gregori et al., 2013].

Partint de la hipòtesi de taxa de secreció variable en línies cel·lulars de càncer sota diferents pertorbacions biològiques, hem desenvolupat un model GLM per comparacions de secretoma cel·lula-a-cel·lula lliures de biaix [Gregori et al., 2014] (veure secció 4.3). Aquest model incorpora: i) una normalització de mostra pel nombre total de SpC, ii) la normalització cel·lula-a-cel·lula emprant la taxa de secreció observada en cada condició, iii) el factor de tractament rellevant per la comparació, i iv) factors per blocs.

El codi R produit en el desenvolupament d'aquests estudis s'ha estructurat en funcions d'ús general i s'ha encapsulat en dos paquets a Bioconductor [Gentleman

et al., 2004]. Un paquet per l'anàlisi exploratòria de dades per detectar valors extrems, efectes batch o la presència de factors de confusió (`msmsEDA`). I un segon que implementa les normalitzacions, els tests i els filtres (`msmsTests`). També s'ha escrit una interfície gràfica (GUI) per cadascun dels dos paquets (`msmsEDA_GUI` i `msmsTests_GUI`).

En els següents apartats es discuteix el que pot faltar, o el que pot desenvoluparse en nous projectes per continuar la línia de recerca encetada.

### 1.6.1 Limitació en l'ús dels SpC

La naturalesa discreta dels SpC complica la interpretació de valors d'expressió molt baixos, i limita la sensibilitat en la detecció de secreció diferencial a aquests nivells. En la mesura en que l'expressió mitjana vagi guanyant entropia podrem millorar la qualitat dels resultats. Això implica que com més baix sigui el nivell d'expressió en què estem interessats major hagi de ser el nombre de rèpliques necessari.

### 1.6.2 Effectes batch en diagnòstic

Malgrat que els efectes batch es puguin detectar, quantificar, i corregir en experiments balancejats [Scherer, 2009; Gregori et al., 2012], queda oberta la qüestió de com tractar mostres aïllades en LC-MS/MS que hagin de classificar-se. En molts cassos el biomarcador pot ser una sola proteïna, o un nombre força limitat de proteïnes, de manera que es pugui mesurar en pacients per tècniques immunoquímiques on els efectes batch tenen escassa o nul·la incidència. Tanmateix quan la combinació d'intensitat del senyal i mida de l'efecte sigui feble el biomarcador pot consistir en una signatura composta per un nombre prou gran de proteïnes com per aconsellar l'ús de LC-MS/MS en diagnòstic. En aquest cas caldrà poder calibrar la influència dels factors no controlats amb mostres control.

El control ha de ser estable en el temps i contenir totes les proteïnes rellevants en les concentracions degudes. Ambdues mostres, control i problema, hauran de tractar-se i mesurar-se en paral·lel, de tal manera que tots els factors no controlats les afectin de manera idèntica. Aquesta qüestió no és gens simple i requereix un acurat estudi.

### 1.6.3 Subdispersió

Malgrat que la preocupació principal en la majoria de mètodes desenvolupats en proteòmica comparativa és encabir la sobredispersió causada per la variabilitat biològica de les mostres (veure secció 1.1.6), en els experiments amb mostres controlades de llisat de llevat hem observat sistemàticament un cert grau de subdispersió a tots els nivells d'expressió (Figura 1.1). En diferents experiments amb secretomes de línies cel·lulars hem observat un fenomen semblant.

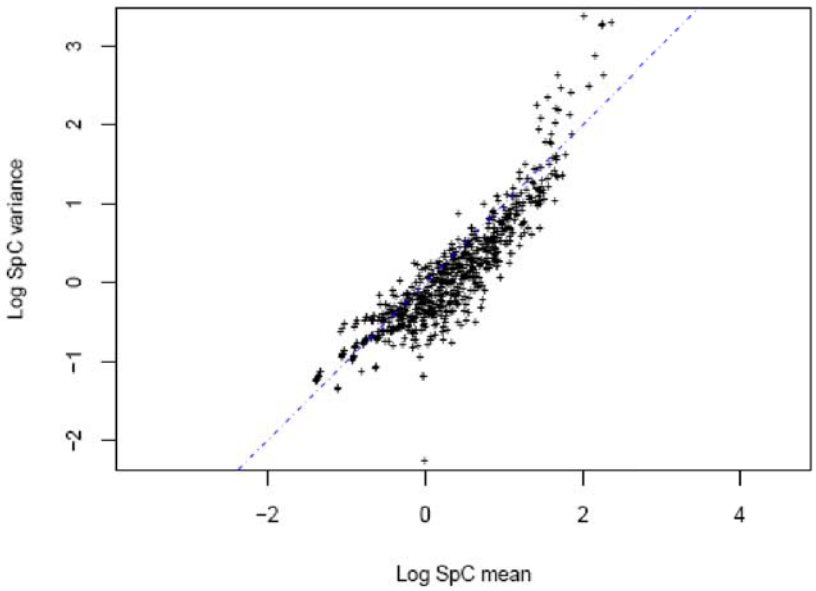



Figura 1.1: Dispersió residual en experiments amb mostres controlades de llevat amb proteïnes humanes afegides. Els punts sota la diagonal mostren subdispersió.

Els models que incorporin la subdispersió es podran beneficiar d'una sensibilitat major que el GLM basat en la distribució de Poisson. El model basat en la binomial negativa (veure equació 1.4) i el model basat en l'extensió de la quasiversemblança (veure equació 1.5) poden explicar tant la subdispersió com la sobredispersió. La quasiversemblança modela subdispersió per valors positius de $\psi$ inferiors a 1, la binomial negativa modela subdispersió per valors negatius de $\phi$ majors que $-1/\mu$ [Agresti, 2002].

L'estimació de la dispersió introdueix un segon paràmetre en el model, fent necessàries un major nombre de rèpliques, que constitueix la major limitació en un laboratori de proteòmica. La solució que proporciona `edgeR` [Robinson et al., 2010], similar a la implementada en `limma` [Smyth, 2005] per microarrays, permet

compartir informació entre proteïnes de nivells semblants d'expressió i fa menys crítica la grandària de la mostra. Un treball interessant podria ser el desenvolupament d'una aproximació bayessiana empírica semblant per la quasiversemblança.

### 1.6.4 Usos possibles en altres *òmiques*

La tesi s'ha desenvolupat sobre conjunts de dades que són matrius esparses de comptatges, semblants a les emprades en almenys altres dues *òmiques*: taules RNA-seq en transcriptòmica [Wang et al., 2009] i taules d'OTUs en metagenòmica [Wooley et al., 2010], ambdues basades en tècniques NGS però responent a preguntes molt diferents. Així l'estudi de la secreció diferencial de proteïnes per línies cel·lulars de càncer representa reptes estadístics semblants a l'estudi de l'expressió genètica diferencial per RNA-seq, o a l'estudi de l'abundància diferencial de microorganismes en microbiomes per seqüenciació del 16S ribosomal. De fet el paquet `edgeR` [Robinson et al., 2010] emprat en el nostre software va estar específicament desenvolupat per anàlisi de dades de RNA-seq.

En totes aquestes *òmiques* la reproductibilitat dels resultats és un tema clau, i la correcció d'efectes batch i l'ús dels filtres post-test descrits aquí podrien contribuir a millorar els resultats en totes elles.

### 1.6.5 Possibles línies de recerca derivades

Amb el model de l'equació 1.7 la proteòmica comparativa sobre secretomes basada en SpC pot considerar-se lliure de biaix sempre que els lots de mostres estiguin balancejats en les condicions a comparar. Això imposa una restricció semblant a la observada en proteòmica amb marcadors. Una possible solució a aquesta mancança pot consistir en l'ús de mostres de control universals, de manera que qualsevol condició pugui mesurar-se respecte a aquest control. Això podria permetre comparacions no esbiaixades entre mostres en lots diferents sempre que tinguin el mateix control. Encara que conceptualment simple, aquesta posibilitat requereix un estudi acurat, i no està clar si caldrien mostres control diferents per cada línia cel·lular. Els efectes batch es podrien corregir per aquelles proteïnes en comú entre les condicions a comparar i els controls, i sempre que el nivell d'expressió sigui d'un ordre semblant. Aquest estudi implica una experimentació extensiva i pot constituir la base per a un nou projecte, juntament amb les consideracions fetes més amunt en diagnòstic (veure 1.6.2).

Un altre tema importat és la validació de signatures moleculars en proteòmica. Quan el descobriment de biomarcadors porta a una simple molècula, o un nombre molt limitat de molècules, la validació es portarà a terme amb mètodes diferents al LC-MS/MS, com ara l'ELISA per exemple. En canvi quan el BD porta a una complexa mescla de proteïnes, a una signatura molecular, el procés de descobriment i el de validació esdevenen part del mateix problema. En aquest cas cal evitar el sobreajust del discriminador al conjunt de dades emprat en el seu descobriment i construcció. Una línia interessant d'estudi seria la implementació en proteòmica de la metodologia general proposada per [Parry et al., 2010] en microarrays a la llum de l'estudi MAQC-II [Shi et al., 2010].

Finalment una línia de recerca paral·lela prodria ser estudiar la incidència dels efectes batch, i els llindars en el filtre post-test necessaris per a millorar la qualitat dels resultats quan la mesura es fa per intensitat de l'ió precursor en comptes de per SpC.

## 1.7 Conclusions

En aquesta tesi s'han explorat aspectes fonamentals en el descobriment de biomarcadors en proteòmica per la tècnica shotgun amb pèptits sense marcar, i mesurats per LC-MS/MS en nombre d'espectres, estudiant secretomes de línies cel·lulars de càncer. En concret s'ha demostrat:

1. Que la normalització entre rèpliques tècniques pel nombre total d'espectres de la mostra proporciona resultats més estables que l'ús d'estàndards interns. Fins i tot quan s'empren múltiples estàndards.

2. Que els efectes batch estan probablement presents en tots els projectes de proteòmica desenvolupats durant més d'un dia d'adquisició de dades.

3. La importància de l'anàlisi exploratòria de dades com a eina per determinar la qualitat d'un conjunt de dades, i per identificar la presència d'efectes batch, de factors de confusió o de valors extrems.

4. Que un disseny experimental completament desbalancejat molt probablement pot comportar biaixos, i aquests seran impossibles de corregir.

5. Que els efectes batch es poden i s'han de corregir en dissenys balancejats.

6. Que l'ús d'un filtre post-test que tingui en compte la intensitat del senyal i la mida de l'efecte, més enllà del nivell de significació estadística, millora la reproductibilitat dels resultats de BD.

7. Que les llistes llargues de DEPs es veuran favorablement escurçades a l'augmentar els llindars del filtre post-test, en comptes de recórrer a nivells més astringents de significació.

8. Que les línies cel·lulars de càncer mostren taxes de secreció diferents en el seu estat basal, o sota pertobacions biològiques.

9. Que sota un model GLM, l'equació 1.7 permet una comparació cel·lula-a-cel·lula, de secretoma en línies cel·lulars, lliure de biaix.

10. Que la taxa de secreció i la de proliferació semblen estar inversament correlacionades en línies cel·lulars de càncer.

11. Finalment, les solucions desenvolupades i el model dissenyat han estat implementats en dos paquets R/Bioconductor, `msmsEDA` i `msmsTests`, i la seva disseminaci i ús per part de no experts ha estat facilitada amb dues interfcies gràfiques `msmsEDA_GUI` i `msmsTests_GUI` lliurement disponibles.

## 1.8 Altres publicacions

Tot seguit es dóna la llista de publicacions de l'autor de treballs en bioestadística/bioinformàtica aliens a la tesi però desenvolupats durant el periode de la tesi, i que han contribuit a la formació del doctorand.

1. *Inference with viral quasispecies diversity indices: Clonal and NGS approaches.*
   **Gregori J**, Salicrú M, Domingo E, Sánchez A, Esteban JI, Rodríguez-Frías F, Quer J
   Bioinformatics. 2014 doi: 10.1093/bioinformatics/btt768

2. *Ultra-deep pyrosequencing (UDPS) data treatment to study amplicon HCV minor variants.*
   **Gregori J**, Esteban JI, Cubero M, García-Cehic D, Perales C, Casillas R, Alvarez-Tejado M, Rodríguez-Frías F, Guardia J, Domingo E, Quer J
   PLoS One. 2013 Dec 31;8(12):e83361.
   doi: 10.1371/journal.pone.0083361

3. *A comparative study of ultra-deep pyrosequencing and cloning to quantitatively analyze the viral quasispecies using hepatitis B virus infection as a model.*
   Ramírez C, **Gregori J**, Buti M, Tabernero D, Camós S, Casillas R, Quer J, Esteban R, Homs M, Rodríguez-Frías F.
   Antiviral Res. 2013 May; 98(2):273-83. Epub 2013 Mar 20.
   doi: 10.1016/j.antiviral.2013.03.007

4. *Identification of host and viral factors involved in a dissimilar resolution of a hepatitis C virus infection.*
   Cubero M, **Gregori J**, Esteban JI, García-Cehic D, Bes M, Perales C, Domingo E, Rodríguez-Frías F, Sauleda S, Casillas R, Sanchez A, Ortega I, Esteban R, Guardia J, Quer J.
   Liver Int. 2013 Oct 17. [Epub ahead of print]
   doi: 10.1111/liv.12362

5. *Extinction of hepatitis C virus by ribavirin in hepatoma cells involves lethal mutagenesis.*

Ortega-Prieto AM, Sheldon J, Grande-Pérez A, Tejero H, **Gregori J**, Quer J, Esteban JI, Domingo E, Perales C.
PLoS One. 2013 Aug 16; 8(8):e71039. PMID: 23976977 Free PMC Article
doi: 10.1371/journal.pone.0071039

6. *Molecular epidemiology and putative origin of hepatitis C virus in random volunteers from Argentina.*
del Pino N, Oubiña JR, Rodríguez-Frías F, Esteban JI, Buti M, Otero T, **Gregori J**, García-Cehic D, Camós S, Cubero M, Casillas R, Guàrdia J, Esteban R, Quer J.
World J Gastroenterol. 2013 Sep 21;19(35):5813-27. PMID: 24124326 Free PMC Article
doi: 10.3748/wjg.v19.i35.5813

## 1.9 Informe factors d'impacte

En aquest informe es descriu el factor d'impacte i el quartil on es situa la revista en el seu camp per cada una de les publicacions que formen part d'aquesta tesi, i en les quals he estat codirector dels projectes.

– Article 1: "**Batch effects correction improves the sensitivity of significance tets in spectral counting-based comparative discovery proteomics**"

  Revista: Journal of Proteomics
  **Factor d'impacte: 4.08**
  Quartil: Q1

– Article 2: "**An effect size filter improves the reproducibility in spectral couting-based comparative proteomics**"

  Revista: Journal of Proteomics
  **Factor d'impacte: 4.08**
  Quartil: Q1

– Article 3: "**Unconventional Secretion is a Major Contributor of Cancer Cell Line Secretomes**"

  Revista: Molecular & Cellular Proteomics
  **Factor d'impacte: 7.25**
  Quartil: Q1

Els directors de la tesi

Alexandre Sánchez i Pla (UB) Josep Villanueva i Cardús (VHIO)

## 1.10 Informe participació en coautoria

En aquest informe es detalla en quines tasques va participar el doctorand en cada una de les publicacions que formen part de la tesi i en les quals he estat codirector dels projectes.

- Article 1: "**Batch effects correction improves the sensitivity of significance tets in specttral counting-based comparative discovery proteomics**"
  En aquest treball el doctorand ha participat en la concepció del projecte, el desenvolupament de la metodologia necessària, l'anàlisi i la interpretació de les dades, així com en lescriptura del manuscrit.

- Article 2: "**An effect size filter improves the reproducibility in spectral counting-based comparative proteomics**"
  En aquest treball el doctorand ha participat en la concepció del projecte, el desenvolupament de la metodologia necessària, l'anàlisi i la interpretació de les dades, així com en lescriptura del manuscrit.

- Article 3: "**A model for cell to cell comparisons**"
  En aquest treball el doctorand ha participat en la concepció del projecte, el desenvolupament de la metodologia necessària, l'anàlisi i la interpretació de les dades, així com en lescriptura del manuscrit.

- Article 4: "**Unconventional Secretion is a Major Contributor of Cancer Cell Line Secretomes**"
  En aquest treball el doctorand ha realitzat l'anàlisi estadístic de les dades de proteòmica quantitativa.

Els directors de la tesi

Alexandre Sánchez i Pla (UB) Josep Villanueva i Cardús (VHIO)

# Chapter 2

# INTRODUCTION

The first decade of this century has seen the emergence of a number of high throughput techniques in the '-*omics*' fields raising each time high expectations of success in the exploration and discovery of potential biomarkers useful in clinics for diagnosis or prognosis. The high throughput represented also high costs per sample, at least in their first stage, which limited the number of samples attainable per study. This limited sample size together with sophisticated protocols, the inherent biological variability, and the high number of variables simultaneously studied, brought with them a new challenge, generically known as the *curse of dimensionality* [Bellman, 1961], or the "*many-genes-few-replicates*" problem. This problem may be described in short as follows: when the dimensionality increases, the volume of the space increases so fast that the available data becomes sparse, and the clusters appear more and more fuzzy; to avoid its effects the sample size should grow exponentially with the dimensionality.

The pioneer of these high throughput techniques, the gene expression microarrays (MA), used to interrogate the expression of a genome, had to deal in first term with the growing disappointment of the medical community when promising discoveries could not be reproduced [Kuo et al., 2002; Tan et al., 2003; Ransohoff, 2004; Marshall, 2004; Dupuy & Simon, 2007; Borst & Wessels, 2010; Baggerly & Coombes, 2009; Potti et al., 2011]. A big consortium of experts led by the FDA was formed to study the causes of this apparent lack of reproducibility in an unprecedented community-wide effort [Shi et al., 2006]. The MAQC-I study (Figure 2.1) involved more than 600 hybridizations, across 7 platforms, including 137 participants from 51 organizations. The lessons were [Shi et al., 2006, 2008]:

1. The need for good experimental design.
2. Better suited statistical tools in discovery.

3. Control of the false discovery rate in the multiple tests.

4. Account for non controllable confounding factors [Luo et al., 2010].

The MAQC-II study [Shi et al., 2010] was devoted to describing good practices in the development and validation of predictors based on microarrays data, and provided a workflow and a strict set of rules [Parry et al., 2010].

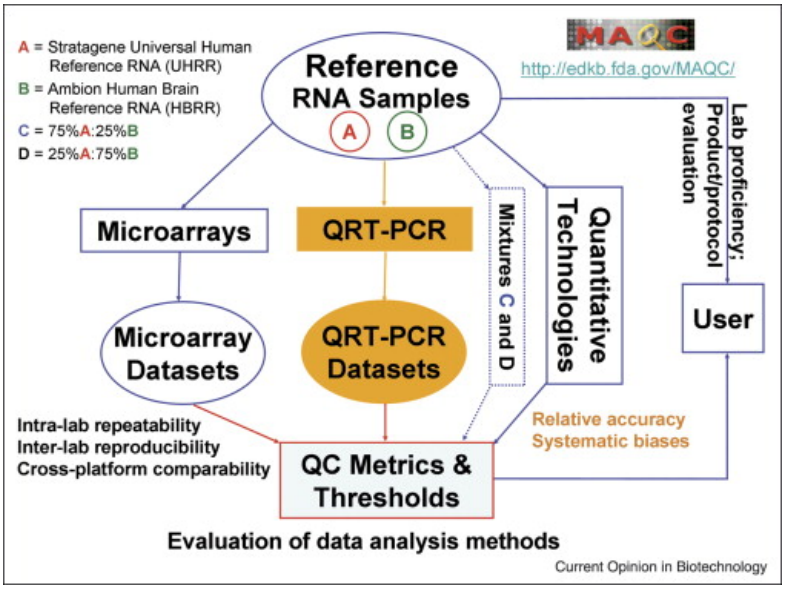



Figure 2.1: MAQC I

**This work intends to explore specific issues in comparative proteomics, and to translate and implement some of the lessons first learned with microarrays in the biomarker discovery stage**.

## 2.1 Biomarkers

A biomarker [Rifai et al., 2006; Madu & Lu, 2010] may be defined as a molecule, or set of molecules, whose level is indicative of some biological state or condition. In clinics a biomarker is relevant as predictive of disease state (diagnostic) or disease evolution (prognostic). Diagnostic biomarkers are baseline measurements which provide information about which patients are likely or unlikely to receive a given treatment. An example of biomarker is the level of expression on the `HER2` gene, a factor which transmits growth signals to breast cancer cells. An overexpression of `HER2` may be suggestive of a treatment with trastuzumab (Herceptin$^{TM}$)

which blocks the HER2 effects. The prognostic biomarkers are pre-treatment measurements informative about the evolution of the disease, the long-term outcome, under no treatment or under a given treatment. The prognostic biomarkers are risk indicators which could recommend more aggressive treatments. An example is the MammaPrint test. Broad patient genotyping (genomic markers) together with appropiate biomarkers are the key points in the paradigm of personalized medicine, the practice which prescribes the best treatment with the least side-effects for everyone.

An ideal molecular biomarker (Figure 2.2) has been defined [Madu & Lu, 2010] as: i) a molecule which is shown to correlate with the interested outcome, ii) quick, consistent, and economical in its determination, iii) quantifiable in an accessible biological fluid or clinical sample, and iv) that is readily interpretable by a clinician. Besides, its expression should be significantly increased (or decreased) in the related disease condition, and no overlap should exist in the levels of biomarker between healthy controls and untreated patients.

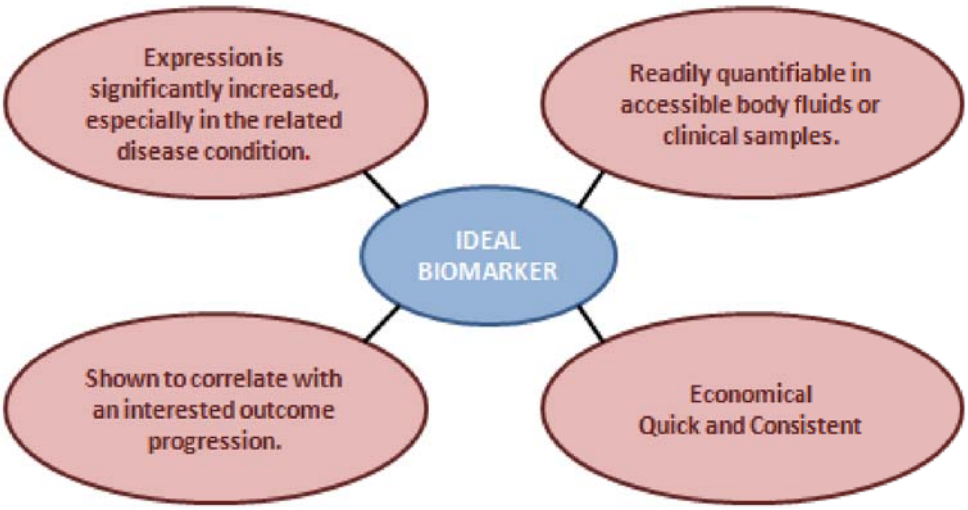



Figure 2.2: Ideal biomarker, taken from [Madu & Lu, 2010]

The '*omics*' high throughput technologies aim to interrogate the full exome, the full transcriptome, or the proteome of a sample at a time, and represent an unprecedented way for biomarker discovery. Nevertheless they must be considered as prospective discovery techniques, whose results should eventually be complemented and validated by more accurate methods.

The initial promises of quick development in the biomarker discovery by the

whole transcriptome analysis provided by the microarrays platforms were not immediately fulfilled. The reasons are multiple [Simon et al., 2003; Ransohoff, 2004; Dupuy & Simon, 2007; Baggerly & Coombes, 2009; Rakha, 2013]:

- High initial costs and scarcity of samples brought to small sample studies. Sometimes with poor experimental design.
- The biological variability within diseases has been found to be much higher than expected [Kim, 2009].
- The end-points resulted not so black-and-white as supposed [Rakha, 2013].
- What was suddenly possible in a conventional biomedical research lab required sophisticated statistic and bioinformatic data analysis.
- A new type of dataset was created, with tens of thousands of variables explained by very few samples.
- A new discipline emerged with it, and countless methods of data analysis were published at a pace difficult to follow, and still new methods are being developed [Sánchez-Pena et al., 2013].

As the initial costs reduced to affordable levels, the availability of clinical samples has been the major drawback in biomarker discovery by microarrays. The recent development of protocols, reagents, and microarrays able to cope with archived formalin-fixed paraffin-embedded (FFPE) samples [Hoshida et al., 2008] could revolutionize the field again. Examples are the cDNA-mediated annealing, selection, extension, and ligation (DASL) technology [Bibikova et al., 2004] adopted by Illumina, or the molecular inversion probe technology [Absalan & Ronaghi, 2007] adopted by Affymetrix.

Besides these improvements, and beyond genetic expression, arrays specific to address the finding of genetic markers, like chromosomal copy number aberrations, loss of heterozygosity, or single nucleotide polymorphism, have been developed and offer interesting and complementary diagnostic and prognostic tools [Ho et al., 2013].

In summary, the tool once developed to quickly uncover single and simple markers in research revealed an unexpected complexity, and because of that and because of new developments it is becoming by itself a tool useful in clinics [Matsui, 2013].

Biomarker discovery (BD), the central topic of this work, represents just the very first step in biomarker development, a long process involving the following steps [Rifai et al., 2006].

- **Discovery**: Identify candidate biomarkers.
- **Qualification**: Confirm differential abundance in the target fluid by a low throughput and highly accurate method.
- **Validation**: Assess sensitivity and specificity with independent samples.
- **Clinical assay development**: Establish sensitivity and specificity in a big cohort with given eligibility rules, and perform assay optimization.

When the biomarker is a signature - a set of genes or proteins, in general - the process of discovery becomes rather complex and involves strict protocols to avoid overfitting the predictor to the training data set. An independent data set is required for the validation of the signature, or in its absence a sufficient cross-validation must be performed. [Parry et al., 2010]. This was the matter of the MAQC-II study on common practices for the development and validation of microarrays-based predictive models [Shi et al., 2010].

As an example of this complex process on a classical molecular biomarker, the prostate-specific antigen (PSA), discovered in 1970, was found to be elevated in men with prostate cancer in 1980, took six years to reach FDA approval for monitoring cancer recurrence, and an extra 8 years for FDA approval for screening in conjunction with a digital rectal exam. However there are still some controversies over PSA screening as no study has successfully shown any correlation between such screening and a decline in mortality rate [Madu & Lu, 2010].

The protein `HER2` provides another example of classical biomarker. The epidermal growth factor receptor (`EGFR`, `HER1`) was discovered in 1978. The proto-oncogene `Neu` (`HER2`, `ERBB2`, `p185`) was discovered in 1982. The `HER2` amplification in breast cancer was discovered in 1985. The amplification of human epidermal growth factor 2 (`HER2`), expressed by the `HER2/neu` oncogene, was associated with a shorter time to relapse and lower survival rate in women with breast cancer in 1987. These findings were extended to ovarian cancer in 1989. The first test for `HER2` overexpression received FDA premarket approval in 1998. ASCO guidelines recommend `HER2` testing for all breast cancers in the same year. In 2002 the FDA

approved inclusion of the FISH (fluorescence in situ hybridation) gene amplification test for HER2 gene in Herceptin$^{TM}$ product labelling, included in the ASCO guidelines [Wolff et al., 2007].

The MammaPrint prognostic test is a gene signature fully developed in the *omics* era and provides an example of success despite some flaws in its development [Tibshirani & Efron, 2002] and recent warnings in its use [Rakha, 2013]. It is based on the Amsterdam 70-gene breast cancer signature discovered by microarrays experiments [van 't Veer et al., 2002]. It is used to analyse early-stage breast cancers under given eligibility criteria (stage I or II, invasive, smaller than 5cm, ER positive or negative), and predicts the risk level (high or low) of breast tumour metastasis within 10 years after diagnosis. It helps physicians to determine whether or not each patient will benefit from chemotherapy to reduce recurrence risk. The signature was discovered in 2002, and FDA-cleared as *in vitro diagnostic multivariate index assay* (IVDMIA) in 2007. It can be used both on FFPE or fresh tissue samples. MammaPrint is the only gene expression breast cancer test currently available in the United States that has met the FDA's IVDMIA criteria.

## 2.2 Proteomics

While DNA contains full information on the functions of a cell, and mRNA works as a messenger with pieces of information sent to the machinery of the cell, proteins are the functional part and more accurately reflect the phenotype. Also because of alternative splicings and post-translational modification a poor correlation has been found between the mRNA and the proteins in the cell [Gygi et al., 1999]. As the protein lays in a higher functional level than mRNA it is expected that proteomics could bring to the discovery of molecular targets and biomarkers more effectively than transcriptomics did [Gygi et al., 1999]. Nevertheless this field presents a few challenges to be solved.

The immediate target would be to study the proteome of tumor cells, nevertheless a very large fraction of the protein lysate corresponds to structural proteins, like cellular organelles, proteosome and proteins related to cellular core functions and protein translation. The fraction of disease specific proteins in the whole cell lysate is negligible.

Looking for non-invasive tests the best source would be blood plasma, the most comprehensive human proteome containing proteins from all tissues and processes, with disease specific secreted proteins. But blood-based biomarker discovery has

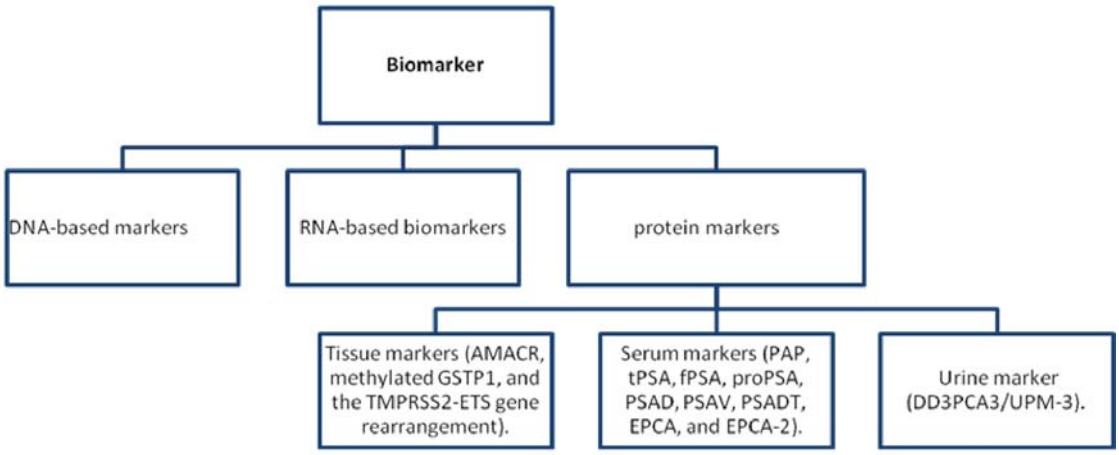



Figure 2.3: Prostate protein biomarkers, taken from [Madu & Lu, 2010]

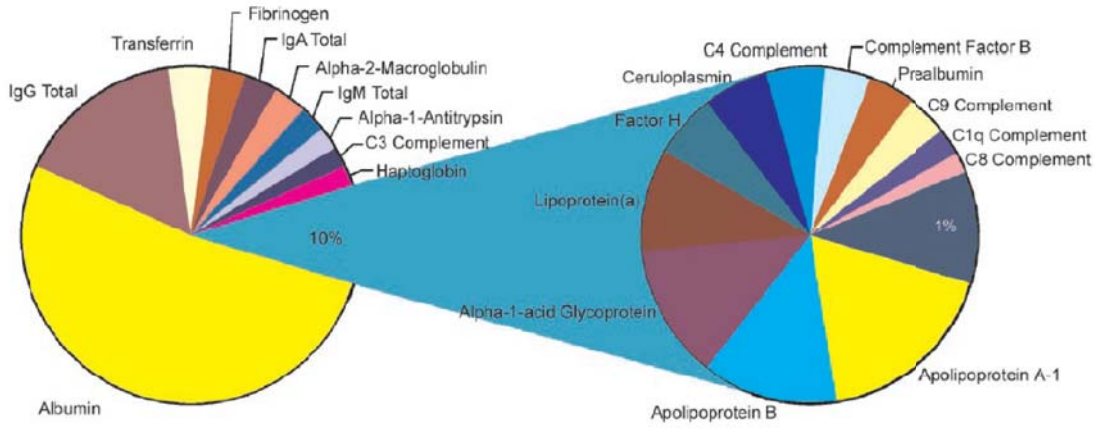



Figure 2.4: Pie chart representing the relative contribution of proteins within plasma. The top 22 proteins account for over a 99% of the proteome. Taken from [Tirumalai et al., 2003]

encountered important limitations. The proteome in plasma shows great complexity and a high dynamic range. With typical proteins like albumin accounting for over 50% of the total of proteome, and the top 22 proteins giving the 99% of the protein content of plasma (Figure 2.4), the relative concentration of disease-specific biomarkers is expected to be very low except in fortuitous cases [Tirumalai et al., 2003]. Many protein biomarkers used currently in clinics as diagnosis have concentrations in blood five to seven orders of magnitude lower than the most abundant proteins, with a total estimated dynamic range of eight orders of magnitude [Shen et al., 2005]. The Human Proteome Organization (HUPO) in its collaborative study to characterize human plasma initially reported 9504 proteins identified with one or more peptides and 3020 with two or more peptides, but reanalysis of the datasets from 18 laboratories led to just 889 proteins identified with a confidence level of at least 95% [States et al., 2006]. Since more abundant proteins interfere with the detection of less abundant proteins, extensive fractionation is required to achieve sufficient depth of coverage. This extensive fractionation may be multidimensional, meaning that the proteines or the constituent peptides are separated through successive electrophoretic and/or chromatographic steps by different physico-chemical properties. At the protein level by size exclusion chromatography (SEC), and at the peptide level by strong cation exchange chromatography (SCX), followed by reverse phase chromatography (RP) in-line with the mass spectrometer. It is estimated that a single blood sample could expand to over 50 fractions, each requiring a single LC-MS/MS experiment, severely limiting the throughput [Shen et al., 2005]. On the other hand, in biomarker discovery, a limited number of samples with a high number of proteins identified with low abundance would lead to low reproducibility and inflated false positives [Rifai et al., 2006]. One possible solution, for the use of plasma in biomarker discovery, is the depletion of the most abundant proteins, but this itself represents a challenge as albumin and other abundant proteins act as carriers and may easily bind proteins of interest that would be depleted along [Tirumalai et al., 2003].

## 2.3 Secretomes

As biomarkers specific for a particular disease arise locally from the affected tissue, it is expected that a fluid closer or in direct contact with the tissue will be enriched in these highly informative molecules. These proximal fluids are local sinks for proteins secreted or leaked from the tissue. Of particular interest is the interstitial

fluid [Rifai et al., 2006].

Disease models such as cell lines or genetically homogeneous animals provide an important alternative to human materials for biomarker discovery. These models provide easy sources of samples to discover biomarker candidates for subsequent assessment. The use of the cancer secretome has recently been proposed to interrogate tissue-proximal fluids and conditioned media of cell lines for biomarker discovery [Stastna & Van Eyk, 2012]. The presence of growth factors and proteases in these fluids, indicates that secretomes might help in monitoring critical aspects of cancer progression such as invasion and metastasis. In fact, a significant fraction of abnormally regulated genes in cancer encode secreted proteins [Gronborg et al., 2006; Lawlor et al., 2009; Mathias et al., 2009]. Cancer proteins contained in the secretome have already been linked to agiogenesis, tumor invasion and metastasis, either through cell autonomous mechanisms or tumor-stroma interactions [Stastna & Van Eyk, 2012].

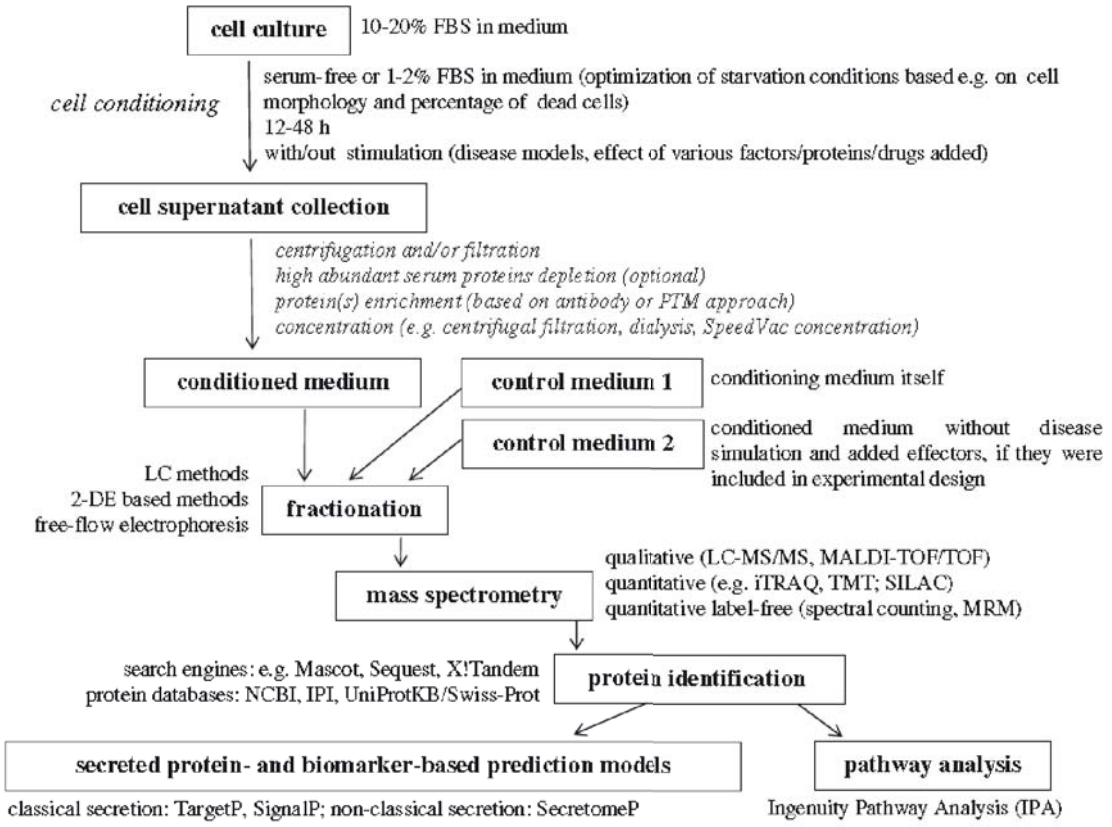



Figure 2.5: A secretome experiment workflow, taken from [Stastna & Van Eyk, 2012]

Despite the apparent simplicity of secretomes in cell lines, the analysis of secreted proteins faces analytical challenges that interfere with the search for true secreted tumor markers. We have to distinguish the secreted proteins from the intracellular proteins arising from cell death and from proteolysis induced during the cell culture handling. The conditioned medium may also contain exosomes and microsomal vesicles, an unconventional way of secretion of proteins from the cell [Stastna & Van Eyk, 2012]. The basic steps in the characterization of secreted proteins in cell conditioned medium are shown in Figure 2.5.

Contamination caused by the required use of serum for cell culture, despite the washes, has to be carefully taken into account. Also the serum-starvation phase may cause apoptosis and cell viability has to be ensured at a high level to avoid contamination by intracellular proteins [Villarreal et al., 2013]. Finally, when the differential expression of secreted proteins has to be referred to a single cell in each condition, the total amount of secreted protein and the number of viable cells which produced the protein need to be accurately measured, excluding any possible source of bias.

**The works in this thesis are addressed to the development of statistic tools that could help in the discovery of tumor biomarkers in cell-line secretomes, emphasizing its reproducibility**.

## 2.4 Elements of LC-MS/MS

LC-MS/MS stands for "Liquid Chromatography Tandem Mass Spectroscopy", a high throughput technique to analyse complex protein samples.

Proteins show a very wide fan of molecular weights and physicochemical properties, and some are just soluble under very specific conditions if any at all. In this respect, handling proteins is much more difficult than peptides, entities of lower molecular weight and generally soluble under varying conditions. In LC-MS/MS once the protein mix sample has been purified it is digested by enzymes (usually trypsin) into a complex mix of peptides. The advantage of trypsin digestion is that it cleaves the proteins very specifically on the carboxy-terminal side of arginine and lysine residues. This specificity greatly helps in the later identification of peptide sequences in the MS/MS.

The peptide digest may be fractionated by different chromatographic or electrophoretic techniques to obtain multiple samples of peptides of similar physico-

chemical properties, depending on the dynamic range and complexity of the proteome under study. Each fraction then is passed through a nanoflow capillary reverse phase chromatographic column to fed the mass spectrometer to identify and quantify each peptide.

Thus a high throughput proteomics experiment consists in the following steps [Steen & Mann, 2004] (Figure 2.6):

1. Isolation and purification of the proteins sample.
2. Digestion to a complex peptide mix.
3. Separation of the peptides in liquid chromatography by physicochemical properties.
4. Electro spray ionization of the eluted peptides.
5. Characterization of co-eluting ions by MS through their mass to charge ratio (m/z) and ion intensity.
6. Isolation and further fragmentation of parent ions by MS/MS to obtain their amino acid sequence.

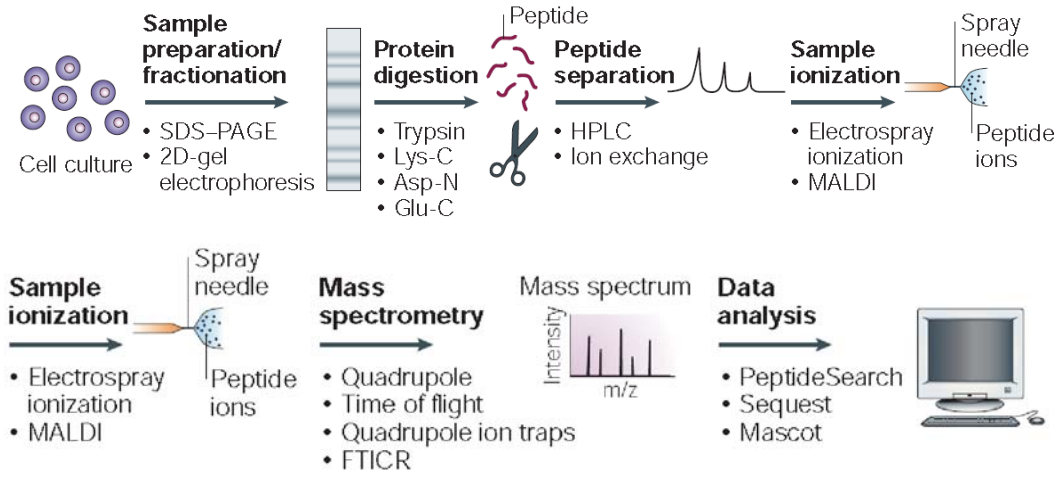



Figure 2.6: A proteomics experiment workflow, taken from [Steen & Mann, 2004]

In a LC-MS/MS run, the first mass spectrometer (MS) identifies the co-eluting ions by their m/z ratio (Figure 2.7). These parent ions are captured and fragmented in a second MS chamber by gas collision to obtain a typical sequence of m/z peaks, which may be identified as amino acid sequences by database searching

and applying sophisticated statistic procedures [Nesvizhskii et al., 2007] (Figure 2.8). The output of a LC-MS/MS run consist in thousands of MS scans and thousands of MS/MS fragmentation spectra that after peptide and protein identification will give rise to a list of identified proteins. Additionally, both the raw mass spectrometric data and the protein identification data contain information that can be used to do relative protein quantification of the proteins present in the sample.

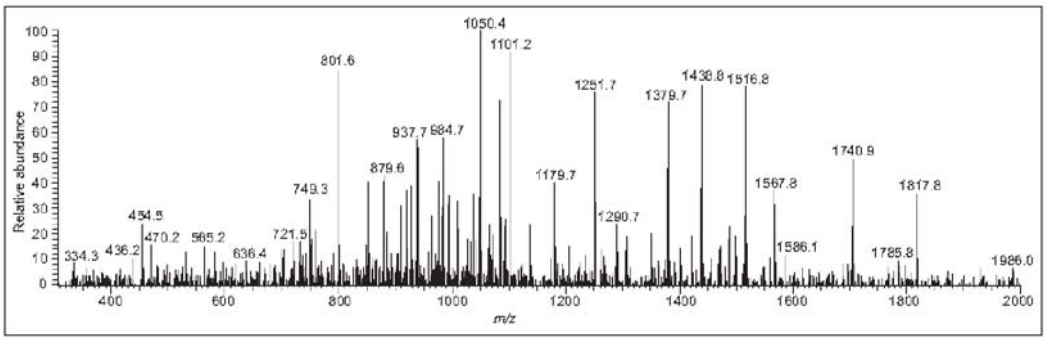



Figure 2.7: Chromatogram showing relative abundances of parent ions. Taken from [Ahn et al., 2007]

The proteins may be quantified by two methods. By ion signal intensity of the MS scan, that is "the signal intensity from the electrospray ionization", or by spectral counts (SpC), that is "the number of peptide MS/MS spectra assigned to each protein". The instrument should be optimized according to the method of choice. Multiple sampling of the chromatographic peak by survey mass spectra at the expense of MS/MS events is required for accurate quantification by ion signal intensities, but this could limit the number of proteins identified. Quantification by SpC depends of the number of MS/MS spectra assigned to peptides with high confidence, and is favoured by a higher number of MS/MS events at the expense of survey mass spectra.

The equipment used throughout this work (Figure 2.9) is formed by an EASY-nLC (Proxeon Biosystems, Thermo Fisher Scientific) on-line nanoflow liquid chromatographer with a two-linear-column system, and a linear ion trap LTQ Orbitrap Velos mass spectrometer (Thermo Fisher Scientific, Bremen, Germany). Ions were generated by applying 1.9 kV to a stainless steel nano-bore emitter (Proxeon, Thermo Fisher Scientific). The instrument was controlled by the software Xcalibur v2.1.0 (Thermo Fisher Scientific, Bremen, Germany). The experimental protocol

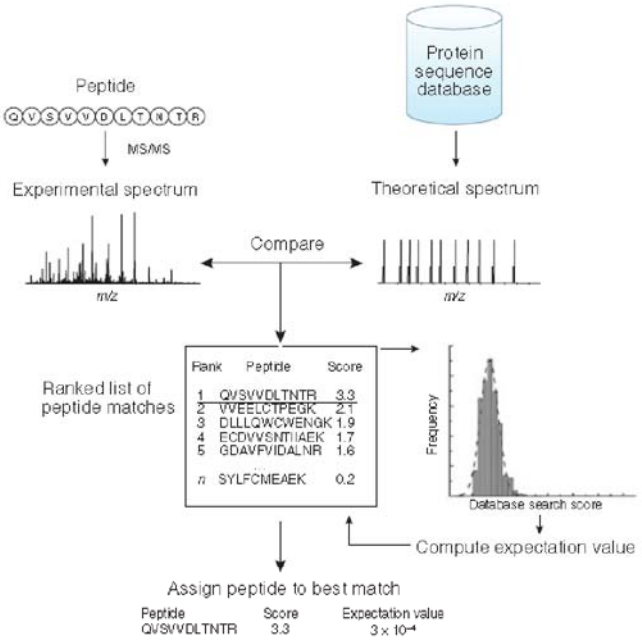



Figure 2.8: Peptide identification by MS/MS database searching, taken from [Nesvizhskii et al., 2007]

involved the following:

1. The LTQ Orbitrap Velos was operated in data-dependent mode.
2. A scan cycle was initiated with a full-scan MS spectrum (from m/z 300 to 1600) acquired in the Orbitrap with a resolution of 30,000.
3. The 20 most abundant ions were selected for collision-induced fragmentation in the linear ion trap when their intensity exceeded a minimum threshold, excluding single charged ions.
4. The maximum ion accumulation time was 500 ms in the MS and 200 ms in the MS/MS mode.

The papers collected in the appendix describe precisely the conditions used in each experiment.

## 2.5 Quantifying proteins by spectral counts

Protein quantification by ion intensity requires of sophisticated data treatment steps such as background signal identification, baseline correction, feature detec-

Figure 2.9: The LTQ Velos-Orbitrap equipment used throughout this work

tion and alignment, and signal normalization [Sandin et al., 2011]. SpC is more straightforward and requires just of a normalization.

A strong correlation between SpC and ion chromatograms in protein quantification has been demonstrated [Old et al., 2005; Gao et al., 2005], with SpC more reproducible and with higher dynamic range [Zybailov et al., 2005]. A recent expert review [Lundgren et al., 2010] considers instrument aspects, inherent limitations of SpC, common normalizations by protein length or mass, and relative or absolute quantification by SpC.

In the early implementation of SpC different normalizations were proposed. The purpose of these normalizations was to account for the protein complexity and the fact that in equal conditions longer proteins are expected to give rise to a higher number of peptides and thus of SpC. Parameters like the molecular weight, the amino acid sequence length, or the number of tryptic peptides were considered. The protein abundance index (PAI) [Rappsilber et al., 2002] divides the number of observed SpC by the number of tryptic peptides of the protein. The Protein Abundance Factor (PAF) [Powell et al., 2004] normalizes the SpC by the molecular weight of the protein. The normalized spectral abundance factor (NSAF) [Zybailov et al., 2006] normalizes the SpC of each protein dividing by the protein length and further dividing by the sum of SpC/L for all proteins identified in the experiment.

Biomarker discovery relays on relative quantification of a single protein in two biological states. Under these circumstances the normalization of SpC based on protein complexity contributes the same to the two values being compared, and for most statistical methods should have no effect. In fact [Lundgren et al., 2010] demostrated that the PLGEM method, originally based on NSAF values, performed better with raw SpC. **The methods developed in this thesis are based on raw SpC, and do not use any transformation to account for protein complexity as described above**.

## 2.6 Label-free differential expression

In biomarker discovery we look for differential expression between two biological conditions, generally disease and control. That is, unbiased relative quantification. To ensure this unbiased comparison, labelled procedures have been employed both in transcriptomics and in proteomics. Labelled procedures allow the early mix of the samples to be compared. Since they are pooled they are processed together

and any non controlled factor affects equally both conditions (Figure 2.10). The labels allow the posterior identification of features belonging to each condition. Two-color microarrays are an example in transcriptomics [Shalon et al., 1996]. In proteomics different approaches have been used. Among the most popular [Patel et al., 2009]: Stable isotope labelling by amino acids in cell culture (SILAC), isotope-coded affinity tags (ICAT), and isobaric tags for relative and absolute quantification (iTRAQ). The later is commercially available with eight isobaric tags.

Despite the advantage of labelled approaches they have potential limitations. Complex preparation steps, requirements for increased sample concentration and incomplete labelling, are the main issues. Labelling usually requires fractionation since all samples are analysed together, causing a drop in sensitivity. The key of this approach is that all samples to be compared are processed together.

Label-free proteomic analysis provide a more flexible and easy alternative. This means that each sample is processed and analysed separately (Figure 2.10). The counterpart is high risk of bias due to uncontrolled factors affecting differently one condition than the other [Neilson et al., 2011; Sandin et al., 2011; Zhu et al., 2010; Patel et al., 2009]. This requires carefully planned and designed experiments, to avoid confounding and bias as much as possible.

**This thesis is devoted to label-free differential expression analysis in cell-line secretome proteomics by LC-MS/MS, using SpC**.

## 2.7 Counts and inference

The result of a label-free proteomics experiment is contained in an expression matrix with SpC, where each row corresponds to a different protein and each column corresponds to a different sample belonging to a given biological condition. In this matrix we may have several technical and/or biological replicates for the same condition. We use this dataset to find those proteins which are differentially expressed in statistical terms, and which may be potential biomarkers. In what follows we give some background about the statistical methods which have been used in biomarker discovery in this field. The last part of this section reviews the state of the art in SpC comparative proteomics.

In the context of this work, a sample, or a replicate, is understood as a MS experiment. That is a run of the LC-MS/MS system, where a fixed amount of total protein is analysed and quantified in its components.

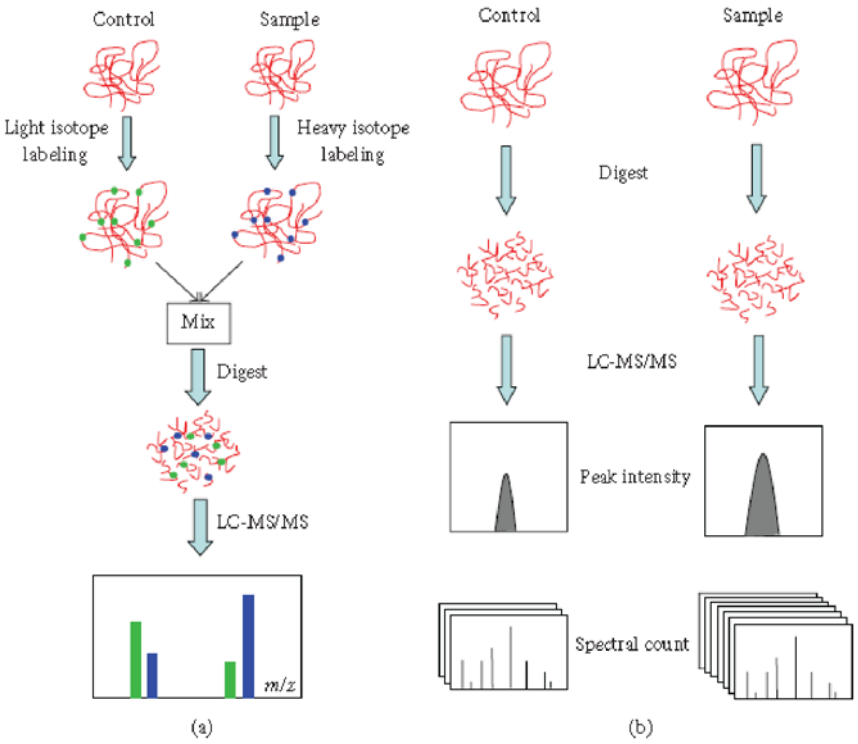



Figure 2.10: General approaches of quantitative proteomics. (a) Shotgun isotope labelling method. (b) Label-free quantitative proteomics. Taken from [Zhu et al., 2010]

Table 2.1: Contingency table

| | $Cond_1$ | $Cond_2$ | ... | $Cond_c$ | Total |
|---|---|---|---|---|---|
| SpC for target protein | $x_{11}$ | $x_{12}$ | ... | $x_{1c}$ | $x_{1o}$ |
| SpC for any other protein | $x_{21}$ | $x_{22}$ | ... | $x_{2c}$ | $x_{2o}$ |
| Total SpC in sample | $x_{o1}$ | $x_{o2}$ | ... | $x_{oc}$ | $n$ |

### 2.7.1 Contingency tables

The most basic approach to inference with SpC considers a single sample (LS-MS/MS run) in each condition to be compared, with no experimental replicates, and forms a contingency table for each identified protein [Zhang et al., 2006]. The table may contain multiple conditions, as in table 2.1, where the notation is given. Testing whether a given protein is differentially expressed between two biological conditions is equivalent to testing the statistical significance of the equality between two proportions. That is, inference on differential expression is done by the Fisher exact test, the Pearson $\chi^2$ test or the $G^2$ test [Agresti, 2002] with the usual restrictions. For instance the Pearson $\chi^2$ and the $G^2$ test require a minimum of 5 counts in each cell of the table.

The Fisher exact test applies to 2x2 tables, and assumes that the row totals and the column totals are fixed. Hence any entry in the table fully determines the others. Under the $H_o$ of independence, conditioning on both sets of marginal totals yields the hypergeometric distribution.

$$P(x_{11} = k) = \frac{\binom{x_{1o}}{k}\binom{x_{2o}}{x_{o1}-k}}{\binom{n}{x_{o1}}} \tag{2.1}$$

It is called exact test because it does not depend of asymptotic properties, and the p-values may be calculated from the exact distribution.

The Pearson's test assesses the null hypothesis of independence, that is $H_o$ : $\pi_{ij} = \pi_{io}\,\pi_{oi}$, $i = 1, 2$, $j = 1, \ldots, c$, where $\pi_{io}$ and $\pi_{oi}$ are the marginal probabilities of the table. The $\pi_{io}$ and $\pi_{oi}$ are estimated by maximum likelihood as the marginals $x_{io}/n$ and $x_{oi}/n$. When $H_o$ is true the expected values of $x_{ij}$, called expected frequencies, are $\hat{\mu}_{ij} = (x_{io}\,x_{oi})/n$. Thus the Pearson statistic, $X^2$, given by

$$X^2 = \sum_{ij} \frac{(x_{ij} - \hat{\mu}_{ij})^2}{\hat{\mu}_{ij}} \tag{2.2}$$

has asymptotically a chi-squared distribution for large samples, with $c-1$ degrees of freedom for two rows contingency tables as in 2.1.

The $G^2$ statistic is a likelihood-ratio statistic comparing the null model against the saturated model, that is

$$\Lambda = \frac{\prod_{ij}(\hat{\mu}_{ij})^{x_{ij}}}{\prod_{ij}(x_{ij})^{x_{ij}}} = \frac{\prod_{ij}(x_{io}x_{oj}/n)^{x_{ij}}}{\prod_{ij}(x_{ij})^{x_{ij}}} = \frac{\prod_{ij}(x_{io}x_{oj})^{x_{ij}}}{n^n \prod_{ij}(x_{ij})^{x_{ij}}} \tag{2.3}$$

and by the $G^2$ statistic becomes:

$$G^2 = -2\, log\, \Lambda = 2 \sum_{ij} x_{ij}\, log(x_{ij}/\hat{\mu}_{ij}) \tag{2.4}$$

The $G^2$ statistic, under the null hypothesis, has asymptotically a chi-squared distribution with $c-1$ degrees of freedom, as established by the Wilks theorem [Wilks, 1938]. The two statistics $G^2$ and $X^2$ are asymptotically equivalent, although the convergence to the $\chi^2$ is quicker for $X^2$ than $G^2$ [Agresti, 2002]. The results of the $G^2$ test are equivalent to a GLM Poisson regression with an intercept (see below).

These tests do not require experimental replicates. A single run (sample) of each condition is sufficient. When having experimental replicates of each condition, the SpC may be added or pooled to form a single contingency table as before. This is however not recommended as it inflates the results towards lower p-values and may result in a number of false postives. The methods given below are better suited to the situations with experimental replicates.

### 2.7.2 Transforming counts

When spectral count data for replicates of each biological condition are available we may wish to transform the counts into normally distributed variables, so that the classical t-test or the ANOVA could be applied for differential expression inference. Counts may be transformed by variance stabilization methods to better resemble a normal distribution when the mean is reasonably high [Kutner et al., 2005].

When the variance of a random variable is proportional to its mean, as is the case for counts, the square root transformations $x' = \sqrt{x}$ or $x' = \sqrt{x} + \sqrt{x+1}$ are helpful. When working with proportions $p_{ij} = x_{ij}/n$, then the arcsin transformation $x' = 2\,arcsin\,\sqrt{x}$ is suitable. The *log* transformation is indicated when the standard deviation is proportional to the mean and it is not recommended for counts [OHara & Kotze, 2010].

In these cases, as the variance has to be estimated from the data a minimum of three replicates are advised. Besides, the mean expression in each condition should be reasonably high for the approximation to be reliable.

### 2.7.3 Generalized Linear Models (GLM)

Most biomarkers are expressed at low concentrations, even in secretomes, so the assumptions required for the use of tests based in normality are very limiting. Instead we may use GLM methods which do not rely on the normal distribution. A GLM [Agresti, 2002] is specified by three components:

1. The response as a random variable $Y$ and its probability distribution.
2. A systematic component given by a linear combination of predictor variables.
3. A link function which relates $E(Y)$ with the linear predictor.

The distribution of $Y$ should belong to the exponential family, which has a probability mass function factorizable as:

$$f(y_i;\theta_i) = a(\theta_i)b(y_i)exp[y_iQ(\theta_i)] \tag{2.5}$$

Examples are the Poisson distribution, and the negative-binomial when the dispersion parameter $\phi$ is given. The link function that transforms the mean to the natural parameter $\theta_i$ is known as the canonical link. The canonical link for the Poisson or the negative-binomial distribution is the natural logarithm, and the regression model using this link is:

$$log\,\mu_i = \sum_j \beta_j\,x_{ij} \tag{2.6}$$

where $x_{ij}$ are the design matrix elements and $\beta_j$ the model parameters.

Dealing with counts brings naturally to the Poisson distribution, a distribution with just one parameter, $\mu$, the mean.

$$Pr(Y = k) = \frac{\mu^k e^{-\mu}}{k!} \tag{2.7}$$

This distribution explains the uncertainty in the number of SpC positively identified as belonging to a given protein, when the expected number of SpC for this protein at a given concentration is $\mu$.

The Poisson distribution has the property that the variance equals the mean. The higher the number of expected counts, the higher is its variance.

$$\mu = E[Y] = Var[Y] \tag{2.8}$$

When the only source of variation comes from the sampling process, as when running technical replicates, the Poisson distribution works very well. Nevertheless when doing biological experiments, to the typical variation in sampling technical replicates we have to add the biological variability expected of individuals belonging to the same biological condition. In these circumstances the Poisson model will underestimate the variance and the inference may bring to false positives in differential expression. This phenomenon is known as overdispersion. [Agresti, 2002] The immediate alternative is the negative-binomial (NB) distribution, which allows for overdispersion. The probability mass function of a NB random variable with mean $\mu$ and dispersion $\phi$ [Agresti, 2002; Robinson & Smyth, 2008] is given by:

$$Pr(Y = k) = \frac{\Gamma(k + \phi^{-1})}{\Gamma(\phi^{-1})\,\Gamma(k+1)} \left(\frac{1}{1+\mu\phi}\right)^{\phi^{-1}} \left(\frac{\mu}{\phi^{-1}+\mu}\right)^{k} \tag{2.9}$$

where the variance is a function of both mean and dispersion.

$$Var(Y) = \mu + \phi\mu^2 \tag{2.10}$$

When $\phi \to 0$ the NB reduces to the Poisson distribution. Values of $\phi$ greater than 0 bring to overdispersed distributions. Strictly speaking any value $\phi > -\mu^{-1}$ is permitted by the model, allowing for subdispersion as well.

An extension of the GLM, making abstraction of the true distribution, considers the mean-variance relationship

$$Var(Y) = \psi\mu_i \tag{2.11}$$

for some constant $\psi$. The case $\psi > 1$ corresponds to overdispersion. The likelihood equations for this model are identical to the Poisson model, and the model parameter estimates are the same, but $\psi$ is not assumed to be fixed at 1 and estimated from the data. This brings to the quasi-likelihood model which fits the Poisson model and multiplies the standard error estimates of the model parameters by the square root of $\hat{\psi}$, thus adjusting inference for overdispersion. [Agresti, 2002]

Still another approach to account for overdispersion is a mixed effects model (GLMM) where the intercept of a Poisson GLM model is a random effects term distributed normally [Agresti, 2002].

### 2.7.4 State of the art in comparative proteomics by SpC

All the methods seen so far have been used in differential expression in proteomics. The first to be explored were the methods based on contingency tables and variations of the t-test, or methods specifically developed for microarrays, Serial Analysis of Gene Expression (SAGE), or Digital Gene Expression (DGE). A short description of a few selected works is given in the following list:

- [Zybailov et al., 2006] found that the natural logarithm of NSAF normalized counts (see above) distributed normally, and applied the t-test to a quantitative profiling of membrane-associated proteins. The data set consisted in three biological replicates of each of two conditions. *S. cerevisae* cultured on $^{14}$N-rich or $^{15}$N-minimal media. Zero spectral count values were replaced by an empirically determined value of 0.16 to help in the log-transformation. The samples were pairwise analysed by LC-MS/MS to minimise bias and technical error, and the results were validated by functional analysis with radioisotope uptake assays for selected proteins.

- [Zhang et al., 2006] compared the Fisher exact test, the $G^2$ test, the $\chi^2$ test, the t-test, the Audic and Claverie test (AC), and the local-pooled-error test (LPE). The AC test [Audic & Claverie, 1997] was developed for DGE and calculates the conditional probability of finding $x_2$ counts in condition 2 when $x_1$ counts have been observed in condition 1, and requires no replicates. LPE is a method developed for microarrays [Jain et al., 2003] which pools proteins with similar counts by percentile intervals and fits a smooth local regression curve to estimate the variance. The test compares the median of the two conditions using the estimated variances, and requires few replicates because

of the pooling method. For the t-test and the LPE tests, a normalization was performed by dividing the protein spectral count in a particular experiment by the average spectral count across all the proteins in that experiment. This is done so that the global average count is the same across all LC-MS/MS experiments. The comparison of methods concluded that for fewer than three replicates the Fisher exact test, the $G^2$ test and the AC test performed similarly. The t-test was better with three or more replicates. The datasets used to compare the results consisted in a yeast lysate spiked with 6 human proteins at 0.25%, 1.25% and 2.5%.

– [Pavelka et al., 2008] used a power-law global error model (PLGEM), previously developed for microarrays analysis [Pavelka et al., 2004], on NSAF normalized counts. Missing values were replaced with zeros, and data were normalized by dividing each value by the mean value of the corresponding column. The authors compare the statistical properties of NSAF values with transcript abundance values from Affymetrix GeneChip data, and conclude that both follow a similar power-law. The method was tested on a dataset consisting of four biological replicates of a yeast cell culture grown in rich medium and harvested in logarithmic phase and in stationary phase. The results were evaluated by functional analysis, by significant enrichment of Gene Ontology (GO) annotation terms or Swiss-Prot keywords among the top ranked 100 proteins.

Since 2008 the works on methods based on GLMs dominate the field, and though some further work could be required these methods seem to be well established and to offer wider and more flexible solutions in that they may incorporate covariates and normalizing factors as offsets. Examples are:

– [Choi et al., 2008] developed a mixed effects generalized linear model (GLMM) where the sampling is modelled through a Poisson distribution and the biological variability is introduced by a normal random effects term. It used a hierarchical Bayes factor for taking into account the small number of replicates in the data, which is achieved by pooling information across proteins. The regression parameters are assumed to have prior normal distribution. The GLMM model incorporates two offset terms as normalizing factors accounting for protein complexity and total sample abundance. The method tends to give false positives (FP) with low signal/size effect and incorporates

a filter which flags those proteins likely to produce FP. The method is known as QSpec.

– [Pham et al., 2010] contrasts the performance of the beta-binomial test against the G-test, the t-test, the t-test with log-transformed SpC, LPE, and QSpec. The SpC are normalized to the total sample abundance. The authors found similar results for the beta-binomial, the t-test with log-transformed SpC and QSpec, where the former compares favourably. Although the beta-binomial allows for overdispersion, as the negative binomial, the former does not belong to the exponential family and cannot be used in a GLM framework nor incorporate covariates.

– [Li et al., 2010] using spiked samples of whole yeast lysate with 48 equimolar human proteins (UPS1, Sigma-Aldrich) at six abundance levels compares the performances of the Fisher's exact test, the Wilcoxon rank test, the Student $t$ test, the Poisson-based GLM, and the quasi-Poisson (quasi likelihood) GLM. The test used with these GLMs is the F-test in an ANOVA comparing the null and the alternative models. According to the authors this confers robustness to the comparisons when we have all zeroes in one condition. The GLMs incorporate an offset as normalizing condition for total sample abundance. At the highest spike level all methods performed similarly in identifying almost all the spiked proteins. At the two lowest spike levels the quasi-likelihood outperformed the other tests, at the cost of an increased number of false positives. The artificial low p-values are related either to 0 SpC in all replicates in one condition, or to non-zero values being all equal in one condition. Such proteins are flagged as to have NA or 0 cv values, and may be subject to separate analysis.

– [Lundgren et al., 2010] in an expert review discusses the main statistical methods applied to proteomics data, with a mention to the LPE test, to the work of Zhang [Zhang et al., 2006] comparing multiple tests, to PLGEM [Pavelka et al., 2008], and QSpec [Choi et al., 2008]. The review reports the finding that the reanalysed results of the Choi *et al.*'s synthetic data sets omitting the two normalizing offsets are indistinguishable of the original. The same was observed with PLGEM using both the raw and the abundance-normalized SpCs.

– [Leitch et al., 2012] compare the GLM and GLMM models with special attention to the GLM with quasi-likelihood or with the negative-binomial distribution, and to the GLMM provided by QSpec. The authors conclude that the quasi-likelihood is less conservative than the QSpec is, and the negative-binomial is more conservative. Both QSpec and quasi-Poisson are not robust to the zero variance problem. The negative-binomial is not robust when there are no observations in one of the groups. One last conclusion is that there is great potential for future research in this field. Of note, the authors used the F-test as proposed by [Li et al., 2010] for the quasi-likelihood [Li et al., 2010], but the Wald test for the negative-binomial GLM.

**In this work we implemented the Poisson GLM, the negative-binomial GLM, and the the quasi-likelihood GLM extension.**

## 2.8 Lessons in biomarker discovery

The issue of reproducibility is the most crucial in BD. The almost 20 years of experience on BD in transcriptomics is worth being considered in depth, in benefit of younger *omics*, particularly proteomics. In this section the main lessons learned in BD (including experiences in proteomics) are introduced. These lessons bring us to the objectives of this work.

### 2.8.1 Normalization

When comparing equal amounts of total substance between two biological conditions, differences in the treatment and measurement process of two samples introduce bias in the relative measures, and normalization attempts to correct this effect. Normalization is then understood as an attempt to compensate globally for systematic technical differences that affect equally all features in a sample. In sample normalization, all features in the sample are submitted to the same transformation. Ideally sample normalization brings the measures in all samples to the same scale, with identical origin.

Quantification by peak intensity requires of sophisticated data pre-processing and normalization steps [Degroeve et al., 2011; Sandin et al., 2011; Callister et al., 2006], the same as microarrays data [Bolstad et al., 2003; Park et al., 2003]. Normalization of SpC data is much simpler. The most extended method assumes that for a given amount of total protein, the sum of SpC should be the same. In general

when GLM models are used [Choi et al., 2008; Li et al., 2010; Leitch et al., 2012] the normalization is implicitly done with the help of an offset term in the model [Agresti, 2002].

$$E[y] = \mu$$

$$log\left(\frac{\mu}{size}\right) = \alpha + \beta x$$

$$log(\mu) = log(size) + \alpha + \beta x \tag{2.12}$$

where $\mu$ is the expected expression of a given protein, $size$ is the normalizing condition, and $\alpha$ and $\beta$ are the model parameters, with x equal to 0 for the control condition, or equal to one for the tumor condition. The term $log(size)$ is the *offset*. Different covariates or blocking factors may be added to this simple model. This is independent of the underlying distribution for the expression level $y$ of this protein.

Differential expression in the '*omics*' field is based on basic assumptions, not always explicitly given. When these assumptions are roughly fulfilled the comparisons between two biological states may be considered as nearly unbiased, provided that a good experimental design is used. These assumptions are usually taken for granted and receive no criticism in most, if not all, studies. In transcriptomics and proteomics, where equal amounts of substance gathered from two biological conditions are measured, it is considered that the cells produce globally an almost equal quantity of total substance. Under this assumption comparing equal amounts of substance corresponds to comparing the substance produced by equal number of cells. And this is a cell to cell comparison, where the cell is the biological unit of interest. Specifically when studying the full proteome of a tissue a very large fraction of intracellular proteins corresponds to structural proteins comprising cellular organelles, protein complexes such as the proteosome, and proteins related to cellular core functions as metabolism and protein translation. Structural and house-keeping proteins account for the vast majority of the proteome, and the assumption of almost equal yield of total protein per cell in the two biological states may be accepted.

**When studying secretomes this assumption deserves a careful consideration, as the number of involved proteins is drastically reduced, and no structural proteins or related to the metabolism, are expected.**

**If it fails, the normalization given in equation 2.12 could be non appropriate. This crucial issue is also investigated and solved in the thesis by means of an specific normalization based on offsets.**

### 2.8.2 Effect size and p-values

In the early times of microarrays, BD relied simply on observed fold-changes. Later the t-test or a rank test was used on transformed and normalized data to infer significance through a p-value assigned to each feature. The statistical methods grew in sophistication and multitest p-value adjustments with control of the false discovery rate were introduced [Allison et al., 2006]. The result of a transcriptomics study was a long list of features ordered by p-values. The top features were the statistically most significant. In this context and with different microarrays platforms commercially available, promising published biomarkers with apparently very high sensitivity and specificity could not be reproduced in different laboratories, or on different platforms. The next quotation exemplifies the situation in the years 2004-2006.

> *The unresolved issue of measurement variability and measuring variability has hampered the great hopes researchers had with the advent of microarrays technology and the human genome sequence project. Since consensus technological, analytical, and reporting processes were (and still are) largely missing, it appeared that not only were gene expression data irreproducible, but also the results were very much dependent on the choice of analytical methods. A lively discussion on the validity of microarrays technology resulted in publications and comments like "Microarrays and molecular research: noise discovery?" (Ioannidis 2005), "An array of problems" (Frantz 2005), countered by "Arrays of hope" (Strauss 2006), and "In praise of arrays" (Ying and Sarwal 2008), and publications which raise questions about the reproducibility of microarrays data (Marshall 2004; Ein-Dor et al. 2006) or showing increased reproducibility (Dobbin et al. 2005b; Irizarry et al. 2005; Larkin et al. 2005).*
>
> Andreas Scherer in Ch. 1 in Scherer [2009]

Under this pressure, regulatory (FDA) and academic authorities, together with commercial institutions constituted the MicroArray Quality Control (MAQC)

Consortium [Shi et al., 2006]. The MAQC brought together more than a hundred researchers at 51 academic, government and commercial institutions to assess the performance of seven microarrays platforms in profiling the expression of two commercially available RNA sample types. Results were compared not only at different locations and between different microarrays formats but also in relation to three more traditional quantitative gene expression assays [Shi et al., 2006]. The Nature Biotechnology issue of September 2006 was fully dedicated to the results of this MAQC-I study. Its editorial emphasized the importance of the project.

> *No technology embodies the rise of 'omic' science more than the DNA microarray. First reduced to practice in the early 1990s, it has since undergone numerous iterations, adaptations and refinements to achieve its present status as the platform of choice for massively parallel gene expression profiling. Today, several thousand papers describing data from microarrays are published each year. Sales of arrayers, array scanners and microarray kits to the academic and industrial R&D community represent a multi-billion-dollar business. The microarray has even made its first forays into the clinic, with the US Food and Drug Administration's approval of the 'AmpliChip' to help physicians tailor patient dosages of drugs that are metabolized differentially by cytochrome P450 enzyme variants. And yet doubts linger about the reproducibility of microarray experiments at different sites, the comparability of results on different platforms and even the variability of microarray results in the same laboratory. After 15 years of research and development, broad consensus is still lacking concerning best practice not only for experimental design and sample preparation, but also for data acquisition, statistical analysis and interpretation. Though problematic for bench research, lack of resolution of these issues continues to even more seriously hamper translation of microarray technology into the regulatory and clinical settings. Indeed, several regulatory authorities have been wrestling with the problem of how and when (and indeed whether) to implement microarray expression profiling data as part of their decision-making processes.*
>
> ...
>
> *Clearly, microarrays have a long way to go before they can be used to support regulatory decision-making or accurate and consistent predic-*

*tion of patient outcomes in the clinic. But the MAQC study has given us a solid foundation from which to build.*
Nature Biotechnolgy editorial in the September 2006 issue, 24 (9) 2006

The conclusions of the MAQC-I [Shi et al., 2006] study may be summarized as follows [Shi et al., 2008]. With careful experimental design and appropriate data transformation and analysis, microarrays data can be reproducible and comparable among different patforms and laboratories. The fold change results from microarray experiments correlate closely with the results from orthogonal assays like quantitative reverse transcription PCR (qRT-PCR).

One goal of the MAQC study was to optimize intra- and inter-platform reproducibility. The approach to achieve the highest degree of reproducibility was to limit the number of transcripts identified as differentially expressed (DEG), and to sort the corresponding genes using fold-change ranking with a nonstringent p-value cutoff.

These results were later questioned [Chen et al., 2007] and then reinforced comparing different gene selection procedures, and with the help of additional simulations [Shi et al., 2008]: *"We recommend the use of FC-ranking plus a nonstringent P cutoff as a straightforward and baseline practice in order to generate more reproducible DEG lists. Specifically, the P-value cutoff should not be stringent (too small) and FC should be as large as possible"* and *"Using FC and P together balances reproducibility, specificity, and sensitivity. Control of specificity and sensitivity can be accomplished with a P criterion, while reproducibility is enhanced with an FC criterion."*

The top genes are not required to be the most statistically significant but those with the highest effect size with a reasonable multi-test adjusted p-value.

**This lesson is implemented in this work in the form of a post-test filter flagging as unlikely reproducible those proteins with low signal and/or low fold-change despite being statistically significant. The list of differentially expressed proteins is ordered with increasing adjusted p-values, with the corresponding flag.**

### 2.8.3 Batch effects

The conclusions of the MAQC-I study are directly linked to the problem of *many-genes-few-replicates*. This suggests that by extending the number of replicates reproducibility may be improved, besides increasing the power in the detection

of DEGs. Unfortunately it has been observed that by increasing the number of replicates the chances of bias increase as well (Speed T. in [Scherer, 2009]).

When the experiments are collected within a long period of time bias may be unavoidable. This is related to the so known *batch effects*. Although good experimental design with appropriate use of randomization and blocking at each step may greatly help, batch effects seem to be unavoidable and ubiquitous.

The batch effect is defined to be systematic an unintentional, in contrast with the experimental noise which is random in nature. It refers exclusively to systematic technical differences when samples are processed and measured in different batches or in different times. These effects may be visualized by multidimensional techniques like Principal Components Analysis (PCA) (Figure 2.11), Singular Value Decomposition (SVD), Hierarchical Clustering (HC), or Heatmaps (HM) [Scherer, 2009; Luo et al., 2010; Chen et al., 2011; Lazar et al., 2013]. Ideally the samples should cluster by treatment level, with independence of the time in which they were treated and measured (See Figure 2.11).

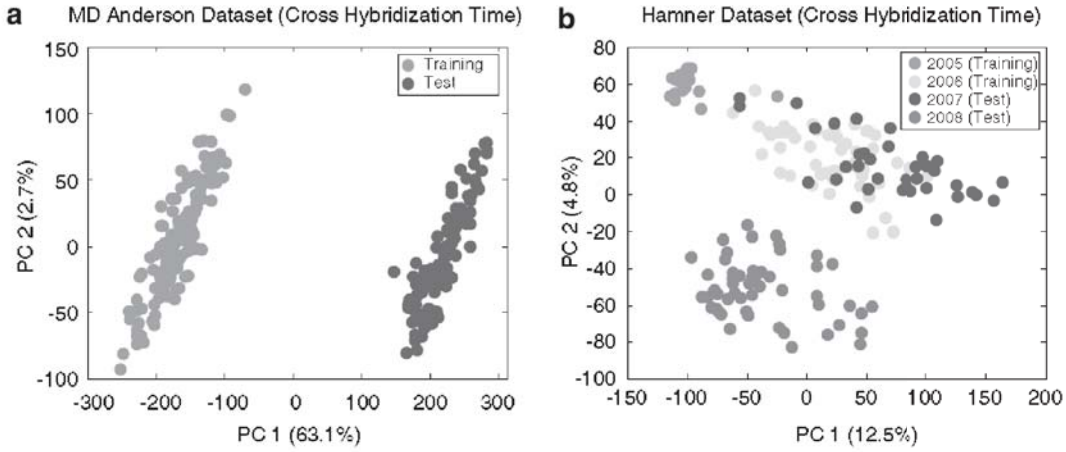



Figure 2.11: Visualization of the bias caused by batch effects, by PCA. Taken from [Luo et al., 2010]. (a) MD Anderson breast cancer data set. Training/test split was performed according to hybridization dates, the first 130 samples assayed were used as training set and the remaining 100 samples were used as test set. (b) Hamner lung carcinogen data set. Two batches in training set hybridized in 2005 and 2006, and two batches in test set hybridized in 2007 and 2008.

Because of the its systematic nature, the worst manifestation of batch effects is bias. And this usually occurs when most disease (or control) samples are processed separately of the others. As an example of the consequences of bias, in 2002 a study reported that a blood test, based on MS signatures, was 100% sensitive and specific

to detect ovarian cancer [Petricoin et al., 2002; Conrads et al., 2004]. A commercial screening test had to be introduced by 2004, but was delayed amid concerns of reproducibility. It was demonstrated that the study was seriously biased [Baggerly et al., 2004, 2005; Ransohoff, 2005b] with bias caused by batch effects. Another example on a test for ovarian cancer based on a microarrays signature is given by [Dressman et al., 2007] and [Baggerly et al., 2008]. Batch effects, together with overfitting in training predictors, are the major causes of flawed biomarkers and signatures [Baggerly & Coombes, 2009].

A mild manifestation of batch effects occurs when the batches are balanced in the number of samples of each biological condition to be compared. This manifestation is an increase of the intra-class variance reducing the power of the statistical tests for differential expression. Different methods of correction of batch effects have been proposed and compared [Luo et al., 2010; Chen et al., 2011; Lazar et al., 2013].

Batch effects are widespread and affect all *omics* [Ransohoff, 2005a; Scherer, 2009; Leek et al., 2010; Auer & Doerge, 2010; Schloss et al., 2011; Valsesia et al., 2013]. In words of one of the fathers of microarrays analysis:

> *Samples might come in one at a time, over months or years, but are commonly collected in batches. However the collection, processing and analysis are conducted, time or batch effects are unavoidable. Important though design is, and it is rightly emphasized in the book, there is in general little chance of entirely eliminating these effects. We must do our best with good design, but we must also plan to be in a position to identify and subsequently correct for those effects we are unable to eliminate by design.*
> Terry Speed in the foreword to the book [Scherer, 2009]

**In the R package `msmsEDA` we implemented multidimensional tools to evidence putative outliers and batch effects, and in the R package `msmsTests` we implemented GLM tools able to cope with block factors to correct potential batch effects.**

# Chapter 3

# OBJECTIVES

The establishment of a new proteomics biomarker discovery lab in the Vall d'Hebron Oncology Institute has offered the opportunity to develop tools and methods in comparative proteomics data analysis, and to contribute to translate some of the experience acquired in the *omics* data treatment with microarrays of over a decade to the field of BD with proteomics. In this respect, the objectives in the thesis were the implementation of tools and methods specifically devised to for:

- Data analysis of label-free shotgun proteomics experiments based on spectral counts.
- Dataset quality evaluation in terms of outliers and confounding factors.
- Modelization and normalization of cell-line secretomes data.
- Filters which could help to reinforce the reproducibility of biomarker discovery results in this field.
- Production of R packages encapsulating the tools developed.
- Production of graphical user interfaces to facilitate the use of these tools in a proteomics lab.

# Chapter 4

# RESULTS

All publications presented in this thesis have been submitted to international peer-reviewed journals. In what follows a summary and a description of the main findings in each paper is given. More than a reiteration of the contents of these papers, the findings and the conclusions relevant to the thesis are highlighted. It is an exercise of rewriting what this author considers as more relevant once the paper has been published.

The publications are divided in methodological, software, and applications. In the methodological papers the main contributions fulfill the objectives of the thesis. In the application papers, the methodology developed is applied to uncover biological responses in cell-line biomarker discovery. Under software, the packages developed along the works and contributed to Bioconductor are presented.

The chapter starts with a list of the publications with impact scores, and a summary of the software produced, then there is a section devoted to each paper and package. At the end of the chapter there is a section with a summary of other publications, in the same period 2011-2014. These publications, although not related to the contents of the thesis, belong to the field of bioinformatics / biostatistics, specifically to the NGS domain, and contributed to this doctorand education during the PhD.

## List of papers

1. *Batch effects correction improves the sensitivity of significance tests in spectral counting-based comparative discovery proteomics.*
   Gregori J, Villarreal L, Méndez O, Sánchez A, Baselga J, Villanueva J.
   J Proteomics. 2012 Jul 16; 75(13):3938-51. doi: 10.1016/j.jprot.2012.05.005. Epub 2012 May 12.
   **Journal impact factor**: 4.08

2. *An effect size filter improves the reproducibility in spectral counting-based comparative proteomics.*
   Gregori J, Villarreal L, Sánchez A, Baselga J, Villanueva J.
   J Proteomics. 2013 Dec 16; 95:55-65.
   doi: 10.1016/j.jprot.2013.05.030. Epub 2013 Jun 11
   **Journal impact factor**: 4.08

3. *Enhancing the Biological Relevance of Secretome-based Proteomics by Linking Tumor Cell Proliferation and Protein Secretion.*
   Gregori J., Méndez O., Katsila T., Pujals M., Salvans C., Villarreal L., Arribas J., Tabernero J., Sánchez A., Villanueva J.
   Submitted to J. of Proteome Research, pending of publication.
   (Journal impact factor 5.06)

4. *Unconventional secretion is a major contributor of cancer cell line secretomes.*
   Villarreal L, Méndez O, Salvans C, Gregori J, Baselga J, Villanueva J.
   Mol Cell Proteomics. 2013 May; 12(5):1046-60.
   doi: 10.1074/mcp.M112.021618. Epub 2012 Dec 26.
   **Journal impact factor**: 7.25

## Software

### R/Bioconductor packages

All software was developed in the R statistical environment and language [R Core Team, 2012], and the following two packages were produced:

1. `msmsEDA` package

Functions for the exploratory data analysis (EDA) of LC-MS/MS SpC datasets. Visual tools to discover outliers and eventual confounding factors. Dispersion analysis by factor, to help in choosing the best model.

2. `msmsTests` package

   Statistical tests for label-free LC-MS/MS data by spectral counts, to discover differentially expressed proteins between two biological conditions. Three tests are available: Poisson GLM regression, quasi-likelihood GLM regression, and the negative binomial of the `edgeR` package. The three models admit blocking factors to control for nuisance variables. To assure a good level of reproducibility a post-test filter is available, where the user may set the minimum effect size considered biologicaly relevant, and the minimum expression of the most abundant condition.

### Graphical user interfaces

Two graphical user interfaces (GUI) have been developed based on the functions in the two R packages, and with the help of the infrastructure provided by the R packages `gWidgets` and `RGtk2` [Verzani, 2012; Lawrence & Verzani, 2012; Lawrence & Temple Lang, 2010].

1. `msmsEDA_GUI`

   Useful in the exploratory data analysis of a LC-MS/MS experiment to assess the data quality, identify putative outliers, and detect potential batch effects in the dataset.

2. `msmsTests_GUI`

   Useful in BD by SpC, where a number of parameters may be easily adjusted to better fit the data specificities and to guarantee a good level of reproducibility.

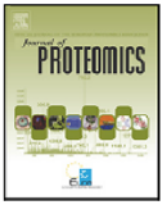
## 4.1 Paper 1: Batch effects

JOURNAL OF PROTEOMICS 75 (2012) 3938–3951

Available online at www.sciencedirect.com

SciVerse ScienceDirect

www.elsevier.com/locate/jprot



# Batch effects correction improves the sensitivity of significance tests in spectral counting-based comparative discovery proteomics

Josep Gregori[a,b], Laura Villarreal[a], Olga Méndez[a], Alex Sánchez[b,c], José Baselga[a], Josep Villanueva[a,*]

[a]Vall d'Hebron Institut of Oncology (VHIO), Barcelona, Spain
[b]Statistics Department, University of Barcelona (UB), Barcelona, Spain
[c]Statistics and Bioinformatics Unit, Vall d'Hebron Institut de Recerca, Barcelona, Spain

ARTICLE INFO

Article history:
Received 6 March 2012
Accepted 2 May 2012
Available online 12 May 2012

Keywords:
Biomarker discovery
Label-free quantitation
Secretome
Experimental design
Spectral counts
Quantitative proteomics

ABSTRACT

Shotgun proteomics has become the standard proteomics technique for the large-scale measurement of protein abundances in biological samples. Despite quantitative proteomics has been usually performed using label-based approaches, label-free quantitation offers advantages related to the avoidance of labeling steps, no limitation in the number of samples to be compared, and the gain in protein detection sensitivity. However, since samples are analyzed separately, experimental design becomes critical. The exploration of spectral counting quantitation based on LC-MS presented here gathers experimental evidence of the influence of batch effects on comparative proteomics. The batch effects shown with spiking experiments clearly interfere with the biological signal. In order to minimize the interferences from batch effects, a statistical correction is proposed and implemented. Our results show that batch effects can be attenuated statistically when proper experimental design is used. Furthermore, the batch effect correction implemented leads to a substantial increase in the sensitivity of statistical tests. Finally, the applicability of our batch effects correction is shown on two different biomarker discovery projects involving cancer secretomes. We think that our findings will allow designing and executing better comparative proteomics projects and will help to avoid reaching false conclusions in the field of proteomics biomarker discovery.



J Proteomics. 2012 Jul 16;75(13):3938-51. doi: 10.1016/j.jprot.2012.05.005. Epub 2012 May 12.

### 4.1.1 Aim

The aim of this work was to study the influence of bacth effects in label-free comparative proteomics based on SpC. It is done on one side by detecting them using multidimensional methods as in transcriptomics [Scherer, 2009]. On the

other, when observed, by assessing the improvement in the BD results obtained by common batch correction methods [Luo et al., 2010; Chen et al., 2011; Lazar et al., 2013].

### 4.1.2 Background

See section 2.8.3 with a description of the importance of batch effects in the '-*omics*'.

### 4.1.3 Findings

A number of technical replicates of yeast lysate with and without spiked human proteins were measured at different dates spanning a few months. Also, to approach real life BD proteomics projects, control/treatment experiments with cancer cell lines secretomes spanning a few months were performed, and the datasets were studied by EDA techniques.

1. The experiments done with technical replicates of yeast lysate, using each time the same quantity of total protein, showed that the most reliable and stable sample normalization was to scale the dataset to the total signal obtained for the sample. This option showed far better results than using internal controls, even when multiple controls were employed.

2. When technical replicates of a yeast digest were run by LC-MS/MS for a few days, differences in the abundance of proteins could be observed among the runs, even after normalization by total SpC. The analysis of the data by HC and PCA revealed a clear sample partitioning by the day of the LC-MS/MS run (Figure 4.1). The influence of the observed batch effects was assessed by a GLM Poisson test between yeast samples run on different days, which gave several significant differences attributable to non controlled variables influencing the experiment (Figure 4.2). Conversely, no significant differences were found when comparing samples run in the same day.

   To further confirm this finding, a number of samples composed of 48 equimolar human proteins (Universal Proteomics Standard Set, UPS1, Sigma-Aldrich®) were run on consecutive days. The results with the UPS1 samples confirmed the trend previously observed with the yeast samples. Furthermore, this data confirms our view of a *sample batch as being the samples run by LC-MS/MS during the period of 24 h*, since this is the smallest fraction of time

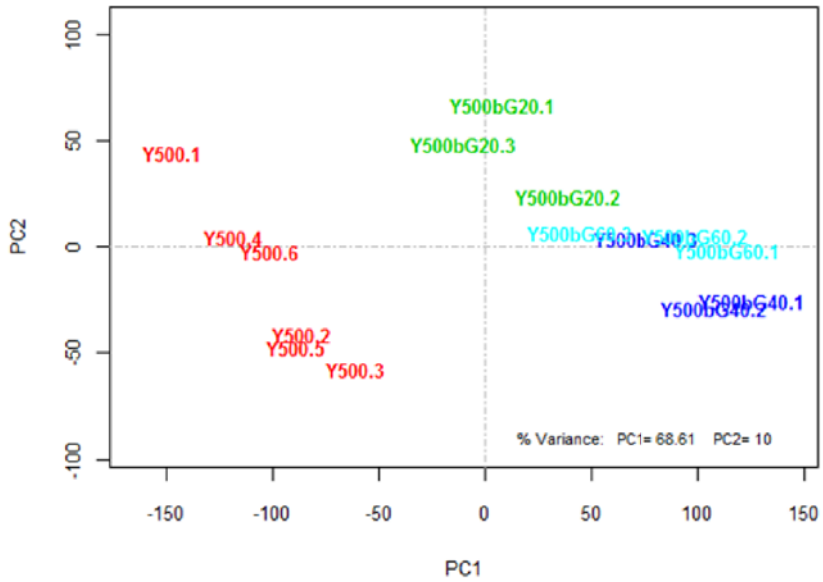



Figure 4.1: Batch effects on yeast lysate. PCA of yeast lysate samples measured in LC-MS/MS runs performed on different dates. Each run is colored differently.

upon sample comparison showed non-significant variability. This data shows that sample batch effects seem to be involved in the day-to-day variability observed in spectral counting-based quantitation.

3. Our limited experimental dataset show that long hydrophobic peptides eluting at the end of reversed-phase chromatography tend to be more sensitive to batch effects.

4. To study the influence of the batch effects and the correction methods, a set of technical replicates of 500ng standard yeast lisate samples spiked with 200 and 600fm of UPS1 were analysed spanning a few weeks, in five balanced runs according to the experimental design in table 4.1.

Table 4.1: Experimental design.

| | Batch | | | | |
|---|---|---|---|---|---|
| Spiking | 12.01 | 13.01 | 20.01 | 03.02 | 06.02 |
| 500ng yeast + 200fm UPS1 | 2 | 2 | 2 | 3 | 3 |
| 500ng yeast + 600fm UPS1 | 2 | 2 | 2 | 3 | 3 |

The effect of two methods of batch effects correction on the performance

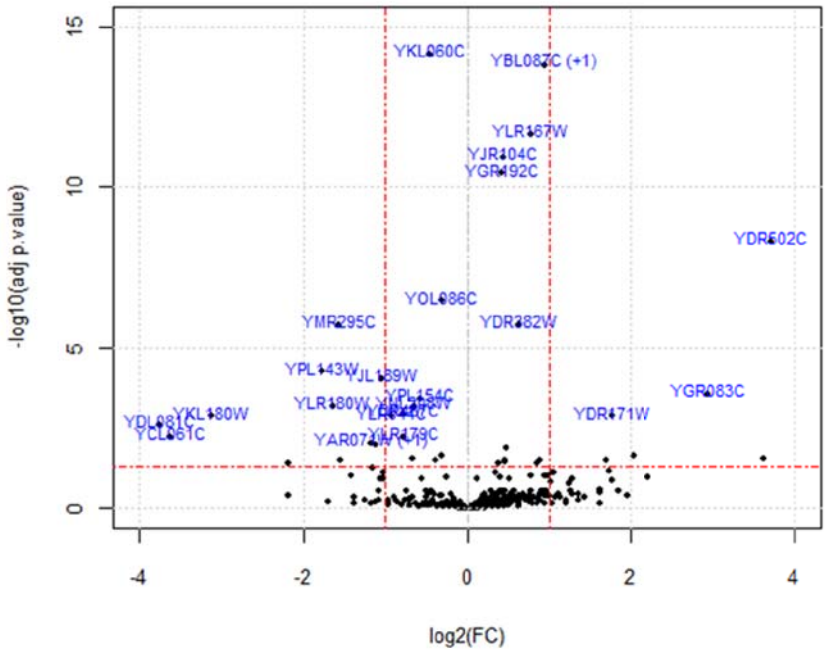



Figure 4.2: Volcano plot showing false positives due to batch effects on yeast lysate.

of two typical statistical tests was evaluated. The statistical tests selected were the square root transformation of SpC followed by ANOVA [Kutner et al., 2005], and the quasi-likelihood with log link (QLLL) test [Li et al., 2010; Agresti, 2002]. The two methods of batch correction selected, were the batch mean-centering [Luo et al., 2010; Lazar et al., 2013], and a blocking factor in the model [Kutner et al., 2005; Agresti, 2002]. The batch mean-centering is an additive correction that brings the center of all batches on the same point [Luo et al., 2010]. On the other hand a blocking factor contributes an additive correction (a shift) for ANOVA, or a multiplicative correction (a scaling) for the QLLL.

Table 4 shows the results of the tests at two typical significance levels, 0.05 and 0.01, when carried out on the raw SpC matrix, after the correction by batch mean centering, and with a batch block factor in the model.

Both the ANOVA and the QLLL test clearly benefit from the batch effect correction in terms of TP. The impact in the number of the FP may be reduced with a more stringent significance level. The QLLL is not robust to very low variances [Leitch et al., 2012] and produces a higher number of FP. The reduced variance is a direct consequence of the batch effects correction.

Table 4.2: Incidence of batch effects correction.

| Test | Data treatment | Significance 0.05 TP | FP | Significance 0.01 TP | FP |
|---|---|---|---|---|---|
| ANOVA | Raw SpC | 22 | 0 | 18 | 0 |
| | Block factor | 44 | 4 | 29 | 1 |
| | Batch mean centering | 45 | 10 | 31 | 4 |
| QLLL | Raw SpC | 22 | 2 | 19 | 0 |
| | Block factor | 47 | 56 | 45 | 21 |
| | Batch mean centering | 42 | 31 | 28 | 5 |

According to the area under the curve (AUC) in the ROC curves, modelling with a blocking factor gives better results than the mean-centering approach with both tests (Figure 4.3). The FP were fully controlled in both tests by excluding the significant features with low signal (<2 mean Spc in the most abundant condition) or low effect size (absolute logFC <1).

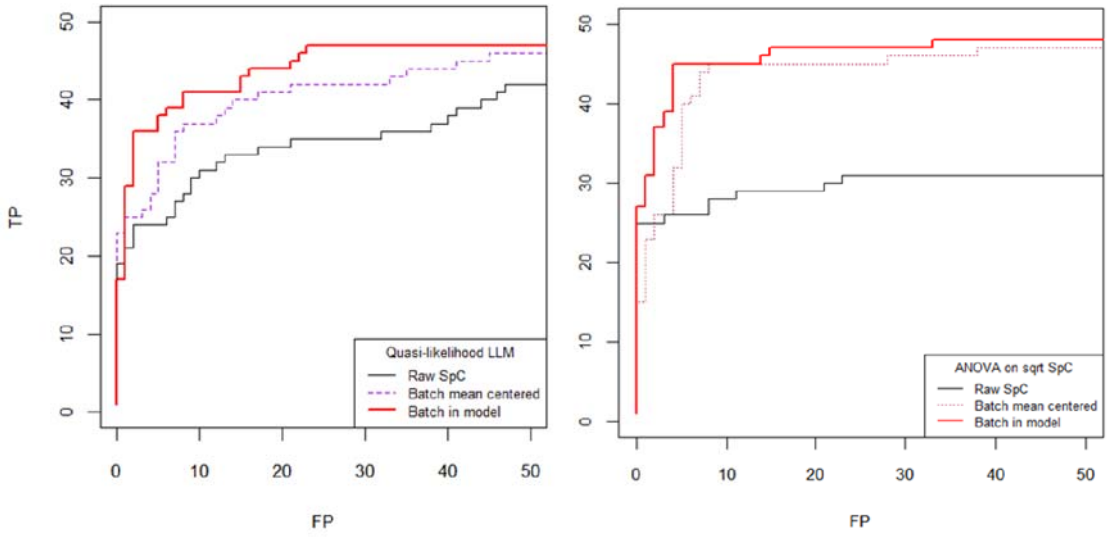



Figure 4.3: ROC curve showing the impact of the batch effects correction. Left) Quasi-likelihood. Right) ANOVA on square root transformed SpC.

5. Besides the spikings, two biological studies on cancer cell lines were evaluated for batch effects by EDA. The two studies represent two different levels of biological signal. The first with a moderate effect (HMEC treatment with TGF $\beta$), and the second with a strong effect (MCF7 versus MDA231). These

evaluations show that *batch effects are probably present on every proteomics project done for more than one day of instrument data acquisition.* It also shows that they can be corrected when using a proper experimental design followed by a batch effects correction method. Despite not having evaluated the significance tests for these projects because the list of true positives is unknown, the data showed that only after batch effect correction could the two conditions in the HMEC with TGF$\beta$ project be separated (Figure 4.4). In the other project (MCF7 versus MDA231), where *the differences in abundance for several proteins in the secretomes are large, batch effects were not able to mask the biological signal between the two conditions, but sample batches were still evident by EDA tools.*

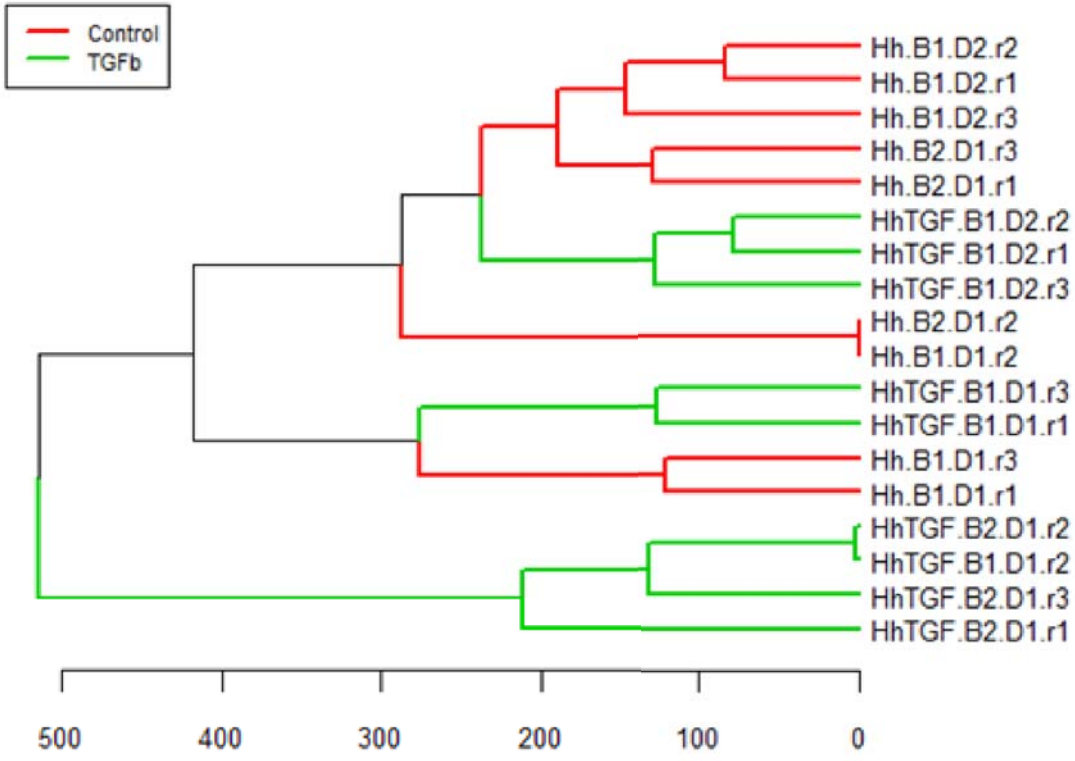



Figure 4.4: Average linkage HC dendrogram showing batch effects in a project on HMEC cells treated with TGF$\beta$.

### 4.1.4 Conclusions

1. Normalizing by total SpC by sample results in a more stable normalization than using internal standards, even when multiple standards are used.

2. Batch effects are probably present on every proteomics project done for more than one day of instrument data acquisition.

3. Fully unbalanced LC-MS/MS runs will likely produce biased results. The eventual bias cannot be corrected by any means.

4. When the LC-MS/MS runs are balanced in the conditions to compare, the batch effects produce a higher intraclass variability, but the results are unbiased.

5. The intraclass variance may be reduced by batch effects correction methods.

6. The reduced variance after batch effects correction may produce a high number of FP, which may be controlled by a signal and effect size filter.

7. The longer and more hydrophobic peptides are more sensitive to the observed batch effects.

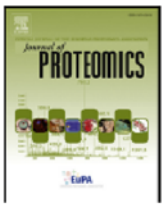
## 4.2 Paper 2: Reproducibility

JOURNAL OF PROTEOMICS XX (2013) XXX–XXX

Available online at www.sciencedirect.com

SciVerse ScienceDirect

www.elsevier.com/locate/jprot



# An effect size filter improves the reproducibility in spectral counting-based comparative proteomics☆

Josep Gregori[a,b], Laura Villarreal[a], Alex Sánchez[b,c], José Baselga[a], Josep Villanueva[a,*]

[a]Vall d'Hebron Institute of Oncology (VHIO), Universitat Autònoma de Barcelona (UAB), Barcelona, Spain
[b]Statistics Department, University of Barcelona (UB), Barcelona, Spain
[c]Statistics and Bioinformatics Unit, Vall d'Hebron Institut de Recerca, Barcelona, Spain

ARTICLE INFO

*Keywords:*
Effect size
Feature filters
Poisson
Quasi-likelihood
edgeR
QSpec

ABSTRACT

The microarray community has shown that the low reproducibility observed in gene expression-based biomarker discovery studies is partially due to relying solely on p-values to get the lists of differentially expressed genes. Their conclusions recommended complementing the p-value cutoff with the use of effect-size criteria. The aim of this work was to evaluate the influence of such an effect-size filter on spectral counting-based comparative proteomic analysis. The results proved that the filter increased the number of true positives and decreased the number of false positives and the false discovery rate of the dataset. These results were confirmed by simulation experiments where the effect size filter was used to evaluate systematically variable fractions of differentially expressed proteins. Our results suggest that relaxing the p-value cut-off followed by a post-test filter based on effect size and signal level thresholds can increase the reproducibility of statistical results obtained in comparative proteomic analysis. Based on our work, we recommend using a filter consisting of a minimum absolute $\log_2$ fold change of 0.8 and a minimum signal of 2–4 SpC on the most abundant condition for the general practice of comparative proteomics. The implementation of feature filtering approaches could improve proteomic biomarker discovery initiatives by increasing the reproducibility of the results obtained among independent laboratories and MS platforms. This article is part of a Special Issue entitled: Standardization and Quality Control.

J Proteomics. 2013 Dec 16; 95:55-65. doi: 10.1016/j.jprot.2013.05.030.
Epub 2013 Jun 11.

### 4.2.1 Aim

As mentioned in the introduction (see section 2.8.2) the issue of reproducibility is very important in BD. As pointed out by the MAQC-I study [Shi et al., 2006, 2008] low p-values are no guarantee of reproducibility. Instead, features with a combination of good effect size and low p-value are seen as reliably reproducible. The aim of this work was to study feature filters with a good balance of reproducibility and sensitivity in our field.

### 4.2.2 Background

Feature filtering is a method commonly used in microarray data analysis [Pounds & Cheng, 2005] to reduce the impact of multiple test p-values adjustment with FDR control. This is because of the huge multiple testing penalty incurred when dealing with tens of thousands of variables.

*A priori* filters - that is filters used before the tests - should be non-specific or unsupervised so that the sample class labels are not seen before the test, as data-based filtering constitutes by itself a statistical test [Bourgon et al., 2010]. Examples of the most used filters are by signal or by variance. Nevertheless when using these filters together with a FDR correction method, a FDR bias may occur despite the filter/test independence. The reason is that FDR methods rely on the assumption that the null p-values follow a uniform distribution. To keep the distribution uniform, the probability to exclude a feature should be independent of the test p-value [Iterson et al., 2010]. Because of all these reasons the use of a priori filters is an active field that deserves further study [Iterson et al., 2010]. In transcriptomics, where tens of thousands of probes are evaluated, the impact of filtering by signal or variance up to a 40-50% seems not very relevant. Instead in cell-line secretomes, where the number of features studied is in the order of very few thousands, this impact might not be negligible.

According to clues observed in the previous work, and to the recommendations of the MAQC-I study [Shi et al., 2006, 2008], an alternative could be to use *a posteriori* or post-test filters together with a less stringent p-value threshold. The increase in FP due to a less stringent p-value could be largely compensated by the filter. The aim is to limit the FPs beyond the nominal FDR value and to increase the reproducibility of declared DEPs.

### 4.2.3 Findings

1. To study the impact of the proposed filters a series of experiments with spikings were done. The experimental setting consisted in a yeast tryptic digest with different controlled amounts of 48 equimolar human proteins (Universal Proteomics Standard Set, UPS1, Sigma-Aldrich®). Each sample consisted in 500 ng of yeast lysate with either 100, 150, 200, 400, 600 or 750 fm of UPS1. Despite being equimolar, the 48 human proteins span the full range of observed SpC because of its different structural complexities and physico-chemical properties (see Figure 4.5), providing a good model for this

study.

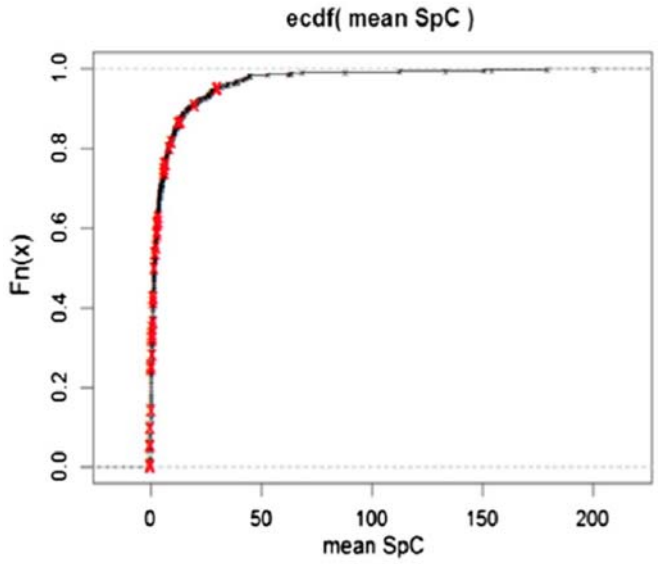



Figure 4.5: UPS1 expression range. The UPS1 proteins in red, and the yeast proteins in black, on the SpC cumulated distribution curve.

Five different statistic tests were used. GLM Poisson [Agresti, 2002], GLM quasi-likelihood [Li et al., 2010], `edgeR` [Robinson et al., 2010], QSpec [Choi et al., 2008], and the t-test on square root transformed SpC [Kutner et al., 2005]. The tests compared pairs of the given conditions. The confusion matrices (truth tables) of the tests with and without post-test filter were built out of the results and compared. The confusion matrix without filter was constructed with the significant proteins at an adjusted p-value threshold of 0.01, to limit the number of FP. The confusion matrix with filter was constructed with the significant proteins at an adjusted p-value threshold of 0.05, thus relaxing the former threshold, and showing a minimum of 2 mean SpC in the most abundant condition, and a minimum absolute logFC of 0.8 (See table 4.3).

The first observation was that the positive predictive value (PPV), measured as the ratio TP/(TP+FP), improved notably with the filter, and that the differences among tests were reduced. The sensitivity also increased with the filter, while the FDR decreased as a general trend.

No test dominates absolutely over the others when considering all comparisons made, although the Poisson, the quasi-likelihood and `edgeR` work rea-

Table 4.3: Results with and without post-test filter

| Comparison | Statistical test | Adjusted $p.val \leq 0.01$ TP/FP | Filter + adj. $p.val \leq 0.05$ TP/FP | TP among 35 top ranked |
|---|---|---|---|---|
| 750 vs 150 | Poisson | 18/2 | 26/1 | 28 |
| 4 repl. 4 | QL | 7/1 | 21/1 | 21 |
| | edgeR | 16/1 | 19/0 | 28 |
| | t-Test | 8/6 | 15/1 | 20 |
| | QSpec | 25/16 (24/1) | 24/1 | 24 |
| 600 vs 150 | Poisson | 14/8 | 20/9 | 20 |
| 6 repl. 4 | QL | 14/12 | 23/4 | 18 |
| | edgeR | 14/5 | 18/8 | 21 |
| | t-Test | 13/3 | 17/3 | 19 |
| | QSpec | 20/35 (17/9) | 17/6 | 16 |
| 600 vs 200 | Poisson | 27/9 | 26/0 | 26 |
| 6 repl. 12 | QL | 28/7 | 29/2 | 28 |
| | edgeR | 23/4 | 26/0 | 28 |
| | t-Test | 21/2 | 22/0 | 29 |
| | QSpec | 29/30 (25/1) | 25/0 | 22 |
| 400 vs 200 | Poisson | 11/6 | 12/1 | 19 |
| 4 repl. 12 | QL | 3/0 | 8/0 | 22 |
| | edgeR | 6/0 | 11/0 | 21 |
| | t-Test | 0/0 | 4/0 | 16 |
| | QSpec | 15/18 (6/0) | 6/0 | 15 |
| 200 vs 100 | Poisson | 1/4 | 1/3 | 9 |
| 12 repl. 4 | QL | 0/2 | 0/0 | 8 |
| | edgeR | 0/3 | 1/2 | 11 |
| | t-Test | 0/0 | 1/0 | 14 |
| | QSpec | 5/20 (1/2) | 1/2 | 7 |

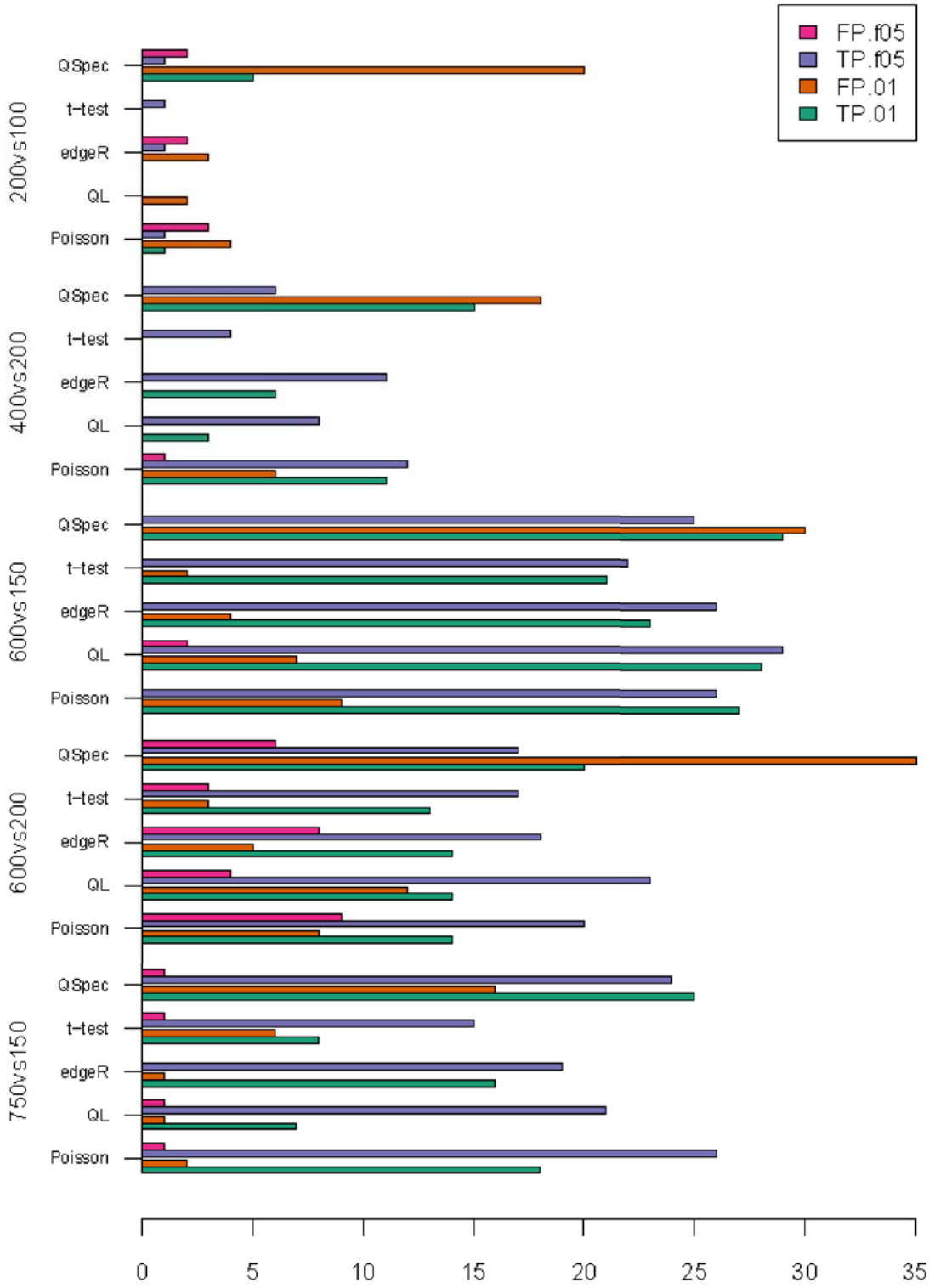



Figure 4.6: Results with and without post-test filter. Bar diagram with the data in table 4.3. **TP.01**: TP $p.val \leq 0.01$, **FP.01**: FP $p.val \leq 0.01$, **TP.f05** TP $p.val \leq 0.05$ and post-test filter, **FP.f05**: FP $p.val \leq 0.05$ and post-test filter.

sonably well in all circumstances. Remarkably the post-test filter minimizes the differences among tests, some of them highly sensitive to problems caused by low signal and/or extremely low variances [Leitch et al., 2012; Lundgren et al., 2010; Choi et al., 2008].

2. Besides the experimental setting, an extensive in-silico simulation was carried out (see Figure 4.7). The Poisson distribution parameter, $\lambda$, of the set of proteins in six replicates of samples of 500 ng of yeast lysate with 600 fm of UPS1 was estimated. For increasing fractions of DEPs, ranging from 1% to 25%, 1000 datasets of 4+4 samples were generated with indices of DEPs randomly generated. The FC for the DEPs were generated from a uniform distribution between 2 and 5, and its signs were obtained from a Bernouilli with probability 0.5. For each fraction of DEPs, the 1000 datasets provided a set of confusion tables from which the empirical distributions of power and FDR were obtained. For both, the test at an adjusted p-value of 0.01, and for the post-test filter with an adjusted p-value relaxed to 0.05. The proteins filtered out were the significant with less than 2 average SpC in the most abundant condition or with absolute LogFC below 0.8. The tests used were the GLM Poisson, the quasi-likelihood extension of the GLM, and the negative-binomial provided by the R package `edgeR`.

   At a fraction of 1 or 2% DEP both distributions span over the full range of 0 to 1, and this is consistent with the difficulty of discovering DEP at very low SpC. As the fraction of DEP increases the importance of those expressed at a low level is limited, because of the uniform distribution of DEP over the estimated lambdas. It was observed how after the post-test filter the sensitivities increased, and that this effect was stronger as the fraction of DEP increased too. It was also observed how the FDR values were pushed up too, although seemingly at a lower extent. In summary, in this simulation, filtering and using a higher p-value threshold brought to an increase in the number of both TPs and FP, but with a better TP/FP ratio.

   To prove the influence of the poor reproducibility of DEP expressed at very low level in the number of FP still observed, the simulations were repeated restricting the DEP to those showing a $\lambda$ above 1 SpC. The variability previously observed on FDR and power was now very much restricted. Indeed, the distribution of the FDR values was now compressed at values near 0, with no visible differences between the full set and the filtered one. The

distribution of sensitivity values showed the same trend as before, although slightly magnified, with higher power for the filtered dataset (see Figure 4.7).

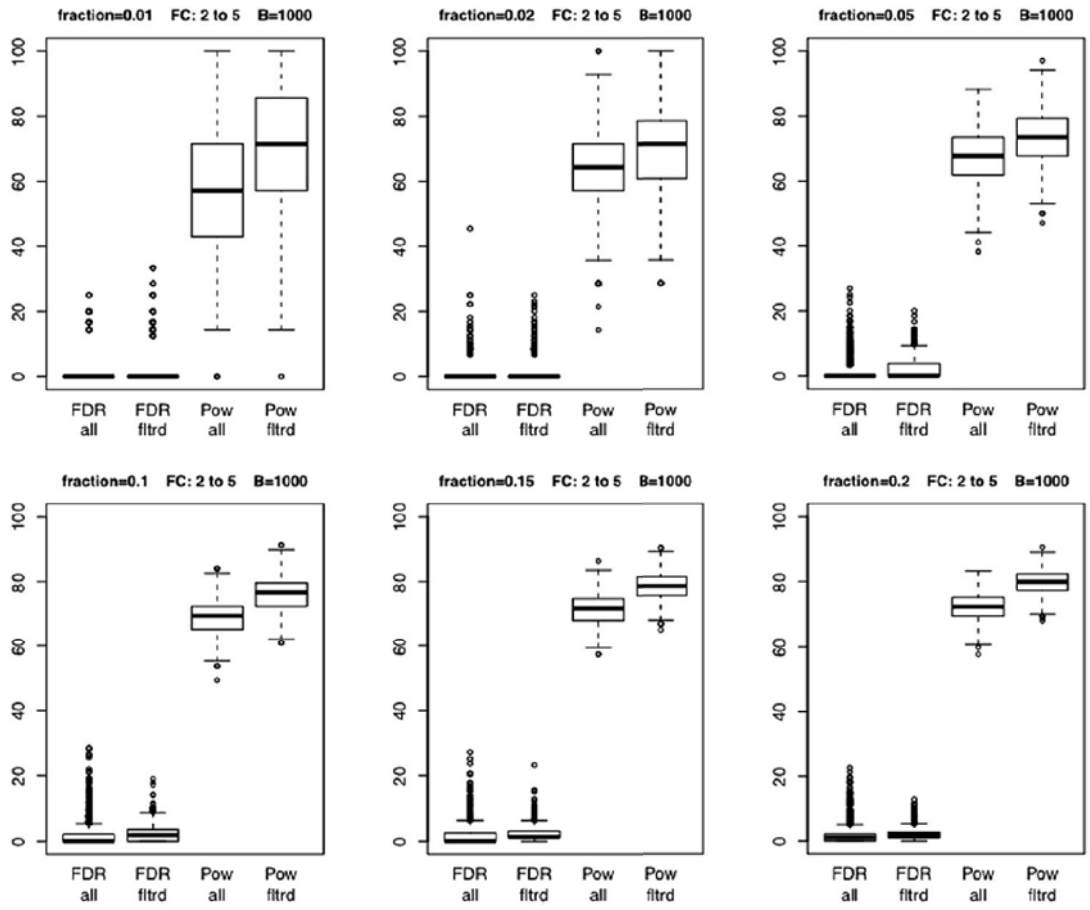



Figure 4.7: Empirical distribution of power and FDR by simulation, when the simulated DEPs are restricted to have a $\lambda \geq 1$. Inference with `edgeR`.

This difference could only be caused by the proteins with $\lambda < 1$, confirming that a low signal leads easily to lack of reproducibility.

### 4.2.4 Conclusions

1. The results obtained in this work prove that a post-test filter with reasonable effect size and signal level thresholds helps to increase the reproducibility of comparative proteomic analysis.

2. The signal and effect size post-test filter improves the results of the most common methods for SpC data analysis, and reduces the differences among them.

3. The extent of this improvement will depend on the exact problem, the statistical model used, and the number of replicates available.

4. Based on this work we recommend using a filter consisting of a minimum absolute $log_2$ fold change between 0.6 and 1, and a minimum mean signal between 2 and 4 SpC in the most abundant condition. The higher the number of replicates the lower the filter thresholds may be, still with good results. With 4 replicates, the thresholds of 0.8 and 2 have given good results.

5. Long lists of DEPs could be favourably shortened by increasing the thresholds in the post-test filter, instead of increasing the significance level.

## 4.3 Paper 3: A model for cell to cell comparisons

*Enhancing the Biological Relevance of Secretome-based Proteomics by Linking Tumor Cell Proliferation and Protein Secretion.*
Gregori J., Méndez O., Katsila T., Pujals M., Salvans C., Villarreal L., Arribas J., Tabernero J., Sánchez A., Villanueva J.
Submitted to J. of Proteome Research, pending of publication.

### 4.3.1 Aim

Normalization is the approach used to guarantee that comparisons are made in the proper scale. The aim of this work was to establish a normalization scheme for cell-line secretomes where we wish to compare the amounts of protein secreted by a single cell in two biological states.

### 4.3.2 Background

As seen in section 2.8.1, when equivalent quantities of total substance from two biological states are compared, equation 2.12 provides an unbiased comparison.

### 4.3.3 Development of a cell-centric normalization

With cell line secretomes we may count the number of cells involved, and we may measure the total quantity of protein secreted by these cells. This allows to formulate a more general model which is valid even when there are serious deviations from the basic assumption (see section 2.8.1).

We intend to compare the proteins secreted by one single cell in each of two given biological states of a cell line, even when one state is globally stimulated or depressed with respect to the other. Suppose we gather $Q_j\,\mu$g of total protein secreted by $n_j$ cells in the $j$-th biological condition, of which $q\,\mu$g are digested and injected into the LC-MS/MS system to be measured.

The total quantity of digested protein measured is the same for the two conditions. The ratio $q/Q_j$ gives the proportion of total secreted protein that gets measured for condition $j$, hence $(q/Q_j)n_j$ is the number of cells which secreted the $q\,\mu$g in the $j$-th biological condition. Then we obtain the number of cells which produced the $q\,\mu$g in the $j$-th condition as in equation 4.1.

$$c_j = \frac{q}{Q_j} n_j \tag{4.1}$$

On the other hand given the expected SpC value $\mu$ of a protein at a given concentration in the total $q\,\mu$g, the expected value per cell is given by equation 4.2

$$E\left[\frac{y}{c_j}\right] = \frac{\mu}{c_j} \tag{4.2}$$

As the total protein measured for the two biological conditions is the same, the factor $q$ contributes equally to both conditions and may be removed, bringing to equation 4.3 where the mass scale is undefined but equal for both conditions.

$$E\left[\frac{y}{c_j}\right] = \frac{\mu}{n_j/Q_j} \tag{4.3}$$

This allows to formulate a GLM model, as in equation 2.12, taking into account the total spectral counts by sample, $size$, the protein production rate of each condition, $Q/n$, the treatment factor, $X$, and a blocking factor to account for non controlled factors leading to batch effects, $Z$, as in equation 4.4.

$$log(\mu) = log\left(\frac{n}{Q}\right) + log(size) + \alpha + \beta x + \gamma z \tag{4.4}$$

With this model the logFC may be estimated from equation 4.5

$$FC_z = \frac{\mu_A/(size_A\, n_A/Q_A)}{\mu_B/(size_B\, n_B/Q_B)} = \frac{exp(\alpha+\beta+\gamma z)}{exp(\alpha+\gamma z)} = exp(\beta)$$

$$\widehat{logFC} = log_2(exp(\hat{\beta})) = 1.44\hat{\beta} \tag{4.5}$$

where the subindex A stands for treatment, with $x = 1$, and subindex B for control, with $x = 0$.

The p-value for differential secretion is obtained from the log likelihood ratio test comparing the model given by 4.4 with the model with $\beta = 0$.

### 4.3.4 Findings

The hypothesis of different secretion rates in different conditions was verified by observing large differences in the global protein secretion among cells experiencing different cellular perturbations, and even among cell lines at the basal state (see Figure 4.8).

Then the general model formulated above was evaluated in two biological situations. First, the global protein secretion in a colorectal cancer cell line (SW48) that has a large dependence on the epitelial growth factor receptor (EGFR) pathway

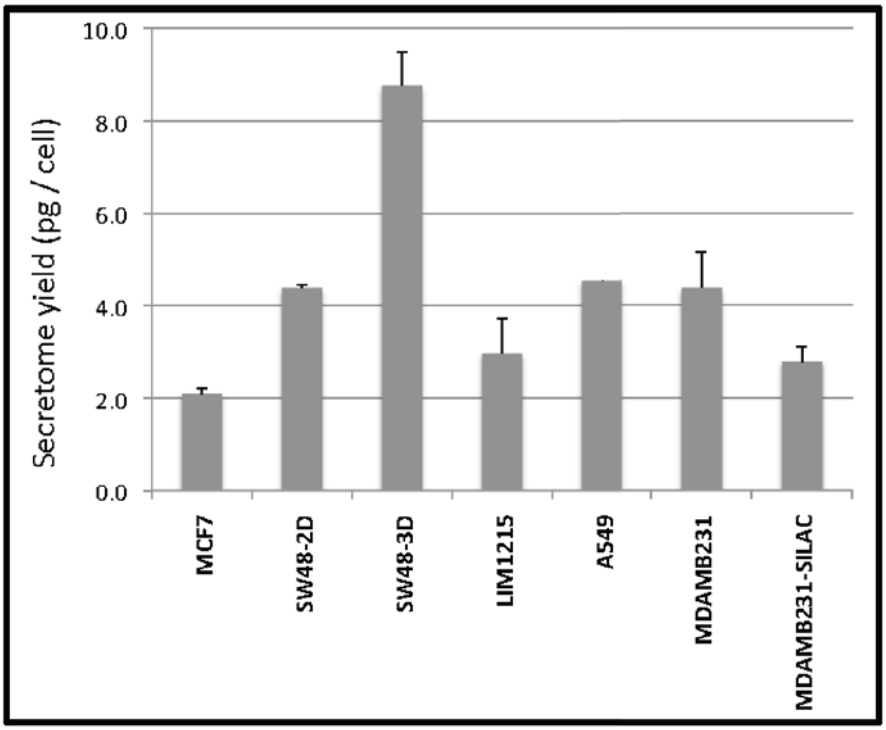



Figure 4.8: Secretion rates of different cancer cell lines in the basal state.

for its proliferation was evaluated. On one hand, the stimulation of the pathway with exogenous EGF induces a high proliferative state in SW48 cells duplicating the number of cells in 48h. On the other hand, the treatment with cetuximab, an anti-EGFR, greatly blocks the proliferation of the cells. The results showed that the high proliferative state induced by EGF yields a lower global protein secretion as compared with the non-treated cells. Conversely, the blockage of proliferation with cetuximab increases the global protein secretion by 3-fold compared to the cells treated with EGF (Figure 4.9).

The second example is the epithelial to mesenchymal transition (EMT) in A549 lung cancer cells. The treatment of A549 cells with transforming growth factor beta (TGFb) change their standard epithelial morphology to a spindle shape typical of mesenchymal cells. In this case, global secretion is also affected. The A549 cells treated with TGFb secrete globally twice as much protein as the control cells. The reason this model system was chosen is because EMT induces a massive change in the cell's secretome and its biology is reasonably well characterized. First, the fold changes of EGF over Cetuximab were estimated for SW48 cells, and TGFb over control for A549 cells. The estimation was done both under the model

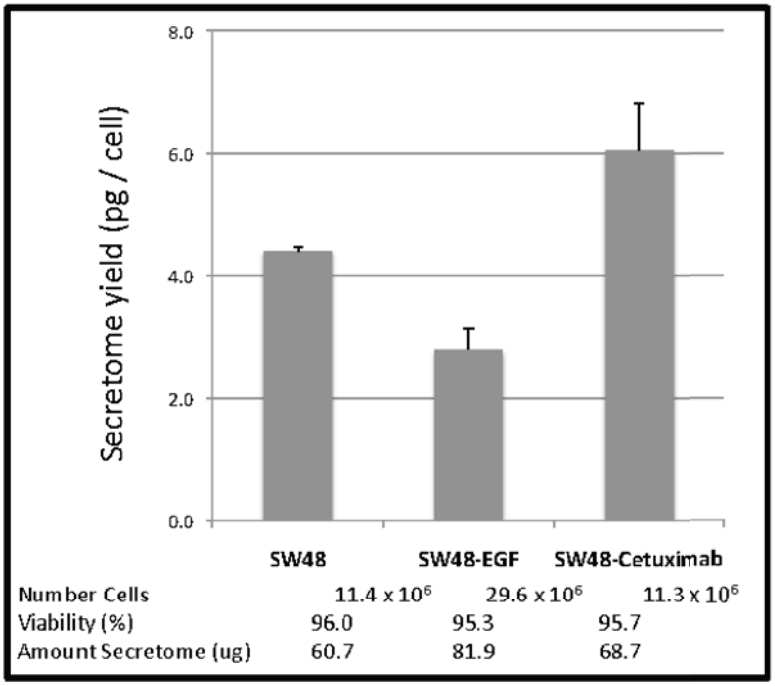



Figure 4.9: Secretion rates of a cell line under different treatments.

normalizing just by total SpC, and under the proposed model. The resulting DEPs were compared. The DEPs significant only under the proposed model have been previously linked to EMT, by either up- or down-regulating vimentin, E-caderin, N-caderin or fibronectin, or related to the EMT for other causes in different studies.

Another observation was that cellular perturbations affecting cellular proliferation had a large effect on the amount of secretome produced by cells. This leads to the hypothesis that global protein secretion and cellular proliferation could be linked. An explanation could be that during mitosis, when the nuclear envelope is dissolved and chromosomes are condensed, transcription is inhibited by the hyper-phosphorylation of the transcription apparatus. Also during mitosis protein transport between the ER and the Golgi apparatus is arrested, since the Golgi stacks are disassembled.

To relate proliferation with secretion rate two of the most widely used breast cancer cell lines, MCF-7 and MDA-MB-231 were taken. The MTT proliferation assay confirmed that MCF-7 cells grow almost twice as fast as MDA-MB-231. Calculated doubling times of 25 h for MCF-7 and 42 h for MDA-MB-231 cells contrast with a global protein secretion of 2.1 pg/cell and 4.3 pg/cell respectively. This result shows the inverse proportionality between proliferation rate and global

protein secretion.

Next, three colorectal cancer cell lines (SW48, LIM1215 and DiFi) were taken and their cell proliferation was manipulated by stimulation with EGF and by blocking with cetuximab (Figure 4.10). The global secretion rate in the cetuximab condition approximately doubles that in the EGF condition. These experiments further confirm the inverse proportionality between proliferation rate and global protein secretion, and suggest that proliferation rate play a role in the global secretion rate of a cell.

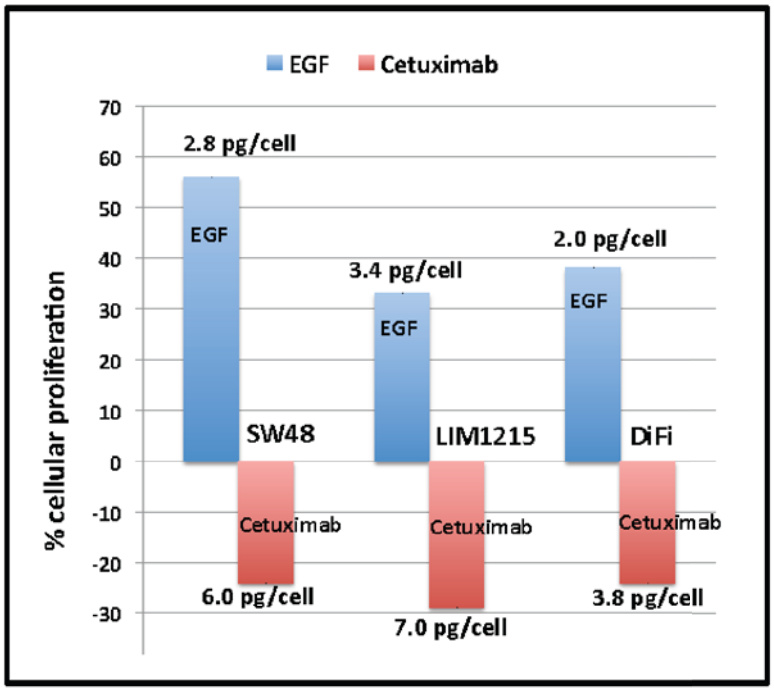



Figure 4.10: Secretion rates vs proliferation.

### 4.3.5 Conclusions

1. Cancer cell lines show different global secretion rates at their basal state, or under different perturbations.

2. Under GLM, equation 4.4 provides an unbiased cell-to-cell comparison in cell-line secretomes.

3. Our observations show that secretion rate and proliferation rate are inversely correlated for a given cell-line.

## 4.4 Software

### 4.4.1 R/Bioconductor packages

All software was developed in the R statistical environment and language [R Core Team, 2012]. Two R packages were produced with the code developed throughout the works which brought to the publication of the above papers. They were further refined to fulfill Bioconductor [Gentleman et al., 2004] requirements, and specifically adapted to work with instances of the `MSnSet` S4 class defined in the `MSnbase` package [Gatto & Lilley, 2012]. The packages are:

- `msmsEDA` for the exploratory data analysis (EDA) of SpC matrices.
- `msmsTests` for inference on SpC matrices, based on GLMs.

The functions provided for the EDA allow for the identification of outliers and potential batch effects or confounding factors. Any BD study should systematically start by an in-depth EDA to validate the model and samples used subsequently in the study. The main `msmsEDA` functions are described in table 4.4.

The BD is conducted by the application of the same model and test to all the proteins in the expression matrix. This means a replicated test for each row in the matrix. The GLM models considered include the Poisson distribution, the extension of the quasi-likelihood, and the negative binomial distribution. For the Poisson and the quasi-likelihood, the test is on the likelihood ratio between the alternative and the null model. For the negative-binomial the empirical Bayes approach provided by the `edgeR` package is used [Robinson et al., 2010]. This package has proven good results with few replicates in RNA-seq and demostrated a good behaviour for SpC in our studies [Gregori et al., 2013]. The main `msmsTests` functions are described in table 4.5.

Both packages include utility functions to help in the interpretation of the results.

#### Availability

The packages have been integrated in the Bioconductor project [Gentleman et al., 2004] and make use of the S4 class `MSnSet` in the `MSnbase` package [Gatto & Lilley, 2012].


Home » Bioconductor 2.14 » Software Packages » msmsEDA

# msmsEDA

## Exploratory Data Analysis of LC-MS/MS data by spectral counts

Bioconductor version: Development (2.14)

Exploratory data analysis to assess the quality of a set of LC-MS/MS experiments, and visualize de influence of the involved factors.

Author: Josep Gregori, Alex Sanchez, and Josep Villanueva

Maintainer: Josep Gregori <josep.gregori at gmail.com>

Home » Bioconductor 2.13 » Software Packages » msmsTests

# msmsTests

## LC-MS/MS Differential Expression Tests

Bioconductor version: Release (2.13)

Statistical tests for label-free LC-MS/MS data by spectral counts, to discover differentially expressed proteins between two biological conditions. Three tests are available: Poisson GLM regression, quasi-likelihood GLM regression, and the negative binomial of the edgeR package.The three models admit blocking factors to control for nuissance variables.To assure a good level of reproducibility a post-test filter is available, where we may set the minimum effect size considered biologicaly relevant, and the minimum expression of the most abundant condition.

Author: Josep Gregori, Alex Sanchez, and Josep Villanueva

Maintainer: Josep Gregori i Font <josep.gregori at gmail.com>

Figure 4.11: The two packages are available on-line at Bioconductor.

| | |
|---|---|
| `pp.msms.data` | data preprocessing to replace NAs by 0 and remove all zero rows. |
| `gene.table` | extract gene symbols from protein description. |
| `count.stats` | summaries of Spc and number of proteins by sample. |
| `counts.pca` | principal components analysis of the SpC matrix. |
| `counts.hc` | hierarchical clustering of samples. |
| `norm.counts` | normalization of spectral counts matrix. |
| `counts.heatmap` | experiment heatmap. |
| `disp.estimates` | dispersion analysis by factor and plots. |
| `spc.barplots` | barplots of the relative normalization divisors. |
| `spc.boxplots` | boxplots showing the distribution of SpC by sample. |
| `spc.densityplots` | density plots showing the distribution of SpC by sample. |
| `filter.flags` | flag features by signal and variability thresholds. |
| `bacth.neutralize` | correct the batch effects in the expression matrix. |

Table 4.4: Functions in the `msmsEDA` package.

| | |
|---|---|
| `msms.glm.pois` | Poisson based GLM regression |
| `msms.glm.qlll` | Quasi-likelihood GLM regression |
| `msms.edgeR` | The binomial negative of `edgeR` |
| `pval.by.fc` | Table of cumulative frequencies of features by p-values in bins of log fold change |
| `test.results` | Multitest p-value adjustement and post-test filter to flag DEPs as likely reproducible. |
| `res.volcanoplot` | Volcanplot of the results. |

Table 4.5: Functions in the `msmsTests` package.

Both packages are available from Bioconductor:
`http://www.bioconductor.org/packages/2.13/bioc/html/msmsEDA.html`
`http://www.bioconductor.org/packages/2.13/bioc/html/msmsTests.html`

**Documentation**

- `msmsEDA` manual
  `http://www.bioconductor.org/packages/2.13/bioc/manuals/msmsEDA/man/msmsEDA.pdf`
- `msmsEDA` vignette: *LC-MS/MS Exploratory Data Analysis*
  `http://www.bioconductor.org/packages/release/bioc/vignettes/msmsEDA/inst/doc/msmsData-Vignette.pdf`
- `msmsTests` manual
  `http://www.bioconductor.org/packages/2.13/bioc/manuals/msmsTests/man/msmsTests.pdf`
- `msmsTests` vignette: *LC-MS/MS post test filters to improve reproducibility*
  `http://www.bioconductor.org/packages/2.13/bioc/vignettes/msmsTests/inst/doc/msmsTests-Vignette.pdf`
- `msmsTests` vignette: *Bocks design to compensate batch effects*
  `http://www.bioconductor.org/packages/2.13/bioc/vignettes/msmsTests/inst/doc/msmsTests-Vignette2.pdf`

### 4.4.2 Graphical User Interfaces

Two graphical user interfaces have been developed to ease the routine lab computations, and to approach the solutions provided by the `msmsEDA` and `msmsTests` packages to the researchers in the proteomics field with no programming skills. The GUIs have been developed based on the functions in the two R/Bioconductor packages, and with the help of the infrastructure provided by the R packages `gWidgets` and `RGtk2` [Verzani, 2012; Lawrence & Verzani, 2012; Lawrence & Temple Lang, 2010].

**msmsEDA_GUI**

Allows an exploratory data analysis of a LC-MS/MS experiment given two files. The metadata in the samples description file (targets), with samples identifiers,

labels and eventual normalizing factors, and a file with the SpC expression matrix, with proteins description and accessions (figure 4.12). A data quality assessment is performed by boxplots and density plots of the observed SpC in each sample, and by barplots of the relative normalization factor by sample. In this GUI, the expression matrix is explored by PCA, HC and heatmaps. Both, with the raw SpC matrix, with the sample size normalized SpC matrix, and after batch effects correction if a batch factor is provided in the metadata. The distribution of informative features according to the main factor are also explored at each data treatment step. Finally the distribution of residual dispersions is explored and plotted. A number of files are generated with the results along with an index html file with their names, description and link (figure 4.13).

#### msmsTests_GUI

Given the metadata and the SpC files as above, this GUI provides great flexibility in BD with controls in the GUI which allow the choice of the normalization method, the statistical test, the multiple test adjustment method, the significance level, and the post-test filter thresholds (figure 4.14). The output control in the GUI shows the development of the computations and the main results. A number of text files and pdf plots with the results are generated.

### Availability and Documentation

The GUIs, and corresponding documentation, are available online at GitHub.com

| | |
|---|---|
| msmsEDA_GUI | https://github.com/JosepGregori/msmsEDA_GUI_repos |
| msmsTests_GUI | https://github.com/JosepGregori/msmsTests_GUI_repos |

A tutorial and the user guides of the two GUIs are attached to the thesis.

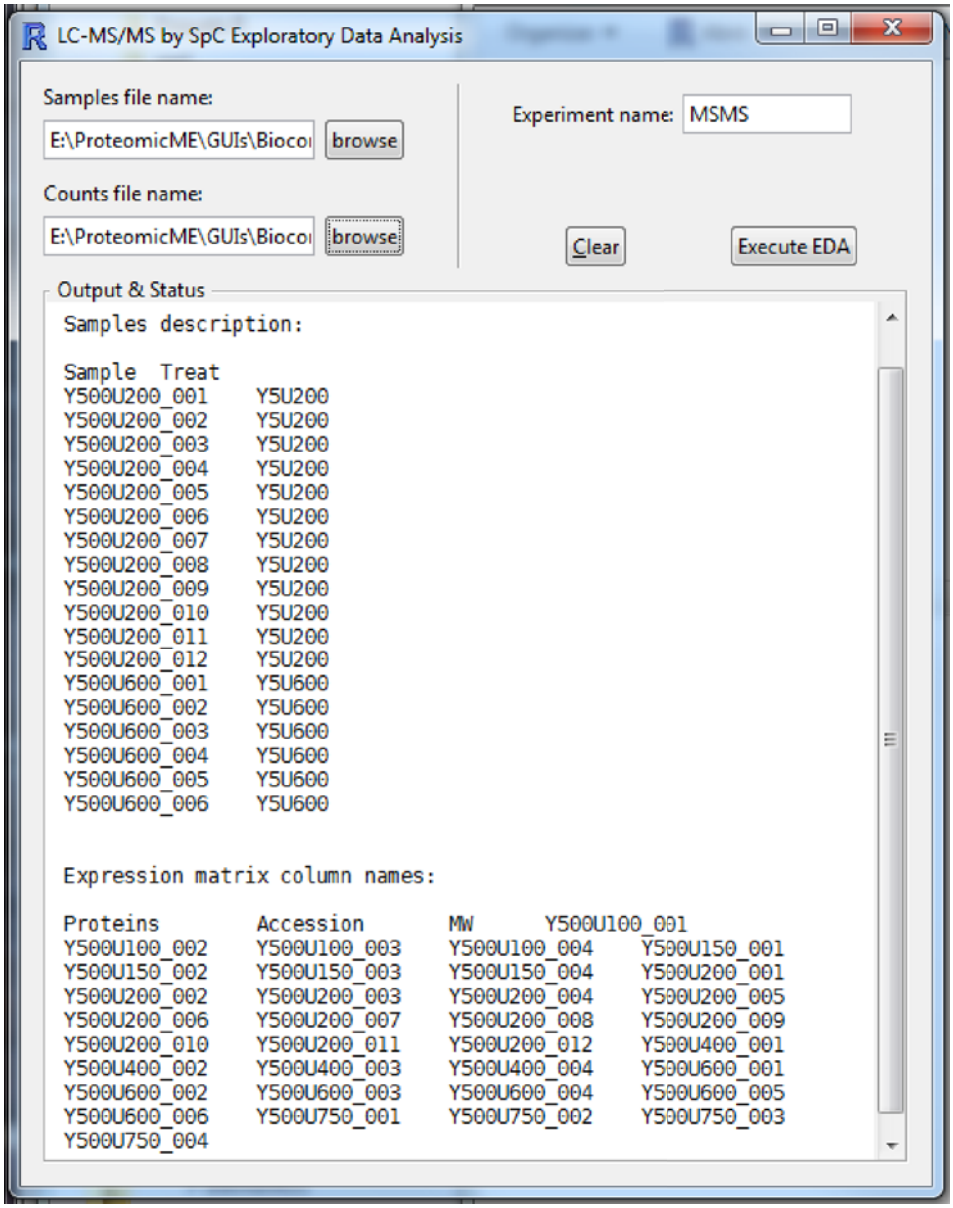



Figure 4.12: The msmsEDA_GUI window.

# msmsEDA Graphical User Interface

**EXPLORATORY DATA ANALYSIS RESULTS - Output files index**

| FILE / LINK | Contents |
|---|---|
| MDA231_EDA.txt | **Text file with statistic summaries at each step** |
| MDA231-DispPlots.pdf | PDF file with plots<br>Final SpC matrix<br>**Residual dispersion plots** |
| MDA231-BatchCorrected-PCA.pdf | PDF file with plots<br>Normalized + batch corrected SpC matrix<br>**Principal components plot** |
| MDA231-BatchCorrected-HCD.pdf | PDF file with plots<br>Normalized + batch corrected SpC matrix<br>**Hierarchical clustering plot** |
| MDA231-BatchCorrected-HeatMap.pdf | PDF file with plots<br>Normalized + batch corrected SpC matrix<br>**Heatmap** |

Figure 4.13: Snapshot of the HTML index file with links to the files generated by the EDA.

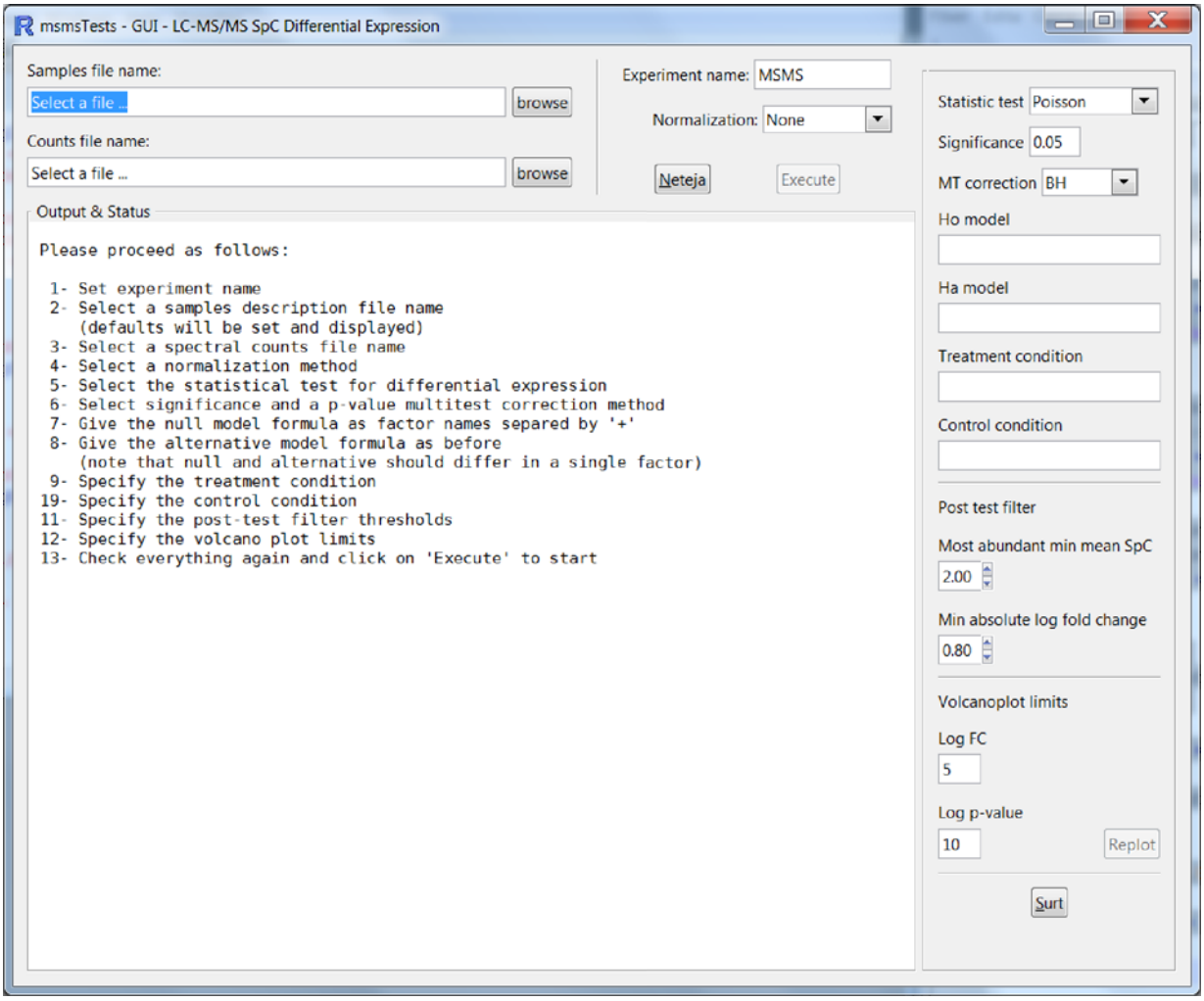



Figure 4.14: The msmsTests_GUI window.

## 4.5 Applications

### 4.5.1 Secretome composition

Research



# Unconventional Secretion is a Major Contributor of Cancer Cell Line Secretomes*

Laura Villarreal‡§**, Olga Méndez‡§**, Cándida Salvans‡§, Josep Gregori‡¶, José Baselga‡, and Josep Villanueva‡§‖

A challenge in achieving optimal management of cancer is the discovery of secreted biomarkers that represent useful surrogates for the disease and could be measured noninvasively. Because of the problems encountered in the proteomic interrogation of plasma, secretomes have been proposed as an alternative source of tumor markers that might be enriched with secreted proteins relevant to the disease. However, secretome analysis faces analytical challenges that interfere with the search for true secreted tumor biomarkers. Here, we have addressed two of the main challenges of secretome analysis in comparative discovery proteomics. First, we carried out a kinetics experiment whereby secretomes and lysates of tumor cells were analyzed to monitor cellular viability during secretome production. Interestingly, the proteomic signal of a group of secreted proteins correlated well with the apoptosis induced by serum starvation and could be used as an internal cell viability marker. We then addressed a second challenge relating to contamination of serum proteins in secretomes caused by the required use of serum for tumor cell culture. The comparative proteomic analysis between cell lines labeled with SILAC showed a number of false positives coming from serum and that several proteins are both in serum and being secreted from tumor cells. A thorough study of secretome methodology revealed that under optimized experimental conditions there is a substantial fraction of proteins secreted through unconventional secretion in secretomes. Finally, we showed that some of the nuclear proteins detected in secretomes change their cellular localization in breast tumors, explaining their presence in secretomes and suggesting that tumor cells use unconventional secretion during tumorigenesis. The unconventional secretion of proteins into the extracellular space exposes a new layer of genome post-translational regulation and reveals an untapped source of potential tumor biomarkers and drug targets. *Molecular & Cellular Proteomics 12: 10.1074/mcp.M112.021618, 1046–1060, 2013.*

Mol Cell Proteomics. 2013 May;12(5):1046-60. doi: 10.1074/mcp.M112.021618. Epub 2012 Dec 26.

**Summary**

A challenge in achieving optimal management of cancer is the discovery of secreted biomarkers that represent useful surrogates for the disease and could be measured noninvasively. Because of the problems encountered in the proteomic interrogation of plasma, secretomes have been proposed as an alternative source of tumor markers that might be enriched with secreted proteins relevant to the disease. However, secretome analysis faces analytical challenges that interfere with the search for true secreted tumor biomarkers. Here, we have addressed two of the main challenges of secretome analysis in comparative discovery proteomics.

The study included the following issues:

- Cell viability - degree of apoptosis
- Contamination by serum proteins
- Types of secretion

**Statistic methods**

All statistical computations were performed using the open-source statistical package R. The data from an MS/MS experiment was assembled in a matrix of spectral counts where the different conditions are represented by the columns, and the identified proteins are represented in the rows of that matrix. The need for normalization was assessed by comparing the total spectral counts (SpC) in technical replicates of each sample. As the quantity of substance for each sample, in each experiment, was the same, any substantial deviation was corrected by normalizing to the median total sample counts. An exploratory data analysis by means of principal components analysis (PCA) and hierarchical clustering (HC) of the samples on the SpC matrix was performed to find potential outliers and patterns in the data. Dealing with counts precludes the use of statistical tests and procedures based on the normal distribution, and restricts the appropriate methods to those in the general frame of the Generalized Linear Models (GLM) with discrete distributions. As no substantial biological variability is expected from cell line data, according to our experience, a Poisson regression was used for significance testing throughout this work. The GLM model based on the Poisson distribution was used as a significance test throughout our work. Finally we used the Benjamini and Hochberg multiple test adjustment [Benjamini & Hochberg, 1995].

**Results**

First, we carried out a kinetics experiment whereby secretomes and lysates of tumor cells were analyzed to monitor cellular viability during secretome production. Interestingly, the proteomic signal of a group of secreted proteins correlated well with the apoptosis induced by serum starvation and could be used as an internal cell viability marker. We then addressed a second challenge relating to contamination of serum proteins in secretomes caused by the required use of serum for tumor cell culture. The comparative proteomic analysis between cell lines labeled with SILAC showed: i) a number of false positives coming from serum, and ii) that several proteins are both in serum and secreted from tumor cells.

A thorough study of secretome methodology revealed that under optimized experimental conditions there is a substantial fraction of proteins secreted through unconventional secretion in secretomes. Finally, we showed that some of the nuclear proteins detected in secretomes change their cellular localization in breast tumors, explaining their presence in secretomes and suggesting that tumor cells use unconventional secretion during tumorigenesis. The unconventional secretion of proteins into the extracellular space exposes a new layer of genome post-translational regulation and reveals an untapped source of potential tumor biomarkers and drug targets.

## 4.6 Other publications

During the development of this thesis I was also working on NGS data, developing tools for the analysis of sequences of amplicons of viral quasispecies. In what follows there is a summary of the publications in international peer-reviewed journals, with title and abstracts.

1. Diversity in quasispecies by CCSS and NGS

**Bioinformatics Advance Access published January 21, 2014**

BIOINFORMATICS **ORIGINAL PAPER** *2014, pages 1–8*
*doi:10.1093/bioinformatics/btt768*

*Sequence analysis* Advance Access publication January 2, 2014

### Inference with viral quasispecies diversity indices: clonal and NGS approaches

Josep Gregori[1,2,3,*], Miquel Salicrú[3], Esteban Domingo[4,5], Alex Sanchez[3,6], Juan I. Esteban[1,4,7], Francisco Rodríguez-Frías[4,7,8] and Josep Quer[1,4,7]

[1]Liver Unit, Internal Medicine Lab Malalties Hepàtiques, Vall d'Hebron Institut Recerca (VHIR-HUVH), 08035 Barcelona, Spain, [2]Roche Diagnostics SL, 08174, Sant Cugat del Vallès, Spain, [3]Statistics Department, Biology Faculty, Barcelona University, 08028, Barcelona, Spain, [4]CIBER de Enfermedades Hepáticas y Digestivas (CIBERehd) del Instituto de Salud Carlos III, 28029 Madrid, Spain, [5]Centro de Biología Molecular Severo Ochoa (CSIC-UAM), Campus de Cantoblanco, 28049, Madrid, Spain, [6]Bioinformatics and Statistics Unit, Vall d'Hebron Institut Recerca (VHIR-HUVH), 08035, Barcelona, Spain, [7]Universitat Autònoma de Barcelona, 08193 Bellaterra, Barcelona, Spain and [8]Biochemistry Unit. Virology Unit/Microbiology Department, HUVH, 08035 Barcelona, Spain

Associate Editor: Michael Brudno

**ABSTRACT**

Given the inherent dynamics of a viral quasispecies, we are often interested in the comparison of diversity indices of sequential samples of a patient, or in the comparison of diversity indices of virus in groups of patients in a treated versus control design. It is then important to make sure that the diversity measures from each sample may be compared with no bias and within a consistent statistical framework. In the present report, we review some indices often used as measures for viral quasispecies complexity and provide means for statistical inference, applying procedures taken from the ecology field. In particular, we examine the Shannon entropy and the mutation frequency, and we discuss the appropriateness of different normalization methods of the Shannon entropy found in the literature. By taking amplicons ultra-deep pyrosequencing (UDPS) raw data as a surrogate of a real hepatitis C virus viral population, we study through in-silico sampling the statistical properties of these indices under two methods of viral quasispecies sampling, classical cloning followed by Sanger sequencing (CCSS) and next-generation sequencing (NGS) such as UDPS. We propose solutions specific to each of the two sampling methods—CCSS and NGS—to guarantee statistically conforming conclusions as free of bias as possible.

**Contact:** josep.gregori@gmail.com

**Supplementary information:** Supplementary data are available at *Bioinformatics* online

2. Minority variants in HCV quasispecies

OPEN ACCESS Freely available online

PLOS | ONE

# Ultra-Deep Pyrosequencing (UDPS) Data Treatment to Study Amplicon HCV Minor Variants

**Josep Gregori[1,2]*, Juan I. Esteban[1,3,4], María Cubero[1,3], Damir García-Cehic[1,3], Celia Perales[3,5], Rosario Casillas[1,6], Miguel Alvarez-Tejado[2], Francisco Rodríguez-Frías[3,4,6], Jaume Guardia[1,3,4], Esteban Domingo[3,5], Josep Quer[1,3,4]**

**1** Liver Unit, Internal Medicine, Lab. Malalties Hepàtiques, Vall d'Hebron Institut Recerca Hospital Universitari Vall d'Hebron (VHIR HUVH), Barcelona, Spain, **2** Roche Diagnostics SL, Sant Cugat del Vallès, Spain, **3** CIBER de Enfermedades Hepáticas y Digestivas (CIBERehd) del Instituto de Salud Carlos III, Madrid, Spain, **4** Universitat Autònoma de Barcelona, Bellaterra, Spain, **5** Centro de Biología Molecular Severo Ochoa (CBM), UAM, Madrid, Spain, **6** Biochemistry Unit, HUVH, Barcelona, Spain

## Abstract

We have investigated the reliability and reproducibility of HCV viral quasispecies quantification by ultra-deep pyrosequencing (UDPS) methods. Our study has been divided in two parts. First of all, by UDPS sequencing of clone mixes samples we have established the global noise level of UDPS and fine tuned a data treatment workflow previously optimized for HBV sequence analysis. Secondly, we have studied the reproducibility of the methodology by comparing 5 amplicons from two patient samples on three massive sequencing platforms (FLX+, FLX and Junior) after applying the error filters developed from the clonal/control study. After noise filtering the UDPS results, the three replicates showed the same 12 polymorphic sites above 0.7%, with a mean CV of 4.86%. Two polymorphic sites below 0.6% were identified by two replicates and one replicate respectively. A total of 25, 23 and 26 haplotypes were detected by GS-Junior, GS-FLX and GS-FLX+. The observed CVs for the normalized Shannon entropy (Sn), the mutation frequency (Mf), and the nucleotidic diversity (Pi) were 1.46%, 3.96% and 3.78%. The mean absolute difference in the two patients (5 amplicons each), in the GS-FLX and GS-FLX+, were 1.46%, 3.96% and 3.78% for Sn, Mf and Pi. No false polymorphic site was observed above 0.5%. Our results indicate that UDPS is an optimal alternative to molecular cloning for quantitative study of HCV viral quasispecies populations, both in complexity and composition. We propose an UDPS data treatment workflow for amplicons from the RNA viral quasispecies which, at a sequencing depth of at least 10,000 reads per strand, enables to obtain sequences and frequencies of consensus haplotypes above 0.5% abundance with no erroneous mutations, with high confidence, resistant mutants as minor variants at the level of 1%, with high confidence that variants are not missed, and highly confident measures of quasispecies complexity.

**Citation:** Gregori J, Esteban JI, Cubero M, García Cehic D, Perales C, et al. (2013) Ultra Deep Pyrosequencing (UDPS) Data Treatment to Study Amplicon HCV Minor Variants. PLoS ONE 8(12): e83361. doi:10.1371/journal.pone.0083361

**Editor:** Oliver Schildgen, Kliniken der Stadt Köln gGmbH, Germany

**Received** September 26, 2013; **Accepted** November 8, 2013; **Published** December 31, 2013



**Funding:** This study has been supported by CDTI (Centro para el Desarrollo Tecnológico Industrial), Spanish Ministry of Economics and Competitiveness (MINECO), IDI 20110115. It has also been supported by MINECO projects SAF 2009 10403, and also by the Spanish Ministry of Health Instituto de Salud Carlos III (FISS) projects PI10/01505. CIBERehd is funded by the Instituto de Salud Carlos III, Madrid. Work at CBMSO was supported by grant BFU2011 23604, FIPSE and Fundación Ramon Areces. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

**Competing Interests:** J. Gregori and M. Alvarez Tejado are employees of Roche Diagnostics Spain. M. Cubero was employer of Roche Diagnostics Spain. All other authors have no conflict of interest to declare. Roche Diagnostics Spain had no role in the study design, data collection, data analysis, data interpretation, or writing of the report, and does not alter the authors' adherence to all the PLOS ONE policies on sharing data and materials.

* E-mail: josep.gregori@vhir.org

3. Minority variants in HBV quasispecies. Cloning vs NGS

Antiviral Research 98 (2013) 273–283

Contents lists available at SciVerse ScienceDirect

Antiviral Research

journal homepage: www.elsevier.com/locate/antiviral

# A comparative study of ultra-deep pyrosequencing and cloning to quantitatively analyze the viral quasispecies using hepatitis B virus infection as a model

Clara Ramírez [a], Josep Gregori [b], Maria Buti [c,d], David Tabernero [a,d], Sílvia Camós [a,d], Rosario Casillas [a,b], Josep Quer [b,d], Rafael Esteban [c,d], Maria Homs [a,d], Francisco Rodriguez-Frías [a,d,*]

[a] Biochemistry Department, Hospital Vall d'Hebron, Universitat Autònoma de Barcelona, Spain
[b] Liver Unit, Research Institute Vall d'Hebron, Universitat Autònoma de Barcelona, Spain
[c] Hepatology Department, Hospital Vall d'Hebron, Universitat Autònoma de Barcelona, Spain
[d] Centro de Investigación Biomédica en Red de Enfermedades Hepáticas y Digestivas (CIBERehd), Instituto Carlos III, Spain

ARTICLE INFO

*Article history:*
Received 16 December 2012
Revised 5 March 2013
Accepted 11 March 2013
Available online 20 March 2013

*Keywords:*
Quasispecies
HBV
Ultra-deep pyrosequencing
Cloning
Polymorphic positions
Haplotypes

ABSTRACT

In this study, the reliability and reproducibility of viral quasispecies quantification by three ultra-deep pyrosequencing (UDPS) methods (FLX+, FLX, and Junior) were investigated and results compared with the conventional cloning technique. Hepatitis B virus (HBV) infection was selected as the model. The pre-Core/Core region, the least overlapped HBV region, was analyzed in samples from a chronic hepatitis B patient by cloning and by UDPS.

After computation filtering of the UDPS results, samples A1 and A2 (FLX+) and sample B (FLX) yielded the same 20 polymorphic positions. Junior yielded 18 polymorphic positions that coincided with the FLX results. In contrast, 50 polymorphic positions were detected by cloning. Quasispecies complexity plotted on graphs showed superimposed patterns and the quantitative parameters were similar between FLX+, FLX, Junior, and the cloning sequences. Twenty-two haplotypes were detected by Junior, and 37, 40, and 39 were detected by FLX A1, A2, and B, respectively. These differences may be attributable to methodological differences between FLX and Junior. By cloning, 47 haplotypes were detected. Eight clones with insertions and deletions that induced *de novo* stop codons were not observed by UDPS because the UDPS filter discarded them.

Our results indicate that UDPS is an optimal alternative to molecular cloning for quantitative study of the viral quasispecies. Nonetheless, specific mutations, such as insertions and deletions, were only detected by cloning. A filter should be designed to analyze cloning sequences, and UDPS filters should be improved to include the specific mutations.

4. HCV infection by sex transmission. A case study

Liver INTERNATIONAL
OFFICIAL JOURNAL OF THE INTERNATIONAL ASSOCIATION FOR THE STUDY OF THE LIVER

Liver International ISSN 1478-3223

VIRAL HEPATITIS

# Identification of host and viral factors involved in a dissimilar resolution of a hepatitis C virus infection

Maria Cubero[1,8], Josep Gregori[1,8], Juan I. Esteban[1,2,3], Damir García-Cehic[1,2], Marta Bes[2,4], Celia Perales[2,5], Esteban Domingo[2,5], Francisco Rodríguez-Frías[2,3,6], Silvia Sauleda[2,4], Rosario Casillas[1,6], Alex Sanchez[7], Israel Ortega[7], Rafael Esteban[1,2,3], Jaume Guardia[1,2,3] and Josep Quer[1,2,3]

1 Liver Unit. Internal Medicine. Lab. Malalties Hepàtiques, Vall d'Hebron Institut Recerca-Hospital Universitari Vall d'Hebron (VHIR-HUVH), Barcelona, Spain
2 CIBER de Enfermedades Hepáticas y Digestivas (CIBERehd) del Instituto de Salud Carlos III, Madrid, Spain
3 Universitat Autònoma de Barcelona, Bellaterra, Barcelona, Spain
4 Banc de Sang i de Teixits, Institut Català de la Salut, Barcelona, Spain
5 Centro de Biología Molecular Severo Ochoa (CSIC-UAM), Campus de Cantoblanco, Madrid, Spain
6 Biochemistry Unit, Virology Unit/Microbiology Department, HUVH, Barcelona, Spain
7 VHIR-HUVH, Unitat Estadística i Bioinformàtica (UEB), Barcelona, Spain
8 Roche Diagnostics SL, Sant Cugat del Vallès, Barcelona, Spain

**Keywords**
chronicity – HCV – NS3 – quasispecies – resolution – UDPS

**Correspondence**
Senior Researcher Josep Quer, PhD, Liver Unit, Internal Medicine, Vall d'Hebron Institute of Research (VHIR), Hospital Universitari Vall d'Hebron (HUVH), Pg Vall d'Hebron 119-129, 08035 Barcelona, Spain
Tel: +34 9 3489 4028/4034
Fax: +34 9 3489 4032
e-mail: josep.quer@vhir.org

Received 23 May 2013
Accepted 13 October 2013

DOI:10.1111/liv.12362

## Abstract

*Background & Aims:* Hepatitis C virus (HCV) transmission from a chronic patient to a susceptible individual is a good opportunity to study viral and host factors that may influence the natural course of hepatitis C infection towards either spontaneous recovery or chronicity. To compare a documented case of a bottleneck event in the sexual transmission of HCV from a chronically infected patient to a recipient host that cleared infection. *Methods:* Host genetic components such as Class I and II HLA and IL28B polymorphism (rs12979860 SNPs) were identified by direct sequencing and LightMix analysis, respectively. Deep nucleotide sequence analysis of quasispecies complexity was performed using massive pyrosequencing platform (454 GS-FLX), and the CD4 specific immune response was characterized by ELISPOT. *Results and Conclusions:* Sequencing analysis and CD4 response highlighted several NS3-helicase domains in which an interplay between amino acid variability and CD4 immune response might have contributed either to chronicity in the donor patient or to viral clearance in the receptor (newly infected) patient.

5. Ribavirin as a HCV mutagen

OPEN ACCESS Freely available online

PLOS | ONE

# Extinction of Hepatitis C Virus by Ribavirin in Hepatoma Cells Involves Lethal Mutagenesis

**Ana M. Ortega-Prieto[1], Julie Sheldon[1], Ana Grande-Pérez[2], Héctor Tejero[1,3], Josep Gregori[5,7], Josep Quer[4,5,6], Juan I. Esteban[4,5,6], Esteban Domingo[1,4]*, Celia Perales[1,4]***

**1** Centro de Biología Molecular "Severo Ochoa" (CSIC-UAM), Consejo Superior de Investigaciones Científicas (CSIC), Campus de Cantoblanco, Madrid, Spain, **2** Instituto de Hortofruticultura Subtropical y Mediterránea "La Mayora" (IHSM-UMA-CSIC), Departamento de Biología Celular, Genética y Fisiología, Universidad de Málaga, Campus Teatinos, Málaga, Spain, **3** Departamento de Bioquímica y Biología Molecular I, Universidad Complutense de Madrid, Madrid, Spain, **4** Centro de Investigación Biomédica en Red de Enfermedades Hepáticas y Digestivas (CIBERehd), Barcelona, Spain, **5** Liver Unit, Internal Medicine Lab. Malalties Hepàtiques, Vall d'Hebron Institut Recerca-Hospital (VHIR-HUVH), Barcelona, Spain, **6** Universitat Autónoma de Barcelona, Barcelona, Spain, **7** Roche Diagnostics, S.L., Sant Cugat del Vallès, Spain

## Abstract

Lethal mutagenesis, or virus extinction produced by enhanced mutation rates, is under investigation as an antiviral strategy that aims at counteracting the adaptive capacity of viral quasispecies, and avoiding selection of antiviral-escape mutants. To explore lethal mutagenesis of hepatitis C virus (HCV), it is important to establish whether ribavirin, the purine nucleoside analogue used in anti-HCV therapy, acts as a mutagenic agent during virus replication in cell culture. Here we report the effect of ribavirin during serial passages of HCV in human hepatoma Huh-7.5 cells, regarding viral progeny production and complexity of mutant spectra. Ribavirin produced an increase of mutant spectrum complexity and of the transition types associated with ribavirin mutagenesis, resulting in HCV extinction. Ribavirin-mediated depletion of intracellular GTP was not the major contributory factor to mutagenesis since mycophenolic acid evoked a similar decrease in GTP without an increase in mutant spectrum complexity. The intracellular concentration of the other nucleoside-triphosphates was elevated as a result of ribavirin treatment. Mycophenolic acid extinguished HCV without an intervening mutagenic activity. Ribavirin-mediated, but not mycophenolic acid-mediated, extinction of HCV occurred via a decrease of specific infectivity, a feature typical of lethal mutagenesis. We discuss some possibilities to explain disparate results on ribavirin mutagenesis of HCV.

**Citation:** Ortega-Prieto AM, Sheldon J, Grande-Pérez A, Tejero H, Gregori J, et al. (2013) Extinction of Hepatitis C Virus by Ribavirin in Hepatoma Cells Involves Lethal Mutagenesis. PLoS ONE 8(8): e71039. doi:10.1371/journal.pone.0071039

**Editor:** Jean-Pierre Vartanian, Institut Pasteur, France

**Received** May 14, 2013; **Accepted** June 26, 2013; **Published** August 16, 2013



**Funding:** This work was supported by grants BFU 2011-23604, SAF2009-10403, PI 10/01505 and ref. IDI-20110115 CDTI (Centro para el Desarrollo Tecnológico Industrial) from Ministerio de Ciencia e Innovación, P09-CVI-5428 and P10-CVI-6561 from Junta de Andalucía, and Fundación Ramon Areces. CIBERehd (Centro de Investigación Biomédica en Red de Enfermedades Hepáticas y Digestivas) is funded at the Instituto de Salud Carlos III. A.M.O. is supported by an FPI contract from MINECO, and J.S. by a Juan de la Cierva contract from CSIC. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

**Competing Interests:** The authors have declared that no competing interests exist.

* E-mail: edomingo@cbm.uam.es (ED); cperales@cbm.uam.es (CP)

World Journal of Gastroenterology

Online Submissions: http://www.wjgnet.com/esps/
wjg@wjgnet.com
doi:10.3748/wjg.v19.i35.5813

World J Gastroenterol 2013 September 21; 19(35): 5813-5827
ISSN 1007-9327 (print) ISSN 2219-2840 (online)


ORIGINAL ARTICLE

# Molecular epidemiology and putative origin of hepatitis C virus in random volunteers from Argentina

Noemí del Pino, José Raúl Oubiña, Francisco Rodríguez-Frías, Juan Ignacio Esteban, María Buti, Teresa Otero, Josep Gregori, Damir García-Cehic, Silvia Camos, María Cubero, Rosario Casillas, Jaume Guàrdia, Rafael Esteban, Josep Quer

## Abstract

**AIM:** To study the subtype prevalence and the phylogenetic relatedness of hepatitis C virus (HCV) sequences obtained from the Argentine general population, a large cohort of individuals was analyzed.

**METHODS:** Healthy Argentinian volunteers ($n$ = 6251) from 12 provinces representing all geographical regions of the country were studied. All parents or legal guardians of individuals younger than 18 years provided informed written consent for participation. The corresponding written permission from all municipal authorities was obtained from each city or town where subjects were to be included. HCV RNA reverse transcription-polymerase chain reaction products were sequenced and phylogenetically analyzed. The 5′ untranslated region (5′UTR) was used for RNA detection and initial genotype classification. The NS5B polymerase region, encompassing nt 8262-8610, was used for subtyping.

**RESULTS:** An unexpectedly low prevalence of HCV infection in the general population (0.32%) was observed. Our data contrasted with previous studies that reported rates ranging from 1.5% to 2.5%, mainly performed in selected populations of blood donors or vulnerable groups. The latter values are in keeping with the prevalence reported by the 2007 Argentinian HCV Consensus (approximately 2%). HCV subtypes were distributed as follows: 1a (25%), 1b (25%), 2c (25%), 3a (5%), and 2j (5%). Two isolates ascribed either to genotype 1 (5%) or to genotype 3 (5%) by 5′UTR phylogenetic analysis could not be subtyped. Subtype 1a sequences comprised a highly homogeneous population and clustered with United States sequences. Genotype 1b sequences represented a heterogeneous population, suggesting that this genotype might have been introduced from different sources. Most subtype 2c sequences clustered close to the 2c reported from Italy and Southern France.

**CONCLUSION:** HCV has a low prevalence of 0.32% in the studied general population of Argentina. The pattern of HCV introduction and transmission in Argentina appears to be a consequence of multiple events and different for each subtype.

# Chapter 5

# GENERAL DISCUSSION

This work focused on three relevant aspects of biomarker discovery (BD): batch effects, reproducibility, and the design of a model for cell-to-cell comparisons for cancer cell-line secretomes.

We studied the incidence of batch effects [Scherer, 2009] on label-free comparative proteomics based on SpC, and fixed the most narrow time window where LC-MS/MS runs are equally influenced by uncontrolled factors in one single day of data acquisition [Gregori et al., 2012]. The batch effects were evidenced by multidimensional techniques like PCA or HC. We then studied possible ways to correct these batch effects [Chen et al., 2011] in balanced comparative experiments, and found that the best correction is the scale correction implemented in a log-link GLM with a batch blocking factor [Quinn & Keough, 2002].

On the basis of the conceptual aspects of the MAQC-I recommendations [Shi et al., 2008] we studied the advantages in using post-test signal and effect size filters to improve the reproducibility of biomarkers discovered by label-free LC-MS/MS with SpC. We found that with these filters we may improve the number of true positives (TP), while restricting the number of false positives (FP), by relaxing the significance level. We found also that the use of this sort of filter gives the additional advantage of improving the overlap in the lists of proteins declared as differentially expressed using different tests [Gregori et al., 2013].

Starting from the most general hypothesis of a variable secretion rate of cancer cell-lines under different biological conditions we developed a GLM model for unbiased cell-to-cell secretome comparisons (see section 4.3). This model incorporates: i) a sample size normalization by the total SpC in the sample, ii) the cell-to-cell normalization using the observed secretion rate, iii) the relevant treatment factor for the comparison, and iv) blocking factors.

The R code produced during these studies was structured in functions of general use and transformed in two Bioconductor [Gentleman et al., 2004] packages. One for the exploratory data analysis that evidences outliers, batch effects and confounding factors (`msmsEDA`). And a second implementing the tests and filters (`msmsTests`). Two graphical user interfaces (GUI) using these two packages were also produced (`msmsEDA_GUI` and `msmsTests_GUI`).

In the next sections we discuss what is missing or what could be developed next.

## 5.1 Limitation in the use of SpC

The discrete nature of the SpC complicates the interpretation of very low levels of expression, where the quantitation by ion intensity could be more reliable (see section 2.5). Or at least not as of the black and white condition.

Either with the tests or when using the post-test filter to improve reproducibility, the quality in the results at this low level will depend of the number of replicates, as long as the average expression within biologic condition loses its discrete nature. That is, in the measure that low mean values acquire more entropy. This means that the lower the level of expression on which we are interested, the higher the number of required replicates.

## 5.2 Batch effects in diagnostics

Despite the fact that batch effects may be visualized, quantified, and corrected in balanced experiments, a question remains in how to deal with isolated samples in LC-MS/MS. In biomarker discovery (BD) there seems to be no alternative to minimally balanced experiments [Scherer, 2009]. The number of samples of each condition in a batch has to be adjusted according to the capabilities of one single day of LC-MS/MS in the lab [Gregori et al., 2012]. And the samples in a batch have to be processed in parallel from the very beginning to the final LC-MS/MS measurement [Quinn & Keough, 2002]. The question to answer is how to measure a new single sample of unknown condition to be classified (diagnosed).

In most cases the biomarker could be a single protein, or a very limited number of proteins, which could be measured by a direct immunochemical test where batch effects are limited or of no concern. In some cases, where the combination of signal

and effect size is not strong enough for any single protein, the biomarker could be a signature composed of a higher number of proteins which could recommend the use of LC-MS/MS in diagnostics.

To answer the question in this latter case we must distinguish between sample normalization, which is a transformation by columns, and the batch effects, which is row dependent because each peptide may have a different response to the global uncontrolled factors. Normalization of a new sample by the total number of SpC, and eventually by the protein production rate of a cell-line, poses no problems provided that the total quantity of protein measured is the same. On the other hand avoiding confusion due to uncontrolled factors will require a control sample measured immediately after. This control sample must be stable in time and contain all the relevant proteins in due concentrations. Both samples must be processed in parallel so that they may constitute a single batch, in that they are equally influenced by all uncontrolled factors [Quinn & Keough, 2002]. This is not at all simple and requires further study and assessment.

## 5.3 Underdispersion

Although the main concern in most methods developed for comparative proteomics is overdispersion caused by biologic variability (see section 2.7.3), in the experiments with controlled samples of yeast lysate we observed a significant degree of underdispersion. This was consistently observed at all levels of expression (Figure 5.1). That is, the observed variance was mainly below the observed mean expression. In different experiments with cell-line secretomes we observed some degree of underdispersion as well.

Any model able to describe this reduced variance could benefit from a higher sensitivity than the Poisson GLM. Both the negative-binomial distribution (see equation 2.10), and the quasilikelihood extension to the GLM (see equation 2.11) may account for underdispersion as well as for overdispersion. The quasilikelihood models underdispersion with positive values of $\psi$ below 1, the negative-binomial models underdispersion for negative values of $\phi$ above $-1/\mu$ [Agresti, 2002].

The estimation of the dispersion introduces a second parameter in the model, requiring then a higher number of replicates, which is the biggest limitation in a proteomics lab, where scarcely more than three replicates are affordable. The solution provided by `edgeR` [Robinson et al., 2010], similar to that implemented in `limma` [Smyth, 2005] for microarrays, allows to share information across proteins,

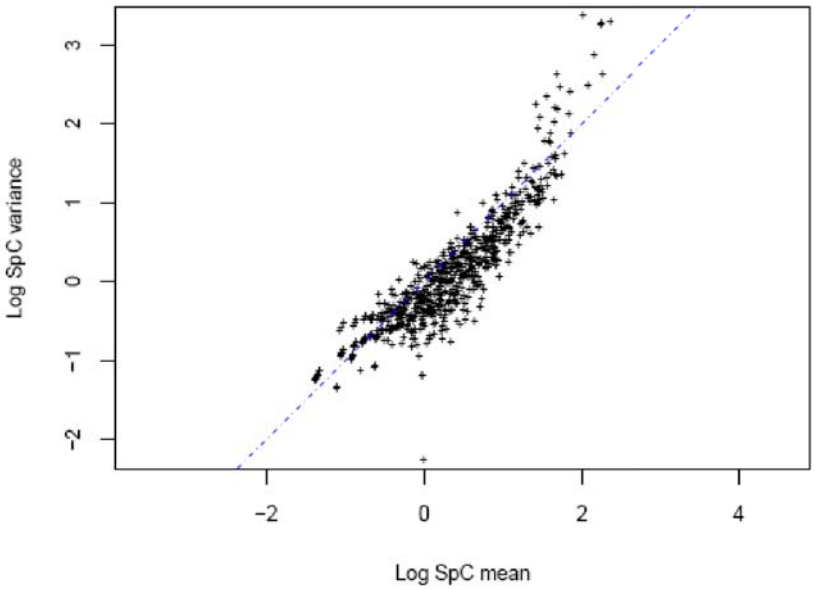



Figure 5.1: Residual dispersion in the experiments with yeast lisate spiked with human proteins. The dots under the diagonal show underdispersion.

and alleviates the need for a higher number of samples. An interesting work could be to study and develop a similar empirical Bayes approach, such that used in the mentioned packages, for the quasilikelihood.

## 5.4 Possible uses in other *omics*

This thesis has been developed on datasets which are sparse matrices of counts similar to those observed in at least other two *omics*: RNA-seq tables in transcriptomics [Wang et al., 2009] and OTU tables in metagenomics [Wooley et al., 2010], both using NGS technologies but answering quite different questions. In this respect, besides specific normalization issues, studying differential protein secretion in cancer cell-lines represents similar statistical challenges as studying differential gene expression by RNA-seq, or differential abundance in microbiomes by 16S rRNA sequencing. Specifically the `edgeR` package [Robinson et al., 2010] used in our software was developed for RNA-seq analysis.

In both *omics* reproducibility is a key issue and the batch effects correction and post-test filters could contribute to improve the results in all of them. The implementation of some of the solutions provided by this work to these other *omics* should not be mimetic because of specificities of each field, but could constitute a

good basis.

## 5.5 What next ?

With the model in equation 4.4, comparative proteomics with secretomes based on SpC may be considered unbiased, provided that sample batches are minimally balanced in the conditions to compare. This imposes a restriction of the sort observed for labelled proteomics. A possible solution to this shortcoming could be to use a universal control sample, so that any condition could be measured against this control. This could allow the unbiased comparison of conditions in different batches having the same control samples. Although conceptually simple this possibility deserves a careful study, and it is not clear whether a different universal control sample could be appropriate for different cell-lines. The batch effects could be corrected just for the proteins in common among the conditions to be compared and the control samples, and provided that the expression level is of the same order. This study requires extensive experiments and attention and may constitute the basis for a new project, along with the considerations in diagnostics briefly pointed out above (see 5.2).

One remaining important issue is the validation of biomarker signatures in proteomics. When the BD process brings to one single or very few molecules, the validation will be done by methods different to LC-MS/MS, as ELISA for instance. Instead, when the BD process brings to a complex mixture of proteins, to a proteomic signature, discovery and validation become part of the same problem, and attention has to be given to avoid the overfitting of the signature to the dataset used in BD. One interesting line of study could be the implementation in proteomics of the general methodology proposed by [Parry et al., 2010] for microarrays on the light of the MAQC-II study [Shi et al., 2010].

Finally, another possible project could be to study the incidence of the batch effects, and the required levels of post-test filters to improve reproducibility when measuring the proteins by ion intensity, instead of by SpC.

# Chapter 6

# GENERAL CONCLUSIONS

This thesis has explored essential aspects of biomarker discovery by spectral counts in label-free shotgun proteomics by LC-MS/MS, studing cancer cell-lines secretomes. Specifically, it has been shown that:

1. Normalizing by total SpC by sample in technical replicates results in a more stable normalization than using internal standards, even when multiple standards are used.

2. Batch effects are likely present on every proteomics project done for more than one day of instrument data acquisition.

3. An exploratory data analysis is a necessary step to assess the quality of a dataset, to identify putative outliers, and eventual batch effects and confounding factors.

4. Fully unbalanced LC-MS/MS runs will likely produce biased results. The eventual bias cannot be corrected by any means.

5. Batch effects may and should be corrected in balanced experimental designs.

6. A post-test filter with reasonable effect size and signal level thresholds helps to increase the reproducibility of comparative proteomic analysis.

7. Long lists of DEPs could be favourably shortened by increasing the thresholds in the post-test filter, instead of using a more stringent significance level.

8. Cancer cell lines show different global protein secretion rates at their basal state, or under different perturbations.

9. Under GLM, equation 4.4 provides an unbiased cell-to-cell comparison in cell-line secretomes.

10. Secretion rate and proliferation rate seem to be inversely correlated for a given cell-line.

11. Finally, the developed solutions and the designed model have been implemented in two R/Bioconductor packages, `msmsEDA` and `msmsTests`, and its dissemination and use by non-experts has been facilitated by two graphical interfaces, `msmsEDA_GUI` and `msmsTests_GUI`.

# Bibliography


Absalan, F. & Ronaghi, M. (2007). Molecular inversion probe assay. In N. Bergman (Ed.), *Comparative Genomics*, volume 396 of *Methods In Molecular Biology* (pp. 315–330). Humana Press.

Agresti, A. (2002). *Categorical Data Analysis*. Wiley Series in Probability and Statistics - Applied Probability and Statistics Section Series. Wiley.

Ahn, N. G., Shabb, J. B., Old, W. M., & Resing, K. A. (2007). Achieving in-depth proteomics profiling by mass spectrometry. *ACS Chem. Biol.*, *2*(1), 39–52. PMID: 17243782.

Allison, D. B., Cui, X., Page, G. P., & Sabripour, M. (2006). Microarray data analysis: from disarray to consolidation and consensus. *7*(1), 5565.

Audic, S. & Claverie, J. M. (1997). The significance of digital gene expression profiles. *Genome Res.*, *7*(10), 986–995. PMID: 9331369.

Auer, P. L. & Doerge, R. W. (2010). Statistical design and analysis of RNA sequencing data. *Genetics*, *185*(2), 405–416. PMID: 20439781.

Baggerly, K. A. & Coombes, K. R. (2009). Deriving chemosensitivity from cell lines: forensic bioinformatics and reproducible research in high-throughput biology. *Ann. Appl. Stat.*, *3*(4), 1309–1334.

Baggerly, K. A., Coombes, K. R., & Neeley, E. S. (2008). Run batch effects potentially compromise the usefulness of genomic signatures for ovarian cancer. *J. Clin. Oncol.*, *26*(7), 1186–1187; author reply 1187–1188. PMID: 18309960.

Baggerly, K. A., Edmonson, S. R., Morris, J. S., & Coombes, K. R. (2004). High-resolution serum proteomic patterns for ovarian cancer detection. *Endocr. Relat. Cancer*, *11*(4), 583–584; author reply 585–587. PMID: 15613439.

Baggerly, K. A., Morris, J. S., Edmonson, S. R., & Coombes, K. R. (2005). Signal in noise: evaluating reported reproducibility of serum proteomic tests for ovarian cancer. *J. Natl. Cancer Inst.*, *97*(4), 307–309. PMID: 15713966.

Bellman, R. (1961). *Adaptive control processes: a guided tour*. Princeton: Princeton University Press.

Benjamini, Y. & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *57*(1), 289300.

Bibikova, M., Talantov, D., Chudin, E., Yeakley, J. M., Chen, J., Doucet, D., Wickham, E., Atkins, D., Barker, D., Chee, M., Wang, Y., & Fan, J.-B. (2004). Quantitative gene expression profiling in formalin-fixed, paraffin-embedded tissues using universal bead arrays. *Am. J. Pathol.*, *165*(5), 1799–1807. PMID: 15509548.

Bolstad, B. M., Irizarry, R. A., Astrand, M., & Speed, T. P. (2003). A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, *19*(2), 185–193. PMID: 12538238.

Borst, P. & Wessels, L. (2010). Do predictive signatures really predict response to cancer chemotherapy? *Cell Cycle*, *9*(24), 4836–4840. PMID: 21150277.

Bourgon, R., Gentleman, R., & Huber, W. (2010). Independent filtering increases detection power for high-throughput experiments. *107*(21), 95469551.

Callister, S. J., Barry, R. C., Adkins, J. N., Johnson, E. T., Qian, W.-J., Webb-Robertson, B.-J. M., Smith, R. D., & Lipton, M. S. (2006). Normalization approaches for removing systematic biases associated with mass spectrometry and label-free proteomics. *J. Proteome Res.*, *5*(2), 277–286. PMID: 16457593.

Chen, C., Grennan, K., Badner, J., Zhang, D., Gershon, E., Jin, L., & Liu, C. (2011). Removing batch effects in analysis of expression microarray data: an evaluation of six batch adjustment methods. *PLoS ONE*, *6*(2), e17238. PMID: 21386892.

Chen, J. J., Hsueh, H.-M., Delongchamp, R. R., Lin, C.-J., & Tsai, C.-A. (2007). Reproducibility of microarray data: a further analysis of microarray quality control (MAQC) data. *BMC Bioinformatics*, *8*, 412. PMID: 17961233.

Choi, H., Fermin, D., & Nesvizhskii, A. I. (2008). Significance analysis of spectral count data in label-free shotgun proteomics. *7*(12), 23732385.

Conrads, T. P., Fusaro, V. A., Ross, S., Johann, D., Rajapakse, V., Hitt, B. A., Steinberg, S. M., Kohn, E. C., Fishman, D. A., Whitely, G., Barrett, J. C., Liotta, L. A., Petricoin, E F, r., & Veenstra, T. D. (2004). High-resolution serum proteomic features for ovarian cancer detection. *Endocr. Relat. Cancer*, *11*(2), 163–178. PMID: 15163296.

Degroeve, S., Colaert, N., Vandekerckhove, J., Gevaert, K., & Martens, L. (2011). A reproducibility-based evaluation procedure for quantifying the differences between MS/MS peak intensity normalization methods. *Proteomics*, *11*(6), 1172–1180. PMID: 21298791.

Dressman, H. K., Berchuck, A., Chan, G., Zhai, J., Bild, A., Sayer, R., Cragun, J., Clarke, J., Whitaker, R. S., Li, L., Gray, J., Marks, J., Ginsburg, G. S., Potti, A., West, M., Nevins, J. R., & Lancaster, J. M. (2007). An integrated genomic-based approach to individualized treatment of patients with advanced-stage ovarian cancer. *J. Clin. Oncol.*, *25*(5), 517–525. PMID: 17290060.

Dupuy, A. & Simon, R. M. (2007). Critical review of published microarray studies for cancer outcome and guidelines on statistical analysis and reporting. *J. Natl. Cancer Inst.*, *99*(2), 147–157. PMID: 17227998.

Gao, J., Friedrichs, M. S., Dongre, A. R., & Opiteck, G. J. (2005). Guidelines for the routine application of the peptide hits technique. *J. Am. Soc. Mass Spectrom.*, *16*(8), 1231–1238. PMID: 15978832.

Gatto, L. & Lilley, K. S. (2012). MSnbase-an R/Bioconductor package for isobaric tagged mass spectrometry data visualization, processing and quantitation. *Bioinformatics*, *28*(2), 288–289. PMID: 22113085.

Gentleman, R. C., Carey, V. J., Bates, D. M., & others (2004). Bioconductor: Open software development for computational biology and bioinformatics. *Genome Biology*, *5*, R80.

Gregori, J., Mendez, O., Katsila, T., Pujals, M., Salvans, C., Villareal, L., Arribas, J., Tabernero, J., Sanchez, A., & Villanueva, J. (2014). Enhancing the biological relevance of secretome-based proteomics by linking tumor cell proliferation and protein secretion. *J. of Proteome Research*, under review.

Gregori, J., Villarreal, L., Méndez, O., Sánchez, A., Baselga, J., & Villanueva, J. (2012). Batch effects correction improves the sensitivity of significance tests in spectral counting-based comparative discovery proteomics. *Journal of Proteomics*, *75*(13), 3938–3951. PMID: 22588121.

Gregori, J., Villarreal, L., Sánchez, A., Baselga, J., & Villanueva, J. (2013). An effect size filter improves the reproducibility in spectral counting-based comparative proteomics. *Journal of Proteomics*, *75*, 55–65.

Gronborg, M., Kristiansen, T. Z., Iwahori, A., Chang, R., Reddy, R., Sato, N., Molina, H., Jensen, O. N., Hruban, R. H., Goggins, M. G., Maitra, A., & Pandey, A. (2006). Biomarker discovery from pancreatic cancer secretome using a differential proteomic approach. *Mol. Cell Proteomics*, *5*(1), 157–171. PMID: 16215274.

Gygi, S. P., Rochon, Y., Franza, B. R., & Aebersold, R. (1999). Correlation between protein and mRNA abundance in yeast. *Mol. Cell. Biol.*, *19*(3), 1720–1730. PMID: 10022859.

Ho, C. C., Mun, K. S., & Naidu, R. (2013). SNP array technology: an array of hope in breast cancer research. *Malays J Pathol*, *35*(1), 33–43. PMID: 23817393.

Hoshida, Y., Villanueva, A., Kobayashi, M., Bruix, J., Friedman, S. L., Kumada, H., Llovet, J. M., & Golub, T. R. (2008). Gene expression in fixed tissues and outcome in hepatocellular carcinoma. *N Engl J Med*, *359*(19), 1995–2004.

Iterson, M. v., Boer, J. M., & Menezes, R. X. d. (2010). Filtering, FDR and power. *11*.

Jain, N., Thatte, J., Braciale, T., Ley, K., O'Connell, M., & Lee, J. K. (2003). Local-pooled-error test for identifying differentially expressed genes with a small number of replicated microarrays. *Bioinformatics*, *19*(15), 1945–1951. PMID: 14555628.

Kim, S.-Y. (2009). Effects of sample size on robustness and prediction accuracy of a prognostic gene signature. *BMC Bioinformatics*, *10*, 147. PMID: 19445687.

Kuo, W. P., Jenssen, T.-K., Butte, A. J., Ohno-Machado, L., & Kohane, I. S. (2002). Analysis of matched mRNA measurements from two different microarray technologies. *Bioinformatics*, *18*(3), 405–412. PMID: 11934739.

Kutner, M., Nachtsheim, C., Neter, J., & Li, W. (2005). *Applied Linear Statistical Models* (5 ed.). Boston, MA: McGraw-Hill.

Lawlor, K., Nazarian, A., Lacomis, L., Tempst, P., & Villanueva, J. (2009). Pathway-based biomarker search by high-throughput proteomics profiling of secretomes. *8*(3), 14891503.

Lawrence, M. & Temple Lang, D. (2010). RGtk2: A graphical user interface toolkit for R. *Journal of Statistical Software*, *37*(8), 1–52.

Lawrence, M. F. & Verzani, J. (2012). *Programming Graphical User Interfaces in R*. Stanford: Chapman & Hall/CRC.

Lazar, C., Meganck, S., Taminau, J., Steenhoff, D., Coletta, A., Molter, C., Weiss-Solís, D. Y., Duque, R., Bersini, H., & Nowé, A. (2013). Batch effect removal methods for microarray gene expression data integration: a survey. *Brief. Bioinformatics*, *14*(4), 469–490. PMID: 22851511.

Leek, J. T., Scharpf, R. B., Bravo, H. C., Simcha, D., Langmead, B., Johnson, W. E., Geman, D., Baggerly, K., & Irizarry, R. A. (2010). Tackling the widespread and critical impact of batch effects in high-throughput data. *11*(10), 733739. PMID: 20838408.

Leitch, M. C., Mitra, I., & Sadygov, R. G. (2012). Generalized linear and mixed models for label-free shotgun proteomics. *Stat Interface*, *5*(1), 89–98. PMID: 22822415.

Li, M., Gray, W., Zhang, H., Chung, C. H., Billheimer, D., Yarbrough, W. G., Liebler, D. C., Shyr, Y., & Slebos, R. J. C. (2010). Comparative shotgun proteomics using spectral count data and quasi-likelihood modeling. *9*(8), 42954305.

Lundgren, D. H., Hwang, S.-I., Wu, L., & Han, D. K. (2010). Role of spectral counting in quantitative proteomics. *Expert Rev Proteomics*, *7*(1), 39–53. PMID: 20121475.

Luo, J., Schumacher, M., Scherer, A., Sanoudou, D., Megherbi, D., Davison, T., Shi, T., Tong, W., Shi, L., Hong, H., Zhao, C., Elloumi, F., Shi, W., Thomas, R., Lin, S., Tillinghast, G., Liu, G., Zhou, Y., Herman, D., Li, Y., Deng, Y., Fang, H., Bushel, P., Woods, M., & Zhang, J. (2010). A comparison of batch effect removal methods for enhancement of prediction performance using MAQC-II microarray gene expression data. *10*(4), 278291.

Madu, C. O. & Lu, Y. (2010). Novel diagnostic biomarkers for prostate cancer. *J Cancer*, *1*, 150–177. PMID: 20975847.

Marshall, E. (2004). Getting the noise out of gene arrays. *Science*, *306*(5696), 630–631. PMID: 15499004.

Mathias, R. A., Wang, B., Ji, H., Kapp, E. A., Moritz, R. L., Zhu, H.-J., & Simpson, R. J. (2009). Secretome-based proteomic profiling of ras-transformed MDCK cells reveals extracellular modulators of epithelial-mesenchymal transition. *J. Proteome Res.*, *8*(6), 2827–2837. PMID: 19296674.

Matsui, S. (2013). Genomic biomarkers for personalized medicine: development and validation in clinical studies. *Comput Math Methods Med*, *2013*, 865980. PMID: 23690882.

Neilson, K. A., Ali, N. A., Muralidharan, S., Mirzaei, M., Mariani, M., Assadourian, G., Lee, A., van Sluyter, S. C., & Haynes, P. A. (2011). Less label, more free: Approaches in label-free quantitative mass spectrometry. *11*(4), 535553.

Nesvizhskii, A. I., Vitek, O., & Aebersold, R. (2007). Analysis and validation of proteomic data generated by tandem mass spectrometry. *Nat. Methods*, *4*(10), 787–797. PMID: 17901868.

OHara, R. B. & Kotze, D. J. (2010). Do not log-transform count data. *Methods in Ecology and Evolution*, *1*(2), 118–122.

Old, W. M., Meyer-Arendt, K., Aveline-Wolf, L., Pierce, K. G., Mendoza, A., Sevinsky, J. R., Resing, K. A., & Ahn, N. G. (2005). Comparison of label-free methods for quantifying human proteins by shotgun proteomics. *Mol. Cell Proteomics*, *4*(10), 1487–1502. PMID: 15979981.

Park, T., Yi, S.-G., Kang, S.-H., Lee, S., Lee, Y.-S., & Simon, R. (2003). Evaluation of normalization methods for microarray data. *BMC Bioinformatics*, *4*, 33. PMID: 12950995.

Parry, R. M., Jones, W., Stokes, T. H., Phan, J. H., Moffitt, R. A., Fang, H., Shi, L., Oberthuer, A., Fischer, M., Tong, W., & Wang, M. D. (2010). k-nearest neighbor models for microarray gene expression analysis and clinical outcome prediction. *Pharmacogenomics J.*, *10*(4), 292–309. PMID: 20676068.

Patel, V. J., Thalassinos, K., Slade, S. E., Connolly, J. B., Crombie, A., Murrell, J. C., & Scrivens, J. H. (2009). A comparison of labeling and label-free mass spectrometry-based proteomics approaches. *J. Proteome Res.*, *8*(7), 3752–3759. PMID: 19435289.

Pavelka, N., Fournier, M. L., Swanson, S. K., Pelizzola, M., Ricciardi-Castagnoli, P., Florens, L., & Washburn, M. P. (2008). Statistical similarities between transcriptomics and quantitative shotgun proteomics data. *Mol. Cell Proteomics*, *7*(4), 631–644. PMID: 18029349.

Pavelka, N., Pelizzola, M., Vizzardelli, C., Capozzoli, M., Splendiani, A., Granucci, F., & Ricciardi-Castagnoli, P. (2004). A power law global error model for the identification of differentially expressed genes in microarray data. *BMC Bioinformatics*, *5*, 203. PMID: 15606915.

Petricoin, E. F., Ardekani, A. M., Hitt, B. A., Levine, P. J., Fusaro, V. A., Steinberg, S. M., Mills, G. B., Simone, C., Fishman, D. A., Kohn, E. C., & Liotta, L. A. (2002). Use of proteomic patterns in serum to identify ovarian cancer. *Lancet*, *359*(9306), 572–577. PMID: 11867112.

Pham, T. V., Piersma, S. R., Warmoes, M., & Jimenez, C. R. (2010). On the beta-binomial model for analysis of spectral count data in label-free tandem mass spectrometry-based proteomics. *Bioinformatics*, *26*(3), 363–369.

Potti, A., Dressman, H. K., Bild, A., Riedel, R. F., Chan, G., Sayer, R., Cragun, J., Cottrill, H., Kelley, M. J., Petersen, R., Harpole, D., Marks, J., Berchuck, A., Ginsburg, G. S., Febbo, P., Lancaster, J., & Nevins, J. R. (2011). Retraction: Genomic signatures to guide the use of chemotherapeutics. *Nat. Med.*, *17*(1), 135. PMID: 21217686.

Pounds, S. & Cheng, C. (2005). Statistical development and evaluation of microarray gene expression data filters. *12*(4), 482495. PMID: 15882143.

Powell, D. W., Weaver, C. M., Jennings, J. L., McAfee, K. J., He, Y., Weil, P. A., & Link, A. J. (2004). Cluster analysis of mass spectrometry data reveals a novel component of SAGA. *Mol. Cell. Biol.*, *24*(16), 7249–7259. PMID: 15282323.

Quinn, G. & Keough, M. (2002). *Experimental Design and Data Analysis for Biologists*. Cambridge University Press.

R Core Team (2012). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. ISBN 3-900051-07-0.

Rakha, E. A. (2013). Pitfalls in outcome prediction of breast cancer. *Journal of Clinical Pathology*.

Ransohoff, D. F. (2004). Rules of evidence for cancer molecular-marker discovery and validation. *Nat. Rev. Cancer*, *4*(4), 309–314. PMID: 15057290.

Ransohoff, D. F. (2005a). Bias as a threat to the validity of cancer molecular-marker research. *5*(2), 142149. PMID: 15685197.

Ransohoff, D. F. (2005b). Lessons from controversy: ovarian cancer screening and serum proteomics. *97*(4), 315319. PMID: 15713968.

Rappsilber, J., Ryder, U., Lamond, A. I., & Mann, M. (2002). Large-scale proteomic analysis of the human spliceosome. *Genome Res.*, *12*(8), 1231–1245. PMID: 12176931.

Rifai, N., Gillette, M. A., & Carr, S. A. (2006). Protein biomarker discovery and validation: the long and uncertain path to clinical utility. *Nat. Biotechnol.*, *24*(8), 971–983. PMID: 16900146.

Robinson, M. D., McCarthy, D. J., & Smyth, G. K. (2010). edgeR: a bioconductor package for differential expression analysis of digital gene expression data. *26*(1), 139140.

Robinson, M. D. & Smyth, G. K. (2008). Small-sample estimation of negative binomial dispersion, with applications to sage data. *Biostatistics*, *9*(2), 321–332.

Sánchez-Pena, M. L., Isaza, C. E., Pérez-Morales, J., Rodríguez-Padilla, C., Castro, J. M., & Cabrera-Ríos, M. (2013). Identification of potential biomarkers from microarray experiments using multiple criteria optimization. *Cancer Med*, *2*(2), 253–265. PMID: 23634293.

Sandin, M., Krogh, M., Hansson, K., & Levander, F. (2011). Generic workflow for quality assessment of quantitative label-free LC-MS analysis. *Proteomics*, *11*(6), 1114–1124. PMID: 21298787.

Scherer, A. (2009). *Batch Effects and Noise in Microarray Experiments: Sources and Solutions*. Wiley Series in Probability and Statistics. Wiley.

Schloss, P. D., Gevers, D., & Westcott, S. L. (2011). Reducing the effects of PCR amplification and sequencing artifacts on 16S rRNA-based studies. *PLoS ONE*, *6*(12), e27310. PMID: 22194782.

Shalon, D., Smith, S. J., & Brown, P. O. (1996). A DNA microarray system for analyzing complex DNA samples using two-color fluorescent probe hybridization. *Genome Research*, *6*(7), 639–645.

Shen, Y., Kim, J., Strittmatter, E. F., Jacobs, J. M., Camp, David G, n., Fang, R., Toli, N., Moore, R. J., & Smith, R. D. (2005). Characterization of the human blood plasma proteome. *Proteomics*, *5*(15), 4034–4045. PMID: 16152657.

Shi, L., et al., & MAQC-Consortium (2006). The MicroArray quality control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements. *Nat Biotech*, *24*(9), 1151–1161.

Shi, L., et al., & MAQC-Consortium (2008). The balance of reproducibility, sensitivity, and specificity of lists of differentially expressed genes in microarray studies. *9 Suppl 9*, S10. PMID: 18793455.

Shi, L., et al., & MAQC-Consortium (2010). The MicroArray quality control (MAQC)-II study of common practices for the development and validation of microarray-based predictive models. *28*(8), 827838. PMID: 20676074.

Simon, R., Radmacher, M. D., Dobbin, K., & McShane, L. M. (2003). Pitfalls in the use of DNA microarray data for diagnostic and prognostic classification. *J. Natl. Cancer Inst.*, *95*(1), 14–18. PMID: 12509396.

Smyth, G. K. (2005). Limma: linear models for microarray data. In R. Gentleman, V. Carey, S. Dudoit, R. Irizarry, & W. Huber (Eds.), *Bioinformatics and Computational Biology Solutions using R and Bioconductor* (pp. 397–420). New York: Springer.

Stastna, M. & Van Eyk, J. E. (2012). Secreted proteins as a fundamental source for biomarker discovery. *Proteomics*, *12*(4-5), 722–735. PMID: 22247067.

States, D. J., Omenn, G. S., Blackwell, T. W., Fermin, D., Eng, J., Speicher, D. W., & Hanash, S. M. (2006). Challenges in deriving high-confidence protein identifications from data gathered by a HUPO plasma proteome collaborative study. *Nat. Biotechnol.*, *24*(3), 333–338. PMID: 16525410.

Steen, H. & Mann, M. (2004). The ABC's (and XYZ's) of peptide sequencing. *Nat. Rev. Mol. Cell Biol.*, *5*(9), 699–711. PMID: 15340378.

Tan, P. K., Downey, T. J., Spitznagel, Edward L, J., Xu, P., Fu, D., Dimitrov, D. S., Lempicki, R. A., Raaka, B. M., & Cam, M. C. (2003). Evaluation of gene expression measurements from commercial microarray platforms. *Nucleic Acids Res.*, *31*(19), 5676–5684. PMID: 14500831.

Tibshirani, R. J. & Efron, B. (2002). Pre-validation and inference in microarrays. *Stat Appl Genet Mol Biol*, *1*, Article1. PMID: 16646777.

Tirumalai, R. S., Chan, K. C., Prieto, D. A., Issaq, H. J., Conrads, T. P., & Veenstra, T. D. (2003). Characterization of the low molecular weight human serum proteome. *Molecular & Cellular Proteomics*, *2*(10), 1096–1103.

Valsesia, A., Mac, A., Jacquemont, S., Beckmann, J. S., & Kutalik, Z. (2013). The growing importance of CNVs: new insights for detection and clinical interpretation. *Front Genet*, *4*, 92. PMID: 23750167.

van 't Veer, L. J., Dai, H., van de Vijver, M. J., He, Y. D., Hart, A. A. M., Mao, M., Peterse, H. L., van der Kooy, K., Marton, M. J., Witteveen, A. T., Schreiber, G. J., Kerkhoven, R. M., Roberts, C., Linsley, P. S., Bernards, R., & Friend, S. H. (2002). Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, *415*(6871), 530–536.

Verzani, J. (2012). *gWidgets: gWidgets API for building toolkit-independent, interactive GUIs*. R package version 0.0-52.

Villarreal, L., Méndez, O., Salvans, C., Gregori, J., Baselga, J., & Villanueva, J. (2013). Unconventional secretion is a major contributor of cancer cell line secretomes. *Mol. Cell Proteomics*, *12*(5), 1046–1060. PMID: 23268930.

Wang, Z., Gerstein, M., & Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet*, *10*(1), 57–63.

Wilks, S. S. (1938). The large-sample distribution of the likelihood ratio for testing composite hypotheses. *Ann. Math. Statist.*, *9*(1), 60–62.

Wolff, A., et al., American Society of Clinical Oncology, & College of American Pathologists (2007). American society of clinical Oncology/College of american pathologists guideline recommendations for human epidermal growth factor receptor 2 testing in breast cancer. *J. Clin. Oncol.*, *25*(1), 118–145. PMID: 17159189.

Wooley, J. C., Godzik, A., & Friedberg, I. (2010). A primer on metagenomics. *PLoS Comput Biol*, *6*(2), e1000667.

Zhang, B., VerBerkmoes, N. C., Langston, M. A., Uberbacher, E., Hettich, R. L., & Samatova, N. F. (2006). Detecting differential and correlated protein expression in label-free shotgun proteomics. *J. Proteome Res.*, *5*(11), 2909–2918. PMID: 17081042.

Zhu, W., Smith, J. W., & Huang, C.-M. (2010). Mass spectrometry-based label-free quantitative proteomics. *J. Biomed. Biotechnol.*, *2010*, 840518. PMID: 19911078.

Zybailov, B., Coleman, M. K., Florens, L., & Washburn, M. P. (2005). Correlation of relative abundance ratios derived from peptide ion chromatograms and spectrum counting for quantitative proteomic analysis using stable isotope labeling. *Anal. Chem.*, *77*(19), 6218–6224. PMID: 16194081.

Zybailov, B., Mosley, A. L., Sardiu, M. E., Coleman, M. K., Florens, L., & Washburn, M. P. (2006). Statistical analysis of membrane proteome expression changes in saccharomyces cerevisiae. *J. Proteome Res.*, *5*(9), 2339–2347. PMID: 16944946.

# Appendix A

# ANNEX DOCUMENTS

Documents attached to the thesis:

1. **Papers**

   (a) Batch effects correction improves the sensitivity of significance tests in spectral counting-based comparative discovery proteomics.

   (b) An effect size filter improves the reproducibility in spectral counting-based comparative proteomics.

   (c) Cell-centric statistical modelling enhances the biological content of secretome-based comparative proteomic studies.

   (d) Unconventional Secretion is a Major Contributor of Cancer Cell Line Secretomes

2. **Software**

   (a) *Tutorial* - msmsEDA and msmsTests: R/Bioconductor packages for spectral count label-free proteomics data analysis.

   (b) *User guide* - GUI for the msmsEDA package: Label-free SpC LC-MS/MS exploratory. data analysis

   (c) *Vignette* - msmsEDA: LC-MS/MS Exploratory Data Analysis.

   (d) *Manual* - Package msmsEDA

   (e) *User guide* - GUI for the msmsTests package: Label-free SpC LC-MS/MS differential. expression

   (f) *Vignette* - msmsTests package: LC-MS/MS post test filters to improve reproducibility.

(g) *Vignette* - msmsTests package: Blocks design to compensate batch effects.

(h) *Manual* - Package msmsTests

3. **Other papers**

Papers in statistics and bioinformatics not related to the thesis and published in international peer reviewed journals during the period of the thesis. Just the first paper page is attached.

(a) Inference with viral quasispecies diversity indices: clonal and NGS approaches.

(b) Ultra-Deep Pyrosequencing (UDPS) Data Treatment to Study Amplicon HCV Minor Variants.

(c) A comparative study of ultra-deep pyrosequencing and cloning to quantitatively analyze the viral quasispecies using hepatitis B virus infection as a model.

(d) Identification of host and viral factors involved in a dissimilar resolution of a hepatitis C virus infection.

(e) Extinction of hepatitis C virus by ribavirin in hepatoma cells involves lethal mutagenesis.

(f) Molecular epidemiology and putative origin of hepatitis C virus in random volunteers from Argentina.