



VNIVERSITAT
E VALÈNCIA

Faculty of Medicine, Biochemistry and Molecular Biology Department

Health Research Institute Hospital La Fe, Molecular, Cellular and Genomic Biomedicine

Group

Development of Artificial Intelligence Methods for Clinical Genomics

A dissertation submitted to the University of Valencia in partial fulfillment of the
requirements for the

Doctoral Programme in Medicine

by

Óscar Bastidas García

Thesis supervisors.

Prof. José Enrique O'Connor Blasco

Dr. Regina Rodrigo Nicolás

Dr. Rafael Vázquez Manrique

Valencia, January 2022

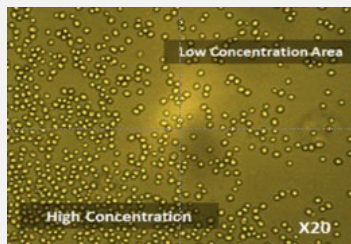
Thesis Title: Development of Artificial Intelligence Methods for Clinical Genomics.



CONTENT WARNING!

The present thesis analyzes and proposes a set of different methodologies in TWO DISTINCT fields.

Part 1: CYTOMETRY



Page 23

Part 2: GENOMICS



Page 108

THESIS SUPERVISORS:

PROF. JOSÉ ENRIQUE O'CONNOR BLASCO

Department of Biochemistry and Molecular Biology,
Faculty of Medicine, University of Valencia

DR. REGINA RODRIGO NICOLÁS

Pathophysiology and Therapy for Vision Disorder
Principe Felipe Research Centre.

DR. RAFAEL VÁZQUEZ MANRIQUE

Molecular, Cellular and Genomic Biomedicine Lab
Health Research Institute Hospital La Fe.

By Óscar Bastidas García.

MBA International INSEAD / Wharton.

M.Sc.EE. Microelectronics, Telecom Bretagne.

B.Sc.EE. Telecommunication, Polytechnic Univ. of Valencia.

THESIS SUPERVISION CERTIFICATE

Professor Dr. **José Enrique O'Connor Blasco**, from the Department of Biochemistry and Molecular Biology, Faculty of Medicine and Odontology, University of Valencia.

Dr. **Regina Rodrigo Nicolás**, from the Pathophysiology and Therapy for Vision Disorders at the Principe Felipe Research Center, Valencia.

Dr. **Rafael Vázquez Manrique**, from the Molecular, Cellular and Genomic Biomedicine Laboratory at the Instituto de Investigación Sanitaria La Fe of Valencia.

CERTIFY:

That **OSCAR BASTIDAS GARCÍA** has performed under our direction the research work for his Doctoral Thesis with the title "**Development of Artificial Intelligence Methods for Clinical Genomics**", here presented with our authorization to the Doctorate Program "Medicine".

Signed in Valencia, on 21 January 2022

Firmado por JOSE
ENRIQUE O'CONNOR
BLASCO -

el día
21/01/2022 con un
certificado emitido por
ACCVCA-120

RODRIGO
NICOLAS
REGINA -

Firmado digitalmente por
RODRIGO
NICOLAS REGINA -
Z
Fecha: 2022.01.21
10:55:53 +01'00'

RAFAEL
PASCUAL|
VAZQUEZ|
MANRIQUE

Firmado digitalmente por
RAFAEL PASCUAL|
VAZQUEZ|
MANRIQUE
Fecha: 2022.01.21
11:38:00 +01'00'

Funding and Support

This research has been supported by the following entities and institutions:

Biochemistry and Molecular Biology Department, Faculty of Medicine, University of Valencia.

Molecular, Cellular and Genomics Biomedicine Group, Health Research Institute La Fe

Genomics Unit, Health Research Institute La Fe

CEEI Valencia.

IDEAS program, StartUPV.

Funding:

Binartis Genomics, S.L.

Celeromics Technologies, S.L.

CDTI – Centro para el Desarrollo Tecnológico e Industrial.

IVACE - Valencian Institute for Business Competitiveness

IVF – Valencian Institute of Finance

Other recognition and support:

Genoma España - Best Project Bioances Entrepreneurs contest.

Entrepreneur Magazine – Best Project Spain.

Stage Two – Top 4 European Technological Start-up from Universities.

Ajuntament de Valencia – Best Start-up Health

Abbreviations

ACMG	American College of Medical Genetics and Genomics
ASCII	American Standard Code for Information Interchange
AEMPS	Agencia Española del Medicamento y Productos Sanitarios.
AI	Artificial Intelligence
BAM	Binary Alignment Map
BWA	Burrow Wheeler Aligner
BLAST	Basic Local Alignment Search Tool
CCD	Charge-coupled Device
CEIC	Clinical Research Ethics Committee
CNC	Computer Numerical Control (machines)
CR	Coefficient of Repeatability
CSV	Comma Separated Values (file format)
CV	Coefficient of Variation
DOR	Diagnostics Odds Ratio
FDA	Food and Drug Administration
FFPE	Formalin-fixed, paraffin-embedded
GATK	Genome Analysis Toolkit.
HGVS	Human Genome Variation Society
IARC	International Agency for Research on Cancer
IDM	Instituto Interuniversitario de Investigación de Reconocimiento Molecular y Desarrollo Tecnológico.
IF	Impact Factor
IMEGEN	Genomics Medicine Institute. (Instituto de Medicina Genómica.)

Abbreviations

INDEL	Insertion Deletion of bases.
IVD	In vitro Diagnosis
LED	Light emitter diode
HGMD	Human Gene Mutation Database
MAF	Minor Allele Frequency
MCR	Matlab Component Runtime.
mCRC	Metastatic Colorectal Cancer
MNP	Multiple Nucleotide Polymorphism
NGS	Next Generation Sequencing.
NN	Neural Network (computer science)
PC	Personal Computer
PCR	Polymerase Chain Reaction
PI	Principal Investigator / Propidium Iodide
PLA	Polylactic Acid
PCR	Polymerase Chain Reaction
UPV	Polytechnic University of Valencia
VC	Variation coefficient
VCF	Variant Calling Format (file format)
VUS	Variant of Uncertain Significance
SD	Standard Deviation
SNP	Single Nucleotide Polymorphism
SSC	Side Scatter (Light)
WES	Whole Exome Sequencing
WGS	Whole Genome Sequencing

Index

DEVELOPMENT OF ARTIFICIAL INTELLIGENCE METHODS FOR CLINICAL GENOMICS	1
ABBREVIATIONS.....	5
INDEX.....	7
RESUMEN.....	11
AUTOMATION FOR CELL ASSAYS IN CLINICAL RESEARCH	22
ABSTRACT	23
1 INTRODUCTION.....	24
1.1 THE SCIENTIFIC METHOD AND ITS CONTRIBUTION TO MEDICINE.....	24
1.2 THE SCIENTIFIC METHOD IN LABORATORY MEDICINE AND BIOMEDICINE.....	25
1.3 PUBLISHING EXPERIMENTAL RESULTS AS AN ESSENTIAL ASPECT OF BIOMEDICINE.....	26
1.4 CELL COUNTING AS A BASIC EXPERIMENTAL PROCEDURE IN BIOMEDICINE.....	27
1.5 MANUAL CELL COUNTING:	28
1.5.1 <i>Automated Cell Counting:</i>	28
1.5.2 <i>Manual versus Automated Cell Counting:</i>	29
1.4.4 <i>Advantages and Limitations of Cell-Counting Systems</i>	29
1.6 THE ROLE OF AUTOMATION AND ARTIFICIAL INTELLIGENCE FOR IMPROVING BASIC EXPERIMENTAL PROCEDURES IN MEDICINE.....	30
1.7 IS THERE A ROLE FOR ARTIFICIAL INTELLIGENCE FOR IMPROVING JOURNAL REVIEWING OF METHODOLOGICAL ASPECTS? .	33
2 SCIENTIFIC HYPOTHESIS	34
3 OBJECTIVES	34
4 METHODS	35
4.1 IDENTIFYING AND QUANTIFYING JOURNAL REQUIREMENTS FOR REPORTING EXPERIMENTAL RESULTS.....	35
4.2 IDENTIFYING POPULAR AND CRITICAL BASIC CELL ASSAY METHODS	38
4.3 DETECTION OF POTENTIAL ERROR SOURCES WHEN USING NEUBAUER COUNTING CHAMBERS.....	38
4.3.1 <i>Determination of Errors Related to the Volume of Sample Examined:</i>	40
4.3.2 <i>Determination of Errors Related to the Chamber-Loading Technique:</i>	41
4.3.3 <i>Determination of Errors Related to the Number of Cells Counted:</i>	43
4.3.4 <i>Design of an Improved Cell Counting System based on Automation and Artificial Intelligence applied to Microscopic Image Analysis</i>	44
4.3.5 <i>Validation of the Improved Cell Counting System</i>	44
4.3.6 <i>Reference Cell-Counting Methods for Validating the Improved Cell Counting System</i>	45
4.3.7 <i>Validating the Automated Configuration of the Cell Counting System: Cell Concentration</i>	47

4.3.8	<i>Validating the Automated Configuration of the Cell Counting System: Cell Confluence.....</i>	48
4.3.9	<i>Validating the Semi-Automated Configuration of the Cell Counting System.</i>	49
4.3.10	<i>Validating the Artificial Intelligence Algorithm of the Cell Counting System.....</i>	49
4.4	AUTOMATED DETERMINATION OF CELL CONCENTRATION FROM PETRI CULTURE DISHES	51
4.4.1	<i>Estimated Cell Distribution in Petri Dishes:</i>	51
4.4.2	<i>Mathematical Modeling of Petri Dish Cell-Distribution and Cell Sampling</i>	51
5	RESULTS	54
5.1	JOURNAL REQUIREMENTS FOR REPORTING OF EXPERIMENTAL RESULTS IN BIOMEDICINE.	54
5.2	IDENTIFICATION OF CELL COUNTING AS A POPULAR AND CRITICAL EXPERIMENTAL PROCEDURE.....	58
5.3	DETERMINATION OF ERRORS WHEN USING A NEUBAUER CHAMBER	58
5.3.1	<i>Errors Inherent to the Sample Volume: Determination of Chamber Height.....</i>	58
5.3.2	<i>Errors due to insufficient number of counted cells:.....</i>	60
5.3.3	<i>Errors due to chamber loading procedure:.....</i>	62
5.4	DESIGN OF AN IMPROVED CELL COUNTING SYSTEM BASED ON AUTOMATION AND ARTIFICIAL INTELLIGENCE APPLIED TO MICROSCOPIC IMAGE ANALYSIS	70
5.4.1	<i>Elements of the Improved Cell Counting System:.....</i>	70
5.4.2	<i>Operation of the Improved Cell Counting System:</i>	71
5.5	VALIDATION OF THE IMPROVED CELL COUNTING SYSTEM.	75
5.5.1	<i>Validating the Automated Configuration of the Cell Counting System: Cell Concentration.....</i>	75
5.5.2	<i>Validating the Automated Configuration of the Cell Counting System: Cell Confluence.....</i>	82
5.6	AUTOMATED DETERMINATION OF CELL CONCENTRATION FROM PETRI CULTURE DISHES.	85
6	DISCUSSION	89
6.1	STATISTICAL REQUIREMENTS OF LIFE SCIENCES JOURNALS.	90
6.2	LIMITATIONS AND ERROR SOURCES OF CURRENT CELL COUNTING METHODS.	93
6.2.1	<i>Limitations of manual counting with Neubauer chambers:.....</i>	95
6.2.2	<i>Limitations in automated counting with Flow Cytometers:</i>	96
6.2.3	<i>Limitations in automated counting with Image-based counters:</i>	96
6.3	PERFORMANCE AND LIMITATIONS OF INNOVATIVE IMAGE-BASED AI DRIVEN TECHNOLOGY.....	97
6.3.1	<i>Performance and Limitations of Micro Counter.</i>	97
6.3.2	<i>Performance and Limitations of Culture Counter.</i>	98
7	CONCLUSIONS.....	100
8	REFERENCES	101
	ARTIFICIAL INTELLIGENCE METHODS FOR DIAGNOSIS OF GENETIC DISORDERS	108
	ABSTRACT	109
1	INTRODUCTION.....	110

1.1	GENETIC TESTING IN HEALTHCARE.....	111
1.2	GENETIC SEQUENCING TECHNOLOGIES.....	113
1.3	THE PROCESSING OF GENETIC DATA USING NGS TECHNOLOGY.....	116
1.3.1	<i>Sequencing</i>	116
1.3.2	<i>Quality Filtering</i>	116
1.3.3	<i>Sequence alignment</i>	118
1.3.4	<i>Artifact detection</i>	120
1.4	VARIANT INTERPRETATION.....	122
1.5	DIFFERENCE IN VARIANT INTERPRETATIONS AMONG LABORATORIES	123
2	HYPOTHESIS	124
3	OBJECTIVES	124
4	METHODS	125
4.1	DESIGN AND VALIDATION OF A METHODOLOGY FOR COLORECTAL CANCER BIOMARKER ANALYSIS FOR CLINICAL APPLICATIONS	125
4.1.1	<i>Expert Panel Survey</i>	125
4.1.2	<i>DNA Microarray Analysis System Targeted Functionality</i>	126
4.1.3	<i>Study Subjects and Inclusion Criteria</i>	126
4.1.4	<i>Sample Size</i>	126
4.1.5	<i>Microarray Automatic Positioning Methodology</i>	127
4.1.6	<i>Fluorescence Image Capturing System</i>	128
4.1.7	<i>System Integration and Validation</i>	129
4.2	VALIDATION OF THE METHODOLOGY FOR BIOMARKERS OF GENETICALLY BASED DISEASES: INTERPRETATION OF GENETIC VARIANTS.....	130
4.2.1	<i>Analysis of the Past and Current Recommendations and Methods for Genetic Variant Interpretation</i>	130
4.2.2	<i>Expert Geneticists Panel Survey</i>	131
4.2.3	<i>Study Subjects and Inclusion Criteria</i>	131
4.2.4	<i>Semi-assisted variant interpretation methodology</i>	131
4.2.5	<i>New Method for Artifact Detection</i>	133
4.2.6	<i>Improved Sensitivity Method for Interpretation of Genetic Variation Based on ACMG Guidelines</i> 135	
5	RESULTS	137
5.1	COLORECTAL CANCER BIOMARKER ANALYSIS SYSTEM	137
5.1.1	<i>Expert Panel Survey</i>	137
5.2	COLORECTAL CANCER DNA MICROARRAY READER SUBSYSTEMS DESIGN AND IMPLEMENTATION.....	140
5.2.1	<i>Automatic Microscopic Stage (XY)</i>	141

5.2.2	<i>Autofocus Z stage positioning</i>	142
5.2.3	<i>Fluorescence sensor based on LED lightning</i>	143
5.2.4	<i>Light Absorbance Sensor Based on LED Lightning</i>	145
5.2.5	<i>Automatic Microarray Spot Value Reading Based on Image Analysis</i>	145
5.3	IMPROVED METHODOLOGIES FOR VARIANT INTERPRETATION	147
5.3.1	<i>Experts Panel Survey</i>	147
5.3.2	<i>BINOME. Semi-assisted Variant Interpretation Methodology</i>	147
5.3.3	<i>New Method for Artifact Classification</i>	151
5.3.4	<i>New Method for Sensitivity Increase in Variant Interpretation</i>	152
6	DISCUSSION	153
6.1	MCRC MUTATION DETECTION WITH DNA MICROARRAY	153
6.2	INTERPRETATION OF GENETIC VARIANTS FOR GENETIC DISEASES DIAGNOSIS.....	155
6.3	THE FUTURE OF VARIANT INTERPRETATION	156
6.4	FUTURE RESEARCH BASED ON THE PRESENT WORK	157
6.5	THE FUTURE OF VARIANT INTERPRETATION AND GENOMICS CLINICAL RESEARCH.....	158
7	CONCLUSIONS	160
8	REFERENCES	161
9	APPENDIX	165
9.1	PCT PATENT FILED : “PARTICLE COUNTING SYSTEM ADAPTABLE TO AN OPTICAL INSTRUMENT”	165
9.2	MOST COMMON STATISTICAL REQUIREMENTS OF LIFE SCIENCE JOURNALS	189
9.3	CELL COUNTING NEEDS AND HABITS INTERVIEW. LIST OF RESEARCHERS AND TECHNICIANS INTERVIEWED	191
9.4	CELL COUNTING NEEDS AND HABITS INTERVIEW. SUMMARY OF QUANTITATIVE AND QUALITATIVE RESULTS.....	192
9.5	TEST RUN DOCUMENT FOR AUTOMATED CELL COUNTER VALIDATION WITH ALTERNATIVE METHODS	194
9.6	CLINICAL GENETIC ANALYSIS SYSTEM USER EXPERT PANEL SURVEY	196
9.7	VARIANT ANALYSIS GENETICISTS EXPERT PANEL SURVEY	198
9.8	VARIANT ANALYSIS GENETICISTS’ SURVEY DETAILED RESULTS EXTRACT.....	201
9.9	ARTIFACT DETECTION WITH NEURAL NETWORKS PYTHON ALGORITHM.	204

Resumen

TÍTULO: DESARROLLO DE METODOLOGÍAS BASADAS EN INTELIGENCIA ARTIFICIAL PARA GENÓMICA CLÍNICA.

INTRODUCCIÓN.

En los últimos 15 años, la genética ha experimentado un progreso vertiginoso gracias a la secuenciación del ADN. En 2003 se secuenció por primera vez el ADN humano en su totalidad con el Proyecto Genoma Humano. Hoy conocemos los genes que causan unas 3.000 enfermedades, y tenemos las herramientas para diagnosticarlas. Gracias en parte a estos avances, la mortalidad por cáncer ha descendido a un ritmo del 0,94% anual en España. Las metodologías de las pruebas genéticas han evolucionado de forma espectacular en las dos últimas décadas para hacer posible estos avances tanto en la investigación como en la práctica clínica. La PCR (Reacción en Cadena de la Polimerasa), los microarrays y la secuenciación masiva de nueva generación (NGS o Next Generation sequencing) han contribuido en gran medida a la asequibilidad de las pruebas genéticas, a reducir su complejidad y a aumentar la cobertura del ADN hasta el 100% en múltiples aplicaciones. Este avance vertiginoso de la genética y la medicina en los últimos años ha venido propiciado en gran medida por la investigación básica y clínica en ciencias de la salud apoyándose en el método científico. Las ciencias médicas estudian procesos en seres humanos que a menudo presentan una gran variabilidad. El estudio de estos procesos necesita generalmente de la realización de múltiples experimentos para estimar las leyes naturales que los rigen y un posterior uso riguroso de la estadística para poder extraer conclusiones válidas sobre los mecanismos biológicos y sus efectos en los humanos. Los experimentos realizados en la actualidad son realizados *in vivo* (organismos vivos), *in vitro* (en un laboratorio a partir de células o tejidos) o *in silico* (simulación computerizada). Muchos de estos experimentos *in vitro* realizados en la comunidad científica presentan numerosas deficiencias. Algunas de estas deficiencias son inherentes a la naturaleza de los experimentos, otras están relacionadas con la metodología de diseño y medición de los resultados. En otros casos las deficiencias están relacionadas con la presión de los científicos por publicar, que choca con la necesidad de verificar los resultados o realizar experimentos complementarios antes de la publicación o la descripción incompleta de la metodología

utilizada. Las consecuencias de estas deficiencias son dramáticas para la ciencia y para la sociedad en general: Sólo el 10-30% de los experimentos publicados son reproducibles. Esto significa que se ha desperdiciado entre el 70% y el 90% de los fondos destinados a la investigación correspondiente. Y lo que es peor, estos resultados engañosos confunden a la comunidad científica, generando un mayor despilfarro de recursos en los grupos que intentan reproducir estos experimentos.

Entre las deficiencias que más destacan en los sistemas de recuento celular encontramos: 1) Falta de precisión en las mediciones realizadas. 2) Baja reproducibilidad de las mediciones realizadas. 3) Baja fiabilidad de las mediciones realizadas. 4) Mala utilización de los sistemas de medición de resultados. 5) Mal diseño del experimento. 6) Incorrecta aleatorización en la selección de las muestras. 7) Diferencias entre distintos operadores de laboratorio y efecto de lote. 8) Subjetividad en algunas mediciones (por ejemplo, en la viabilidad celular, el límite entre célula viva, apoptótica y muerta no siempre está claro). 9) Errores estadísticos y mal uso de las herramientas estadísticas.

HIPÓTESIS.

Nuestra hipótesis de partida es que las metodologías de recuento celular y de análisis genético utilizadas en la práctica sanitaria e investigaciones clínicas presentan ciertas limitaciones y existe un margen de mejora. En el sector de la investigación clínica este hecho puede tener un impacto significativo en los resultados de los experimentos, lo cual contribuye a la falta de calidad y reproducibilidad de la producción científica. En el sector sanitario las consecuencias pueden ser mucho más graves. Estas limitaciones pueden provocar que los pacientes tengan un diagnóstico o pronóstico no adecuados, y reciban un tratamiento sub-óptimo que en el peor de los casos podría provocarles una situación de morbilidad o incluso la muerte. Las metodologías antes descritas podrían ser optimizadas mediante nuevas técnicas basadas en la automatización y la inteligencia artificial para hacer que los experimentos con células y pruebas genéticas sean más fáciles, fiables y completas. En el mejor de los casos, estas mejoras podrían contribuir a hacer mejor ciencia, salvar la vida de los pacientes y evitar el desarrollo de enfermedades mortales de origen genético.

OBJETIVOS.

Nuestro objetivo inicial fue analizar ciertas deficiencias y posibles malas prácticas detectadas en los sistemas de recuento celular utilizados en los laboratorios, cuantificar su impacto para posteriormente proponer metodologías alternativas que eliminaran o minimizaran los errores y sesgos detectados. También se pretendía perseguir en esta fase la reducción de costes de operación de los sistemas, aumentar la precisión, reproducibilidad y robustez de los métodos existentes.

En una segunda parte nuestro objetivo fue analizar puntos de mejora y procesos sub-óptimos en los sistemas de análisis genéticos clínicos. El objeto de este análisis inicial era obtener la información relevante que nos permitiera proponer metodologías alternativas para minimizar o eliminar errores, aumentar la precisión, reproducibilidad y robustez del proceso reduciendo los costes, la complejidad o el mantenimiento de los sistemas existentes. En primer lugar, se previó una metodología basada en microarrays de ADN y un sistema de análisis de imágenes para el pronóstico del cáncer colorrectal (CCR). Se esperaba que la metodología redujese la complejidad y los costes de este tipo de análisis. La inteligencia artificial sería utilizada para el posicionamiento automático de la platina del microscopio y el análisis de los puntos de los microarrays. También se persiguió el objetivo de desarrollar una metodología basada en la tecnología NGS dirigida a mejorar la interpretación de las variantes genéticas gracias a la consulta automatizada de las bases de datos clínicas, y automatizar las tareas tediosas y repetitivas de los genetistas; a la postre estas mejoras deberían traducirse en una mejor comprensión de los resultados por parte de los investigadores y profesionales sanitarios para acabar revirtiendo en mejores tratamientos para los pacientes.

MÉTODOS

Recuento celular. Se diseñó una estrategia metodológica, como sigue: 1) Inicialmente, realizamos un minucioso estudio bibliográfico destinado a investigar el tipo y número de requisitos específicos relacionados con la calidad de los datos experimentales en una amplia gama de revistas científicas del área de la Biomedicina. 2) Con el fin de determinar la naturaleza y el impacto de los errores que podrían dar lugar a un mal rendimiento de los métodos experimentales básicos, identificamos el recuento de células como un procedimiento experimental popular y crítico en Biomedicina, mediante entrevistas sistemáticas a muchos técnicos de laboratorio y científicos. 3) Llevamos a

cabo experimentos de recuento celular manual y automatizado en cámaras de recuento Neubauer para determinar los principales factores que contribuyen a los errores de recuento celular. 4) Posteriormente, diseñamos dos instrumentos mejorados ("Simple Counter" y "Culture Counter") basados en la Inteligencia Artificial aplicada al análisis de imágenes microscópicas, con el objetivo de reducir los errores y aumentar la precisión, exactitud y reproducibilidad del recuento celular. 5) Los datos obtenidos con nuestros nuevos sistemas se compararon con los obtenidos con metodologías robustas de recuento celular, incluyendo la citometría de flujo y con sistemas alternativos basados en análisis de imagen. 6) Guiados por los resultados comparativos, ejecutamos iteraciones de mejora sobre las nuevas metodologías para aumentar su usabilidad y reducir la dependencia del usuario.

Detección de mutaciones en el CCR con un microarray de ADN. La primera metodología incluyó el diseño y la validación de un sistema automatizado de platina microscópica, imágenes de fluorescencia, un sensor microscópico, un sustrato innovador de microchip de PCR, la selección de las mutaciones de CCR más prevalentes en los pacientes y el estudio de las metodologías existentes utilizadas en los sectores clínico y de investigación. La entrada del sistema propuesto es una muestra tumoral FFPE (Formalin-Fixed Paraffin-Embedded) y la salida es (son) el tipo de mutación(es) genética(s) presente(s) en la muestra.

Inicialmente se elaboró un prototipo de todo el sistema para obtener resultados preliminares con muestras de cinco pacientes y en una segunda etapa se realizó una validación ampliada con 20 pacientes y métodos alternativos (Sistema Cobas de Roche Diagnostics, y secuenciación NGS con interpretación de análisis de variantes de Sophia Genetics). Las muestras de los pacientes se obtuvieron del Servicio de Oncología Médica del Hospital Universitario y Politécnico La Fe - Valencia. Se utilizó inteligencia artificial para el posicionamiento automatizado de las etapas y el análisis de las imágenes de los spots de los microarrays.

Interpretación de variantes genéticas para el diagnóstico de enfermedades .

Esta metodología pretende mejorar el análisis terciario del flujo de trabajo del análisis NGS estándar. Las muestras humanas se analizaron utilizando dispositivos de secuenciación NGS habituales (Illumina, Agilent, etc.) y el análisis de datos primario y secundario (producción de lecturas de secuencias, alineación de secuencias) se realizó

utilizando herramientas bioinformáticas comunes (BWA, GATK). Estos pasos iniciales se consideran fuera del ámbito de este proyecto. Las entradas de la metodología propuesta fueron una lista de variaciones genéticas (variantes) que oscilan entre 10 y 120.000 por muestra, junto con los datos médicos relevantes del paciente y los síntomas o la sospecha de patología proporcionados por el médico. La salida del sistema son las variantes genéticas que se consideran responsables del estado clínico del paciente. El sistema automatizado se programó en lenguaje Python y se probó en el sistema operativo Linux Ubuntu 18. Inicialmente se realizó una prueba de concepto con diez muestras de pacientes de la Unidad de Genómica del Instituto de Investigación Sanitaria Hospital La Fe (IIS La Fe) y contra los resultados de Agilent Cartagena/Alissa Software y el análisis manual.

RESULTADOS.

Hemos cuantificado con un alto grado de precisión la magnitud de los errores introducidos por los sistemas de recuento de células más populares. Según nuestros experimentos, la distribución desigual de las células en una cámara de recuento Neubauer puede introducir errores de hasta el 50%. Los sistemas automatizados de recuento de células basados en el análisis de imágenes pueden introducir errores de hasta el 30%-40% para concentraciones celulares bajas (1×10^4 células/ml) y hasta el 5-10% para concentraciones celulares más altas (1×10^6 células / ml). Las principales causas de error identificadas fueron: 1) bajo volumen de muestra analizada. 2) imperfecciones de la cámara de recuento de células. 3) malas prácticas de pipeteo. 4) agregación de células. Con la ayuda de nuestra metodología mejorada basada en el análisis de imágenes de la Inteligencia Artificial fuimos capaces de mantener el error de medición por debajo del 5% incluso para una baja concentración de células.

También hemos concebido un sistema innovador de microarray con elementos y tecnologías sustancialmente diferentes a los utilizados por soluciones alternativas, como iluminación por LED en lugar de láser y captura de imágenes por cámara óptica CCD. Se pusieron en marcha con éxito varios subsistemas para que todo el sistema funcionara (platina automatizada, sistema de iluminación, autoenfoco, posicionamiento automático del microarray y análisis de imagen automatizado para las manchas del microarray). Otros subsistemas, como las imágenes de fluorescencia basadas en la

iluminación LED, no cumplieron los requisitos mínimos de sensibilidad necesarios para un dispositivo de uso clínico.

En el ámbito de la interpretación de variantes genéticas con secuenciación NGS, hemos analizado con éxito el estado actual de los métodos y tecnologías existentes mediante entrevistas personales a 21 expertos, dónde los inconvenientes más comunes encontrados por éstos en los sistemas existentes fueron: 1) dependencia de Internet, 2) bases de datos no actualizadas, 3) sistemas no totalmente automáticos, 4) resultados y clasificación de variantes deficientes, 5) falta de integración de bases de datos, 6) necesidad de intervención humana, o 7) necesidad de utilizar diferentes bases de datos y herramientas. Cuando se les preguntó por las características más importantes de un sistema de interpretación de variantes, los encuestados destacaron las siguientes características en el siguiente orden: 1) garantizar la seguridad de los datos genéticos del paciente, 2) fiabilidad, 3) formación para el uso del sistema, 4) reproducibilidad, 5) soporte técnico telefónico y por correo electrónico, 6) especificidad, 7) sensibilidad y, 8) conexión con sistemas de información propios. También se analizaron las tasas de rendimiento de diagnóstico más comunes de los laboratorios clínicos, el grado de correlación entre los resultados de los distintos laboratorios y hemos identificado varios puntos en los que se podrían mejorar los sistemas existentes, como la detección de artefactos, la flexibilidad del análisis, la simplicidad de la operación y de los informes de resultados, y la sensibilidad. Con nuestra metodología propuesta pudimos aumentar el rendimiento del diagnóstico en un 5-10% (a expensas de la disminución de la especificidad), automatizar más del 80% de las tareas repetitivas realizadas por los genetistas, como consulta de la base de datos, filtrado de variantes de alta prevalencia, detección de artefactos, etc. Con este sistema se espera que el tiempo total dedicado por el genetista se reduzca entre un 50% y un 80%, dependiendo de la muestra.

DISCUSIÓN.

Hemos identificado con éxito los principales tipos de errores introducidos en los ensayos con células, los hemos cuantificado y hemos propuesto una metodología mejorada que puede utilizarse en la mayoría de los laboratorios científicos que trabajan con células. También hemos demostrado que esta metodología puede implementarse mediante sistemas automatizados que contribuyen aún más a la calidad y

reproducibilidad de los resultados y que pueden utilizarse tanto en entornos de investigación como clínicos.

En la metodología de detección de mutaciones de cáncer colorrectal con microarray de ADN, el rendimiento global del sistema en términos de sensibilidad y especificidad no se consideró aceptable para ser utilizado en un entorno clínico, y esta línea de investigación se interrumpió. En caso de querer seguir esta línea de investigación en el futuro para obtener un sistema con características suficientes para uso clínico, recomendamos aumentar la potencia del sistema de iluminación LED y concentrar el haz de luz mediante lentes para ganar potencia lumínica en el área de análisis.

En el área de interpretación de variantes, comprobamos que únicamente existe una concordancia en la clasificación de variantes del 34% entre laboratorios, lo cual apoya nuestra hipótesis inicial de que existe un margen considerable de mejora en cuanto a fiabilidad y reproducibilidad. Hemos propuesto una metodología innovadora y mejorada para la interpretación de variantes en el análisis terciario de NGS, adecuada para el diagnóstico clínico y el cribado genético en medicina preventiva.

Con nuestras metodologías propuestas integradas en un sistema llamado BINOME automatizamos más del 80% de las tareas repetitivas realizadas por los genetistas para algunas aplicaciones clínicas específicas de análisis genético. Además, predecimos la probabilidad de que una variante sea un artefacto, y definimos un método que aumentó la cantidad de variantes de alto riesgo reportadas en un 7,7%. Estimamos que este nuevo método tiene el potencial de aumentar el rendimiento del diagnóstico equivalente en un 5-15% al aumentar la sensibilidad en comparación con la aplicación estricta de las directrices actuales sugeridas por el Colegio Americano de Genética Clínica (ACMG). En una prueba de concepto con 10 muestras de pacientes con una media de 6010 variantes (SD 1535) tras aplicar el conjunto de procesos automatizados realizados por el sistema BINOME, el resultado fue un subconjunto de las variantes de entrada con promedio 4 variantes (SD 2,62). Por término medio, el sistema automático BINOME filtró el 99,93% de las variantes introducidas, dejando un 0,07% de variantes que debían ser revisadas manualmente por los genetistas. En la cohorte de la muestra seleccionada, el sistema mostró una sensibilidad del 100% junto con una especificidad del 99,95%.

También se ha desarrollado una metodología para clasificación automática de artefactos con sistema de Inteligencia Artificial formado por una red neuronal de cuatro capas y

una dimensión de entrada de ocho. La primera capa oculta estaba compuesta por funciones tangentes hiperbólicas con una dimensión de 16. La segunda capa oculta estaba compuesta por funciones de activación lineal rectificadas (ReLU) con una dimensión de 8. La capa de salida estaba compuesta por una función sigmoidea que emite “1” si la red neuronal considera que la muestra es un artefacto o “0” en caso contrario. El método propuesto mostró concordancia con la clasificación humana en el 97,7% de los casos en una muestra de 45 variantes extraídas al azar de la muestra original de 158 variantes. La precisión alcanzó el 94,11% con una recuperación del 100%.

En este sentido, hemos logrado mejoras significativas que podrían aplicarse los flujos de trabajo de numerosos laboratorios genéticos para la interpretación de variantes. Para la metodología de clasificación de artefactos, sería deseable entrenar el sistema con un mayor número de muestras para aumentar la precisión del sistema de IA. Sugerimos que este método se utilice sólo para la priorización de variantes, evitando utilizarlo para filtrar variantes en una línea de interpretación de variantes clínicas estándar, ya que al hacerlo podríamos estar reduciendo la sensibilidad del sistema. Con el enfoque propuesto para la priorización de variantes la capacidad general de detección del sistema no se vería afectada y se conseguiría un ahorro de tiempo sustancial.

Asimismo, sugerimos aplicar con prudencia el método propuesto para el aumento de la sensibilidad. Este método se produce a expensas de una mayor tasa de falsos negativos. Las directrices del ACMG recomiendan no informar de cualquier variable de significado incierto (VUS) para el diagnóstico clínico. Los médicos podrían interpretar erróneamente la VUS notificada como patógena (cuando no lo es), y podrían tomar una decisión médica equivocada prescribiendo un tratamiento no adecuado al paciente.

Nuestra recomendación sería utilizar este método de sensibilidad aumentada sólo en las siguientes situaciones específicas:

- 1) Cuando el genetista y el médico con formación genética coinciden en la probabilidad de patogenicidad de la VUS de alto riesgo.
- 2) Cuando el tratamiento que se vaya a administrar al paciente no tenga consecuencias negativas aunque la variante notificada como patógena resulte ser un falso positivo.
- 3) Para aplicaciones de medicina preventiva, en las que no suele haber prescripciones perjudiciales para los consultantes.

Dado el estado actual de desarrollo de la tecnología NGS, la caída histórica de los precios de la secuenciación genética y la constante evolución de las bases de datos clínicas y de las metodologías de análisis recomendadas, consideramos que la metodología propuesta podría optimizar varios aspectos de los sistemas de análisis genéticos existentes mejorando el rendimiento diagnóstico de los laboratorios, la reproducibilidad y reducir el número de informes sub-óptimos generados por los genetistas.

Sugerimos continuar esta investigación a través de los siguientes pasos: 1) aumentar el número de muestras analizadas, hasta 300 muestras, y compararlas con sistemas alternativos. 2) centrarse en algunas patologías o condiciones clínicas específicas, como el cáncer, las enfermedades cardiovasculares o las enfermedades raras, que pueden requerir ajustes específicos de filtrado. 3) desarrollar una interfaz gráfica que facilite la experiencia del usuario. De acuerdo con la evolución de los sistemas de IA integrados en el método de interpretación de variantes, también consideramos seguir explorando futuras mejoras con Redes Neuronales Gráficas (GNNs), que intuimos pueden encajar en el proceso de interpretación de variantes que se estudia en este documento. En futuras líneas de investigación, el trabajo actual realizado con la integración de bases de datos genéticos y predictores *in silico* podría adaptarse a sistemas que integren miles de muestras de pacientes junto con su correspondiente historial de salud para fines de investigación clínica. En caso de que las muestras disponibles para el entrenamiento y las pruebas aumenten a decenas/cientos de miles, podríamos adaptar nuestros sistemas al aprendizaje profundo a las redes neuronales convolucionales (CNN) y a las redes neuronales recurrentes (RNN), que suelen ser consideradas adecuadas para inferir conocimiento a partir de grandes y complejas cantidades de datos genómicos.

Prevedemos una prometedora línea de investigación para seguir optimizando la interpretación de variantes genéticas con algoritmos de IA, especialmente en la clasificación y priorización como extensión de la presente investigación y la adaptación de las pautas a aplicaciones clínicas específicas como el diagnóstico y pronóstico de enfermedades cardíacas, cáncer y enfermedades raras. Consideramos que el reto de la interpretación de variantes genéticas se adapta especialmente bien a las capacidades de los modernos sistemas de IA. La interpretación de variantes requiere expertos bien formados, automatización, gran cantidad de análisis de datos y médicos con conocimientos de genética para que los pacientes reciban el mejor tratamiento posible.

Sin embargo, definir unas reglas claras que puedan ser seguidas inequívocamente por humanos o máquinas es un proceso engorroso y complejo para esta aplicación. Todas estas son características donde los sistemas de IA tienen el potencial de mejorar y optimizar las capacidades humanas. Por lo que respecta a la investigación clínica, estoy completamente convencido de que con la cantidad adecuada de potencia informática, capacidad de almacenamiento y un conjunto suficientemente grande de muestras humanas (genotipo-fenotipo) ningún reto en investigación genómica se resistirá a la inteligencia artificial en un futuro próximo.

CONCLUSIONES

1. Entre las 727 revistas biomédicas analizadas, las revistas médicas tienen un número significativamente mayor de requisitos estadísticos que las revistas no médicas. Dentro de las revistas no médicas, las situadas en los cuartiles Q1 y Q2 tienen un mayor número de requerimientos estadísticos que las situadas en los cuartiles Q3 y Q4.
2. A raíz de una encuesta realizada entre los técnicos de laboratorio, los métodos de recuento de células más populares resultaron ser el recuento manual de células en dispositivos de tipo Neubauer, la citometría de flujo y el recuento automatizado de células basado en imágenes, en ese orden.
3. El recuento manual de suspensiones celulares en cámaras de Neubauer o en placas de Petri puede dar lugar a errores importantes, debido a especificaciones de volumen erróneas, a un número insuficiente de células puntuadas por campo o a una distribución celular heterogénea por campo del microscopio.
4. Hemos diseñado y construido dos innovadores sistemas de recuento celular basados en algoritmos de Inteligencia Artificial para el análisis automatizado de imágenes microscópicas de células en suspensión (el Micro Counter) o en cultivos monocapa (el Culture Counter).
5. El sistema Micro Counter mejora la precisión y la reproducibilidad con respecto a otros procedimientos basados en imágenes, al aumentar el número de campos del microscopio y de células analizadas. Sin embargo, tiene menos reproducibilidad y precisión que los citómetros de flujo.

6. El sistema Culture Counter permite realizar mediciones precisas y reproducibles de la concentración de células directamente en placas de Petri y frascos de cultivo, sin necesidad de interrumpir el proceso de cultivo.
7. Tras dos encuestas realizadas a 21 expertos en genética, las características más solicitadas para el sistema de análisis genético que se utilizará en las instalaciones clínicas fueron: 1) sensibilidad, 2) especificidad, 3) tiempo de análisis, 4) cobertura, 5) reproducibilidad, 6) capacidad para trabajar con pequeñas cantidades de muestra.
8. En una encuesta realizada a 9 oncólogos, los genes que se utilizan actualmente como biomarcadores en la práctica clínica con un fuerte consenso fueron KRAS, NRAS y BRAF. Otros 25 genes son utilizados actualmente por diferentes oncólogos entrevistados de forma independiente sin un claro consenso entre ellos.
9. Hemos diseñado y construido desde cero dos innovadores sistemas de análisis genético basados en IA para el diagnóstico y pronóstico clínico. Un sistema dirigido a la detección de las mutaciones más prevalentes del cáncer colorrectal utilizando un microarray de PCR multiplex (ONCOMARKER), y el segundo diseñado para realizar el análisis terciario de un flujo de trabajo de análisis genético NGS estándar (BINOME)
10. El lector de microarrays ONCOMARKER incorpora una platina XY automática para el posicionamiento de los microarrays, un sistema de iluminación y un software basado en IA para la colocación automática de los microarrays y el análisis de las muestras.
11. El sistema BINOME es capaz de ahorrar entre el 50% y el 80% del tiempo de análisis práctico del genetista al automatizar el acceso a la base de datos y los predictores in silico. Incorpora un algoritmo de detección de artefactos basado en IA que fue probado con un 94,11% de precisión y un 100% de recuperación. Incluye un filtro de sensibilidad mejorado para detectar VUS de alto riesgo que tiene el potencial de aumentar el rendimiento del diagnóstico en un 5%-15%. Este nuevo método se considera adecuado para casos de diagnóstico específicos y aplicaciones de medicina preventiva.

PART 1

Automation for Cell Assays in Clinical Research

"The cell concept is the axis around which the whole of the modern science of life evolves."

Paul Ehrlich

"Science is the father of knowledge, but opinion breeds ignorance. "

Hippocrates

Abstract

(Part 1)

Many of the *in vitro* experiments performed in the scientific community present numerous shortcomings. Some of them are inherent in the nature of the experiments; others are related to the design methodology and measurement of results, or to the pressure on scientists to publish. The consequences of these deficiencies are dramatic for science and for society in general, wasting valuable time and resources.

We started from the hypothesis that a significant part of the deficiencies are due to methodological causes and have significant impact on the quality of the experiments results. Our objectives were to analyze certain deficiencies and potential malpractices detected in cell-based assays and measurement systems and quantify their impact, and to propose alternative methodologies to minimize those flaws.

We analyzed the reporting requirements of 727 scientific journals in medical and life sciences, identified the most popular cell counting methods used in laboratories and conducted experimental comparative studies of 5 different cell counting methods. We performed an in-depth analysis of the Neubauer cell counting chamber with experiments specifically designed to detect its limitations. Afterwards we designed two improved methodologies based on Artificial Intelligence applied to image analysis to reduce cell counting errors increasing precision and reproducibility.

According to our results, uneven distribution of cells on Neubauer counting chambers can introduce errors as high as 50% in regular laboratory setups, while image analysis automated cell counting systems can introduce errors ranging from 30% to 40% with low cell concentrations (1×10^4 cells/ml). This error is reduced to 5-10% with higher cell concentration (1×10^6 cells / ml). The main causes of error that we identified were: 1) low volume of sample analyzed 2) imperfections of the cell counting chamber 3) pipetting malpractices 4) aggregation of cells. With the help of our improved methodology based on Artificial Intelligence image analysis we were able to maintain the measurement error below the 5% even for low cell concentration. The proposed methodology could contribute to improve the quality and reproducibility of cell based experiments, and it is suitable for both research and clinical environments.

1 Introduction

1.1 The Scientific Method and its Contribution to Medicine.

Homo sapiens appeared in Africa 200.000 years ago (Henry, 2019). Genetic evolution provided *Homo sapiens* with the ability to develop tools, culture and language. It is believed that very early a medicinal knowledge base developed and passed between generations, through emulation of the behavior of fauna. Even Neanderthals may have engaged in medical practices (Spitkin et al., 2018)

The first known predecessors of medical doctors were the shamans that appeared about 30.000 years ago in Europe and Middle East. Shamans were related to the spiritual world, religion, divination and healing in indigenous and tribal societies. Shamans were attributed the ability to communicate with the spirit world and treat sickness caused by malevolent spirits.

Hippocrates of Kos (460 B.C) is credited as the first person to believe that diseases have a natural origin, and did not happen because of superstition or the gods. (Jones, 1868; Adams, 1891). Hippocrates is traditionally considered the “Father of Medicine” in recognition of its contribution to the field such as the concept of prognosis, clinical observation and the categorization of diseases. Since then, the scientific approach has always been an essential part of Western Medicine.

According to W.F Bynum (2008), Medicine and its relationship with the scientific method has evolved across history, through five main paradigms that are still valid to understand modern Medicine:

a) Bedside Medicine, when doctors of ancient times visited patients at home. Bedside medicine has its modern counterpart in primary care.

b) Library Medicine, which is associated with the scholastic approach of the Middle Age when the knowledge was stored in libraries, often related to the clergy. The problem of information retrieval still surfaces in computer and internet age.

c) Hospital Medicine, which origins on 19th century French Medicine. In this paradigm, diagnostic and therapeutic functions are provided by a Central Hospital. Modern hospitals have become a hub of resources, care and teaching.

d) Social Medicine, that is about social and individual prevention. Currently, social medicine is strongly linked to the concepts of lifestyle and its impact on health.

e) Laboratory or Experimental Medicine has its current representation in research establishments, critical for the creation of medical knowledge: universities, research centers, and health organizations that set the standards for both medical science and scientific medicine.

Despite the fact that experiments in Medicine have been performed since remote times, as attested by Galen and other ancient authors, it was Claude Bernard in mid-1800 who established the principles of experimentation in the Life Sciences, advancing beyond the vitalism and indeterminism of earlier physiologists to become one of the founders of Experimental Medicine (Barona, 1989).

1.2 The Scientific Method in Laboratory Medicine and Biomedicine.

The scientific method is an empirical process for acquiring knowledge. It involves careful observation and applying skepticism about what is observed. The performance of the Scientific Methods involves:

a) Observation

b) Formulating hypothesis based on the observation.

c) Experimental and measurement-based testing of deductions drawn from the hypothesis.

d) Refinement (or elimination) of the hypothesis based on the experimental finding.

Most scientific experiments in life sciences and biomedicine are performed according to three levels of complexity:

a) *In vivo Experiments*: These experiments are carried out in a complete living organism, as opposed to a partial living being or dead organism. Research using living animals or clinical trials are the most common in-vivo experiments performed. Its main advantage is that the whole set of reaction of the living being can be observed. Although sometimes the conclusions drawn could be mistaken, or short-term benefits could be prioritized without considering long term damage for patients.

b) *In vitro*. These experiments are performed in a controlled environment outside of a living being. The main weakness of these kinds of experiments is that the mechanisms of

living beings are not always faithfully reproduced. Living cells behave differently in an in-vitro artificial setup than in their natural environment.

c) *In silico*. These experiments are executed using a computer or through computerized simulation. This type of experiment is relatively recent. The term was first used in 1989. (Trisilowati and Mallet, 2012)

1.3 Publishing experimental results as an essential aspect of Biomedicine

Experiments are the main pillar of the scientific method. Lack of rigor and convenience designing, executing and analyzing these experiments will render the scientific method useless for building valuable scientific knowledge. This is of critical importance in Clinical and Biomedical sciences, where the results provided by the scientific community can potentially impact the lives of millions individuals (patients or not) worldwide.

But the reality is that many of the *in vitro* experiments performed in the scientific community present numerous shortcomings. Some of these shortcomings are inherent to the nature of the experiments, others are related to the methodology of design and measurement of results, and others are related to the pressure on scientists to publish, which clashes with the need to verify results or perform complementary experiments before publication. (Sarewitz, 2016; Begley and Ellis, 2012)

The consequences of these deficiencies are dramatic for science and for society in general: Only 10-30% of published experiments are reproducible (Pritsker, 2012). This means that 70%-90% of the funds allocated for the corresponding research have been wasted. Worse, these "irreproducible" or "misleading" (but published as true) results confuse the scientific community, generating a further waste of resources in groups trying to reproduce these experiments. Additionally, these irreproducible results are cited by other publications as true, making it much more complicated for the scientific community to determine which experiments are reproducible and which are not.

As it could be expected, the concerns about the validity of published scientific findings, has fostered ongoing discussions on proper use and interpretation of statistical methods in published biomedical research (Fernandes-Taylor et al., 2011; Hardwicke and Goodman, 2020). Most proposals focus on improving study design, but little attention has addressed specifically on how journals themselves can improve their performance,

although some systematic surveys on this issue are available (Hardwicke and Goodman, 2020).

Among the shortcomings associated with experimental methodologies or measurement systems in cell-based research, the following stand out (Vaux D, 2012):

- Lack of precision in the measurements performed.
- Low reproducibility of the measurements performed.
- Low reliability of the measurements performed.
- Poor use of the results measurement systems.
- Poor design of the experiment.
- Incorrect randomization in the selection of samples.
- Differences between different laboratory operators and batch effect.
- Subjectivity in some measurements (e.g. in cell viability, the boundary between live, apoptotic and dead cell is not always clear).
- Statistical errors and misuse of statistical tools (Baker M, 2016).

Scientific claims in biomedical publications should base on statistical data analysis. However, misunderstanding and misuse of statistical methods is too frequent in biomedical research (Hardwicke and Goodman, 2020). Statistical practices used in top journals, influence the statistical methods used by prospective contributors to those journals and the general scientific community. Accordingly, some biomedical leading journals such as the Lancet, the British Medical Journal, Annals of Internal Medicine, and the Journal of American Medical Association have adopted statistical review since at least the 1970s. Leading all employ statistical review research (Hardwicke and Goodman, 2020).

1.4 Cell Counting as a Basic Experimental Procedure in Biomedicine.

The use of cells in Biomedicine, for drug discovery as well as therapeutic and diagnostic applications has been increasing exponentially. Because of this, cell counting is a critical and routine part of many cell culture laboratory workflows. The practice usually involves directly counting individual cells in a small volume in order to get an estimate of how many cells are in a larger volume. This can be useful for culture maintenance,

setting up assays, assessing the health or viability of cells in a culture, and clinical or diagnostic applications.

The hemocytometer, consisting of a thick glass slide with an etched grid and a sample chamber, has traditionally been used to count cells in a defined volume under the microscope, to get a measure of cell concentration. However, this manual method of cell counting takes time, effort, and is prone to error and contamination. Because of such limitations, laboratories switch to more automated methods for cell counting, based on techniques such as impedance, flow cytometry, or bright field or fluorescence microscopy. These techniques differ in their speed, sample handling, and customization.

1.5 Manual Cell Counting:

The conventional method of cell counting is done manually at a benchtop microscope using a special type of chamber engraved with a grid known as a hemocytometer or counting chamber. The researcher follows a standardized procedure to count individual cells and can use this to calculate the cell concentration in the original culture or any suitable starting sample.

Additionally, the number of live and dead cells can be determined by adding a membrane exclusion dye such as trypan blue. Intact cells do not take up trypan blue, while unhealthy cells with compromised cell membranes will uptake the dye and can be identified under the microscope. This can be useful in getting more accurate counts and assessing the health of a culture and is also a technique employed in automated cell counting systems.

1.5.1 Automated Cell Counting:

Automated systems generally rely on either an impedance-based approach or an imaging-based approach. Impedance-based systems measure electrical resistance to determine the number of cells. Imaging-based systems use a microscope and camera that will capture an image and then use an algorithm to count the cells. Cell counters that use the image-based approach can further be divided into either fluorescent or brightfield imaging. Some cell counters offer more than counting and they may also provide information about cell viability, diameter, live/dead counts, images of cultures, and fluorescence intensity.

Automated cell counting works across most cell types but usually requires that cells are suspended. However, non-mammalian type cells, like bacteria or yeast cell counting, may become challenging to automated systems. Clumped cells also cause problems for both image-based and impedance-based systems, thus requiring interfacing with software algorithm that performs accurate cell counting of clumped cells.

Automated counting systems vary in how samples are inserted. Some use special disposable slides while others can take many samples at once in tubes or microplates. Automated sampling reduces subjective and time-consuming human judgment.

1.5.2 Manual versus Automated Cell Counting:

The main considerations between manual and automated cell counting are cost, labor, and accuracy. Regarding the cost, manual cell counting only requires a one-time purchase of a few items. On the other hand, automated counters will run on the order of a few thousand to a few tens of thousands of dollars for the system alone and will also require consumable purchases and maintenance costs.

The primary advantage in accuracy for automated systems is that it removes the variability from user to user or lab to lab. This reproducibility can be very important for assays or protocols that rely on having a certain number of cells as input.

A secondary accuracy advantage is that automated systems usually use a larger field of view than a hemocytometer. When counting lower numbers of cells or lower concentrations of cells, larger field of view has a benefit to a conventional manual method.

1.4.4 Advantages and Limitations of Cell-Counting Systems:

It is also desirable to improve additional dimensions of the existing methodologies such as:

- a) Eliminate need for reagents or consumable material, reducing the cost of operation and maintenance.
- b) Improve the cell counting range. So that the method could also be used to analyze a wider amount of cells and microscopic particles.
- c) Increased accuracy, reproducibility and robustness of the method.
- d) Improved cell counting quick visual validation of the results.

As basic, translational and event therapeutic cell studies gain relevance, cell counting is becoming stringent and will be required to strictly adhere to Good Manufacturing Practices (GMP) standards. The International Organization for Standardization (ISO) has released two sets of guidance that relate to cell counting. ISO 20391-1:2018 defines terms related to cell counting for biotechnology. It describes counting of cells in suspension (generally cell concentration) and cells adhered to a substrate (generally area density of cells). It provides key considerations for general counting methods (including total and differential counting, and direct and indirect counting) as well as for method selection, measurement process, and data analysis and reporting (ISO, 2022a). ISO 20391-1:2018 is applicable to the counting of all cell types, mammalian and non-mammalian (e.g. bacteria, yeast) cells. ISO 20391-2:2019 refers to the Experimental design and statistical analysis to quantify counting method performance (ISO, 2022b). This document provides a method for evaluating aspects of the quality of a cell counting measurement process for a specific cell preparation through a set of quality indicators derived from a dilution series experimental design and statistical analysis. The quality indicators are based on repeatability of the measurement and the degree to which the results conform to an ideal proportional response to dilution. ISO 20391-2:2019 is most suitable during cell counting method development, optimization, validation, evaluation and/or verification of cell counting measurement processes.

Besides cell counting, there are a number of other parameters such as cell purity, cell morphology, cell size, cell health, apoptosis, and population analysis that can now be measured using some of the newer instruments. All the images and the reports from these measurements can be analyzed before the experiment is done to eliminate problems downstream.

1.6 The role of Automation and Artificial Intelligence for improving basic experimental procedures in Medicine.

In 1866 Gregor Mendel published the article “Experiments in Plant Hybridization” which is considered the origin of Genetics. But it was in 2003 when the first Human Genome was sequenced opening a world of possibilities for science and medicine. Whereas in

2003 the cost of sequencing the first human genome was more than €280 million, by 2015 the cost had dropped below €1000 making genomic sequencing available to researches worldwide. (Figure 1)

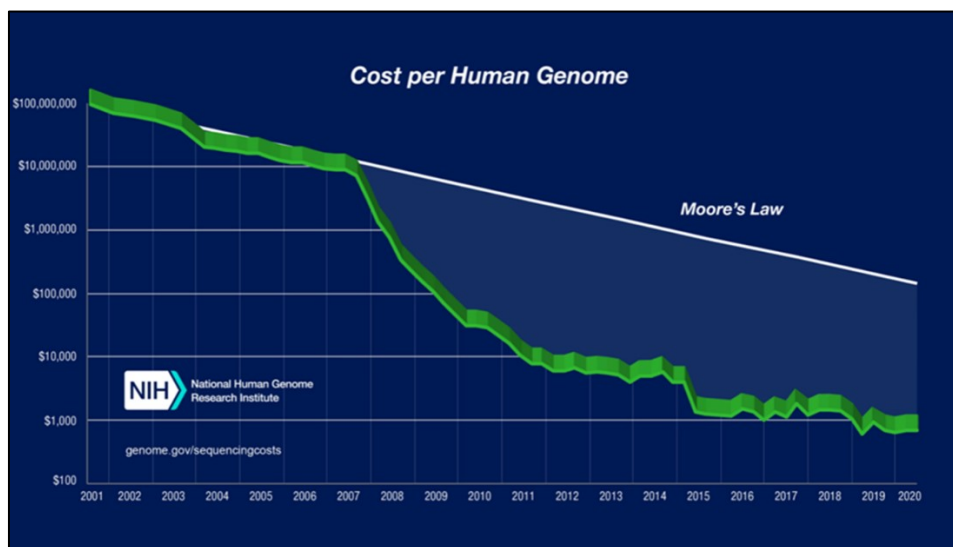


Figure 1: Evolution of sequencing cost of a human genome. (Source : National Human Genome Research Institute)

In parallel with these advances, computer power has been doubling since 1970 following Moore's Law (Figure 2) allowing breakthrough discoveries and advances in artificial intelligence.

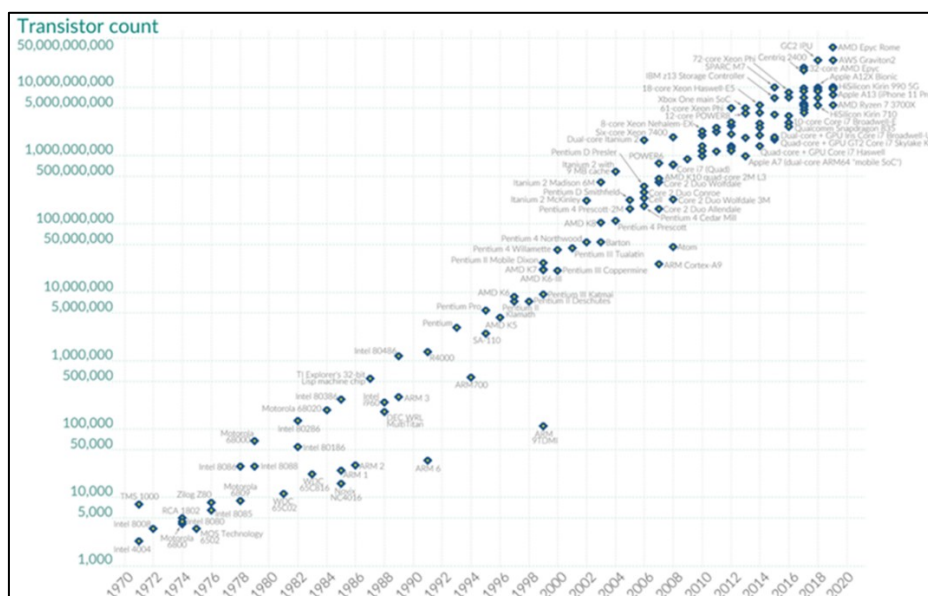


Figure 2: Evolution of computer power from 1970 to 2020. (Source: OurWorldInData.org)

In recent times Artificial Intelligence systems have beaten human intelligence in chess (1996) and Go (2015). Tesla motors pioneered automatic car driving using AI systems, and the AlphaFold AI system provided a solution to a 50-year-old problem by being able to accurately estimate the 3D structure of most proteins.

AI systems have been a major contributor to important discoveries in many areas in life sciences in the last years, and the contribution of Artificial Intelligence to medical practices, while initially limited to some specific areas, such as detection of atrial fibrillation, epilepsy seizures and hypoglycaemia, or diagnosis based on automatic examination of medical imaging (Briganti and Le Moine, 2019) is now revolutionizing medical sciences, thanks to new advances in Deep Learning.

Deep Learning is a class of machine learning algorithms that use multiple layers to extract progressively higher-level features from the raw input. For example, in image processing, lower layers may identify edges, while higher layers may identify the concepts relevant to a human such as digits or letters or faces. Most modern Deep Learning models are based on artificial neural networks, specifically convolutional neural networks, although they can also include propositional formulas or latent variables organized layer-wise in deep generative models.

Deep Learning has been shown to produce competitive results in Biomedicine (Rajpurkar et al., 2022) for medical applications such as cancer cell classification (Stiff, 2022; Capobianco, 2022), lesion detection (Joseph, 2022), organ segmentation and image enhancement (Litjens et al., 2017)

Drug discovery and Toxicology research has repeatedly applied Deep Learning to predict biomolecular targets, off-targets, and toxic effects of environmental chemical in nutrients, household products and drugs (Tian et al., 2021; Kolluri, 2022; Munawar et al., 2022). Generative neural networks were used to produce molecules that were validated experimentally all the way into mice (Zhavoronkov, 2019), while gGraph neural networks have allowed to predict relevant properties of molecules in large data sets (Xia, 2020).

1.7 Is there a Role for Artificial intelligence for Improving Journal Reviewing of Methodological Aspects?

In a recent paper, Hardwicke and Goodman discussed on what role could Artificial Intelligence play in improving review of methodological aspects of biomedical publications, given that human expertise is always scarcer (Hardwicke and Goodman, 2020). According to these authors review, there have been limited attempts to develop programs that examine statistical aspects of a paper, mostly checking whether the reported degrees of freedom and F or chi-square statistic is consistent with a reported p-value, which is mainly of value in the psychological literature, which has a structured way to present such information rarely used in biomedical publications. Even some publishers would be experimenting with software that evaluates the use of reporting standards, but other functionality is unclear (Hardwicke and Goodman, 2020).

Given that methodological reviewers ideally provide an integrated assessment of the research question, design, conduct, analysis, reporting and conclusions, Hardwicke and Goodman conclude, rather pessimistically that it is highly unlikely that Artificial Intelligence applications will be able to be relevant to help methodological reviewing in short- or medium term.

2 Scientific Hypothesis

We started from the hypothesis that a significant part of the deficiencies in the performance of genetic experiments with cells are due to methodological causes and that these deficiencies have significant impact of the experiments results. This fact contributes to the lack of quality of the experiments results and their reproducibility and also impacts on the quality and relevance of the scientific production and scientific articles. Minimizing them will contribute to improving the quality of published scientific experiments and the reproducibility of experiments.

3 Objectives

Our initial objective was to analyze certain deficiencies and potential malpractices detected in cell-based assays and measurement systems and quantify their impact. Our secondary objective was to propose alternative methodologies that minimize or eliminate the errors or biases detected while reducing the operations costs whenever possible (equipment, reagents, etc.). The targeted methodologies would increase accuracy, reproducibility and robustness of the existing methods.

1. Determine what are the experiments reporting requirements of the top journals in life sciences and medicine, and determine if there is a certain correlation between the amount of statistical requirements and the journal prestige.
2. Analyze certain deficiencies and potential malpractices detected in cell-based assays and clinical measurement systems and quantify their impact when possible.
3. Propose alternative methodologies that minimize or eliminate the errors or biases detected.
4. Validate a high-throughput automated system (High-throughput) with cells in suspension and adherent cells. The aim is to validate a cell analysis system that incorporates some of these methodologies.

4 Methods

According to the scientific hypothesis and the objectives of this Thesis, we have designed a methodological strategy, as follows:

- 1) Initially, we performed a thorough bibliographical study aimed to investigate the type and number of specific requirements related to the quality of experimental data in a wide range of scientific journals of the Biomedicine area.
- 2) In order to determine the nature and impact of errors that might result in poor performance of basic experimental methods, we identified cell counting as both a popular and critical experimental procedure in Biomedicine, by means of systematic interviews to many laboratory technicians and scientists.
- 3) We conducted manual and automated cell counting experiments on Neubauer counting chambers to determine the main contributors to cell counting errors.
- 4) Afterwards, we designed two improved instruments (“Simple Counter” and “Culture Counter”) based on Artificial Intelligence applied to microscopic image analysis, aimed to reduce errors and to increase precision, accuracy and reproducibility of cell counting.
- 5) The data obtained with our new systems were compared with those obtained with robust cell-counting methodologies, including flow- and image cytometry.
- 6) Guided by the comparative results, we executed improvement iterations over the new methodologies in order to increase their usability and reduce the dependency from the user.

4.1 Identifying and Quantifying Journal Requirements for Reporting Experimental Results

Most scientific work in the academic community is expected to be published in scientific journals. The majority of researchers would like to be published in a journal that is most valued and cited by others for future research. The better the scientific journal reputation the higher the number of competing manuscripts that in turn will translate into better research published by the journal.

Our initial intuitive perception was that the best research journals had a higher degree of stringency with statistical requirements and reporting as a part of their process to select their best manuscripts. We quantify this perception by performing a deep analysis of the scientific journal reporting requirements from a mathematical and statistical perspective.

727 journals reporting requirements were investigated with a thorough analysis of their guides to authors that are made public online. (See

[Table 1](#)). All categories related to Life Sciences and Medicine were selected using the Journal Citation Reports classification (JCR) from Clarivate Analytics, currently part of the Web of Science (WoS) (Clarivate Analytics, n.d.)





The following tasks were undertaken in order to perform a quantitative analysis of the journal requirements

1. Find the journal website.
2. Save a) Guide for authors b) Guide for reviewers c) FAQs
3. Analyze the 3 previous documents looking for mathematical and statistical reporting requirements.
4. Gather all the requirements in a single list
5. Sort all requirements specifications in different reporting areas (samples, statistical tests, etc.)
6. Extract the percentage of Journals with mathematical and statistical requirements based on its Impact Factor (IF) percentiles : Q1, Q2, Q3, Q4
7. Statistical analysis of the gathered data. Classification of Journal requirements in groups, depending on the number of requirements (See [Table 2](#)).

Table 1: List of journals' categories and number of journals analyzed

Category	Number of journals analyzed
Multidisciplinary Science	64
Cell Biology	190
Methods	78
Cell Tissue	21
Developmental Biology	41
Medicine, General and Internal	155
Medicine, Research	128
Pharma	50
TOTAL	727

Table 2: Classification of journals in subsets depending on their authors' requirements.

Colour Code		Description	Example
	No requirements whatsoever	Anything about mathematics	
	Other maths requirements	Maths requirements not related with statistics	Equations formats, derivations, computer algorithms
	Non-specifics statistical requirements	If there are statistical requirements but in general	This (Results) should include the findings of the study including, if appropriate, results of statistical analysis which must be included either in the text or as tables and figures.
	Specific statistical requirements	If there is how to report all the statistic data	For continuous variables, distributions should be described using graphical displays such as scatterplots, boxplots, or histograms or by reporting measures of central tendency (e.g. mean or median) and dispersion (e.g. SD, interquartile range)

4.2 Identifying Popular and Critical Basic Cell Assay Methods

Before performing any measurement of methodology optimization, we identified and enumerated the most popular cell assay methods and techniques that are utilized in current life sciences and clinical research environments.

Thereafter, we interviewed a set of 17 laboratory technicians for a deeper understanding of their specific needs and their level of comprehension and concerns regarding cell assay methodologies employed. The technicians were active in the laboratories where most of our experimental work was performed or where practical demonstrations of industrial products were conducted, including the Laboratory of Cytomics of the University of Valencia, the Service of Cytomics of the Prince Felipe Research Center, the Service of Cytometry of the Central Research Unit UVEG-INCLIVA, AINIA, Cavanilles Institute and other centers. (See [Appendix 9.3](#) for a detailed list of interviews and [Appendix 9.4](#) for a summary of the survey results)

4.3 Detection of Potential Error Sources when Using Neubauer Counting Chambers.

In order to determine which variables affected the precision and accuracy of cell counting on Neubauer chamber, we performed more than 150 different experiments to evaluate different brands of Neubauer chambers, different cellular materials, chamber loading techniques, and cell-count calculation approaches.

As a general approach, in our different experiments, each independent variable was modified individually, while keeping the remaining independent variables constant. In this way, the effect of the change produced by each independent variable could be easily associated with the outcome on the dependent variable. In order to maximize the potential impact of result dissemination, we prioritized endpoints that could be verified objectively. Thus, the reproducibility of the effect was estimated by visual inspection and verified by automatic cell counting, with at least three technical replicas.

Precision and accuracy of cell counting in the different conditions have been assessed mostly by scoring the homogeneity of cell counting measurements on different chamber areas or microscopic fields. Whenever two different areas showed a significantly different cell concentration it was assumed that the distribution was uneven, and the independent variable that was modified on that experiment should be considered as responsible for the heterogeneity of the cell distribution.

The different independent variables considered for this series of experiments were as follows:

- a) Neubauer cell-counting chambers: Unbranded chambers (white label) obtained from Chinese manufacturer and Neubauer improved (dark lines) chambers obtained from Marienfeld (Germany).
- b) Cell- and particle suspensions: Pressed yeast (Lesaffre Group Iberica, Spain), Jurkat cells (European Cell Culture Collection, ECCAC), N13 cells (Laboratory of Cytoomics, University of Valencia), AccuCheck Counting Beads (ThermoFisher, USA)
- c) Dilution: Tap water, distilled water.
- d) Pipettes: Automatic pipettes (Eppendorf, Germany), Glass Pasteur pipettes.
- e) Chamber loading technique: On chamber, on valley, on top.
- f) Time interval between sample mixing and pipette loading: 1 minutes, 2 minutes, 5 minutes, 10 minutes, 15 minutes.

- g) Time interval between pipette loading and expulsion: 1 minutes, 2 minutes, 5 minutes, 10 minutes, 15 minutes.
- h) Pipetting steadiness when loading chamber: Steady pipetting, Interrupted pipetting (1-2 sec pause).
- i) Coverslip movement after chamber loading: No shifting, Shifting coverslip 1-3 mm.
- j) Chamber cleaning product: Dish washer soap (Fairy), Ethyl Alcohol.
- k) Microscope for image acquisition: Inverted Microscope (Optika, 10x/20x), Image based automated cell counter with stage automation (Microcounter 3100, Celeromics Technologies, Spain)

4.3.1 Determination of Errors Related to the Volume of Sample Examined:

In Neubauer chambers, the nominal height of the loading chamber is 100 μm . We performed an experimental study aimed to measure the real height of the loading chamber, and thus the real calculated examination volume. We tested 20 non-calibrated, white-label chambers manufactured in China and 3 non-calibrated Neubauer chambers purchased from Marienfeld (Germany). The height of the chamber was measured using a Computerized Numerical Control (CNC) machine and weight sensor control.

A CNC machine is an industrial precision machine that is able to move a milling cutter tool in a XYZ (Figure 3-a) axis and it is designed to mill metallic parts for engineering work. Instead of placing a milling cutter, we used a plastic tip (Figure 3-b) to gently touch the Neubauer chamber surface without damaging it (Figure 3-c). At the lower part of the machine we place a precision scale (Figure 3-d) with the cell counting chamber on top (Figure 3-e). The machine was initially calibrated with a Z-axis reference, and afterwards was positioned at the top of the cell-counting chamber (which was in turn on top of the precision scale). As soon as the machine tip touched the chamber, the precision scale sensed the extra weight so that we acknowledge that the machine tip was touching the chamber and we wrote down the height registered at the machine Z coordinate (Figure 3-f). We repeated this process four times for the counting chamber area at different positions (Figure 3-h) and four times for the cover slip area (Figure 3-g) which was expected to be 100 μm higher than the chamber. Afterwards we averaged the four chamber height measurements and the four cover slip holder height measurements and subtracted both results. The final figure was considered a reliable estimate of the

cell chamber height (Figure 3-i). Since the CNC machine had a $2\mu\text{m}$ positioning precision, we estimate that our measurements had $\pm 2\mu\text{m}$ error.

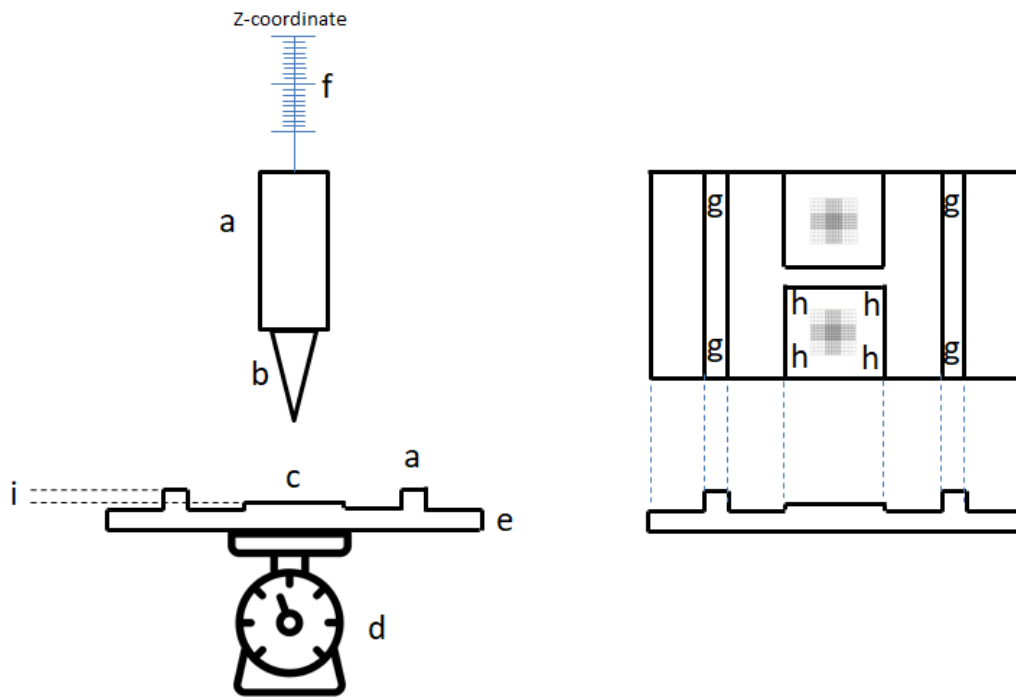


Figure 3: System setup to estimate a cell counting height with $2\mu\text{m}$ precision. The above drawing is not at scale and is for illustrative purposes only.

4.3.2 Determination of Errors Related to the Chamber-Loading Technique:

During our preliminary survey, we had identified three different chamber-loading techniques used in laboratories, according to the position of the pipette when loading the chamber.

A) “On chamber” technique:

This is the most popular technique to load a cell-counting chamber. After cleaning the chamber, the coverslip is placed on top of the chamber. Then the loaded pipette tip is positioned next to the coverslip at the center of the cell counting rectangular area (Figure 4). Afterwards, the pipette is unloaded by pressing the push button in a slow and steady manner until the whole counting rectangle is filled with the aliquot. We called this method *on chamber technique* because the unloading of the micropipette is performed directly on the cell-counting chamber. When the liquid is expelled from the

pipette tip, it enters the space between the coverslip and the chamber by capillary action.

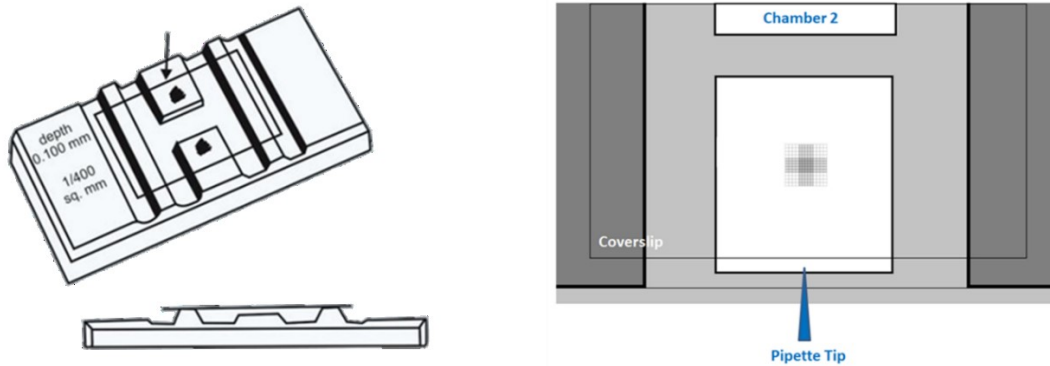


Figure 4: Pipette placement using on chamber loading technique. Side view and zenithal view.

B) On valley technique:

This procedure is similar to the previous one, but instead of placing the pipette tip at the center of the cell counting rectangular area, pipette tip is placed at the valley (also known as the overflow area) of the chamber (Figure 5). This pipetting technique is less popular than the “on chamber” method, and many laboratory technicians consider this an incorrect way to load the counting chamber.

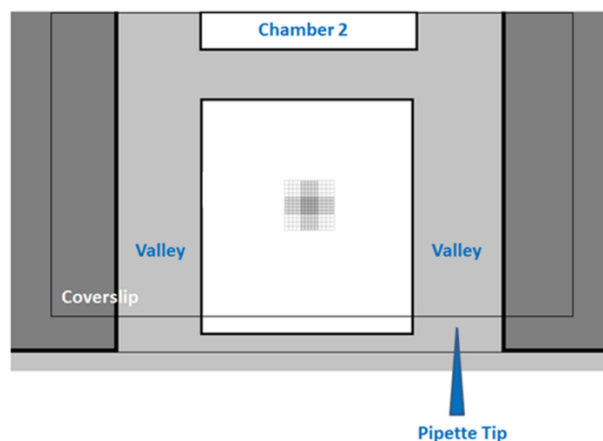


Figure 5: Pipette placement using on valley loading technique. Zenithal view.

C) On top technique:

With this technique, the pipette is placed on top of the cell counting rectangular area before placing the coverslip (Figure 6). After pipetting at the center of the counting area leaving a big drop of liquid at the top of the chamber, the coverslip is placed over the liquid, spreading the liquid all over the chamber. Very few laboratories have reported to use this technique, as most laboratory technicians consider that it does not distribute the cells evenly on the chamber.

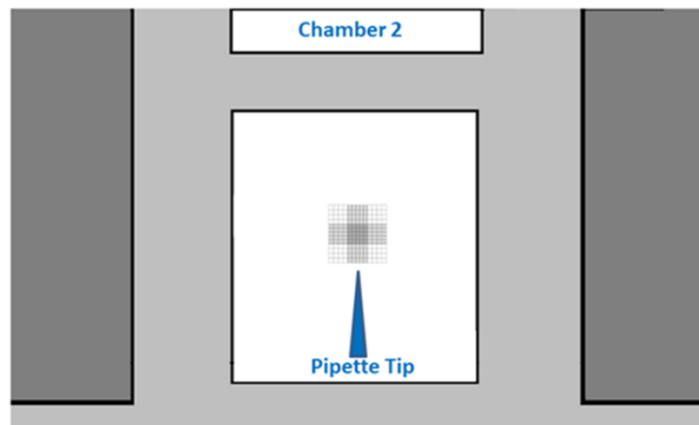


Figure 6: Pipette placement using on valley loading technique. Zenithal view.

4.3.3 Determination of Errors Related to the Number of Cells Counted:

In order to show the effect on accuracy and precision of counting low number of cells, we first estimated theoretically the maximal error when counting with Neubauer or equivalent chamber by means of the formula (Lund JWG et al., 1958):

$$\text{Error}_{\max} = \frac{\pm 200}{\sqrt{n}} \%$$

Thereafter, and based on this formula, we determined the actual error by repeating 20 separate measurements in a Neubauer chamber of low numbers (average of 8 cells counted per field) from the same original suspension in several experiments performed with cultures of different cell types, including Jurkat cells, N13 cells, tomato microspores, and calibrated FlowCount fluorescent microbeads.

4.3.4 Design of an Improved Cell Counting System based on Automation and Artificial Intelligence applied to Microscopic Image Analysis

Based on the results of the survey among laboratory technicians and our own cell counting experience, we developed a novel automated cell counting system aimed to improve the current systems performance, reliability, maintenance costs and usability. The novel cell counting system is based on the innovative approach of interfacing an existing laboratory microscope to a high-resolution camera and to an image processing unit for automatic analysis of microscopic fields.

On the course of this Ph.D. Thesis a prototype of a fully operational cell counting system has been implemented and patented (PCT/EP2013/057164; Main Inventor: Oscar Bastidas). The system later was developed by the company Celeromics Technologies (Valencia, Spain) as two different commercial versions: The Micro Counter system, aimed to automated counting based on cell- or particle suspensions and the Culture Counter system, designed for cell counting based on monolayer cell cultures. Both systems are currently in use in more than 100 laboratories in Europe, USA and Mexico.

[Appendix 9.1](#) includes a detailed description of the invention and its technical details. Here we provide a summary of the most relevant methodological aspects of its configuration, calibration and validation, as performed during the experimental part of this Ph.D. Thesis.

4.3.5 Validation of the Improved Cell Counting System

The process of validation of the automated cell counting system was aimed to determine the error introduced by the system and to decide if this error was acceptable for a cell-counting instrument in research and clinical practice applications.

In order to validate and further optimization our new cell counting system, we reviewed the parameters to assess the performance of a laboratory instrument (Steiner and Norman, 2006). Accordingly, the most relevant quality concepts to our validation were:

- a) Error: The difference between the true or actual value and the measured value.
- b) Accuracy: The closeness of a measurement is to the correct value for that magnitude.
- c) Precision: The degree of agreement is between measurements repeated under the same conditions.

4.3.6 Reference Cell-Counting Methods for Validating the Improved Cell Counting System

In order to determine the error introduced by the measurement system and its accuracy, we needed to introduce a cell-counting method that provided a reference truth. Manual cell counting with the Neubauer chamber is considered as the golden standard, and many research groups choose *de facto* this method to evaluate new cell counting systems. In addition, it has the advantage of being very intuitive to understand and easy to work with. However, Neubauer chamber is prone to systematic errors, such as structural chamber miscalibration, and may introduce human error in cell counting, pipetting or mathematical calculations. Because of this possible uncertainty, we extended the comparison of the new Improved Cell-Counting System to calibrated Neubauer chambers (Marienfeld, Germany) and other robust and well established reference instruments of automated cell counting, including flow cytometer (Cytomics FC500, Beckman Coulter), hand-held impedance-based cell counter (Scepter, Merck-Millipore) and image-based automated cell counters (Countess, Life Sciences Technologies; TC20, Bio-Rad Laboratories).

According to the instrument and the particular experimental design, different cell types or synthetic microbeads (AccuCheck Counting Beads, ThermoFisher) were used as source of reference particle suspensions to be counted. The specific material used for calibration procedures will be indicated when describing the corresponding experiment.

4.3.6.1 Cytomics FC500 Flow Cytometer:

The Cytomics FC 500 Flow Cytometer (Beckman-Coulter) allows automated tube-based acquisition for single-cell assays in suspension. This system (Figure 7) has the capacity to conduct 5-color analysis using a dual 488 nm/635 nm (blue/red) laser. Accurate cell- or particle counting from standard test tubes is usually performed by identifying single objects of interest by their light scatter and fluorescence properties. Count calibration may be achieved by running suspensions of synthetic particles at known concentration (FlowCount Beads, Beckman-Coulter).



Figure 7: Cytomics FC500 Flow Cytometer (Beckman Coulter)

4.3.6.2 Scepter 3.0 hand-held impedance-based cell counter:

Scepter™ 3.0 Handheld Cell Counter (Merck Millipore) is based on the Coulter impedance principle and provides automated counts in seconds (Figure 8). The Scepter™ accurately counts particles as small as 3 μm in diameter. With its microfabricated, precision-engineered 40 μm sensor, the Scepter™ can count samples with concentrations as high as 1,500,000 cells/mL.



Figure 8: Scepter 3.0 hand-held impedance-based cell counter (Merck-Millipore)

4.3.6.3 Countess image-based automated cell counter:

The Invitrogen Countess™ Automated Cell Counter (Thermo Fisher Scientific) includes automated lighting, focus, capturing, counting, and saving (Figure 9). The procedure requires only inserting a proprietary slide with a cell sample. Then, by advanced machine-learning image analysis algorithms it delivers accurate cell counts and viability in less than 30 seconds.



Figure 9: Countess image-based automated cell counter (Invitrogen Thermofisher)

4.3.6.4 TC20 image-based automated counter:

The TC20 Automated Cell Counter (Bio-Rad) counts mammalian cells in a broad range of cell sizes and types in one simple step (Figure 10). Upon insertion of a counting slide, by using auto-focus technology and cell counting algorithm it produces accurate cell counts in less than 30 seconds. Cell size gates allow user to select a population of interest in complex samples.



Figure 10: TC20 image-based automated counter (Bio-Rad).

4.3.7 Validating the Automated Configuration of the Cell Counting System: Cell Concentration

In this setup, we evaluated the performance of the new improved cell-counting system in fully automatic operation. In preliminary experiments, the cell density of suspension cultures of HepG2 cells was determined in parallel with a Micro Counter (Celeromics) and four independent systems : a) Countess image-based automated cell counter (Life Technologies), b) TC20 image-based automated counter (Bio-Rad), c) hand-held

impedance-based cell counter (Millipore) and d) Cytomics FC500 flow cytometer (Beckman-Coulter). In parallel, aliquots from each same original cell suspension were counted by independent operators in a calibrated Neubauer counting chamber (Marienfeld, Germany).

Each measurement was repeated 20 times shaking the original suspension before sampling, and repeated three times.

4.3.8 Validating the Automated Configuration of the Cell Counting System: Cell Confluence.

For assessing the determination of confluence with Culture Counter, flasks of growing HepG2 cell cultures were placed on an inverted microscope. For each culture, 20 photographs were taken for measurements in the Culture Counter. Monolayers were trypsinized and cells resuspended in culture medium. For each suspension, cell concentration was measured with the same three independent systems indicated. Each measurement was repeated 20 times, shaking the original sample before sample extractions (Figure 11).

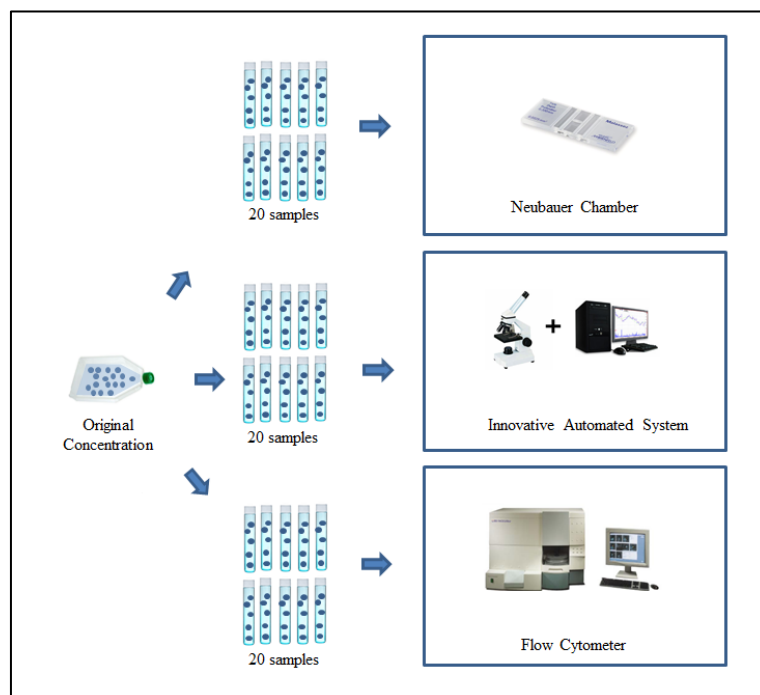


Figure 11: Typical Validation Experiment performed with 3 different cell counting systems.

4.3.9 Validating the Semi-Automated Configuration of the Cell Counting System.

Under this testing approach, the Automatic cell-counting Mode was deactivated and the system was run on Manually-assisted cell counting Mode. This procedure dissociates the two main blocks of the Automatic Cell Counting Method: the Size- and Volume-calibration system and the Automatic Cell-counting system based on image analysis. In this way we can estimate specifically the error introduced by the image analysis system in charge of counting cells present in a microscope field. For this purpose, operators manually loaded calibrated Neubauer chambers with samples from different suspensions of AccuCheck Counting Beads (6.4 μm diameter, original concentration: $10^6/\text{mL}$) and allowed for fully-automatic or semi-automatic cell counting of each suspension in the Cell-counting system.

4.3.10 Validating the Artificial Intelligence Algorithm of the Cell Counting System

With this approach only the AI automatic cell counting algorithm was put under test. It was used to test the system against a large database of 2500 reference microscopic images from 38 different laboratories. The method involved counting manually all cells present in a microscopic image and store the result as a ground truth for future testing (Figure 12).

Afterwards, an automatic program analysed these images with the AI algorithm to be tested, and compared the algorithm result with the previously counted number of cells (ground truth). The result of each analysis along with the difference from the ground truth (estimated error) was stored in a computer file for later analysis. Two types of regression tests databases were generated, one for suspension cells and a second one for adherent cells.

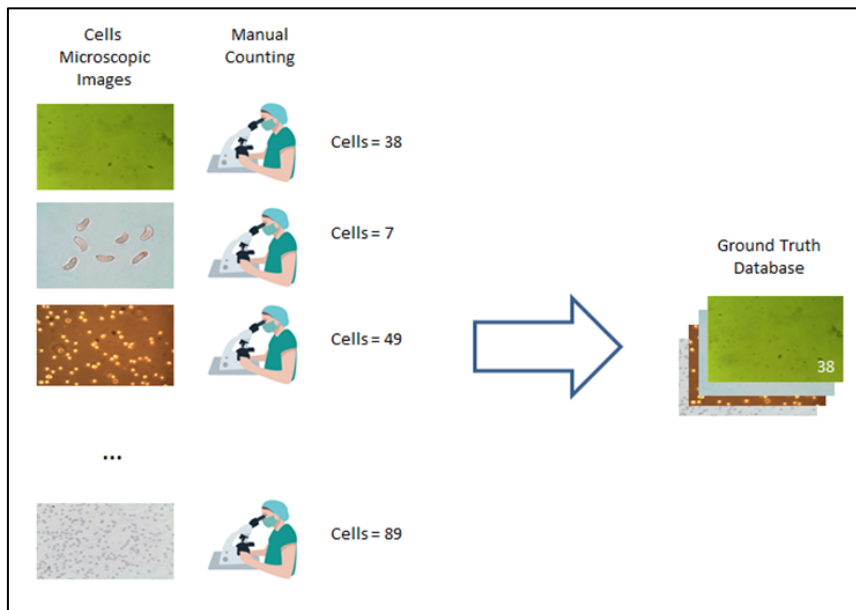


Figure 12: Generation of the ground truth database. The database is generated by manually counting cells present in microscopic images.

This method allowed for the development of a robust image analysis algorithm with a heterogeneous database of different kind of images. Automated regression tests could be run automatically without the need of laboratory personnel (Figure 13).

See [Appendix 9.5](#) for the document that was used as a template to conduct the validation of the prototype system where the methodology was implemented.

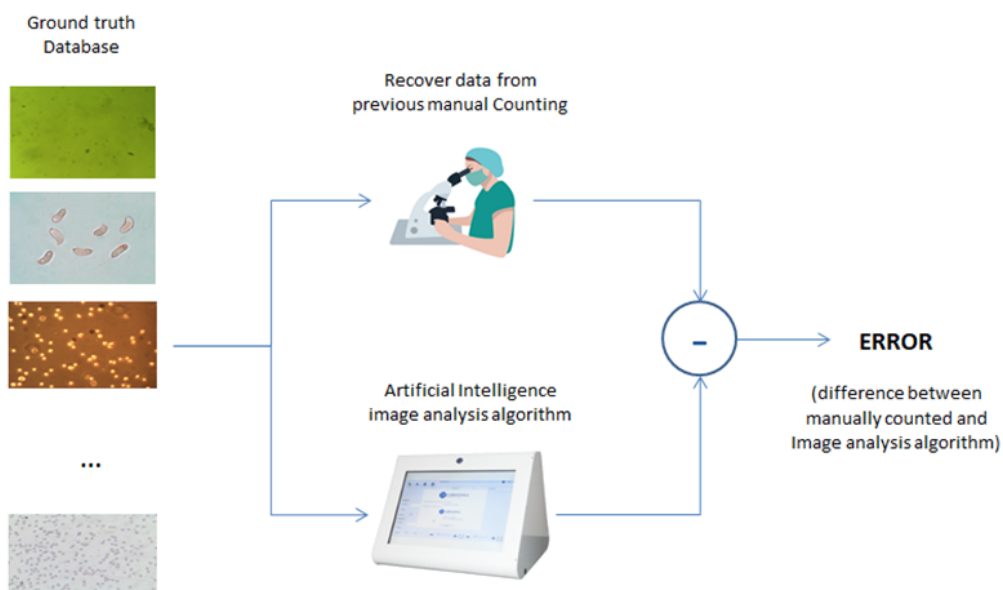


Figure 13: Running of a regression tests. Manually counted images are compared against the AI image analysis algorithm.

4.4 Automated Determination of Cell Concentration from Petri Culture Dishes

4.4.1 Estimated Cell Distribution in Petri Dishes:

To estimate the distribution of cells on Petri Dishes, four different types of cells and particles were analyzed on plastic 30mm Petri Dishes: Jurkat cells (ECCAC), HepG2 cells (ECCAC), tomato microspores (Provided by AINIA, Valencia), and Flow-Count™ Fluorospheres (Beckman Coulter). Flow-Count Fluorospheres were diluted in a 1:2 proportion with distilled water. An estimate of 500.000 cells or particles was introduced in each Petri dish.

Cell distribution was estimated using the new system MicroCounter. The system was configured using a 20x lens, with an automatic sampling protocol called "VERTICAL". This protocol captures a straight line of adjacent microscopic fields following the Petri Dish vertical diameter with no overlapping. The size of each microscopic field captured by the CCD system camera was 386 μm (W) x 290 μm (H). After the initial image-capturing phase the system performs an automatic image analysis of each field, and reports the total cell concentration and the number of cells found on each field.

16 Petri dishes of 30-mm diameter were analyzed using this method. A first set of 4 dishes (one per cell or particle type) was analyzed immediately after leaving the incubator or being prepared. A second set of 4 dishes was analyzed after a light rotary movement performed manually by the same laboratory technician. A third set of 4 dishes was analyzed after shaken for 3 minutes in a Sartorius Stedin shaker, at 120 rotations/minute. A fourth set of 4 dishes was analyzed after shaken for 5 minutes in a Grant Boakel BFR25 platform shaker, configured in oscillatory mode, performing 25 oscillations per minute.

4.4.2 Mathematical Modeling of Petri Dish Cell-Distribution and Cell Sampling.

For the numerical model of the Petri dish, we used the R programming language (R Foundation). Based on the results of the analysis of the 16 Petri dishes, cell distribution was modeled using a mathematical formula that mimicked the empirical cell distribution measured in the worst-case scenario of highest heterogeneity.

We called this mathematical formula the cell number function. This function received X and Y coordinates and returned the total number of cells that would be counted if we would position ourselves on those coordinates of the Petri dish, and analyzed an area of 386 μm (W) x 290 μm (H). This is the equivalent area of a microscope field captured using a 20x lens.

Samplings or cell counts were simulated by generating a list of specific or random coordinates of the Petri dish numerical model, and calling the cell number function to obtain the number of cells found on those coordinates. The total number of cells on the dish can be estimated by the following formula:

$$\bar{X} = \frac{\sum \text{cells}}{A_{\text{fields}}} \times A_{\text{dish}}$$

where,

\bar{X} = Estimated total number of cells on the dish.

$\sum \text{cells}$ = Total cells on the fields analyzed.

A_{fields} = Area of all the fields analyzed.

A_{dish} = Total area of the Petri dish.

The real cell population was calculated by scanning the full numerical model of the Petri dish, and adding the cell number found on all fields. The real cell concentration was calculated dividing the real cell population by the total area of the dish. The cell counting error for a given sampling was calculated by subtracting the sampling result to the total cell population. The main advantages of using a numerical model were the possibility to measure the total cell population, as well as performing all types of cell

counting simulations in a very short time span, when compared to real experimental cell counting. (E.g. analyzing the whole Petri dish with the MicroCounter 3100 system at 20x would take more than 10 hours). Using a numerical model avoided the mathematical and statistical complexities of using a purely analytical method (Sandgren & Robinson, 1984).

5 Results

5.1 Journal Requirements for Reporting of Experimental Results in Biomedicine.

727 journals from seven different categories of the JCR classification in the Biomedicine general area (Biochemical Methods; Cell Biology; Cell & Tissue; Developmental Biology; Medicine: General & Internal; Medicine: Research & Experimental; Multidisciplinary Science), were divided the journals in two broad groups, according to their degree of involvement in clinical or basic biomedical sciences:

- 1) Medical journals (Medicine General & Internal, Medicine Research & Experimental)
- 2) Non-medical journals. (Biochemical Methods, Cell Biology, Cell & Tissue, Developmental Biology, Multidisciplinary Science)

Then, we classified all the journals analysed in 4 groups, according to the statistical and/or mathematical reporting requirements for publication that are made public online, as follows:

- 1) Journals with specific statistical requirements.
- 2) Journals with non-specific statistical requirements.
- 3) Journals with mathematical requirements, but no statistical requirements.
- 4) Journals with neither statistical nor mathematical requirements.

As seen in [Figure 14](#), statistical requirements were significantly higher on Medical Journals. On average, 60.74 % of medical journals stated specific statistical requirement on their guides to authors. The higher percentage corresponded to the category *Medicine, General & Internal*, with 72.26 % of their journals having specific statistical requisites.

Non-Medical Journals scored an average of 24.74 %. Less than one fourth of all non-medical journals had specific statistical requirements, almost 3 times less than medical journals.

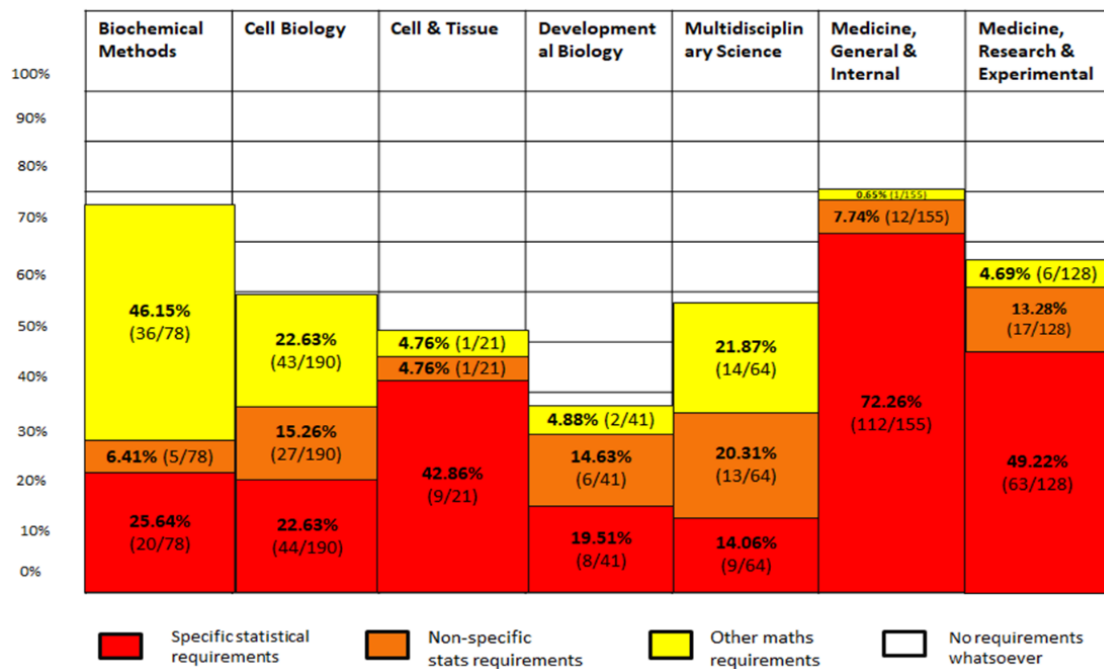


Figure 14: Statistical and mathematical requirements to authors in 727 journals of Biomedicine area, classified in seven categories, as defined by the Journal Citation Reports classification.

In order to visualize the correlation between the Impact Factor (IF) of the analyzed journals and their reporting requirements, we generated a color table in which all the journals were classified in four Quartiles (Q-1 to Q-4), according to their IF, and then highlighted in different colors according to their statistical and mathematical reporting requirements to authors (Figure 15).

In this color map of all analyzed categories ranked by IF, we visually appreciate a higher concentration of red (specific statistical requirements) on Q1 and Q2 in 5 out of the 7 categories analyzed: *Biochemical Methods*, *Cell Biology*, *Cell & Tissue*, *Developmental Biology*, and *Multidisciplinary Science*. All of these are non-medical categories. On the other hand statistical requirements in medical categories were more evenly distributed all over the ranking table.

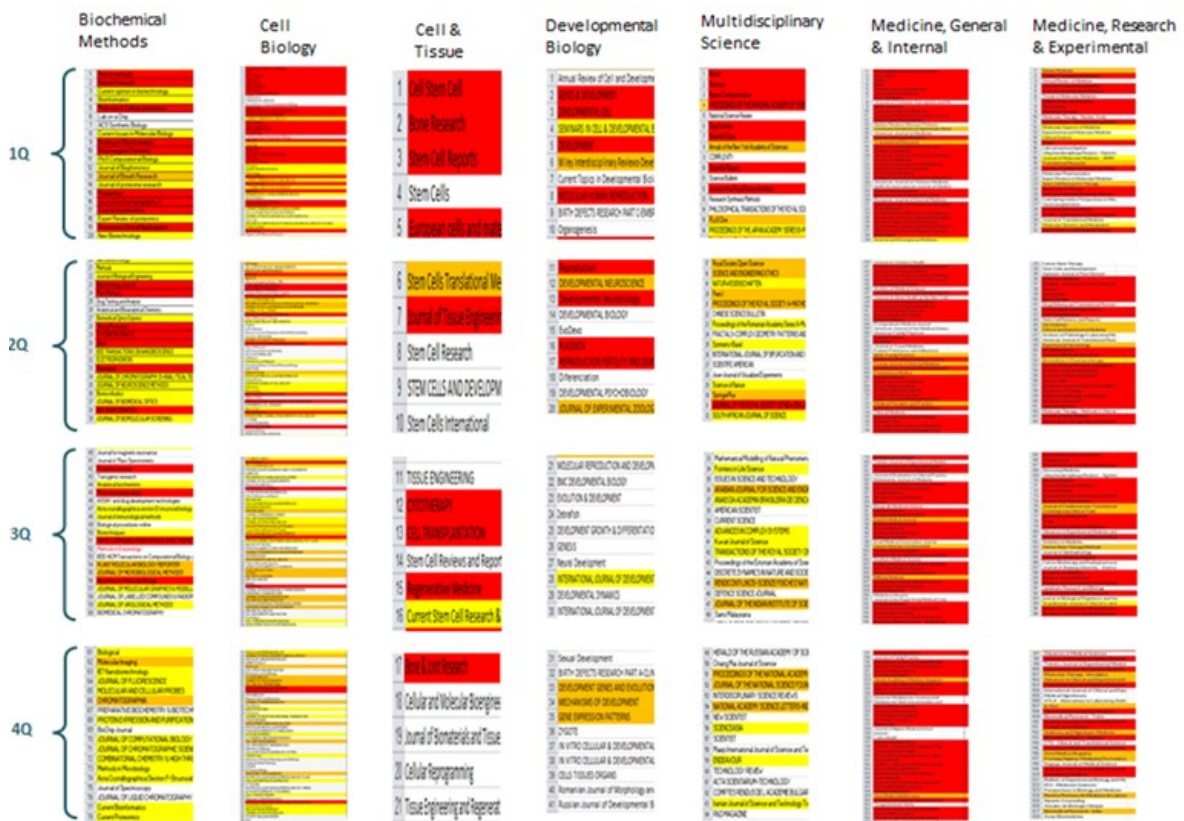


Figure 15: Color Map of Journals Statistical Requirements in relationship with Journal Position in specific areas based on journal Impact Factor. See Figure 14 for an explanation of the color code.

The correlation between IF and number of specific statistical requirements revealed by Figure 15 was confirmed mathematically by calculating Pearson (Table 3) and Spearman (Table 4) correlations.

Table 3: Pearson correlation between journal IF and the number of statistical requirements.

	Stats Correl.	p-value	Correl.Strenght	Math Correl.	p-value	Correl. Strenght
Multidisciplinary Science	-0.4259	0.0004501	MEDIUM	0.1001694	0.431	WEAK
Cell Biology	-0.4901924	7.04E-13	MEDIUM	-0.05056	0.4884	WEAK
Methods	-0.3448155	0.001991	MEDIUM	-0.1060481	0.3554	WEAK
Cell Tissue	-0.51327	0.01733	MEDIUM	-0.18024	0.4343	WEAK
Developmental Biology	-0.4215	0.006055	MEDIUM	-0.050464	0.754	WEAK
Medicine, General & Internal	-0.1262	0.1174	WEAK	-0.1365	0.09015	WEAK
Medicine, Research	-0.067	0.4522	WEAK	-0.286472	0.001045	WEAK

We noted that medium degree correlation was found between statistical requirements and impact factor on non-medical journals with all p-values lower than 0.05.

No significant correlation was found between mathematical (non-statistical) requirements and impact factor: weak correlation with p-value > 0.05 on most categories.

Table 4: Spearman correlation between journal IF and the number of statistical requirements.

Spearman Correlation						
	rho coeff	p-value	Correl.Strenght	Math Correl.	p-value	Correl. Strenght
Multidisciplinary Science	-0.526	0.000008022	MEDIUM	0.089475	0.482	WEAK
Cell Biology	-0.49022	7.02E-10	MEDIUM	-0.02342	0.7484	WEAK
Methods	-0.4072334	0.0002154	MEDIUM	-0.07140591	0.5344	WEAK
Cell Tissue	-0.4526	0.03936	MEDIUM	-0.1802	0.4343	WEAK
Developmental Biology	-0.4906258	0.001127	MEDIUM	-0.05229174	0.7454	WEAK
Medicine, General & Internal	-0.1066357	0.1866	WEAK	-0.1365827	0.09015	WEAK
Medicine, Research	-0.081864	0.3583	WEAK	-0.286472	0.001045	WEAK

Moreover, as shown in Figure 16, we performed a Wilcoxon test to prove that the average number of statistical requirements in Q1+Q2 quartiles H1 was different from H2 (Q3+Q4). The Wilcoxon test rejected the null hypothesis ($H_0: \mu_{H1} = \mu_{H2}$) in all non-medical categories, except for *Cell & Tissue*, thus proving that the number of statistical requirements are higher in journals from the first and second quartile in most of the bioscience journals that we included in the non-medical categories.

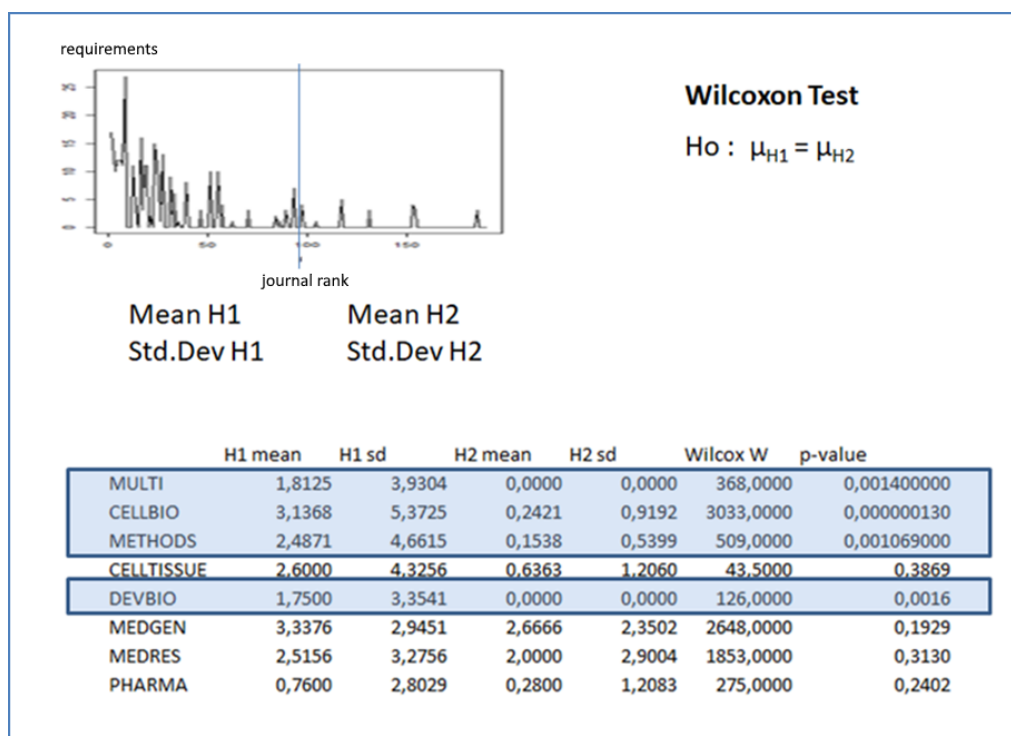


Figure 16: Wilcoxon test performed to prove that the number of statistical requirements is higher in journals in Q1 and Q2 quartiles.

[Appendix 9.2](#) enumerates the most common requirements found on guides to authors by life science journals.

5.2 Identification of Cell Counting as a Popular and Critical Experimental Procedure.

Most technicians and researchers consulted agreed that cell count is a necessary step in most cell assays and in-vitro cell biology experiments, in order to:

1. Determine the initial cell population when starting the experiment.
2. Determine the effect of the experiment.

Based on the interviews specifically conducted with 17 laboratory technicians and, more informally along laboratory demonstrations (over 150) the most popular cell counting systems found in laboratories, as per 2017, were:

1. Cell counting chambers (Neubauer Improved, Thoma, etc.)
2. Flow Cytometers.
3. Image Based Automated Cell Counters.
4. Specialized counting methods directly examining cell cultures on Falcon flasks, Petri dishes, Microscope chambers, Multiwell plates, etc.

The survey also determined that the most popular magnitude measured in cell assays and experiments is the **cell count** or **cell concentration**. Given the fact that living cells need a liquid medium to grow and survive, most cell count measurements are determined with an indirect measurement based on the cell concentration of one or several samples of the base cell population.

5.3 Determination of Errors when Using a Neubauer Chamber

5.3.1 Errors Inherent to the Sample Volume: Determination of Chamber Height.

In a first series of experiments, we wanted to determine precisely the height of the counting chamber, in order to provide exact calculations of the actual volume examined for a theoretical height of the chamber = 100 μm . For this purpose, we tested 20 calibrated (Marienfeld, Germany) and 20 non-calibrated (Unbranded, Chinese origin)

Neubauer counting chambers, using a CNC machine and weight sensor control that was able to perform height measurements with $2\mu\text{m}$ precision.

As seen in [Figure 17](#), only 5 non-calibrated chambers out of 20 had a height between 80 and $120\mu\text{m}$. We found errors as high as 81% or 118%.

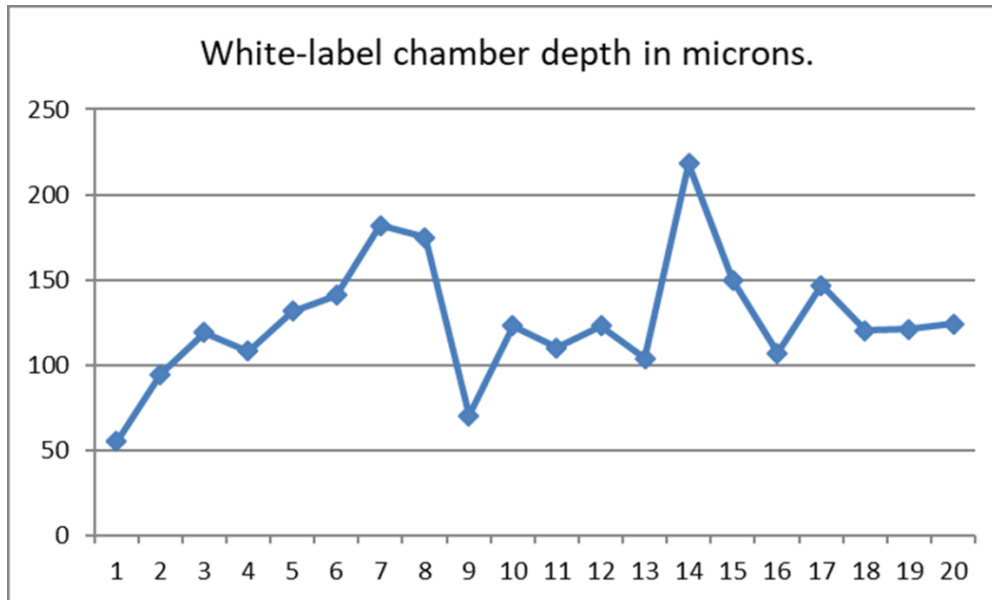


Figure 17: Distribution of the depth (in μm) of 20 non-calibrated Chinese white-label Neubauer chambers, as measured with a $2\mu\text{m}$ precision system.

On the other hand, Marienfeld Neubauer chambers were shown to have more consistent depths between $95\mu\text{m}$ and $120\mu\text{m}$, with errors ranging from 5%-20% ([Figure 18](#)). However, it must be noted that these errors are outside the 2% tolerance indicated by the manufacturer, although significantly closer to the nominal depth of the chamber than the white label chambers.

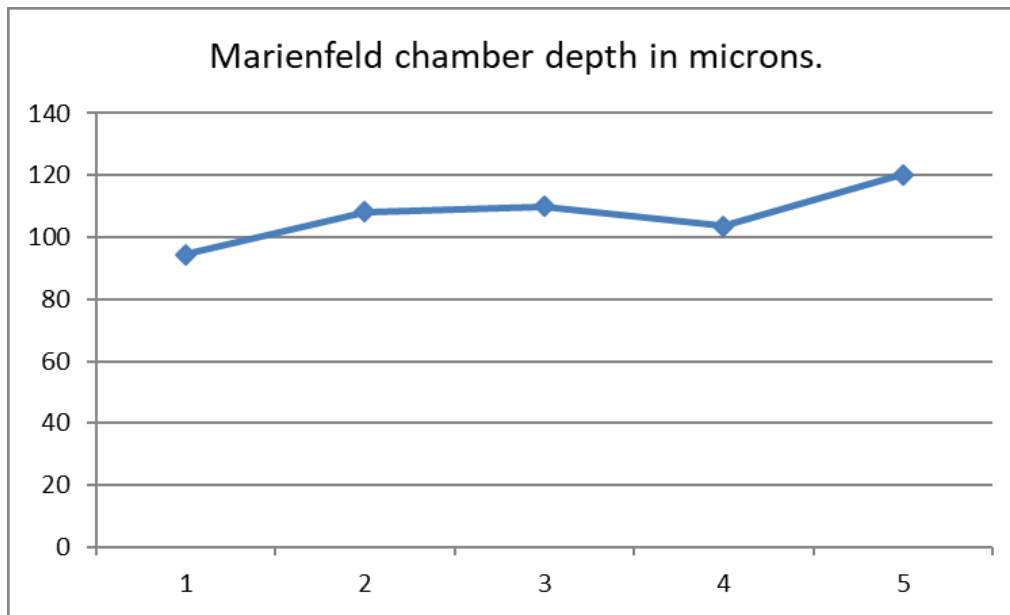


Figure 18: Distribution of the depth (in μm) of 5 non-calibrated Marienfeld Neubauer chambers, as measured with a $2\mu\text{m}$ precision system.

The caveats found on this structural analysis on Neubauer chambers might be extrapolated to other similar types of counting chambers, such as Neubauer Improved, Thoma, Howard, Nageotte, McMaster, Sedgewick Rafter, etc.

5.3.2 Errors due to insufficient number of counted cells:

The most relevant formula to estimate the error when counting with Neubauer or equivalent chamber is:

$$Error\ max = \pm \frac{200}{\sqrt{n}} \%$$

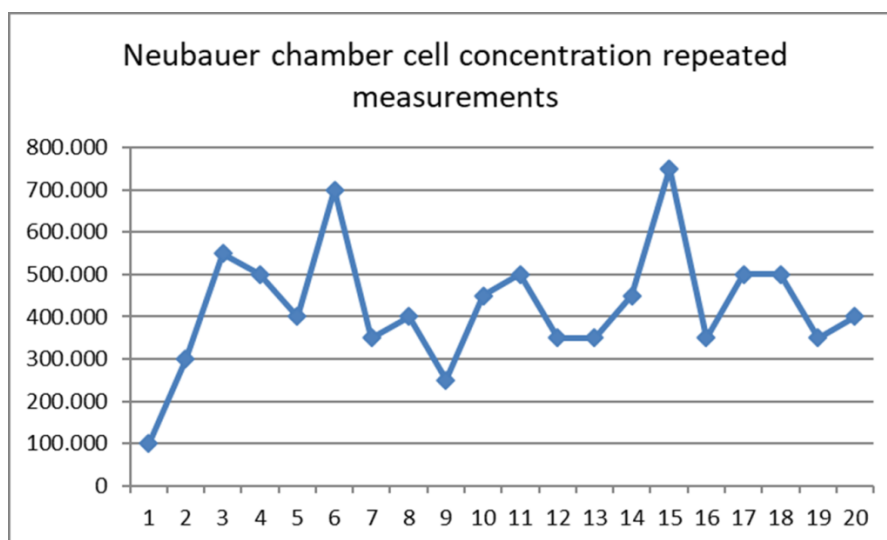
Based on the previous formula, we can determine the amount of cells that should be counted in order to be accurate enough for each experiment. As seen in [Table 5](#), cell counters may introduce an error of up to 44%. Even inside the optimum concentration, range reported by the manufacturer the system may introduce an error of up to 31%.

Table 5: Mathematical estimation of error in cell counting due to the concentration of cell suspension and the actual number of examined cells in a Neubauer counting chamber.

Concentration	Cells in 0.4 μL	Error
5×10^4 cells/mL	20	44%
1×10^5 cells/mL	40	31%
1×10^6 cells/mL	400	10%
1×10^7 cells/mL	4000	3.1%

The common habit of some laboratories of counting at least 100 cells in a Neubauer chamber will introduce an error of $\pm 20\%$, which may be considered insufficient for a wide range of experiments. In order to reduce the estimated error to $\pm 10\%$ a total amount of 400 cells should be analysed. None of the external laboratories consulted had internal protocols that required counting 400 cells.

In order to further show the risk of counting a low number of cells, we estimated the error repeating 20 separate measurements (average of 8 cells counted per field) from the same original concentration in a Neubauer (Figure 19).

**Figure 19:** Estimation of error in cell counting in Neubauer chambers due to a low number of examined cells. Data represent 20 measurements (average of 8 cells counted per field) on a Neubauer cell counting chamber from the same original concentration.

The resulting data are tabulated in [Table 6](#), and show that this procedure may result in both low precision (70% maximal error) and low accuracy (CV of 34.24%)

Table 6: Error, imprecision and inaccuracy in cell counting in Neubauer chambers due to a low number of examined cells. Data represent 20 measurements (average of eight cells counted per field) on a Neubauer cell counting chamber from the same original concentration.

Repeats	Cells counted (Average)	Maximal Error	Mean Cell Concentration (Cells/mL)	Standard Deviation	Coefficient of Variation (%)
20	8	70%	425.000	145.548	34,24%

5.3.3 Errors due to chamber loading procedure:

Loading chamber techniques tend to be customary. It has been observed during this research that different European countries tend to favor one method over the others, but from the laboratories consulted none of them had performed any formal analysis regarding cell homogeneity depending on the chamber loading method used as the one conducted during this research. We had not found any similar cell distribution study on the analyzed bibliography either.

Several phenomena were observed when performing the experiments while modified several independent variables when loading the Neubauer chamber. [Table 7](#) summarizes the most significant results leading to higher heterogeneity in cell count distribution.

Table 7: Most common heterogeneous cell distribution effects found

Phenomenon observed	White-label chamber	Marienfeld chamber
The column	YES	YES
The walking cells	YES	NO
Waves	YES	YES
Marbles	YES	YES
Zebra	YES	NO

In order to better describe the effects and identify the areas where cell concentration might be found uneven; we coded the different zones of the cell counting chamber as shown in **Figure 20**

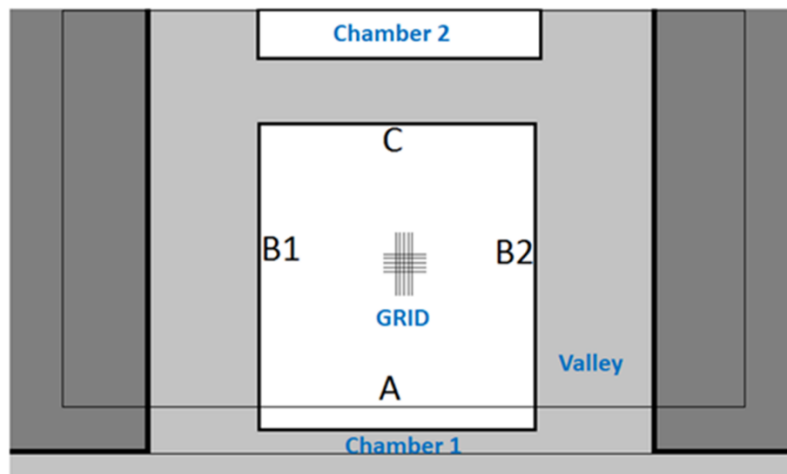


Figure 20: Identification of the different areas on a Neubauer counting chamber for describing errors due to chamber loading procedures.

5.3.3.1 The “column” effect:

Description: A higher concentration of cells are located just in front of the pipetting on zone A (See Figure 21 and Figure 22 for representative images of this effect).

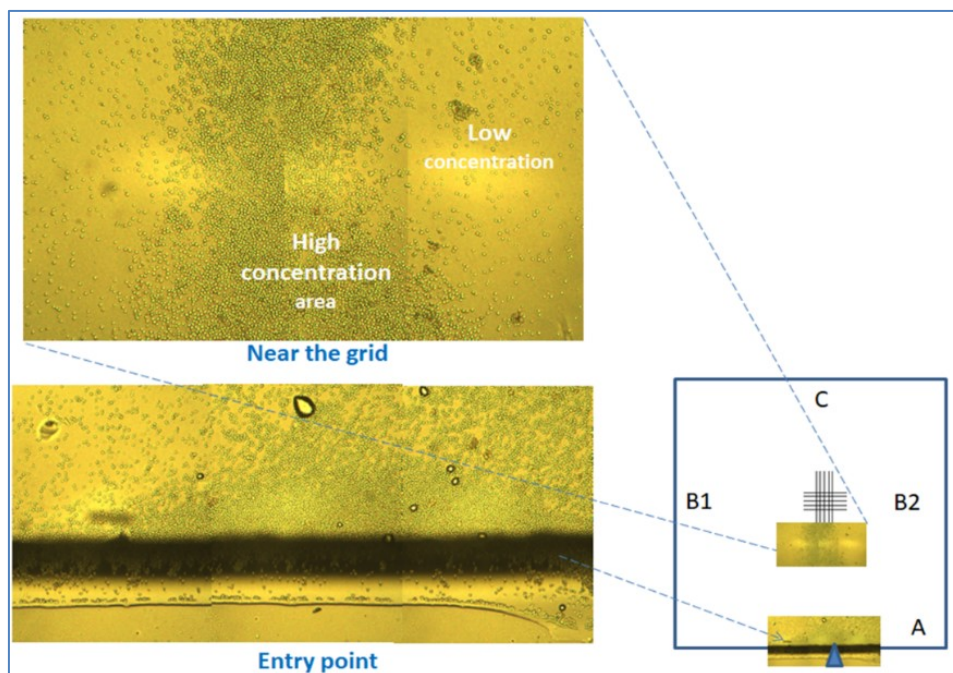


Figure 21: Visual microscope capture of the “column” effect. Composition of images with 10x lens magnification.

When does it happen?: This effect was found when cells settled at the bottom of the microcentrifuge tube (eppendorf tube) for 15-20 minutes or more or when cells remained 10-15 minutes or more loaded into the pipette.

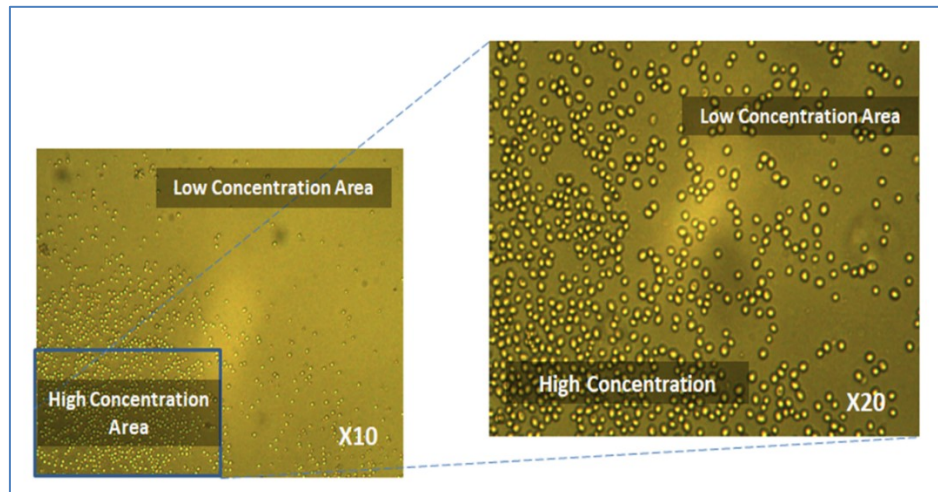


Figure 22: The “column” effect as seen with 10x and 20x microscope magnification lenses.

Quantification: We measured cell concentration in low and high concentration areas, in order to determine the magnitude of the effect. The *effect* was observed in both counting chambers with a similar pattern. The white label chamber introduced a higher heterogeneity in the cells distribution. For higher cell concentration we detected differences as high as 74 times on the same chamber (1% confluence vs. 74% confluence).

5.3.3.2 The “Walking cells” effect.

Description: After loading the chamber, there was a flow of cells from the C side of the chamber. The cells tended to accumulate near the point where they were introduced by the pipette. (See Figure 23 and Figure 24 for representative images of this effect).

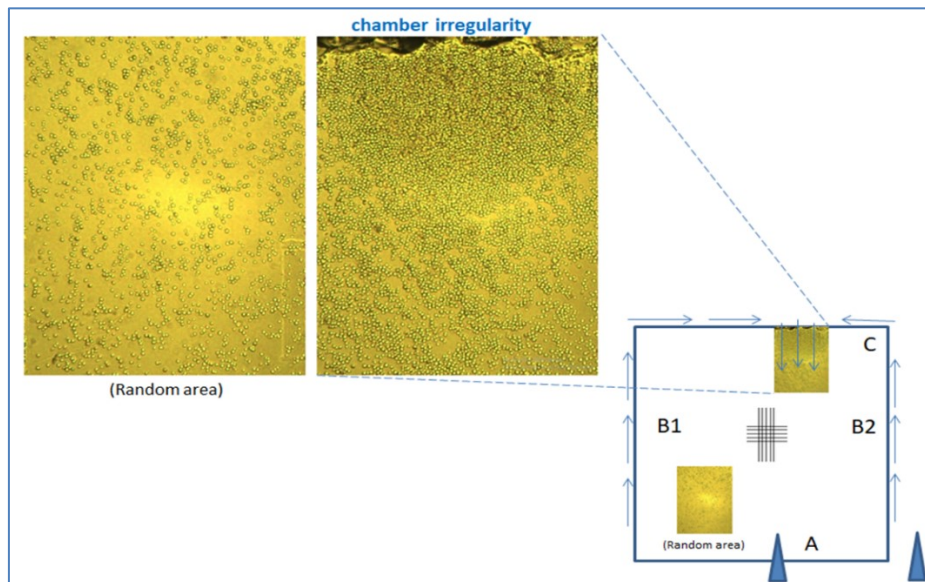


Figure 23: Visual microscope capture of the “Walking Cells” effect. Composition of images with 10x microscope magnification lens.

When does it happen?: At high and low concentrations, and independently of the chamber loading technique or pipette used. When using the Marienfeld chamber cells walked around the chamber slightly, but in fewer quantities and did not re-enter the chamber afterwards. At the white label chamber cells re-entered the chamber after moving around the chamber.

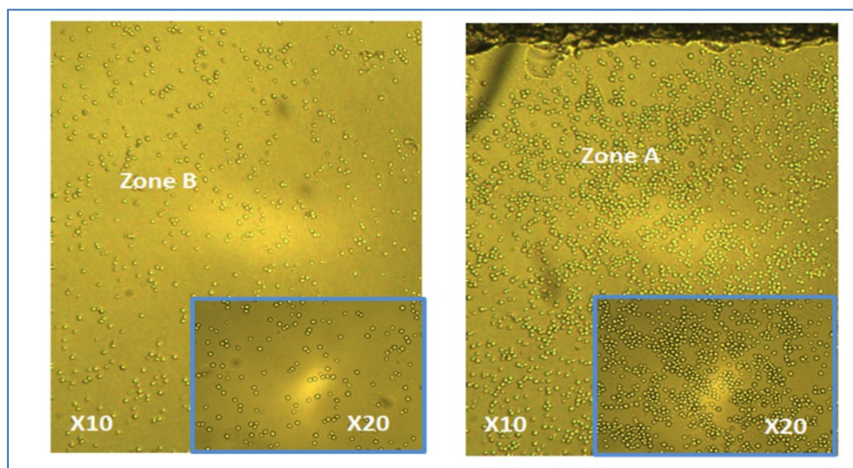


Figure 24: The “Walking Cells” effect as seen with 10x and 20x microscope magnification lenses.

Quantification: The magnitude of this effect varied greatly with each chamber load. The effect was observed mainly on white label chambers; Marienfeld chambers presented it also with smaller magnitude. Sometimes this effect reaches the grid. Even when it does

not reach the grid the heterogeneous distribution produced biases on the results that introducing error some degree of error. For higher cell concentration we detected confluences at 2x difference (51% confluence vs. 26% confluence on the same chamber).

5.3.3.3 The “Waves” effect:

Description: One or several areas of stratification or *waves* appear in the cell distribution inside the chamber. (See Figure 25 and Figure 26 for representative images of this effect).

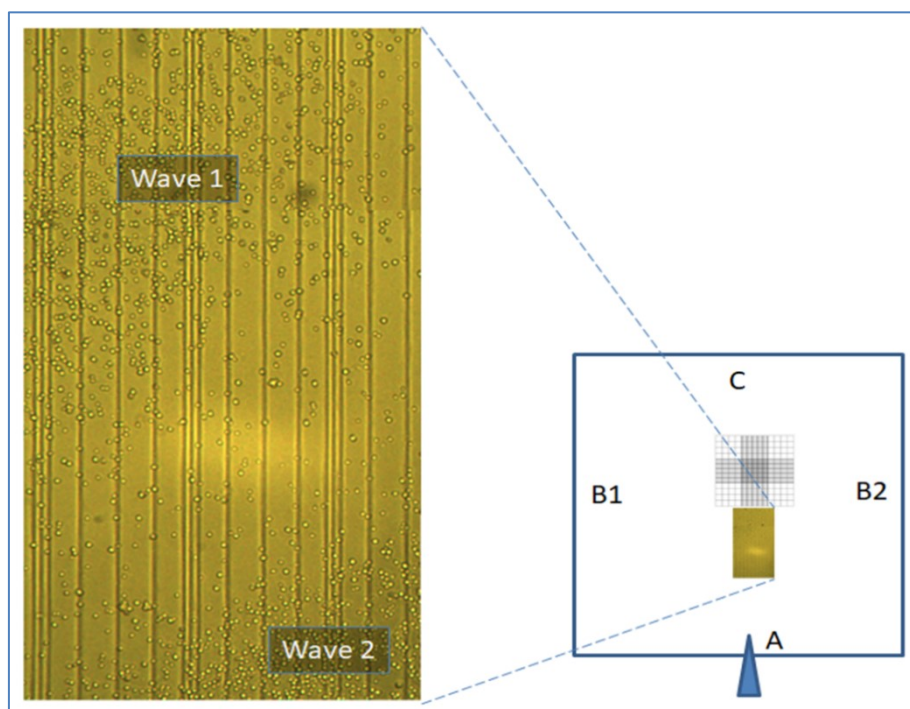


Figure 25: Visual microscope capture of the “Waves” effect with a 10x magnification lens.

When does it happen?: This effect appears when using high concentration samples, when the aliquot is not homogenous (we didn’t mix properly or waited for too long before pipetting) or if we interrupt the pipetting for 1-2 seconds, with a 1 – 10 μ l automatic pipette and *on chamber* pipetting technique.

Quantification: We measured cell confluence in low and high concentration areas, in order to determine the magnitude of the effect. The largest effect observed implied confluence of 98% (Zone A) vs. 56% (Zone B). This suggests that this effect could introduce errors in the order of 50%-100% of the magnitude measured. The effect was observed in both white label chamber and Marienfeld.

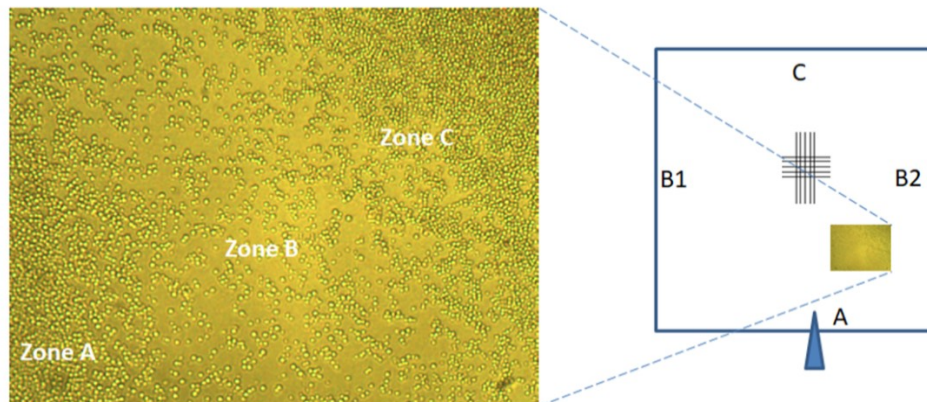


Figure 26: The “Waves” effect as seen with 10x microscope objective lenses.

5.3.3.4 The “Marbles” effect:

Description: When the chamber is loaded, we see at the limit of the coverslip a greater cell concentration. This effect could be due to the surface tension of the liquid that is formed at the coverslip-chamber interface. (See Figure 27 and Figure 28 for representative images of this effect).

When does it happen?: When loading with on chamber technique using a 1-10 μ L automatic pipette or Pasteur pipette. With higher cell concentration this effect is more remarkable.

Quantification: We could not quantify this effect by confluence of cell concentration because the system is unable to determine precisely the objects because of different focal planes.

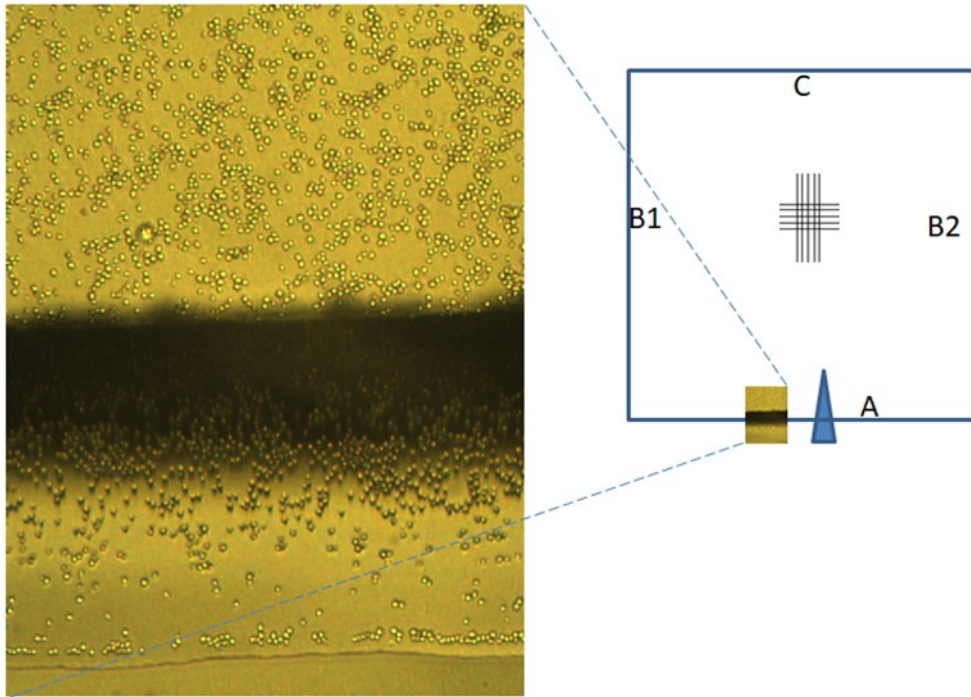


Figure 27: Visual microscope capture of the “Marbles” effect on a Marienfeld chamber as seen with a 10x microscope magnification lens.

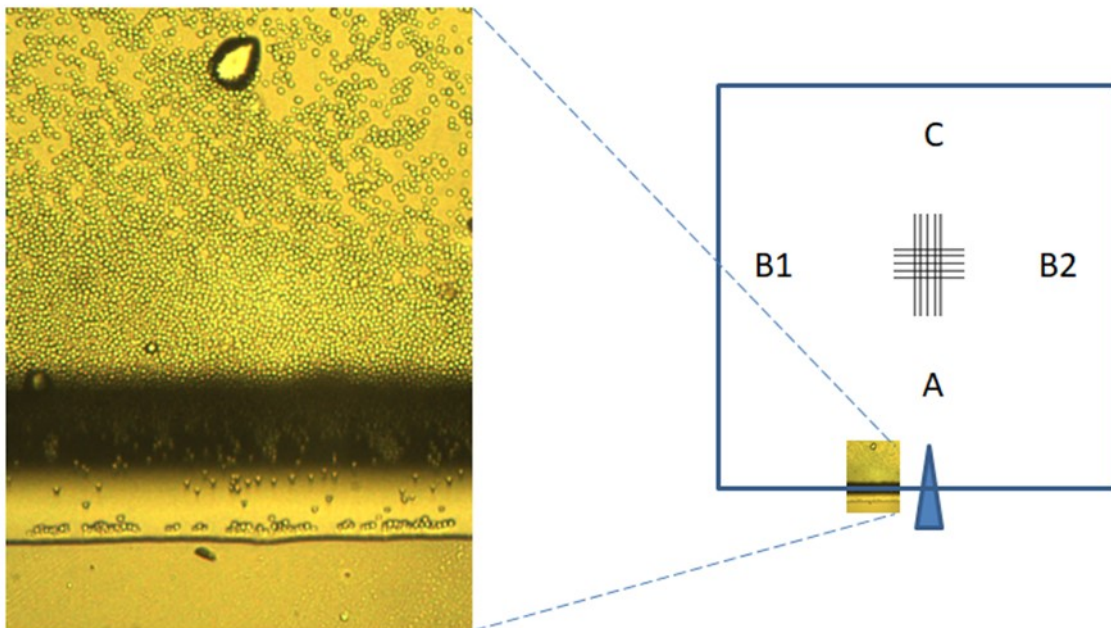


Figure 28: Visual microscope capture of the “Marbles” effect on a white-label chamber as seen with a 10x microscope magnification lens.

5.3.3.5 The “Zebra” effect:

Description: We observed higher concentration of cells on the sides of the chamber, decreasing as we move to the center of the chamber.

When did it happen?: It was observed only once, when loading the chamber with on chamber technique using a Pasteur pipette. (See Figure 29 for a representative image of this effect).

Quantification: We were not able to reproduce this unique effect.

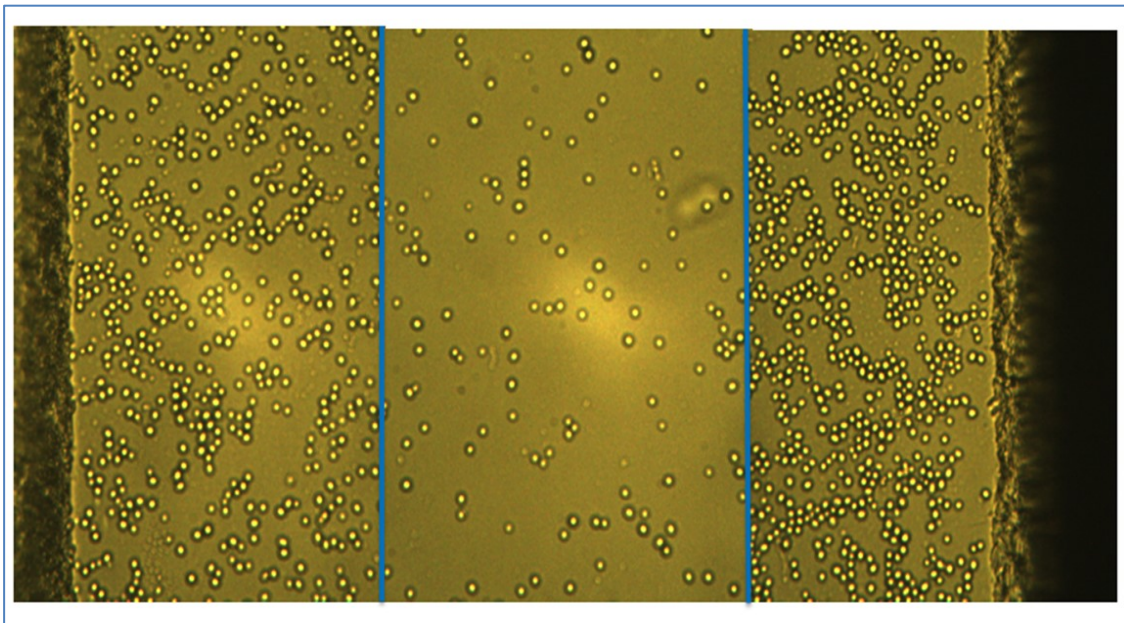


Figure 29: Visual microscope capture of the “Zebra” effect produced by loading with a glass Pasteur pipette. Composition of images with a 20x microscope magnification lens.

5.4 Design of an Improved Cell Counting System based on Automation and Artificial Intelligence applied to Microscopic Image Analysis

Based on the survey to laboratory technicians and on our own cell counting experience, we developed two new cell counting methodologies based on microscope image analysis and in Artificial Intelligence algorithms. The goal of such developments is to improve the current systems performance, reliability, maintenance costs and usability (Table 8).

Table 8: Improved cell counting methodologies proposed.

New Cell Counting Methodologies	Type of cell assay	Innovative Approach
Micro-Counter Microscope based suspension cell counting method	Cell counting on Neubauer chamber or equivalent cell counting chamber	AI Image Analysis + Optical Microscope adaptation
Culture-Counter Inverted microscope base adherent cell counting method	Cell counting on Flask, Petri dish and well plate.	AI Image Analysis + Inverted Optical Microscope adaptation

All the proposed methodologies are based on our innovative approach of using an already existing laboratory microscope and attaching a camera plus an image-processing unit for automatic microscope-field analysis.

5.4.1 Elements of the Improved Cell Counting System:

The basic configuration of the system comprises the following elements (Figure 30):

- Laboratory microscope: Straight (Micro Counter) or Inverted (Culture Counter) Microscope (Optika, 10x / 20x lenses) (Figure 30 a)
- Microscopic camera (Figure 30 b) and connection cable.
- Data analysis unit, based on a small-size personal computer (Mini 11" Desktop, Windows PC software) (Figure 30 c-d).

- Touch screen to configure, operate the system and visualize the results (Standard 15" Touchpad) (Figure 30 c-d).
- Neubauer or equivalent reusable cell-counting chamber (Figure 30 e).

Any transparent plastic- or glass cell container adjustable to microscope stage (Falcon flasks, Petri dishes, multiwell plates, multichamber plates)



Figure 30: Elements of the improved cell counting method.

5.4.2 Operation of the Improved Cell Counting System:

The proposed cell counting method requires the following steps:

5.4.2.1 Initial size and volume calibration:

Calibration of the system is needed to determine the relationship between physical distances and the size in pixels on the screen (Figure 31). The pixel-to microns ratio (pmr) is calculated using the data collected during the system calibration. When a

segment of 100 microns is defined on the screen, the system stores the amount of pixels. The pmr is calculated by dividing the amount of pixels selected by 100.



Figure 31 : The process of system calibration is performed by selecting a segment of 100 μm on the screen

By this initial calibration, the system calculates real distances on the captured images and the volume of liquid on each microscope field (area of the image), thus making unnecessary is no need to use a grid for counting. The determination of the volume of liquid being analysed is necessary to calculate cell concentration of the sample.

5.4.2.2 Cell type configuration:

In order to configure the automatic detection of cells in the microscope field, user must indicate the following parameters for each type of cell to be analyzed:

- Maximum cell size (diameter in microns): All cells above this threshold are rejected.
- Minimum cell size (diameter in microns): All cells under this threshold are rejected.
- System sensitivity: Determines how sensitive the system is to the edges of the cells on the screen. Higher sensitivity allows for the detection of softer cell edges.

5.4.2.3 Automated Cell Counting:

As a final step, the user focuses the microscope on the sample, capturing one or several microscopic fields that are stored as images in the system. Counting is performed by an Artificial Intelligence algorithm that analyses each image (i.e. each microscope field) and determines the amount of cells present on each field without the need to count them one by one. (Figure 32).

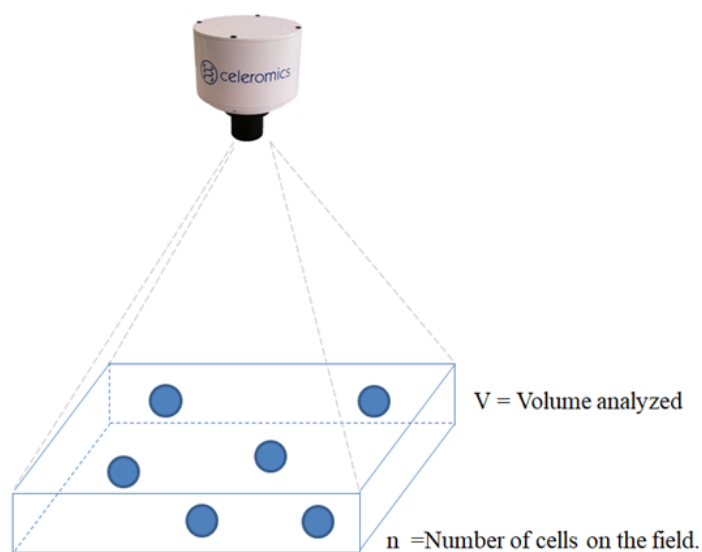


Figure 32 : The process of calculating cell concentration of the sample.

5.4.2.4 Cell Concentration Calculation

Based on the initial calibration the following operations are performed in order to calculate the volume under the area of analysis.

The cell concentration is calculated with the formula **Cell Concentration = n/V** , where **n** is the number of cells counted in a microscope field and **V** is the volume of liquid on the microscope field.

The volume of liquid covered by a microscope field is calculated with the formula **Volume = $A \times h$** , where **A** is the area of the microscope field and **h** is the height of the counting chamber.

The area of the microscope field is calculated with the related formulas $A=w \times l$, where w is the microscope field width in microns and l is the microscope field length in microns.

Width and length of the microscope field are calculated with the formulas:

$w = wp \times pmr$ and $l = lp \times pmr$ where wp is the microscope field width in pixels, lp is the microscope field length in pixels and pmu is the pixel to microns ratio.

Example of internal calculations conducting to cell concentration calculation:

- Distance / pixels ratio : $100 \mu\text{m} = 120 \text{ pixels} > 1.2 \text{ pixels}/\mu\text{m} > 0,833 \mu\text{m}/\text{pixel}$
- Dimension of microscope field (pixels) = $640 \times 480 \text{ pixels}$.
- Dimension of microscope field (μm) = $533 \mu\text{m} \times 400 \mu\text{m}$
- Area of a microscope field (μm^2) = $213.200 \mu\text{m}^2$
- Nominal depth of the counting chamber = $0,1 \text{ mm} = 100 \mu\text{m}$
- Volume covered by a microscope field = $213.200 \mu\text{m}^2 \times 100 \mu\text{m} = 21,32 \times 10^6 \mu\text{m}^3$
- $1 \mu\text{m}^3 = 10^{-12} \text{ mL}$
- Volume covered by a microscope field = $2,13 \times 10^{-5} \text{ ml}$
- Number of cells in a microscope field = 50
- Cell concentration = cells / Volume = $50 / 2,13 \times 10^{-5} \text{ ml} = 2,34 \times 10^6 \text{ cells/ml}$.

5.4.2.5 Implementation of Artificial Intelligence in the System.

The system involves a robust image analysis algorithm constructed with the aid of a heterogeneous database of different kind of images (See [Section 4.3.10](#)). Automated regression tests were automatically ran without the need of laboratory personnel in order to perform the different AI algorithm optimization iterations.

This technique allowed for incremental improvement of the system while it was tested and installed in new laboratories. The final result of this process is that the final AI algorithm ([Figure 33](#)) can be adapted different kinds of optical devices, cell types and lightning systems with very high degree of robustness.

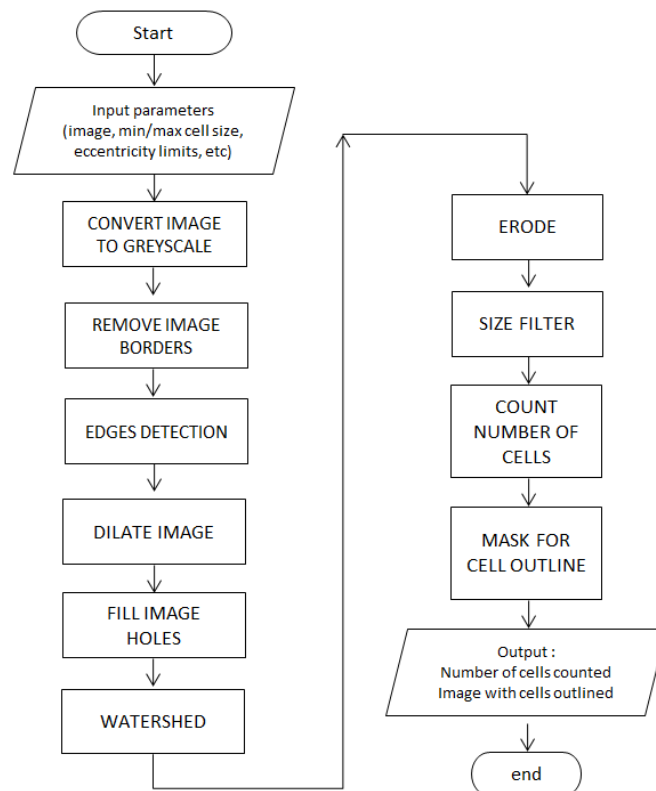


Figure 33 : Top level description of the AI image analysis algorithm for an automated cell counting system that can be adapted to any microscope.

5.5 Validation of the Improved Cell Counting System.

5.5.1 Validating the Automated Configuration of the Cell Counting System: Cell Concentration.

In this setup, we evaluated the performance of the new improved cell-counting system MicroCounter in its fully automatic operation by comparison with other available automated counting systems based in static image analysis or in flow cytometry.

As an starting point, and in order to establish an appropriate experimental design for the comparison, we have calculated the theoretical margin of error introduced by each system given its technical specifications at different cell concentrations, according to the previously used formula $\text{Error max} = \pm 200/\sqrt{n} \%$ (See [Table 9](#))

Table 9: Theoretical margin of error introduced at different cell concentrations by each counting system, as dependent on its technical specifications.

System	Sample volume analyzed	Concentration = 50.000 cells / ml		Concentration = 250.000 cells / ml		Concentration = 1.000.000 cells/ml	
		Cell in Sample	Error Max	Cells in sample	Error max	Cells in sample	Error max
Countess	0.4µl	20	44.7%	100	20%	400	10%
TC-20	0.4µl	20	44.7%	100	20%	400	10%
Scepter	50µl	2500	4%	12,500	1.8%	50,000	0,9%
Manual Counter (Hemocytometer)	0.4µl	20	44.7%	100	20%	400	10%
Micro Counter (5 fields analyzed)	2µl	100	20%	500	8.9%	2,000	4,47%
Micro Counter (20 fields analyzed)	8µl	400	10%	2,000	4.5%	8,000	2.2%

As theoretically calculated in [Table 9](#), most systems behave correctly for concentrations higher than 10^6 cells/mL, with standard errors below 5%. However, most image-based cell counting systems analyzing less than 4 µL sample are not suitable for work with cell concentrations below 500,000 cells/mL, introducing errors higher than 40%. Thus, in order to obtain reliable measurements in that range, the only valid systems are those that analyze samples higher than 4 µL, providing results with errors below 20%.

According to these data, we have performed a series of comparative analyses by performing several cell counts with the different systems from the same original sample. In preliminary experiments, the cell density of suspension cultures of Jurkat cells was determined in parallel with a MicroCounter (Celeromics) and three counting systems, operated by independent experts: 1) Neubauer chamber; 2) Flow cytometer (Cytomics FC500, Beckman-Coulter); 3) Hand-held impedance-based cell counter (Scepter, Merck

Millipore). Each measurement was repeated 20 times, shaking the original sample before sample extractions, and replicated 3 times (Figure 34)

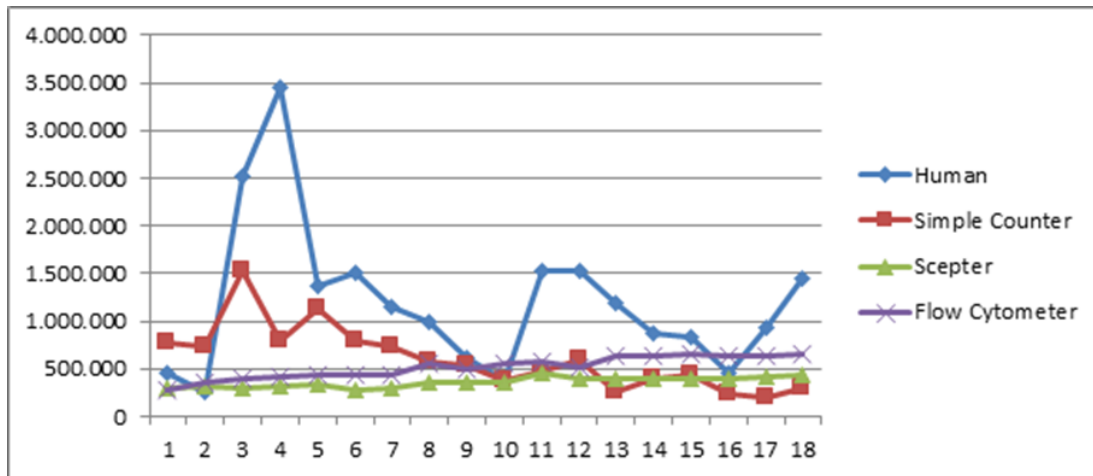


Figure 34 : Precision and accuracy of the automated Micro Counter system on suspensions of Jurkat cells as compared with human cell counting on Neubauer chamber and with two automated counting systems (Scepter and Flow Cytometer). Data represent 20 measurements (average of 37 cells counted per field).

As seen in Figure 34, and as expected from our theoretical calculations, the performance of Micro Counter at a relatively low number of counted cells (average 37 cells/field) was quite comparable to that of both automated systems, Scepter and Flow Cytometer, and much better than the performance of the human-calculated cell count in Neubauer chamber.

The previous experiments were extended to additional counting systems and cell types. As shown in Table 10, HepG2 cell cultures were analyzed alternatively with 4 different instruments : 1) Countess (Life Technologies), 3) Micro Counter 4) Scepter (Millipore), 5) Cytomics FC500 (Beckman-Coulter), and a human performing manual counting.

Table 10: Precision and accuracy of the automated Micro Counter system on suspensions of Hep-G2 cells as compared with human cell counting on Neubauer chamber and with three automated counting systems (Countess, Scepter and Flow Cytometer). Each measurement was repeated 20 times, shaking the original sample before sample extractions, and replicated 3 times.

System	Concentration measured (cells/ml)	Standard Deviation (cells/ml)	Variation Coefficient	Error. Max (%)
Countess	7.2×10^5	1.6×10^5	16%	32%
TC-20	N/A	N/A	N/A	N/A
Scepter	3.6×10^5	0.52×10^5	14.55%	29.1%
Manual Counter (Hemocytometer)	7.19×10^5	2.05×10^5	28.6%	57.2%
Flow Cytometer	5.75×10^5	4.4×10^4	7.7%	15.4%
Microcounter (5 fields analysis)	6.6×10^5	1.6×10^5	14.9%	29.9%
Microcounter (20 fields analysis)	6.8×10^5	4.8×10^4	9.2%	18.4%

In another series of experiments performed on THP-1 cell suspensions we calculated systematically the more relevant parameters of cell counting distribution on Micro Counter system as compared with an automated image-based system (Countess, Life Technologies) and the human-operated Neubauer chamber (Figure 35). Consistent with our previous data obtained with different cell types, the accuracy and precision of Micro Counter were shown to be better than the compared alternatives, as judged by objective parameters as S.D. and confidence intervals.

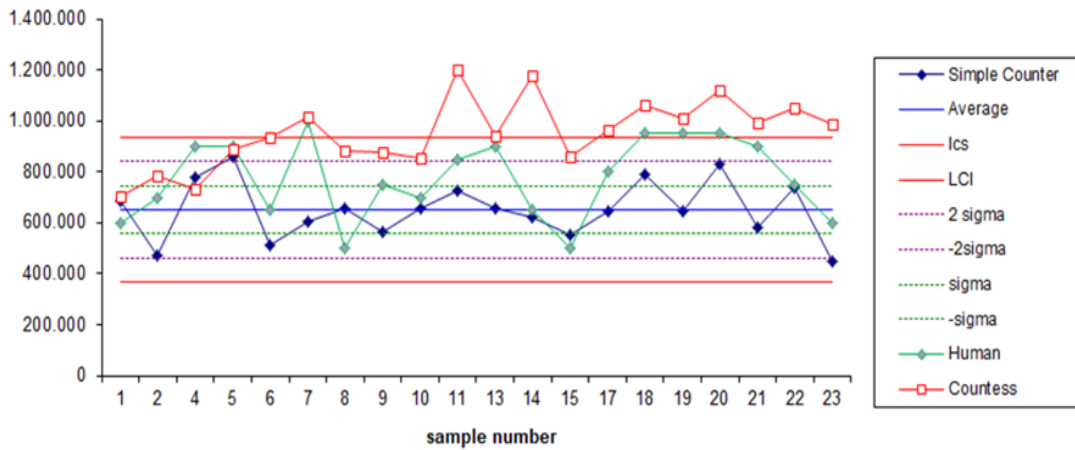


Figure 35 : Main indicators of cell counting distribution on THP-1 cell suspensions in the automated Micro Counter system as compared with human cell counting on Neubauer chamber and an automated counting systems (Countess). Data represent 20 measurements

Moreover, the histograms elaborated with the different cell counting data (Figure 36) showed a trend towards a Gaussian distribution, which is compatible with a random distribution of error.

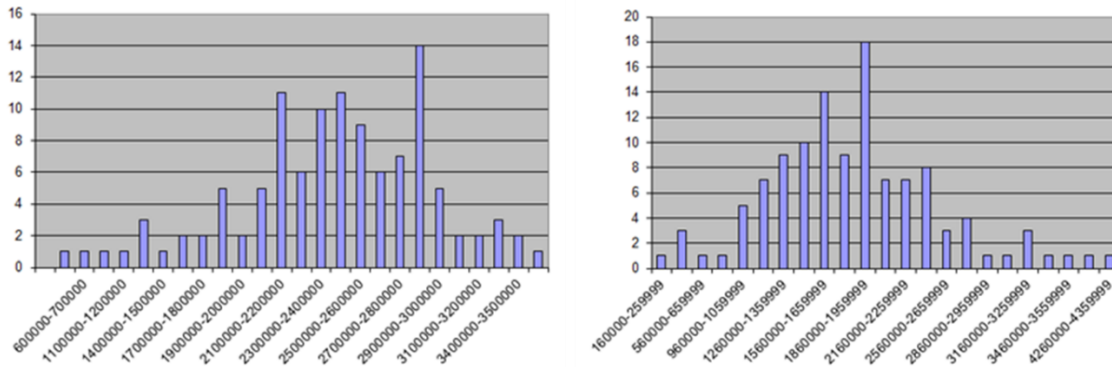


Figure 36 : Representative examples of cell distribution histograms of the replicated cell countings of THP-1 cell suspensions in the automated Micro Counter system.

In all the experiments conducted, whenever manual counting was performed on low concentration samples and a low number of cells were analyzed ($n < 100$), cell distribution showed significant higher concentration variability than automatic cell counters, as it was expected.

We then designed a series of experiments to assess from a statistical point of view the reproducibility and concordance between methods. We determined the concentration of two spores cultures at day 3 and 18 by Micro Counter as compared to two different manual counting procedures (Neubauer chamber and direct microscopic counting by eye on Petri dish) and to a flow cytometer (CyFlow Ploidy Analyzer, Partec) in three different counting protocols (unstained spores, Propidium-stained spores and side-scatter detected spores). These experiments were performed at the COMAV-UPV research center by comparing each method on microspore cultures after 3 and 18 days culture growth. The results of such experiments are summarized in [Figure 37](#).

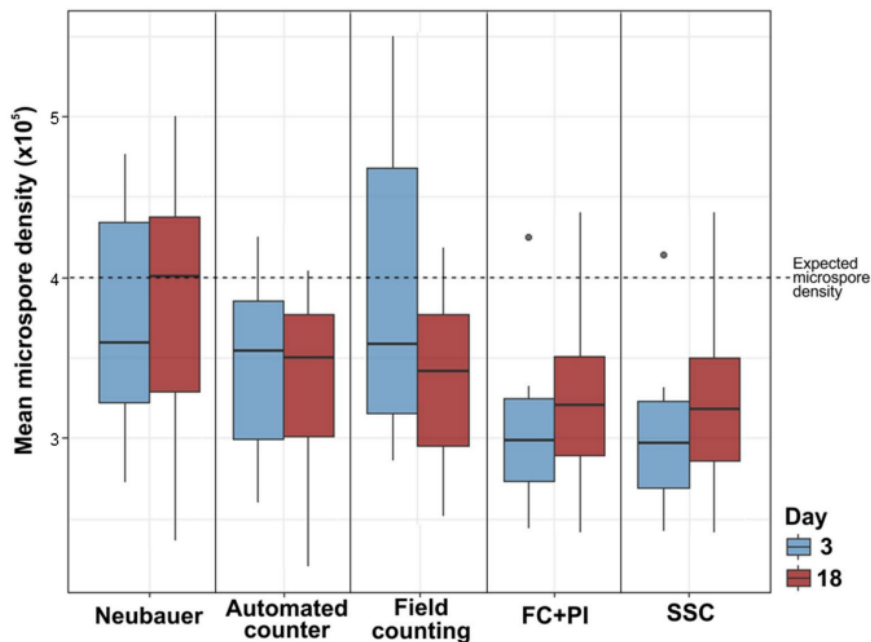


Figure 37 : Box-and-whiskers plots for mean densities of 17 different cultures measured at days 3 and 18. (Camacho-Fernández C et al., 2018)

The numerical results and the dispersion values for these experiments are presented in [Table 11](#)

Table 11: Mean and Standard deviation (first row), median and 1-3 quartile (second row) and percentage of deviation from the initial value and coefficient of variation (third row) of each counting method.

PI=Propidium Iodide, SCC = Side Scattered Light. (Camacho-Fernandez C et al., 2018)

Method	Days 3	Days 18
Neubauer	371.8 (65.2)	383.3 (79.3)
	359.4 (322.2–434.2)	400.8 (328.6–437.2)
	7.1/17.5%	4.2/20.7%
FC unstained	220.3 (90.1)	255.2 (108.4)
	197.8 (173.7–247.6)	255.2 (216.9–293.6)
	44.9/40.9%	36.2/42.5%
Flow cytometry + PI	309.0 (60.2)	332.0 (60.5)
	298.8 (273.3–324.6)	320.6 (289.2–350.4)
	22.8/19.5%	17.0/18.2%
Flow cytometry + SSC	305.4 (57.4)	329.6 (60.5)
	296.5 (269.2–323.3)	317.8 (285.9–350.2)
	23.7/18.8%	17.6/18.4%
Automated counter	349.2 (50.3)	335.5 (52.9)
	354.0 (299.0–385.5)	349.5 (301.2–376.5)
	12.7/14.4%	16.0/15.9%
Field counting	386.9 (81.5)	336.4 (53.9)
	358.1 (315.8–467.9)	341.4 (294.9–377.2)
	3.3/21.1%	15.9/16.0%

In order to assess from these data the reproducibility of the tested methods, counting differences between measurements were represented by Bland-Altman plots (Figure 38) and the coefficient of repeatability (CR) was calculated for each method, and shown in Table 12.

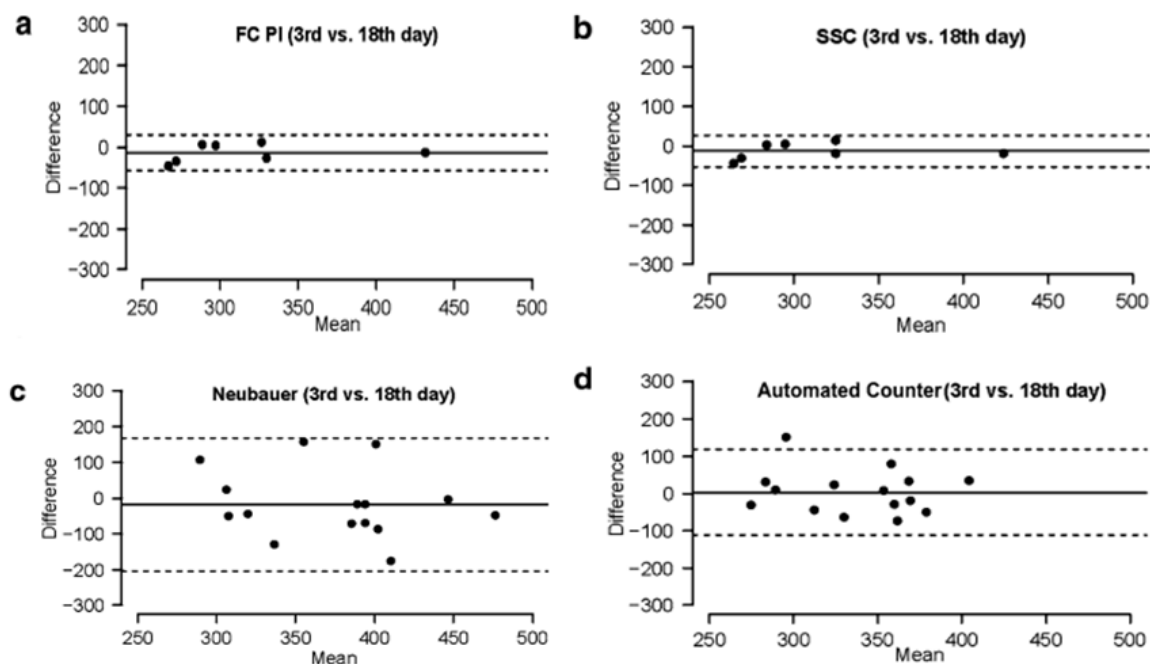


Figure 38: Bland-Altman comparisons of reproducibility of method by comparing 3 and 18-day culture. (Camacho-Fernández C et al., 2018)

Results showed that flow cytometry (FC PI & SSC) were the most reproducible methods, showing narrow limits of agreement and lower coefficient of repeatability (CR) values. The Neubauer method appeared as the less reproducible and the Micro Counter system being validated as moderately reproducible.

Table 12: Assessment of the repeatability and reproducibility of each of the methods tested, expressed by the coefficient of repeatability (CR) and p-value of ANOVA analysis. (Camacho-Fernandez C et al., 2018)

Method	CR (95% CI)	p-value ANOVA
Neubauer chamber	175.9 (100.6, 233)	0.4976
Flow cytometry + PI	39.2 (18.1, 53.6)	0.6597
Flow cytometry + SSC	36.3 (19.5, 50.8)	0.6652
Automated counter	109.6 (66.7, 156.8)	0.8236
Field counting	148.4 (113.3, 177.5)	0.1479

5.5.2 Validating the Automated Configuration of the Cell Counting System: Cell Confluence.

For assessing the determination of confluence with Culture Counter, flasks of growing HepG2 cell cultures were placed on an inverted microscope. For each culture, 20 photographs were taken for measurements in the Culture Counter (Figure 39).

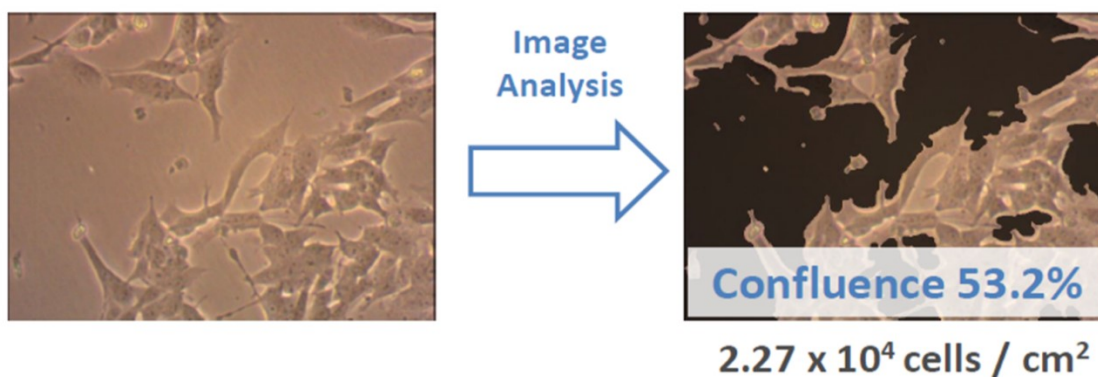


Figure 39: Schematic representation of the image analysis to determine cell confluence with the Culture Counter system.

Monolayers were trypsinized and cells resuspended in culture medium. For each suspension, cell concentration was measured with the same three independent systems indicated. Each measurement was repeated 20 times, shaking the original sample before sample extractions (Figure 40).

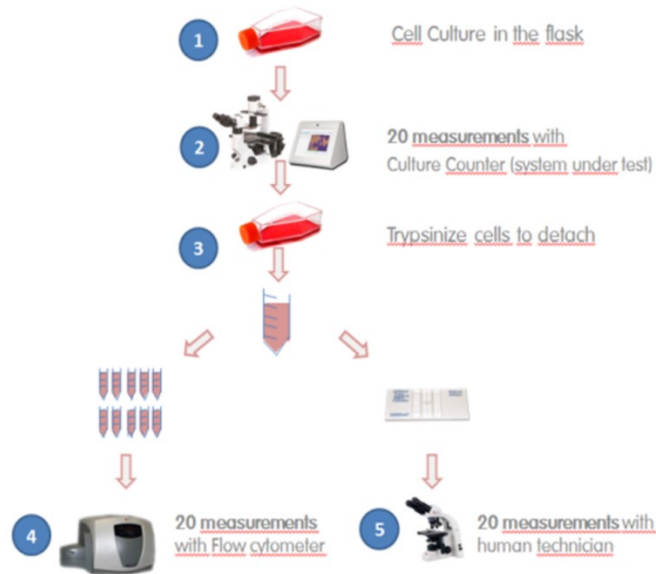


Figure 40 : Simplified representation of the adherent monolayer cell concentration estimation method to be used on flasks and petri dishes.

A high degree of correlation and linearity was found between the confluence measured by Culture Counter system and the actual cell counts per flask measured with three well established methods (Figure 41).

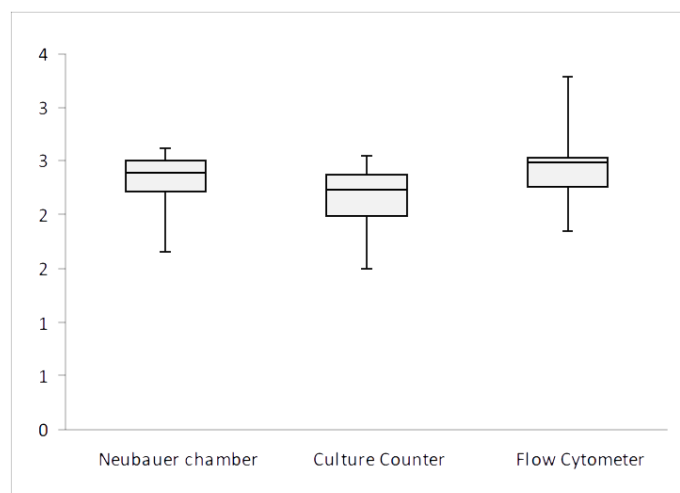


Figure 41 : Box-and-whiskers plots for cell concentration measured with the adherent monolayer cell concentration estimation method (Culture Counter), Flow Cytometer and Neubauer manual cell counting.

Reproducibility was assessed using the variation coefficient and found at a similar level that the flow cytometer and manual counting: 7.85% vs. 12,42% (Flow Cytometer) and 4.22% (Neubauer). Our results showed a cell concentration error of 7,01% (Table 13), slightly larger than the estimated with flow cytometry.

Table 13: Comparison between adherent monolayer cell concentration estimation method (Culture Counter), Flow cytometry and Manual-Neubauer cell concentration measurements. * Error when compared to average concentration of all Manual-Neubauer countings.

System	Mean (total flask cells)	Std.Dev.	Var.Coeff	Error(*)
Culture Counter	2.27×10^6	0.18×10^6	7.85%	7,01%
Flow Cytometer	2.46×10^6	0.30×10^6	12.42%	1,30%
Manual counting	2.44×10^6	0.10×10^6	4.22%	N/A

In addition, as seen in Figure 42, the repeated measurements performed with the Culture Counter system had a normal distribution.

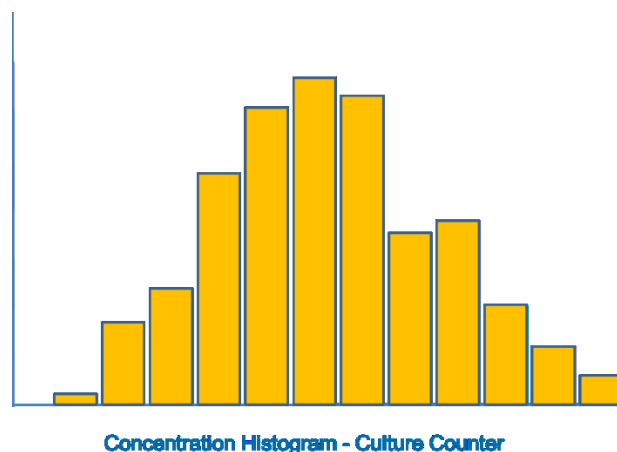


Figure 42 : Distribution of the repeated measurements performed with the Culture Counter System.

5.6 Automated Determination of Cell Concentration from Petri Culture Dishes.

Although Petri dishes are widely used for microbiology cultures and colony count, most researchers decline to measure cell concentration directly inside the Petri dish using an inverted microscope because they assume a non-random cell distribution that will eventually generate biased results.

A first step consisting in experimentally measuring the distribution of mammal cells (cell line 661W) and Flow-Count Fluorospheres (Beckman-Coulter) in suspension on standard plastic Petri dishes showed that in order to achieve acceptable cell concentration measurements with an error below 10%, more than 5,000 cells need to be sampled. This means that for a reference concentration of 1 million cells per ml, approximately 25 fields of the Petri dish need to be analyzed using a 10x lens.

In a second step, we performed an in-silico simulation based on data collected on the first step. We modeled a cell counting process simulating more than one million cell counting on the Petri dish using a mathematic model programmed in R. The main goal of these simulations was to determine how many microscopic fields of the Petri dish needed to be sampled in order to achieve a certain level of accuracy. Our results showed that in order to achieve similar error levels with a Neubauer chamber, where cells are distributed evenly, and only 400 cells need to be analyzed.

The applicability of our model requires the cell distribution across a Petri dish to be homogeneous. Thus, we utilized the new Culture counter system to analyze 16 sequential cross sections of Petri dishes in order to determine the degree of homogeneity in cell distribution profiles on a dish ([Figure 43](#)).

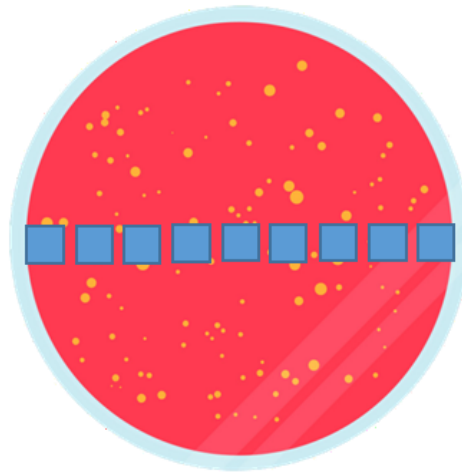


Figure 43: Design of cross-sectional distribution of microscope fields analyzed to determine cell- or particle distribution within a plastic Petri dish.

Figure 44 shows a representation of the typical profiles found: X-axis represents the microscopic field number, and Y-axis represents the number of cells found along a cross section of the dish. 12 of the 16 analyzed distributions presented a high degree of symmetry (panels A and B), three presented a low level of symmetry (panel D), and one presented no symmetry at all (C).

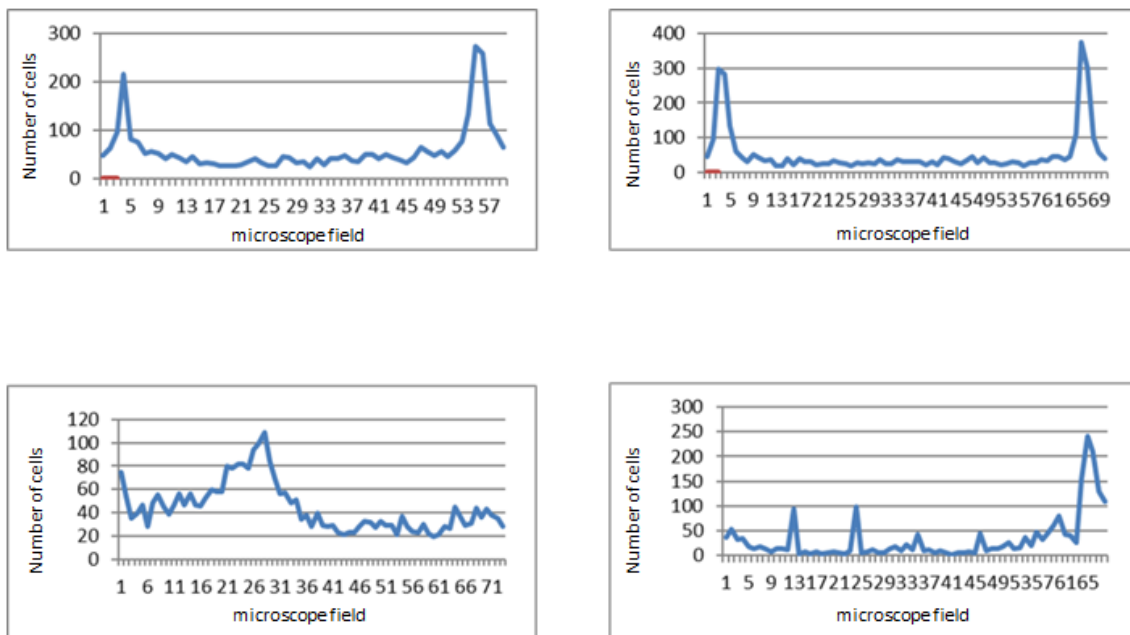


Figure 44 : Representative examples of particle distribution in sequential microscope fields examined with the Micr-Counter system. (A) Distribution of Flow Count microspheres. (B) Distribution of Jurkat cells. (C) Asymmetric distribution of HepG2 cells. (D) Distribution of tomato microspores.

We could not observe a reproducible pattern on cell distributions. Most of them had a very important degree of heterogeneity, with areas of low and high cell concentration. Typically, areas of high cell concentration accounted for 10-15% of the total fields analyzed and the estimated ratio of low- to high- cell concentration ranged from 1:5 to 1:20.

We also realized that the distribution with the highest heterogeneity corresponded to the 12 dishes where the cells accumulated near the periphery wall after being shaken or manually rotated. These types of distributions generally presented radial symmetry. Previous research has confirmed the higher cell concentration near the periphery on Petri dishes (Sandgren and Robinson, 1984).

Based on the above mentioned results, we selected a cell number function that mimicked the highest heterogeneity found on all cell distributions. The worst case would correspond to a distribution where most cells were concentrated around the periphery of the dish. To do so we implemented a mathematical model with radial symmetry and with two Gaussian distributions close to the borders of the dish, similar to Figure 44 (A) and (B) :

$$f(r) = 20 + 200 \times e^{-((r-25)^2/3.5)}$$

where r is the distance to the center of the dish. See Figure 46 for a 3D and 2D cross section representation of the chosen function.

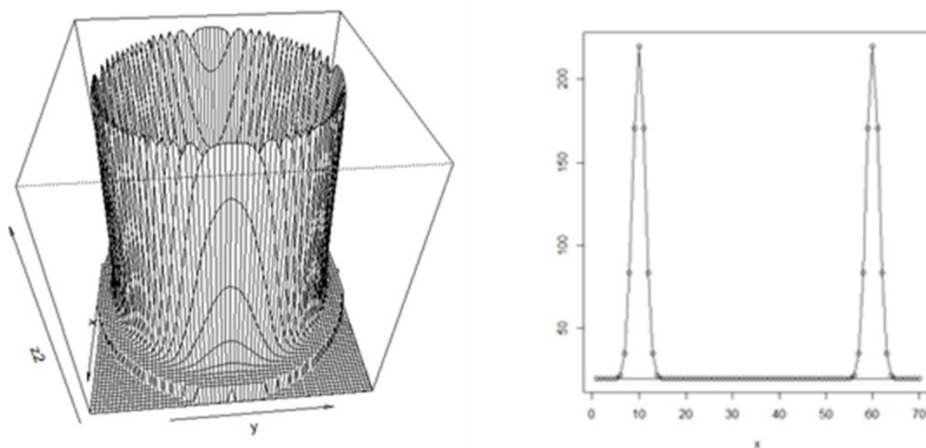


Figure 45: 3D representation of the numeric estimation of the cell distribution measured with the highest heterogeneity, according to previous experimental results (left). 2D cross section of the same distribution (right).

Once the cell number function was modeled into the simulator, we were able to run cell counting simulations.

For each time the numeric simulator was ran it provided:

- 1) **Total cell population data.** It numerically scanned the whole virtual Petri dish.
- 2) **Sampled data.** It simulated cells counting on random microscopic field-areas on the virtual disk.
- 3) **Cell analysis error.** The difference between the known real cell concentration of the cell population on the dish and the sampled data performed with our virtual cell counting.

As it was expected, the average sampling error is reduced when we increase the number of fields analyzed. (Table 14)

Table 14: Average error measured depending on the number of fields. 100.000 cell-counting simulations were performed for the determination of each ERROR figure.

Number of fields analyzed (20x lens)	Average ERROR
5	43,54%
10	29,97%
15	24,43%
50	13,43%
100	9,42%
200	6,70%
1000	2,95%
2000	2,07%

Our data show that it is possible to determine cell concentration directly from Petri dishes. However, due to the heterogeneous nature of cell distribution, a significant extra effort should be made in order to obtain accurate measurements when compared against the homogenous cell distribution.

6 Discussion

The first study in this Thesis is driven by the well-established fact that many of the *in vitro* experiments performed in the scientific community present numerous shortcomings. Some of them are inherent to the nature of the experiments, but others are related to the experimental design and the measurement procedures. The amount of data in high-impact journals has doubled over 20 years (Cordero et al., 2016), and basic-science papers are increasingly expected to include evidence of how results will translate to clinical applications. It must be kept on mind that the pressure on scientists to publish may collide with the need to verify results or to perform complementary experiments before publication. An article in a top journal may represent the work of years by several scientists. However, a low percentage of published experiments are reproducible (Amaral and Neves, 2021). As an example, the “Reproducibility Project: Cancer Biology” has so far managed to replicate the main findings in only 5 of 17 highly cited articles (Errington et al., 2014), and a replication of 21 social-sciences articles in *Science* and *Nature* had a success rate of between 57 and 67% (Camerer et al., 2018). The consequences of these deficiencies may be dramatic for science and for society.

We started from the hypothesis that a significant part of the deficiencies in the performance of cellular or molecular experiments with cells are due to methodological causes and that these deficiencies have significant impact of the experiments results. This fact contributes to the lack of quality of the experiments results, their reproducibility and impacts on the quality and reproducibility of scientific production.

With this in mind, our initial objective was to evaluate to which degree the statistical and mathematical criteria for data reporting in biomedical journals could be met by improving the performance of a basic (and critical) experimental procedure, as it is cell counting. For this, potential deficiencies and malpractices in cell-based assays should be detected and their impact on measurement systems quantified. Thereafter, a secondary objective was to propose new alternative methodologies that could minimize or eliminate the errors or biases detected while reducing the operations costs whenever possible (equipment, reagents, etc.). Such targeted methodologies should increase accuracy, reproducibility and robustness of the existing methods.

Initially, we analyzed what are the more than 700 scientific journal requirements, and analyzed the most critical and popular cell assays performed in laboratories. We conducted manual and automated cell counting experiments on Neubauer counting chambers to determine the main contributors to cell counting errors. Afterwards we designed several improved methodologies based on Artificial Intelligence applied to image analysis targeted to reduce the observed cell counting errors, and increase cell counting precision and reproducibility. Finally, we performed incremental improvement iterations over the proposed methodologies in order to increase the usability, and reduce the dependency from the user conducting the assay.

6.1 Statistical requirements of life sciences journals.

We have performed an in-depth analysis of 727 journals in seven different categories based on the amount of statistical criteria that authors should fulfill when submitting their manuscripts. Our results showed that journals from medical categories had a much higher probability to require specific statistical requirements than other life science and biology journals. Even a methodological category such as *Biochemical Methods* has a lower degree of statistical requirements than medical categories.

Our findings are in line with the reported data by two similar surveys, in 1985 (George, 1985) and 1998 (Goodman et al., 1998), aimed to evaluate systematically policies and practices of biomedical journals concerning statistical review. Goodman and colleagues reported that only 33% of 114 biomedical journals surveyed in 1998 demanded statistical review for all original research manuscripts, while additional 46% applied statistical review under editor discretion. Since then, concerns about poor statistical practices as a relevant factor in low reproducibility of research results have risen markedly. In spite of this reality, an expanded survey conducted recently by Hardwicke and Goodman (Hardwicke and Goodman, 2020) has revealed that these numbers have changed little since the survey of Goodman and colleagues (1998). In their recent survey, Hardwicke and Goodman received suitable responses from 107 of 364 (28%) journals surveyed, across 57 fields, mostly from editors-in-chief. According to this report, 34% (36/107) rarely or never use specialized statistical review, 34% (36/107) used it for 10–50% of their articles and only 23% used it for all articles.

It is widely acknowledged that most editors of life science journals are experts on their specific fields with a limited background of statistics knowledge (Harrington D et al.

2019). Top life sciences journals rely on external consultants in order to verify the statistical quality of the received manuscripts. This fact goes in line with our finding that most life science (non-medical) journals (75.05%) do not have any specific statistical requirements on their guides to authors and those journals that do, very often have non-specific or vague statistical requirements such as suggestions to consult a statistician.

However, according to the comments by Hardwicke and Goodman in their survey (Hardwicke and Goodman, 2020) most of editors considered statistical review an added value to regular peer review and were little concerned about potential increases in reviewing time, cost, and difficulty identifying suitable statistical reviewers.

According to Romano (2013) statistics represents the heel of Achilles for the modern researcher in life sciences and medicine. For example, Clarke (2011) analyzed the fifteen most common mistakes encountered in clinical research, where eight of them were directly associated with poor statistical reporting. Medicine often uses probabilistic statistics that are far from the scientific method. In 2018 the New England Journal of Medicine (NEJM), the highest prestige medicine journal according to IF, corrected five papers due to flawed statistics and retracted a sixth due to weak or flawed statistics (Couzin-Frankel J, 2018) manifesting a weak statistical review process. The crisis ended up with the NEJM changing its guides to authors. (Harrington et al. 2019).

In conclusion, our results point out the lack of specific reporting guidelines for authors in the majority of life science journals analyzed that could be related to lack of statistical expertise by editors or a weak statistical review protocol from the journal.

A potential way to complete this research could be to publish a set of guidelines for authors that gather all statistical recommendations from all journals analyzed, as described in [Appendix 9.2](#) that may serve a useful reference checklist for researchers. In fact, already in 1979, the group now known as the International Committee of Medical Journal Editors first published a set of uniform requirements for preparing manuscripts to be submitted to their own journals. These uniform requirements have been revised several times (International Committee of Medical Journal Editors, 1982), and have been widely adopted by other biomedical journals. In the 1988 revision (International Committee of Medical Journal Editors, 1988), the Committee added guidelines for presenting and writing about statistical aspects of research. The purpose of these

guidelines is to assist authors in reporting statistical aspects of their research in ways that will be responsive to the queries of editors and reviewers and helpful to readers. One very important point in these guidelines, which we have introduced as a main indicator of our cellular studies, as discussed later, is that findings should be quantified and presented with appropriate indicators of measurement error or uncertainty.

Having more specific statistical reporting guidelines to authors may contribute to increase the level of statistical knowledge of both editors and researchers, better define the journal statistical review process and increase the quality of the science generated. Since methodology sections in peer-reviewed articles not always provide all the critical data necessary to accurately reproduce results, efforts must be taken to help improve reproducibility and consistency. Minimum information guidelines for reporting experiments have found broad-based support across biological and technological domains (Taylor et al. 2008).

In this regard, and related to the cellular aspects of our studies, two such efforts are the development and use of the MIFlowCyt standard and sharing data using the FlowRepository. For flow cytometry data, The Minimum Information about a Flow Cytometry Experiment (MIFlowCyt) effort is now an approved International Society for the Advancement of Cytometry (ISAC) standard and has been adopted by journals, including Cytometry A. MIFlowCyt provides a checklist covering details including experimental overview, sample description, instrumentation, reagents, and data analysis. Almost all articles now published in Cytometry A follow this recommendation (Spidlen et al., 2012).

Data sharing is also widely recognized as critical by funders and journals including Nature, PLOS and NIH. The FlowRepository is primarily for sharing data associated with peer-reviewed publications annotated according to MIFlowCyt data annotation requirements. The FlowRepository operates under the auspices of ISAC with guidance provided by ICCS and ESCCA. Together MIFlowCyt and FlowRepository provide a mechanism for researchers to access, review, download, deposit, annotate, share and analyze flow cytometry datasets (Spidlen et al., 2012).

Other ways to assess the robustness of scientific findings may rely on synthesizing the published literature, drawing on results from studies by different research groups (Amaral and Neves, 2021). This is already so for most clinical guidelines, which are

typically derived from meta-analyses of existing evidences. A potentially better approach is to organize confirmatory experiments that are specifically designed to assess robustness and generalizability. These will ideally incorporate multiple methods and experimental models (such as mouse strains or cell types) in different laboratories. Coordination between groups can standardize data collection and guarantee access to results, thus facilitating synthesis and eliminating publication bias (Amaral and Neves, 2021). Diverse types of collaboration have been set up across various areas of science. Such initiatives are intensive in terms of cost and labor, and cannot be conducted for every published finding. Still, they are a more efficient way to confirm key phenomena than waiting for data to accrue from uncoordinated efforts. Moreover, investing effort to increase rigor in selected confirmatory projects is probably more feasible than demanding that every biomedical publication be replicable, generalizable and clinically relevant bias (Amaral and Neves, 2021).

6.2 Limitations and Error Sources of Current Cell Counting Methods.

A preliminary survey that we conducted among laboratory technicians of the Valencian area, confirmed that Neubauer and other cell counting chambers continue to be the most popular cell counting device in clinical research laboratories, due to its low cost, flexibility and portability.

Flow cytometry, the second most popular counting methodology in our survey developed two decades after Coulter technology thanks to advances in the fields of optics and fluorochrome discovery, and became rapidly spread thanks to its ability to distinguish cells based on multiple parameters, speed and precision (Vembadi, 2019).

Finally, with the introduction of image-based cell counting devices based on Artificial Intelligence, the automatic cell counting systems have become more affordable and easy to use than flow cytometers and faster and more convenient to operate than cell counting chambers. They represent a midpoint between manual cell counting and flow cytometry and proved to be suitable for many clinical research applications, since they lack IVD validation.

JWG Lund stated as early as 1958 that “There is no one method of estimate cell number which is the best under all circumstances and for all purposes”. In concordance with this consideration, we made clear that the Neubauer chamber, considered as the golden

standard testing system, did present some drawbacks and limitations, although it was very useful for the initial testing of the cell counting methodology. According to our experiments, uneven distribution of cells in a Neubauer counting chamber can introduce errors as high as 50%. On the other hand, existing automated cell counting systems based on image analysis can introduce errors of up to 30%-40% for low cell concentrations (10^4 cells/mL) and up to 5-10% for higher cell concentrations (10^6 cells/mL). The best performance in terms of precision of accuracy was shown by flow cytometers, which usually gave error below 10% in cell counting.

As expected, the actual performance of the different counting methods was worse than their theoretical performance that we had previously calculated in terms of the maximal error, which depend only on the number of cells examined in each sampling. The main causes of error identified when using the Neubauer chamber were imperfections of the cell counting chamber, especially in low-cost, unbranded devices, which led to wrong assumption of sample volume for calculation low volume of sample analyzed. Operator malpractices in pipetting and chamber loading, and cell clumping phenomena had a negative impact in the accuracy and precision of measurements with Neubauer chambers. Since this device was used as a reference in several experiments, its errors were also added to the system under comparison, therefore making virtually impossible to find what the true concentration of cells was in a given suspension.

The main goal of our counting experiments was to determine when and under which circumstances uneven distribution of cells on the Neubauer chamber happened. In this experiment the dependent variable was the homogeneity of cell distribution on the chamber. For the cell counting chamber being used accurately, cell concentration should be homogenous. In order to maximize the number of experiments performed and variables tested we focused on distribution effects which could be detected visually. When a heterogeneous cell distribution was visually detected, it was later quantified by automated cell counting devices. This strategy saved us considerable counting time, and helped to change the external factors until clear effects on cell distribution were achieved.

From a qualitative point of view, in the course of our comparative cell counting experiments we could describe the main deficiencies identified in each type of cell counting strategy that we evaluated, as summarized here:

6.2.1 Limitations of manual counting with Neubauer chambers:

The main drawbacks that we could detect were:

- a) The number of cells counted is often too low for accurate cell counting.
- b) Low-cost chambers do not meet volume specifications.
- c) Most loading chamber protocols do not guarantee homogenous cell distribution.
- d) The technique requires time and effort of laboratory personnel.

Starting cell concentration and the number of cells counted per field are frequent issues for low reproducibility. In all the experiments conducted, whenever manual counting was performed on low-concentration samples and/or a low number of cells were analyzed ($n < 100$), cell distribution showed significant higher concentration variability than automatic cell counters, as it was expected.

We consider also a crucial issue with cell counting chambers their nominal volume. Based on our results, low cost Neubauer chambers should not be used for applications where a specific concentration of cells is required, since they can introduce errors higher than 100% due to its lack of volume calibration. In cases where a precise cell concentration is required, it is necessary to use IVD / calibrated cell counting chambers. Even the reputed German manufacturer Marienfeld provides cell counting chambers that fall out of the technical specifications provided and introduce higher errors than expected on cell concentration measurements.

However, these chambers can be used for applications where introducing concentration error at the beginning and the end of the experiment is not relevant, such as determining a given effect to cells, provided the same counting chamber is used or all measurements.

Pipetting and chamber-loading malpractices have been shown by us to be important sources of errors, as they introduce heterogeneity on cell distribution in cell counting chambers. Users should avoid keeping cells waiting too long before loading the chamber or loaded into the micropipette; mixing cells with a small micropipette volume (1-10 μl); unloading the pipette in a non-continuous way or using a low quality cell counting chamber. Some of these effects can be detected easily if the introduced heterogeneity in cell distribution may be acknowledged visually. Smaller effects and distortions may be

introduced easily with other practices investigated by us that remain undetected. In this way, many laboratories could be systematically introducing unwanted errors to their research based on these potential flaws.

6.2.2 Limitations in automated counting with Flow Cytometers:

The main drawbacks that we could detect were:

- a) Flow cytometers are costly instruments that require expensive maintenance.
- b) The instruments are difficult to configure and to use routinely
- c) Fluorescent reagents are typically required.
- d) Measured cells are not visualized.
- e) Reproducibility may be heavily affected by rapid changes in cell health.

This last biological limitation of flow cytometry should be well kept on mind. Indeed, in some of our comparative studies, when cells were left on purpose out of incubators for long periods of time (>1,2 hours) the cells suffered major degradation, and the accuracy of the flow cytometer was seriously compromised. In those cases, flow cytometers could introduce errors as high as 50%.

6.2.3 Limitations in automated counting with Image-based counters:

The main drawbacks that we could detect were:

- a) These systems tend to loss accuracy at low cell concentrations, below 10^5 cells/mL)
- b) The size range of cells or particles that can be analyzed is limited.
- c) There is lack of transparency of the counting process.
- d) This technique has additional costs because of using disposable chambers or tips.

We have determined that most image based cell counting systems are not suitable for research work with sample concentration below 500,000 cells / mL. In order to obtain reliable measurements in that range, the only valid systems are those that analyze samples higher than 4 μ L.

6.3 Performance and Limitations of Innovative image-based AI driven technology.

In this Thesis, beyond identifying and quantifying the main types of errors introduced on existing cell counting assays, we have proposed an improved methodology that can be used in most scientific laboratories working with cells. The process of validation of such new automated cell counting systems was aimed to determine the error introduced by the systems and to decide if this error was acceptable for a cell-counting instrument in research and clinical practice applications.

With the help of our improved methodology based on microscope-image analysis by Artificial Intelligence we have been able to maintain the measurement error below the 5% even for low cell concentration. We have also demonstrated that this methodology can be implemented with automated systems that contribute even further to the quality and reproducibility of the results and that can be used in both research and clinical environments.

Two innovative image-based AI-driven cell counting system have been developed and tested: one for cell suspensions and another one for adherent monolayer cells:

a) Micro Counter system: It involves an upright optical microscope and AI algorithms for automated cell counting of cells or particles in suspension.

b) Culture Counter: It involves an inverted microscope and AI algorithms for calculating cell densities from cell monolayers directly on culture transparent supports.

In our opinion, further improvements of the methods presented could be achieved with full automation of the sample focus and microscopic stage positioning so that less user operation is required, and higher reproducibility is achieved.

6.3.1 Performance and Limitations of Micro Counter.

Our results have shown that the Micro Counter is more accurate and precise than traditional Neubauer chambers for suspension cells, but less precise than flow cytometers. Another drawback pointed out was the limitation of analyzing only the focus plane. All cells that are not completely in focus may be ignored eventually by the system therefore introducing a bias in the measurement.

One of the main limitations of image-based automated systems is the limited field of view (FOV), which depends on the magnification and it is usually in the range of hundreds of micrometers (Green and Wachsmann-Hogiu, 2015). However, the Micro Counter system can work with several sets of microscopic lenses, increasing the cell counting range from 1-1000 μm , while most image-based cell counters are suitable for cells and particles ranging from 4 to 25 μm . This capability makes it also suitable for enumeration of bacteria, algae, and large cells such as adipocytes or hepatocytes.

Our results have shown that it is possible to determine cell concentration directly from Petri dishes with the Micro Counter. However, due to the heterogeneous cell distribution observed in some cases (and mathematically modeled by us), a significant extra effort should be made in order to obtain accurate measurements. Camacho-Fernández and coworkers reported automated counting on Petri dish to be at *least as correct as human observations*, but pointed out that this could be due to the fact that microspores and fluorospheres do not distribute homogeneously in the culture dish (Camacho-Fernández et al., 2018). Thus, in order to achieve acceptable cell concentration measurements with an error below 10%, more than 5,000 cells need to be sampled. This means that for a reference concentration of 1 million cells per ml, approximately 25 fields of the Petri dish need to be analyzed using a 10x lens. In order to achieve similar error levels with a Neubauer chamber, if cells are distributed evenly, only 400 cells need to be analyzed.

Additional advantages of the automated suspension cells system is that it does not need any reagent or consumable material to be operated, thus reducing maintenance and operating costs. Once the system is installed and calibrated, it is straightforward and easy to use. In fact, after its initial setup in this Thesis, the Micro Counter has been tested in more than 100 laboratories over the world. Its unique AI image analysis engine is robust enough to work under any type of microscope, lens or lighting system.

On the downside, the proposed system is often seen as more user-dependent and complex to operate than other image-based fully-automated cell counting systems.

6.3.2 Performance and Limitations of Culture Counter.

The Culture Counter for adherent monolayer cell counting has been also tested in more than 10 laboratories with successful results. However, adoption of these kind of devices in regular laboratory practice has proved to be harder than regular image-based

counters since it requires a complete change in the habits of researchers that are not used to perform experiments measuring directly in the culture medium, and the usage of an inverted microscope.

This new methodology can be especially useful when the cell culture should not be compromised for performing the measurement. Also, when implemented in a fully automated system, where time and effort required to analyze large amounts of cells are no longer a significant issue.

7 Conclusions

1. Among the 727 Biomedical journals analyzed, Medical journals have significantly higher number of statistical requirements than non-Medical journals. Within non-Medical journals, those ranged in Q1 and Q2 quartiles have a higher number of statistical requirements than those in Q3 and Q4 quartiles.
2. Following a survey among laboratory technicians, the most popular cell counting methods resulted to be manual cell counting on Neubauer-type devices, flow cytometry and image-based automated cell counting, in that order.
3. Manual counting of cell suspensions in Neubauer chambers or in Petri dishes may lead to significant errors, because of wrong volume specifications, insufficient number of cells scored per field or heterogeneous cell distribution per microscope field.
4. We have designed and built up two innovative cell counting systems based on Artificial Intelligence algorithms for automated analysis of microscope images of cells in suspension (the Micro Counter) or in monolayer cultures (the Culture Counter).
5. The Micro Counter enhances accuracy and reproducibility over other image-based procedures, by increasing the number of microscope fields and cells analyzed. However, it has less reproducibility and precision than flow cytometers.
6. The Culture Counter allows precise and reproducible measures of cell concentration directly on Petri dishes and culture flasks, without needing to disrupt the culture matrix.

8 References

- Adams F**, *The Genuine Works of Hippocrates*, William Wood and Company. 1891.
- Amaral OB, Neves K**. Reproducibility: expect less of the scientific paper. *Nature*. 2021;597(7876):329-331. doi:[10.1038/d41586-021-02486-7](https://doi.org/10.1038/d41586-021-02486-7)
- Anonymous**. Error prone. *Nature*. 2012;487(7408):406-406. doi:[10.1038/487406a](https://doi.org/10.1038/487406a)
- Anonymous**. Repetitive flaws. *Nature*. 2016;529(7586):256-256. doi:[10.1038/529256a](https://doi.org/10.1038/529256a).
- Anzar M**, Kroetsch T, Buhr MM. Comparison of different methods for assessment of sperm concentration and membrane integrity with bull semen. *J Androl*. 2009;30(6):661-668. doi:[10.2164/jandrol.108.007500](https://doi.org/10.2164/jandrol.108.007500)
- Baker M**. Statisticians issue warning over misuse of P values. *Nature* 531, 151. (10 March 2016) doi:[10.1038/nature.2016.19503](https://doi.org/10.1038/nature.2016.19503)
- Barona JL**, Bernard C, *Antología*. Península; 1989.
- Begley CG**, Ellis LM. Raise standards for preclinical cancer research. *Nature*. 2012;483(7391):531-533. doi:[10.1038/483531a](https://doi.org/10.1038/483531a)
- Bland JM, Altman DG**. Measuring agreement in method comparison studies. *Stat Methods Med Res*. 1999;8(2):135-160. doi:[10.1177/096228029900800204](https://doi.org/10.1177/096228029900800204)
- Bland JM, Altman DG** . Statistical Methods for Assessing Agreement Between Two Methods of Clinical Measurement. (1986) *Lancet*, I, 307-310
- Briganti G, Le Moine O**. Artificial Intelligence in Medicine: Today and Tomorrow. *Frontiers in Medicine*. 2020;7:27. doi:[10.3389/fmed.2020.00027](https://doi.org/10.3389/fmed.2020.00027)
- Bynum W**, *The history of Medicine: A very Short Introduction*, OUP Oxford, 2008
- Camacho-Fernández C**, Hervás D, Rivas-Sendra A, Marín MP, Seguí-Simarro JM. Comparison of six different methods to calculate cell densities. *Plant Methods*. 2018;14:30. doi:[10.1186/s13007-018-0297-4](https://doi.org/10.1186/s13007-018-0297-4)

- Capobianco E.** High-dimensional role of AI and machine learning in cancer research. *Br J Cancer*. Published online January 10, 2022. doi:[10.1038/s41416-021-01689-z](https://doi.org/10.1038/s41416-021-01689-z)
- Camerer CF, Dreber A, Holzmeister F, et al.** Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015. *Nat Hum Behav*. 2018;2(9):637-644. doi:[10.1038/s41562-018-0399-z](https://doi.org/10.1038/s41562-018-0399-z)
- Clark GT, Mulligan R.** Fifteen common mistakes encountered in clinical research. *J Prosthodont Res*. 2011;55(1):1-6. doi:[10.1016/j.jpor.2010.09.002](https://doi.org/10.1016/j.jpor.2010.09.002)
- Collins CE, Young NA, Flaherty DK, Airey DC, Kaas JH.** A Rapid and Reliable Method of Counting Neurons and Other Cells in Brain Tissue: A Comparison of Flow Cytometry and Manual Counting Methods. *Front Neuroanat*. 2010;4:5. doi:[10.3389/neuro.05.005.2010](https://doi.org/10.3389/neuro.05.005.2010)
- Cordero RJB, León-Rodríguez CM de, Alvarado-Torres JK, Rodríguez AR, Casadevall A.** Life Science's Average Publishable Unit (APU) Has Increased over the Past Two Decades. *PLOS ONE*. 2016;11(6):e0156983. doi:[10.1371/journal.pone.0156983](https://doi.org/10.1371/journal.pone.0156983)
- Couzin-Frankel J.** Journals under the microscope. *Science*. Published online September 21, 2018. doi:[10.1126/science.361.6408.1180](https://doi.org/10.1126/science.361.6408.1180)
- Couzin-Frankel J.** Following charges of flawed statistics, major medical journal sets the record straight, 2018. Accessed January 9, 2022. doi: [10.1126/science.aau4689](https://doi.org/10.1126/science.aau4689)
- Dung G.** Statistical Evaluation of measurement errors: Design and analysis of reliability studies. 2nd ed. London: Arnold, 2004.
- Errington TM, Iorns E, Gunn W, Tan FE, Lomax J, Nosek BA.** An open investigation of the reproducibility of cancer biology research. Rodgers P, ed. *eLife*. 2014;3:e04333. doi:[10.7554/eLife.04333](https://doi.org/10.7554/eLife.04333)
- Fernandes-Taylor S, Hyun JK, Reeder RN, Harris AH.** Common statistical and research design problems in manuscripts submitted to high-impact medical journals. *BMC Research Notes*. 2011;4(1):304. doi:[10.1186/1756-0500-4-304](https://doi.org/10.1186/1756-0500-4-304)
- García-Armesto MR, Prieto M, García-López ML, Otero A, Moreno B.** Modern microbiological methods for foods: colony count and direct count methods. A review. *Microbiologia*. 1993;9(1):1-13.

- George SL.** Statistics in medical journals: A survey of current policies and proposals for editors. *Medical and Pediatric Oncology*. 1985;13(2):109-112.
doi:[10.1002/mpo.2950130215](https://doi.org/10.1002/mpo.2950130215)
- Goodman SN, Altman DG, George SL.** Statistical reviewing policies of medical journals. *J GEN INTERN MED*. 1998;13(11):753-756. doi:[10.1046/j.1525-1497.1998.00227.x](https://doi.org/10.1046/j.1525-1497.1998.00227.x)
- Green R, Wachsmann-Hogiu S.** Development, History, and Future of Automated Cell Counters. *Clinics in laboratory medicine*. 2015;35:1-10.
doi:[10.1016/j.cll.2014.11.003](https://doi.org/10.1016/j.cll.2014.11.003)
- Hardwicke TE, Goodman SN.** How often do leading biomedical journals use statistical experts to evaluate statistical methods? The results of a survey. *PLOS ONE*. 2020;15(10):e0239598. doi:[10.1371/journal.pone.0239598](https://doi.org/10.1371/journal.pone.0239598)
- Harrington D, D'Agostino RB, Gatsonis C, et al.** New Guidelines for Statistical Reporting in the Journal. *New England Journal of Medicine*. 2019;381(3):285-286.
doi:[10.1056/NEJMe1906559](https://doi.org/10.1056/NEJMe1906559)
- Henry JP.** Génétique et origine d'Homo sapiens. *Med Sci (Paris)*. 2019;35(1):39-45.
doi:[10.1051/medsci/2018311](https://doi.org/10.1051/medsci/2018311)
- Huang LC, Lin W, Yagami M, et al.** Validation of cell density and viability assays using Cedex automated cell counter. *Biologicals*. 2010;38(3):393-400.
doi:[10.1016/j.biologicals.2010.01.009](https://doi.org/10.1016/j.biologicals.2010.01.009)
- İnce FD, Ellidağ HY, Koseoğlu M, Şimşek N, Yalçın H, Zengin MO.** The comparison of automated urine analyzers with manual microscopic examination for urinalysis automated urine analyzers and manual urinalysis. *Practical Laboratory Medicine*. 2016;5:14-20. doi:[10.1016/j.plabm.2016.03.002](https://doi.org/10.1016/j.plabm.2016.03.002)
- International Committee of Medical Journal Editors.** Uniform requirements for manuscripts submitted to biomedical journals. *Ann Intern Med*. 1982;96(6 Pt 1):76671.
- International Committee of Medical Journal Editors.** Uniform requirements for manuscripts submitted to biomedical journals. *Ann Intern Med*. 1988;108:25865.

International Standard Organization (2022a) <https://www.iso.org/standard/68879.html>, last accessed 15 January 2022

International Standard Organization (2022b) <https://www.iso.org/standard/67892.html> last accessed 15 January 2022

Jones W.H.S., Hippocrates Collected Works I, *Cambridge*: Harvard University Press, 2006.

Joseph GB, McCulloch CE, Sohn JH, Pedoia V, Majumdar S, Link TM. AI MSK clinical applications: cartilage and osteoarthritis. *Skeletal Radiol.* 2022;51(2):331-343. doi:[10.1007/s00256-021-03909-2](https://doi.org/10.1007/s00256-021-03909-2)

Journal Impact Factor - Journal Citation Reports. Web of Science Group. Accessed January 8, 2022. <https://clarivate.com/webofsciencegroup/solutions/journal-citation-reports/>

Kolluri S, Lin J, Liu R, Zhang Y, Zhang W. Machine Learning and Artificial Intelligence in Pharmaceutical Research and Development: a Review. *AAPS J.* 2022;24(1):19. doi:[10.1208/s12248-021-00644-3](https://doi.org/10.1208/s12248-021-00644-3)

Li Z, Cui Z. Three-dimensional perfused cell culture. *Biotechnol Adv.* 2014;32(2):243-254. doi:[10.1016/j.biotechadv.2013.10.006](https://doi.org/10.1016/j.biotechadv.2013.10.006)

Litjens G, Kooi T, Bejnordi BE, et al. A survey on deep learning in medical image analysis. *Med Image Anal.* 2017;42:60-88. doi:[10.1016/j.media.2017.07.005](https://doi.org/10.1016/j.media.2017.07.005)

Lund JWG, Kipling C, Le Cren ED. The inverted microscope method of estimating algal numbers and the statistical basis of estimations by counting. *Hydrobiologia.* 1958;11(2):143-170. doi:[10.1007/BF00007865](https://doi.org/10.1007/BF00007865)

Morillon M, Maslin J, De Pina JJ, Louis FJ, Martet G. Evaluation of the monocyte counting by the ABX Vega. Comparison with the manual method and fluoro-flow cytometry. *Hematol Cell Ther.* 1999;41(2):47-50. doi:[10.1007/s00282-999-0047-1](https://doi.org/10.1007/s00282-999-0047-1)

Munawar HS, Mojtahedi M, Hammad AWA, Kouzani A, Mahmud MAP. Disruptive technologies as a solution for disaster risk management: A review. *Sci Total Environ.* 2022;806(Pt 3):151351. doi:[10.1016/j.scitotenv.2021.151351](https://doi.org/10.1016/j.scitotenv.2021.151351)

Pritsker M. Why Less Than 30% of Science Articles are Reproducible. JoVE. Published May 3, 2012. Accessed January 7, 2022. <https://www.jove.com/blog/scientist->

[blog/studies-show-only-10-of-published-science-articles-are-reproducible-what-is-happening/](https://www.nature.com/news/blog/studies-show-only-10-of-published-science-articles-are-reproducible-what-is-happening/)

R development Core Team. An Introduction to R 2004. (First Edition) ISBN 0954161742.

Rajpurkar P, Chen E, Banerjee O, Topol EJ. AI in health and medicine. *Nat Med*. Published online January 20, 2022:1-8. doi:[10.1038/s41591-021-01614-0](https://doi.org/10.1038/s41591-021-01614-0)

Romano R, Gambale E. Statistics and Medicine: the Indispensable Know-How of the Researcher. *Transl Med UniSa*. 2013;5:28-31.

Sandgren CD, Robinson JV. A stratified sampling approach to compensating for non-random sedimentation of phytoplankton cells in inverted microscope settling chambers. *British Phycological Journal*. 1984;19(1):67-72. doi:[10.1080/00071618400650071](https://doi.org/10.1080/00071618400650071)

Sarewitz D. The pressure to publish pushes down quality. *Nature*. 2016;533(7602):147-147. doi:[10.1038/533147a](https://doi.org/10.1038/533147a)

Spidlen J, Breuer K, Brinkman R. Preparing a Minimum Information about a Flow Cytometry Experiment (MIFlowCyt) compliant manuscript using the International Society for Advancement of Cytometry (ISAC) FCS file repository (FlowRepository.org). *Curr Protoc Cytom*. 2012;Chapter 10:Unit 10.18. doi:[10.1002/0471142956.cy1018s61](https://doi.org/10.1002/0471142956.cy1018s61)

Spikins P, Needham A, Tilley L, Hitchens G. Calculated or caring? Neanderthal healthcare in social context. *World Archaeology*. 2018;50(3):384-403. doi:[10.1080/00438243.2018.1433060](https://doi.org/10.1080/00438243.2018.1433060)

Stiff KM, Franklin MJ, Zhou Y, Madabhushi A, Knackstedt TJ. Artificial Intelligence and Melanoma: A Comprehensive Review of Clinical, Dermoscopic, and Histologic Applications. *Pigment Cell Melanoma Res*. Published online January 17, 2022. doi:[10.1111/pcmr.13027](https://doi.org/10.1111/pcmr.13027)

Streiner DL, Norman GR. "Precision" and "Accuracy": Two Terms That Are Neither. *Journal of Clinical Epidemiology*. 2006;59(4):327-330. doi:[10.1016/j.jclinepi.2005.09.005](https://doi.org/10.1016/j.jclinepi.2005.09.005)

- Suresh K.** An overview of randomization techniques: An unbiased assessment of outcome in clinical research. *J Hum Reprod Sci.* 2011;4(1):8-11.
doi:[10.4103/0974-1208.82352](https://doi.org/10.4103/0974-1208.82352)
- Taylor K,** Gordon N, Langley G, Higgins W. Estimates for worldwide laboratory animal use in 2005. *Altern Lab Anim.* 2008;36(3):327-342.
doi:[10.1177/026119290803600310](https://doi.org/10.1177/026119290803600310)
- Tian J,** Song X, Wang Y, et al. Regulatory perspectives of combination products. *Bioact Mater.* 2021;10:492-503. doi:[10.1016/j.bioactmat.2021.09.002](https://doi.org/10.1016/j.bioactmat.2021.09.002)
- Trisilowati,** Mallet DG. In silico experimental modeling of cancer treatment. *ISRN Oncol.* 2012;2012:828701. doi:[10.5402/2012/828701](https://doi.org/10.5402/2012/828701)
- Uniform requirements** for manuscripts submitted to biomedical journals. International Committee of Medical Journal Editors. *Ann Intern Med.* 1982;96(6 Pt 1):766-771.
doi:[10.7326/0003-4819-96-6-766](https://doi.org/10.7326/0003-4819-96-6-766)
- Uniform requirements** for manuscripts submitted to biomedical journals. International Committee of Medical Journal Editors. *Br Med J (Clin Res Ed).* 1988;296(6619):401-405.
- Vaux D.** Know when your numbers are significant | Nature. 492, 180-181 (2012).
<https://doi.org/10.1038/492180a>
- Vembadi A,** Menachery A, Qasaimh MA. Cell Cytometry: Review and Perspective on Biotechnological Advances. *Frontiers in Bioengineering and Biotechnology.* 2019;7:147. doi:[10.3389/fbioe.2019.00147](https://doi.org/10.3389/fbioe.2019.00147)
- Verso ML.** Some nineteenth-century pioneers of haematology. *Med Hist.* 1971;15(1):55-67.
- Xia X,** Hu J, Wang Y, Zhang L, Liu Z. Graph-based generative models for de Novo drug design. *Drug Discov Today Technol.* 2019;32-33:45-53.
doi:[10.1016/j.ddtec.2020.11.004](https://doi.org/10.1016/j.ddtec.2020.11.004)
- Young EWK,** Beebe DJ. Fundamentals of microfluidic cell culture in controlled microenvironments. *Chem Soc Rev.* 2010;39(3):1036-1048.
doi:[10.1039/b909900j](https://doi.org/10.1039/b909900j)

Zhavoronkov A, Ivanenkov YA, Aliper A, et al. Deep learning enables rapid identification of potent DDR1 kinase inhibitors. *Nat Biotechnol.* 2019;37(9):1038-1040.
doi:[10.1038/s41587-019-0224-x](https://doi.org/10.1038/s41587-019-0224-x)

PART 2

Artificial Intelligence Methods for Diagnosis of Genetic Disorders

"I'm fascinated by the idea that genetics is digital. A gene is a long sequence of coded letters, like computer information. Modern biology is becoming very much a branch of information technology."

Richard Dawkins

"Genetics is where we come from. It's deeply natural to want to know."

Ellen Ullman.

Abstract

(part 2)

In the last 15 years, genetics has undergone dizzying progress thanks to DNA sequencing. Thanks in part to these advances cancer mortality has been falling at a rate of 0.94% per year in Spain. New methodologies such as DNA microarrays and Next Generation Sequencing have contributed largely to the affordability of genetic testing.

Our starting hypothesis is that genetic analysis methodologies used in clinical practice still present some limitations and room for improvement, which can derive in patients receiving suboptimal treatment in specific cases. These methodologies could be optimized by new techniques based on automation and artificial intelligence. Our objective was to analyze specific deficiencies and propose alternative methodologies to increase accuracy, specificity and robustness.

In a first stage, we developed a methodology based on DNA microarrays coupled with an image-based analysis system for metastatic Colorectal Cancer prognosis. The method was targeted to reduce complexity and costs of existing systems. After its implementation it was compared to the Cobas System (Roche) and NGS analysis along with Sophia Genetics variant analysis interpreter. In a second stage, we developed a new variant interpretation methodology to improve the tertiary analysis of the standard NGS analysis workflow. The proposed methodology was based on current ACMG variant interpretation guidelines, and developed using Python under a Linux operating system. The system was tested against manual geneticists' analysis and against Agilent's Cartagenia/Alissa.

In our microarray system, several subsystems were successfully developed (automated stage, lightning system, autofocus, microarray positioning, etc.). However, overall system performance was determined not sufficient for clinical applications. We estimate that the NGS variant interpretation methodology developed could increase diagnosis yield by 5-10% (at the expense of decrease specificity) reducing geneticists analysis time by 50%-80%. We considered that the proposed methodology could improve diagnostic yield making it more suitable for preventive medicine. In diagnostic applications the new method could increase results reproducibility and reduce the number of suboptimal reports generated by geneticists, and ultimately to save some patients' lives.

1 Introduction

In the last 15 years, genetics has undergone dizzying progress thanks to DNA sequencing. In 2003, the first human DNA was sequenced in its entirety for the first time with the Human Genome Project. (Lander ES, 2001; Venter JC, 2001)

Today we know the genes that cause some 3,000 diseases, and we have the tools to diagnose them. Thanks in part to these advances cancer mortality has been falling at a rate of 0.94% per year in Spain. However, the number of cancer cases is expected to increase by 50% in the coming years, mainly due to the aging of the population, according to the World Health Organization.

The human genome comprises all the information stored in the DNA. This information is encoded in 3.2 billion base pairs, which are distributed in about 22,000 genes in 23 pairs of chromosomes. The exome is the "useful" part of the human genome, which is synthesized into proteins or RNA, which in turn carry out the fundamental functions of the organism. Most of the genetic variations responsible for known diseases are found in the exome and are classified in databases of pathogenic variations that are accessed by genetic analysts for interpretation.

Sequencing consists of determining the order of the A, C, G and T bases in a DNA fragment. The area of sequencing has recently undergone a revolution, going from Sanger-type sequencers that allowed sequencing of a maximum of 96 sequences of 800 nucleotides to the sequencing of millions of DNA fragments with second generation equipment (Next Generation Sequencing or NGS). NGS technology has drastically changed the way in which the human genome is sequenced, greatly reducing sequencing costs. In 2005, sequencing a complete genome cost about 18 million euros, in 2010 it cost 40,000 euros, and in 2018 it dropped below 1000 euros.

This unprecedented drop in cost of sequencing has enabled great advances at the scientific level in the identification of genes responsible for diseases and a democratization of sequencing for medical use, which has immediately leveraged scientific advances for clinical use. An example of the impact that these advances are having on society has been the improvement in survival in childhood cancer: only a few years ago only three mutations causing pediatric cancer were known (sequencing of

100-150 base pairs); at present 160 complete genes related to the disease are sequenced (500,000 base pairs). Thanks to these advances, the survival rate of pediatric cancer has gone from 54% in 1980 to 80% today.

1.1 Genetic Testing in Healthcare

Genetic testing in the clinical sector is generally performed in a centralized manner in medium/large hospitals. Smaller hospitals and health centers refer patients requiring this type of analysis to the larger hospitals, which due to their volume and budget can afford the necessary equipment for the analysis and corresponding genetic specialists such as oncologists, cardiologists, neurologists, pediatricians, etc. (See [Table 15](#))

In a clinical environment, the physician determines the relevance of a genetic testing. She or he is responsible for generating the genetic test prescription, and for receiving and interpreting the results that will eventually impact the patient treatment. (See [Figure 46](#))

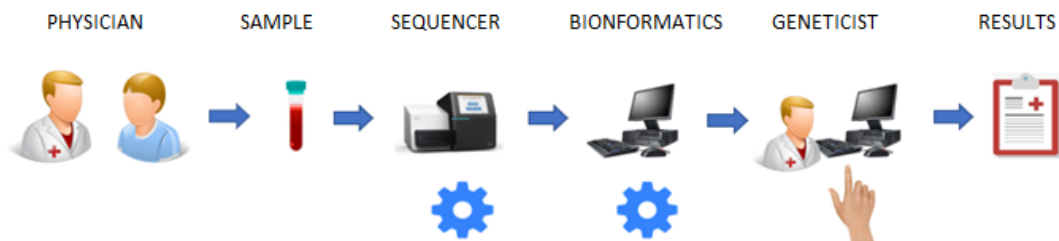


Figure 46: Simplified scheme of the necessary steps required in genetic testing for diagnosis purposes with NGS (Next Generation Sequencing)

Table 15. NGS genetic testing in healthcare. Professionals and actions involved.

Personnel Involved	Action	Type of sample	Tools
Medical practitioner Geneticist	The physician, in the process of diagnosis or prognosis of a patient, considers that the pathology (disease) suffered by a patient may have its origin in a genetic alteration, and requests a genetic analysis.	All types.	Genetic analysis request document Usually done with a blood test, a biopsy (in case of tumor) or other body fluids.
Nursing staff	Drawing of blood from the patient (or other biological sample).	Blood other	Insertion into tube with additives for preservation.
Surgeon	Intervention to remove the tumor / surrounding lymph nodes.	Biopsy.	Biological specimen inserted in tube (with additives / paraffin) for preservation.
Anatomical Pathology Technician	Section of the tumor to remove the non-tumorous part, and cellular analysis.	Biopsy.	Cutting tool / microtome / cellular stains / kerosene transport plate
Sequencing Technician	Preparation of biological sample for sequencing.	All types.	Genetic sequencer.
Bioinformatician	All types.	All types.	Servers / Computers.
Genetic Specialist	The genetic analyst analyzes the list of genetic variations and determines which one (or ones) is responsible for the patient's pathology.	All types.	- List of variants in VCF file - Variant sorting and filtering protocol (Excel, analysis software, etc.). - Access to mutation databases. - Report of results.
Medical Doctor and/or Geneticist.	The medical practitioner analyzes the geneticist's report on the responsible mutation, as well as the bibliography related to the treatment, and prescribes the corresponding treatment (if any). Family or reproductive genetic counseling	All types.	Report of responsible genetic variations.

1.2 Genetic Sequencing Technologies

Genetic sequencing definition. The process of determining the order of nucleotides – adenine (A), thymine (T), cytosine (C) and guanine (G) – along a DNA strand is called genetic sequencing or DNA sequencing (See [Figure 47](#)). The sequencing process is determined using a variety of laboratory techniques and technologies. The DNA base sequence carries the information that the human body needs to assemble protein and RNA molecules. The DNA sequence information is a starting necessary step for scientists investigating the functions of genes, geneticists providing evidence for diagnosis in medical setups. It also allows healthcare professionals to make educated guesses about the predisposition of apparently healthy individual to develop genetic disorders in the future such as cancer, rare diseases or specific cardiopathies.

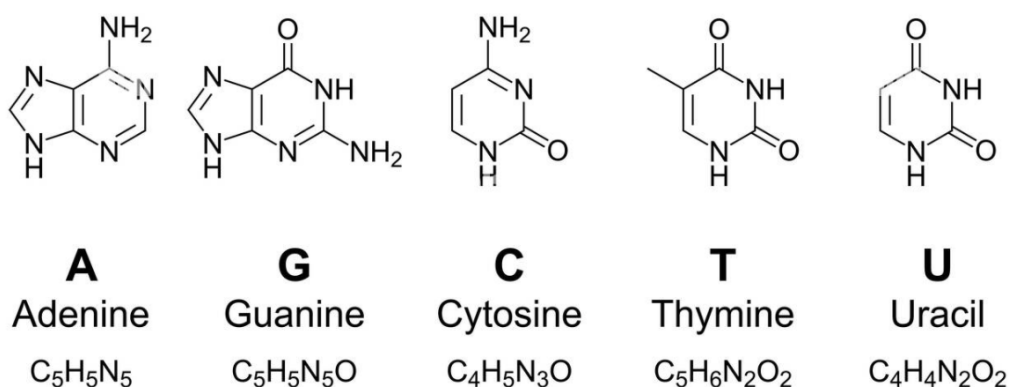


Figure 47: Chemical structure of nucleotides, the building blocks of human DNA.

All human traits information is stored coded as a sequence of nucleotides. (Source: Alamy Stock Photos)

Sanger sequencing. The Sanger sequencing method is based on the selective incorporation of chain-terminating dideoxynucleotides by DNA polymerase during in vitro DNA replication. (Sanger F, 1975; 1977). Sanger sequencing was first developed by Frederick Sanger and colleagues in 1977 and it became the most widely used sequencing methods for 40 years. (See [Figure 48](#)). Nowadays Sanger sequencing has been widely replaced by NGS, especially for large scale research and diagnosis purposes. However, Sanger sequencing remains in use for smaller projects and as a gold standard for validation of NGS results.

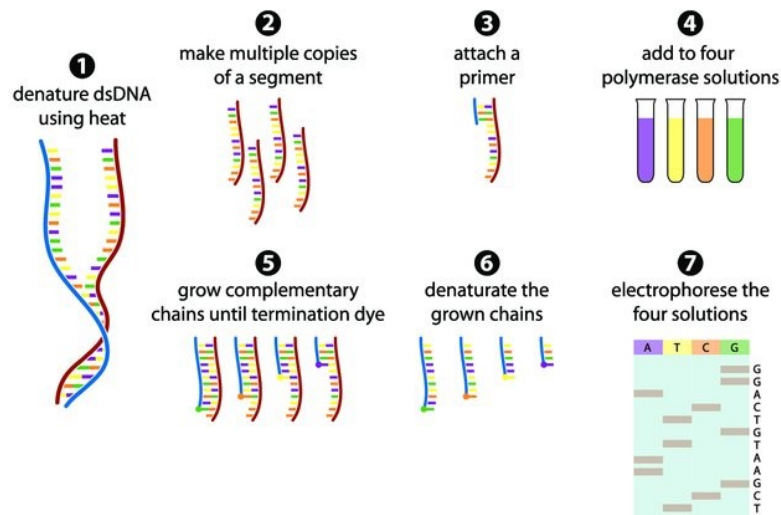


Figure 48: The Sanger sequencing method in seven steps. Source: Gauthier MG, 2007

DNA microarray. It is a collection of microscopic DNA spots attached to a solid surface. Microarrays are used by scientists and diagnosis laboratories to measure the expression levels of large number of genes or to genotype multiple regions of a genome. Each DNA spot contains small amounts (picomoles) of a specific DNA sequence, known as probes. These probes hybridize (attach) to a short section of a gene or genetic sequence. When the genetic sequence is found, the probes hybridizing are detected by a high resolution camera by the use of a fluorophore. The end result is a matrix of genetic sequences that are detected and appear bright on the picture, or not, and appear as a dark spot.

NGS sequencing. It is the technology that has revolutionized genomic research. Using NGS an entire genome can be sequenced within a single day, compared to 10 years that required the sequencing of the first human genome using Sanger. NGS sequencing is becoming the standard for many clinical applications since it can capture a broader spectrum of mutations than Sanger sequencing (See Table 16). (Behjati S and Tarpey PS, 2013).

Table 16: Comparison between DNA microarrays, Sanger & targeted NGS.

	Benefits	Disadvantages
DNA Microarrays	<ul style="list-style-type: none"> • Easy to use. • Once it is design, can be mass produced easily. • Easy to interpret. • Proven technology 	<ul style="list-style-type: none"> • Lack of flexibility. Arrays are design for a fixed set of variants. • Difficult to design and fine tune.
Sanger Sequencing	<ul style="list-style-type: none"> • Fast, cost-effective sequencing for low numbers of targets (1–20 targets) • Familiar workflow 	<ul style="list-style-type: none"> • Low sensitivity (limit of detection ~15–20%) • Low discovery power • Not as cost-effective for high numbers of targets (> 20 targets) • Low scalability due to increasing sample input requirements
Targeted NGS	<ul style="list-style-type: none"> • Higher sequencing depth enables higher sensitivity (down to 1%) • Higher discovery power* • Higher mutation resolution • More data produced with the same amount of input DNA • Higher sample throughput 	<ul style="list-style-type: none"> • Less cost-effective for sequencing low numbers of targets (1–20 targets) • Time-consuming for sequencing low numbers of targets (1–20 targets)

Source: Illumina. Inc.

1.3 The Processing of Genetic Data Using NGS Technology

In the different phases of the genetic analysis process, data are generated and processed by each specialist.

1.3.1 Sequencing

In this phase, the sequencer generates the files with the raw DNA sequences read. NGS systems usually read sequences of between 100 and 600 base pairs. Much of the extracted information is redundant but necessary to achieve reliable sequencing and to be able to detect variants or mutations at very low frequency especially for cancer applications, where high sensitivity is required. (See [Figure 49](#))



Figure 49: Illumina MiSeq. DNA Next-Generation Sequencer. Source: Illumina.

1.3.2 Quality Filtering

In this phase, low quality sequences will be filtered. Low quality means that the system has not been able to accurately determine that the sequence read is correct. In some cases, pieces of low-quality sequences are also trimmed so that they are not completely discarded.

Some sequencers have their own proprietary quality encoding system, but most have adopted *Phred-33 encoding*, where each nucleotide has a quality score associated representing the probability of an incorrect basecall at that position (See [Table 17](#))

Table 17: Correspondence between Phred quality score and nucleotide base sequencing accuracy.

Phred Quality Score	Probability of incorrect base call	Base call accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1000	99.9%
40	1 in 10,000	99.99%
50	1 in 100,000	99.999%

A quality value Q is an integer representation of the probability p that the corresponding base call is incorrect.

$$Q = -10 \log_{10} P \quad \rightarrow \quad P = 10^{-\frac{Q}{10}}$$

Quality scores started as numbers (0-40) but have since change to an ASCII encoding to reduce file size and make data management easier for bioinformaticians. See and [Figure 50](#) for an example of quality score interpretation.

Base Sequence Reads
↙

Nucleotide = C

```
CCTCCTGCTTAAAACCCAAAAGGTCAGAAGGATCGTGAGGCCCGCTTTC
+
CCOFFFFFHGHJJIJJJJJJI@HGIJJJIIIJGIGIHIJJJIIIIJJ
```

Quality = F
F -> P_error=0.0002
Q = 37

Figure 50: Interpretation of quality score in a FASTQ file format.

Q20 is generally considered a good cut-off score for most research and clinical purposes.

The bioinformatics data processing takes into account the quality scores of the sequences in the workflow following steps. If this process is not performed correctly it can be a source of artifacts. These artifacts could end up being reported as a genetic variant that does not really exist in the biological specimen, unless they are filtered out in the final variant interpretation phase.

1.3.3 Sequence alignment

The filtered sequences are aligned in the correct order, taking a human DNA as a reference. Each alignment is also assigned a quality score, which will be useful in the subsequent variant filtering phase (See [Figure 51](#)).

```

GenBank Accession # AF139335.1 Microcystis aeruginosa strain UWOC RID-1
microcystin synthetase (mcyA) gene, partial cds
Length=1319

Score = 754 bits (408), Expect = 0.0
Identities = 408/408 (100%), Gaps = 0/408 (0%)
Strand=Plus/Plus

Query 1  AATTACAGGCAAACATCGGCAGATTCTCAAGGGATATTTAATATTGTTGGCTGGAATAG 60
          |||
Sbjct 141 AATTACAGGCAAACATCGGCAGATTCTCAAGGGATATTTAATATTGTTGGCTGGAATAG 200

Query 61  TAGTTACACGGGGGAACCTATCCCGGTTGCTCAGATGCGAGAATGGCTAGATGATAAAGT 120
          |||
Sbjct 201  TAGTTACACGGGGGAACCTATCCCGGTTGCTCAGATGCGAGAATGGCTAGATGATAAAGT 260

Query 121 TAAGGTTATTCTCGCTCAAAAACCGAAAAAGTTCTGGAAATAGGTTGTGGAACCGGGTT 180
          |||
Sbjct 261  TAAGGTTATTCTCGCTCAAAAACCGAAAAAGTTCTGGAAATAGGTTGTGGAACCGGGTT 320

Query 181  AATATTATTC AAGTTGCTCCCCATTGCCAGTGTTATTGGGGAACCGATATTTTCATCAGT 240
          |||
Sbjct 321  AATATTATTC AAGTTGCTCCCCATTGCCAGTGTTATTGGGGAACCGATATTTTCATCAGT 380

Query 241  AGCCTTAGACCATATTCAGCGAATTAATCAAGAAGGGCCTCAGCTAGAGCAAGTCAGGCT 300
          |||
Sbjct 381  AGCCTTAGACCATATTCAGCGAATTAATCAAGAAGGGCCTCAGCTAGAGCAAGTCAGGCT 440

Query 301  ATTGCATAGCACAGCCGATAATTTTGAGGGTTTGGAGTCAGAAGGATTCGATACAATTAT 360
          |||
Sbjct 441  ATTGCATAGCACAGCCGATAATTTTGAGGGTTTGGAGTCAGAAGGATTCGATACAATTAT 500

Query 361  CCTTAACTCGGTTGTGCAGTATTTCCCCATATAGATTACTTACTGAG 408
          |||
Sbjct 501  CCTTAACTCGGTTGTGCAGTATTTCCCCATATAGATTACTTACTGAG 548

```

Figure 51: Example of a perfect sequence alignment with a 100% match. Alignment performed by BLAST

Variant calling. In this part the genetic variants of each patient are identified with respect to a reference. The most common types of genetic variations are SNPs, MNPs, INDELS and combinations of these.

Variant annotation. Biological information is added to each of the variants: Name, protein change, related diseases, etc.

Variant prioritization and interpretation. The variants related to the patient's pathology are prioritized, discarding the rest. The final variants will be subsequently interpreted by the genetic analyst.

The different phases of the genetic analysis are summarized in

Table 18, including the amount of data generated in each part of the process.

1.3.4 Artifact detection

In genetics, an artifact is defined as a result that does not represent the true biological material or function but arises from a technical, often artificial process. Artifacts can lead to erroneous results from sequencing, and should be thoroughly investigated to avoid providing incorrect patients results data.

During the NGS primary analysis sequence reads and quality scores are produced at the sequencer machine, generally with sequencing machine built-in software. In a secondary data analysis the alignment and assembly of DNA/RNA segments takes place, followed by variant calling and eventual data visualization. However, an NGS technology introduces a certain amount of artifacts that are not always removed by the bioinformatics pipeline at the secondary data analysis. In a NGS variant interpretation application, an artifact is defined as a set of nucleotides that were mistakenly read by the sequencer device creating a false variant on the VCF file analyzed. If this false variant was not identified in the process, it could generate a false positive report result, providing the physician with the wrong information to treat or diagnose the patient.

Table 18: Summary of NGS genetic analysis process steps, including level of automation, tools and total amount of data generated on each step.

Process	Responsible	Automation level	Tools	File Formats	Estimated file size WES	Estimated file size WGS
Sequencing	Laboratory technician	HIGH	Proprietary from manufacturer	FASTQ, XSEQ, unaligned BAM, or FASTA	15 Gb (2Gb targeted)	100Gb (15Gb per file)
Quality filtering	Bioinformatician	HIGH	FASTQC		1.5Gb	13 Gb
Alignment	Bioinformatician	HIGH	Burrows-Wheeler Aligner, bowtie, SOAP2, Map,	SAM, BAM	2 Gb	10 -15 Gb
Variant calling	Bioinformatician	HIGH	GATK, SamTools / BCFTools	VCF	200 Mb	2.3 Gb
Annotation	Bioinformatician	HIGH	ENSEMBL/ VEP ANNOVAR,	CSV /other	200 Mb	2.3 Gb
Variant prioritization	Geneticist	VERY LOW / LOW.	Excel, Cartagenia/Alissa	CSV / other	50-60 Mb	600-720 Mb
Variant interpretation	Geneticist	VERY LOW / LOW.	Cartagenia /Alissa, Sophia Genetics, Varsome	CSV / other	5-20 Mb	60-230 Mb

Source: Genomics Unit, Health Research Institute La Fe.

1.4 Variant Interpretation

In a clinical setting, the main goal of geneticists is to determine if the patient carries a pathogenic or likely pathogenic variant which is responsible for the patient disorder. This knowledge can positively influence the care and treatment of the patient.

Variant interpretation or variant classification is the evaluation of the pathogenicity of variants found in the patient with clinically relevant characteristics. (Ku CS et al., 2012; Hoskinson DC et al., 2017). Variant interpretation is performed in the called tertiary data analysis. (See [Figure 52](#)). It is the last step that provides a genetic analysis report that will be generated for the physician.

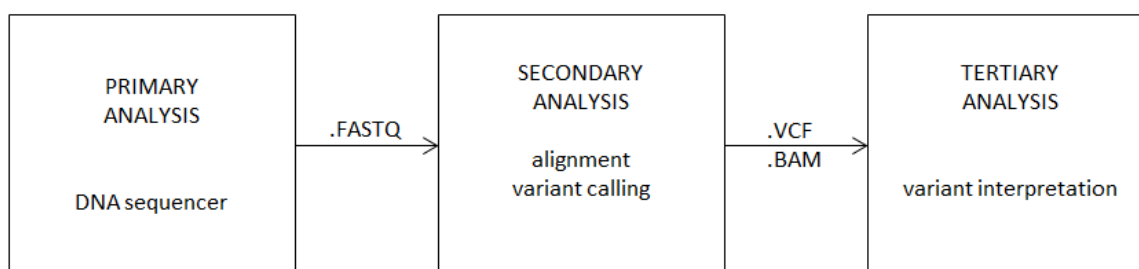


Figure 52: Types of analysis in a DNA analysis pipeline, including file formats for each analysis output.

Most laboratories classify variants in three or five different types:

Pathogenic. It is generally recognized as a cause of disorder.

Likely Pathogenic. It has not been previously reported or recognized as a cause of disorder, but it is of the type which is expected to cause the disorder.

Variant of Uncertain Significance (VUS). These types of variants have not been previously reported and may or may not be causative of the disorder.

Likely Benign. Previously unreported and is probably not causative of the disorder.

Benign. Previously reported and is a recognized neutral variant.

Originally, the International Agency for Research on Cancer (IARC) defined likely pathogenic and likely benign variants as a $\geq 95-98\%$ probability of being causative of disease. Later, the ACMG/AMP guideline expanded this definition to $\geq 90\%$ certainty. (Hoskinson et al., 2018; Richards S et al., 2015)

1.5 Difference in Variant Interpretations among Laboratories

The interpretation of variants requires evaluation of a large amount of evidence to arrive to a single descriptor of pathogenicity. The evidence used to interpret the variants is complex and very often uncertain.

It is not surprising that the same variants presented to different laboratories will not be classified the same way. A study published in 2016 by the American Society of Human Genetics performed in nine different laboratories, showed that there was only 34% inter-laboratory concordance for variant interpretation and showed disagreement in 64% of the decisions. In 22% of cases the disagreement could affect medical management of the patient. Amendola LM (2016) states that *“The Assessment of pathogenicity of genetic variation is one of the more complex and challenging tasks in the field of clinical genetics”*.

Based on these premises and perceived systems limitations our research work focused on the study and optimization of two main clinical genetic analysis technologies available: microarrays and NGS variant interpretation.

2 Hypothesis

Genetic analysis methodologies currently used in clinical practice present limitations and points for improvement. These limitations could - in certain cases - cause that the patient to receive a suboptimal diagnosis or treatment. These methodologies could be optimized by means of advanced analysis techniques based on process automation and AI.

3 Objectives

1. To analyze the current systems and methods for genetic variant interpretation targeted to clinical diagnosis.
2. To analyze certain deficiencies and limitations detected in clinical genetic analysis and measurement systems according to experts, and quantify their impact when possible.
3. Propose alternative methodologies that reduce systems cost, reagents or maintenance. minimize or eliminate the errors or biases, increase accuracy, reproducibility and robustness of each method,
4. To design and validate an automated microarray reading system for advanced Colorectal Cancer prognostic biomarker reading applications. Proof of concept of the microscopic analysis system and its validation.
5. Design and validate an automated system for interpretation of genetic variants. Proof of concept of the analysis system and validation of the same.

4 Methods

4.1 Design and Validation of a Methodology for Colorectal Cancer Biomarker Analysis for Clinical Applications

This part of the research was conducted in the framework of a joint project between two private companies (Celeromics Technologies and IMEGEN), and two research institutions (Health Research Institute Hospital La Fe, and UPV-IDM). The project was called ONCOMARKER and its main goal was to design and develop a platform for the analysis of oncological biomarkers to establish the prognosis and response to treatment in advanced or metastatic colorectal cancer (mCRC). The goal was to build a complete solution adapted to the needs of genetic analysis laboratories and hospital centers, comprising consumable analysis microarrays, a reader for these microarrays, and a kit of reagents. Our work was focused on the design and validation of microarray reader.

In this line we had to design a methodology intended to be used in clinical applications and the usage of patient samples was required. According to current legislation and consultation performed to the AEMPS we had to comply with the following requirements:

1) All tests were to be performed in a hospital setting. In our case they were performed at La Fe Hospital in Valencia. 2) No decisions regarding patient treatment was to be made on the basis of the systems under test. 3) An authorization had to be obtained from the hospital's own Clinical Research Ethics Committee (CEIC) from the same hospital. 4) It was necessary for the patient to sign an informed consent form authorizing the use of his/her biological samples for the research described.

4.1.1 Expert Panel Survey

As a previous step, we conducted a survey to 15 experts involved in the usage of genetic clinical diagnosis systems. The goal of this survey was to identify current technologies used for clinical genetic analysis, current drawbacks and pitfalls, and where those systems could be optimized. (See [Appendix 9.6](#) for the specific form that was used for the interviews)

Of the 15 experts surveyed 60% (nine of them) were oncologists and 40% (six) were head of a genomics laboratory. They belonged to Health Research Institute and Hospital La Fe (Valencia) (4 experts), CNIO (Madrid) (2 experts), H. Clínico Carlos III (Madrid), Hospital La Paz (Madrid), Hospital General (Ciudad Real), Hospital Princes (Madrid), HGUE (Elche), Hospital 12 Octubre (Madrid), Incliva (Valencia), Hospital Gregorio Marañón (Madrid) and Fundación Gimenez Díaz (Madrid). Experts were consulted using personal or telephonic interviews, or with an equivalent online survey in case they were not available for a personal interview.

4.1.2 DNA Microarray Analysis System Targeted Functionality

Based on the results of the survey to experts and potential users, a system for DNA microarray analysis was planned for development. The system was targeted to address some of the limitations of current systems and methodologies: 1) up to 13 genetic variations detected with a single analysis 2) High sensitivity (1% of cancer tissue vs. 99% healthy tissue) 3) High specificity 100% 4) Easy to use 5) Small sample requirements: 15ng DNA 6) Automatic analysis results. 7) Lower instrument cost.

4.1.3 Study Subjects and Inclusion Criteria

This study used samples from patients diagnosed with mCRC and attended by the Medical Oncology Service of the Hospital Universitario y Politécnico La Fe de Valencia. Solid tumor samples were previously processed by the Hospital's Anatomic Pathology Service and had a minimum of 50% tumor cells. Clinical-biological data, histopathological characteristics of the tumor, evolutionary parameters and response to treatment will be collected from each patient.

4.1.4 Sample Size

We determined initially a group of 100 tumors (primary or metastases) embedded in kerosene (FFPE) from patients diagnosed with mCRC. The sample size has been determined according to a preliminary study for validation of the biomarker reading system. The sample size was selected in order to allow a preliminary validation of the system, but it was probably not sufficient for a further validation of the system as an in-vitro diagnostic (IVD) device.

4.1.5 Microarray Automatic Positioning Methodology

The automatic positioning subsystem incorporated artificial intelligence technology reproducing an intelligent human behavior as a computer vision system combined with the automated positioning stage. The system automatically positioned the microarray under the CCD camera sensor and extracted the relevant biomarker information.

The system was constructed from scratch incorporating the following elements:

1. Commercial microscopic stage (from Optika BC-159 microscope)
2. Aluminum box.
3. CCD fluorescence camera.
4. Motorization system (stepper motors, mechanic adaptors, control board)
5. Control board and stepper motor control software : Arduino Uno + Stepper Motor control driver shield for Arduino (See [Figure 53](#))
6. Personal computer, image analysis and positioning system.

A microscopic stage was adapted with three stepper motors in order to obtain a fully motorized stage at the XYZ axis. The prototype is able to automatically move the sample microarray in the three dimensions (XYZ), automatically focus and capture microarray images. A CCD camera was placed on the upper side of a closed chamber.

At the personal computer two different and independent systems were implemented to perform the automatic microarray reading. 1) Autofocus system 2) Microarray positioning system.

Both systems were programmed in Microsoft Visual C# running under Windows 7. The communications between the PC and the stepper control boards were performed using a proprietary communications protocol designed ad-hoc for this application. The communication protocol and command interpreter was programmed in C and placed inside an Arduino Uno microcontroller board. Mechanical parts to adapt the stepper motors to the microscope were designed using Alibre Design (Alibre LLC, Texas) and 3D-printed using Polylactid acid (PLA) with a Prusa i3 3D printer.

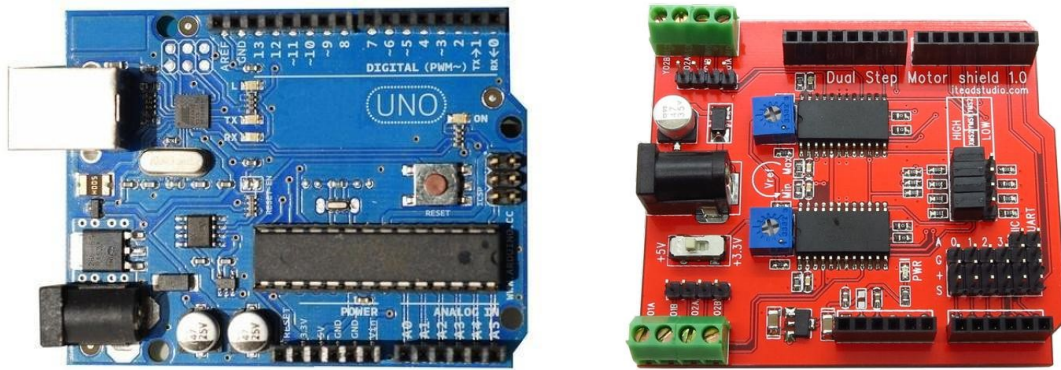


Figure 53: Arduino Uno microcontroller board (left) and Dual Stepper motor driver shield (right)

The two software subsystems were designed, implemented and tested individually, since their operation did not need to run simultaneously.

Autofocus system. Several algorithms for autofocus were tried and tested with the whole system prototype setup. A final version of the system focus algorithm was selected that was optimized for both focus speed and reliability. The flow diagram described in [Figure 54](#) indicates the operations executed by the software control system to perform the autofocus procedure.

4.1.6 Fluorescence Image Capturing System

The system included a fluorescence image capture system. Our goal was to develop this subsystem with LED lighting, instead of the laser current industry standard that was more expensive. We tried different configurations of high sensitivity fluorescence CCD cameras (Optikam B5 with sensor MT9P006, DMK23UXZ249 with sensor Sony IMX249LL, DMK42BUC02 with sensor MT9M021 and Atik 314L+ with sensor ICX-285AL) with different exposition times, two fluorophores (Cy5 and FITC) (See [Figure 62](#)), along with different red LED wavelengths and power (0,5W, 1W, 5W, 30W)

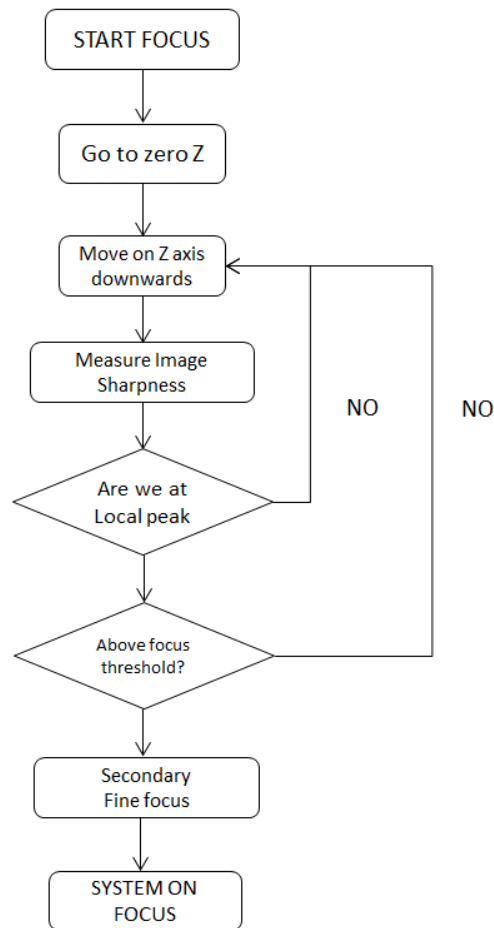


Figure 54: Flow diagram of the tasks performed by the autofocus system.

4.1.7 System Integration and Validation

When all elements of the systems were tested and integrated, we performed a final validation of the system comparing the ONCOMARKER system results against two alternate equivalent systems 1) COBAS (Roche) 2) NGS sequencing + Sophia Genetics variant interpreter. The whole ONCOMARKER system was tested for sensitivity, specificity, accuracy, positive predictive value and negative predictive value. (This final testing and results are outside of the scope of the present research).

4.2 Validation of the Methodology for Biomarkers of Genetically Based Diseases: Interpretation of Genetic Variants

4.2.1 Analysis of the Past and Current Recommendations and Methods for Genetic Variant Interpretation

In this part we analyzed and summarized the evolution of variant interpretation for clinical applications by the world leading variant harmonization organism (ACMG). The ACMG guideline establishes a *de facto* standard and reference for variant interpretation and genetic analysis reporting for clinical applications. The analysis of this evolution helped us to understand 1) how variant interpretation and reporting has evolved through history 2) which inputs should an eventual AI system should be taken into account to provide an automatic or semi-automatic variant interpretation. This part of the research was performed as a review of relevant bibliography in this field.

We analyzed a set of scientific articles published in the Journal Genetics in Medicine where ACMG described the recommended processes that should be followed to perform variant interpretation. (See [Table 19](#))

Table 19. History of ACMG recommendations for variant interpretation.

Year	Reference
2000	"ACMG Recommendations for Standards for Interpretation of Sequence Variations." (ACMG, 2000)
2007	"ACMG Recommendations for Standards for Interpretation and Reporting of Sequence Variations: Revisions 2007" (Richards CS et al., 2008)
2015	"Standards and Guidelines for the Interpretation of Sequence Variants: A Joint Consensus Recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology." (Richards S et al., 2015)

Additional bibliography that detailed or modified some specific areas of the above mentioned guidelines were also studied (Gelb BD et al., 2018; Ghosh R et al., 2018) as well as specific adaptation or recommendations of the guidelines for **cardiopathies** (Kelly MA et al., 2018; Hershberger RE et al., 2018; Morales A et al., 2020), **cancer** (Li MM et al., 2017 ; He MM et al., 2019) and **secondary findings reporting** (Green RC et al., 2013; Kalia SS et al., 2016 ; Miller DT et al., 2021)

4.2.2 Expert Geneticists Panel Survey

We interviewed six expert geneticists with experience in clinical diagnosis and variant interpretation from La Fe Hospital, Health Research Institute La Fe, IMEGEN, INCLIVA and Hospital General (Valencia). ([Appendix 9.7](#) shows the survey form employed). All geneticists were surveyed with a personal interview.

4.2.3 Study Subjects and Inclusion Criteria

The presented methodologies were initially tested with 10 patients' samples from Hospital Universitario y Politécnico La Fe (Valencia, Spain), 35 samples from different collaborating research centers: Institut National d'Hygiène de Rabat (Morocco), Centre for Genetics and Inherited Diseases of Taibah University Madinah (Saudi Arabia), Hospital Universitario Reina Sofía de Córdoba (Spain), Instituto CSS Mendel (Italy) and Istanbul University (Turkey) with different patient disorders, and 13 samples (WES) from the company Binartis Genomics (Valencia, Spain) from consultants interested in preventive medicine with no apparent disorder at the moment of performing the genetic analysis.

Samples were processed at the research center of origin with several sequencing technologies and diverse primary and secondary bioinformatics pipelines. For the purpose of the research only the .VCF and .BAM file along with the clinical history of the patient was received from each entity as the input for the present research. Data received from Binartis Genomics also included the family health history from each consultant.

4.2.4 Semi-assisted variant interpretation methodology

We analyzed the existing methods recommended by ACMG for variant interpretation in clinical laboratories and some proposed improvements (Nykamp K et al., 2017) as well

as several analysis pipelines at Genomics Unit at Health Research Institute Hospital La Fe, data from the geneticists panel survey and the review described in [Section 4.2.1](#).

Based on this broad data analysis we designed a methodology that was intended to 1) save geneticists analysis time 2) increase diagnosis yield and Diagnostics Odds Ratio (DOR) (Šimundić AM, 2009) when compared to current automatic and manual systems. The latter implies improving both sensitivity and specificity. ([Table 20](#))

The system was implemented using Python programming language running under Linux. It uses or integrates the following elements:

- 1) Libraries: Selenium, pymysql, xlswriter, CSV, pyEnsembl, samtools
- 2) Databases integrated:, Clinvar and dbnsfp.
- 3) Web databases connection : OMIM, HGMD, HGMD free, OMIM
- 4) In-silico predictors : Polyphen2, Mutation Taster, Provean, Sift

Table 20. Improvements targeted and expected clinical benefits.

Improvement targeted	Impact on overall system
Save geneticists time	Increases geneticists capabilities Increases laboratory throughput.
Increase analysis sensitivity	More potentially pathogenic variants are detected. Patients could eventually have better treatments based on these results.
Increase analysis specificity	Reduce the number of false negatives. Saves physician time, and avoid misdiagnosis by physicians. Better patient care.

The specific samples used for this method validation were sequenced using Illumina NextSeq 500 System with TruSight Inherited Disease panel for the analysis of 552 genes (#FC-121-0205). FASTQ files from the sequencer was processed according the regular

laboratory workflow and the standard pipeline BWA/GATK (GATK-Haplotypecaller v3.7). The output of this process was 10 VCF files v4.2 format that were introduced to the Binome system.

4.2.5 New Method for Artifact Detection

We developed an AI based system for automatic classification of variants. The system was able to classify variants as real variants or as technology artifacts.

Our initial step was to prepare a dataset to train the AI system with real data that had been previously classified manually and that we could use as a reference.

A total of 200 variants were extracted from a set of ten samples of patients coming from the Health Research Institute La Fe Genomics Unit. These variants were a combination of real sequence variants and artifacts that were previously classified by two different human expert geneticists. A total of 80 variants were classified as real sequence variants with agreement between the two geneticists. 78 were classified as artifacts, and 42 were declared unknown by the geneticists, or a consensus about their classification was not reached.

With the selected variants and artifacts a training set of 158 variants was created to feed the AI system. For each variant, the elements used as an input of the system were: 1) sequencing depth 2) total number of reference reads 3) total number of alternative reads 4) total forward reference reads 5) total reverse reference reads 6) total forward alternative reads 7) total reverse alternative reads 8) whether the variant is a section of repetition. (See [Table 21](#))

The system was programmed using Python programming language and NumPy, samtools and scikit-learn libraries. Python scripts were programmed to extract the required parameters from .VCF and .BAM files and to generate the training elements for the AI system, as well as the chromosome number and position of each variant. (See

[Table 22](#))

Table 21. Input element features extracted to feed the AI algorithm

Artificial Intelligence Input Parameters	Definition
Ref	Total reads of the reference nucleotide at the selected position
Ref-forward	Number of reads of the reference nucleotide read in the forward DNA strand.
Ref-reverse	Number of reads of the reference nucleotide read in the forward DNA strand.
Alt	Total number of read of the alternative nucleotide (variant) at the selected position.
Alt-forward	Total number of read of the alternative nucleotide at the selected position in the forward DNA strand.
Alt-reverse	Total number of read of the alternative nucleotide at the selected position in the reverse DNA strand.
Repetitive section	This flag indicates if the selected position is found in a repetition section. A section where a sequence of one or more nucleotides or sets repeats more than three times.

Table 22. Samples of the system input data with the feature variables (independent) and the target variable we try to estimate (dependent) = column ARTIFACT.

	GENE	POSITION	REF-F	REF-R	REF	ALT-F	ALT-R	ALT	TOTAL	REPETITIVE	ARTIFACT
0	chr1:977330		0	0	0	56	16	72	72	0	0
1	chr1:978762		22	47	69	28	51	79	148	0	0
2	chr1:978775		50	98	148	0	1	1	149	0	1
3	chr1:981931		0	0	0	79	52	131	131	0	0
4	chr1:981934		74	50	124	5	0	5	129	1	1

Pre-processing. Before introducing the data to system we standardize it. Standardization is also called Z-normalization, a scaling technique that when applied to numeric features they will be rescaled to acquire the properties of a standard normal distribution with mean $\mu = 0$ and standard deviation $\sigma = 1$.

Data splitting. We divided the dataset into 75% to train, and 25% to test its performance. Afterwards the system was trained.

See [Appendix 9.9](#) for the Python source code used for pre-processing, data splicing, standardization, data splicing and training of the artifact detection subsystem.

4.2.6 Improved Sensitivity Method for Interpretation of Genetic Variation Based on ACMG Guidelines

We implemented an innovative method to increase the sensibility of the system by complementing the proposed ACMG guidelines and reclassifying part of the Variants of Uncertain Significance (VUS) as high risk VUS that could be considered for reporting in case no pathogenic or likely pathogenic variants are found.

Each ACMG criteria was assigned a specific score depending on the severity of the evidence. For instance, PSV1 include a score associated with the strength of the evidence. In this case, the evidence strength was calculated using a combination of the distance to the start codon, the presence or not of splicing-sites, the number of pathogenic variants in the same gene, the GnomAD o/e ratio and the particular mechanism of pathogenicity along with the inheritance pattern for the disease.

Once evidence-based criteria were calculated and scored, the final 5-tier classification was also scored based on the addition of the individual scores assigned to each criteria.

The final result 5-tier label remained unaffected, but the list of variants was prioritized based on this additional associated score. Based on this scoring system, VUS were classified in three subclasses: high risk, medium risk and low risk of pathogenicity (See [Figure 55](#)). VUS classified as high risk were manually reviewed by two expert geneticists. The method has been preliminary tested in 13 samples coming from Binartis Genomics.

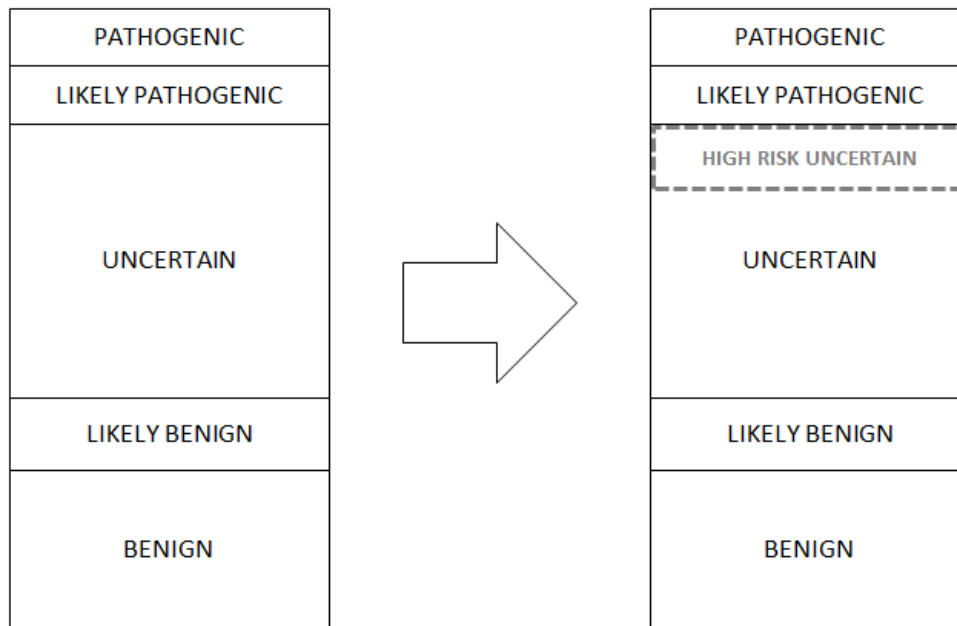


Figure 55: Novel method for ACMG guidelines sensitivity increase. Part of the uncertain variants are reclassified as “High risk uncertain” to be reviewed in case no pathogenic variant is found.

5 Results

5.1 Colorectal cancer biomarker analysis system

5.1.1 Expert Panel Survey

The results of the survey guided the following research based on the improvements suggested by the experts consulted.

When experts were consulted about current systems limitations, the most common answer was 1) system sensitivity, followed by 2) time of analysis required and 3) limited number of mutations detected. (See [Figure 56](#))

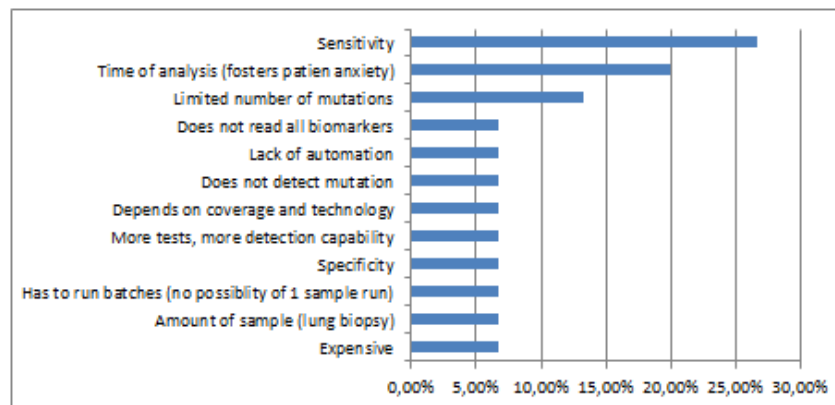


Figure 56: Limitations perceived by experts of current Genetic Analysis systems for clinical applications.

For open questions, the most important characteristics for the experts were 1) sensitivity 2) specificity and 3) coverage of mutations. (See [Figure 57](#)). Experts provided similar answers when questioned in a guided manner and a set of potential characteristics was suggested (See [Figure 58](#))

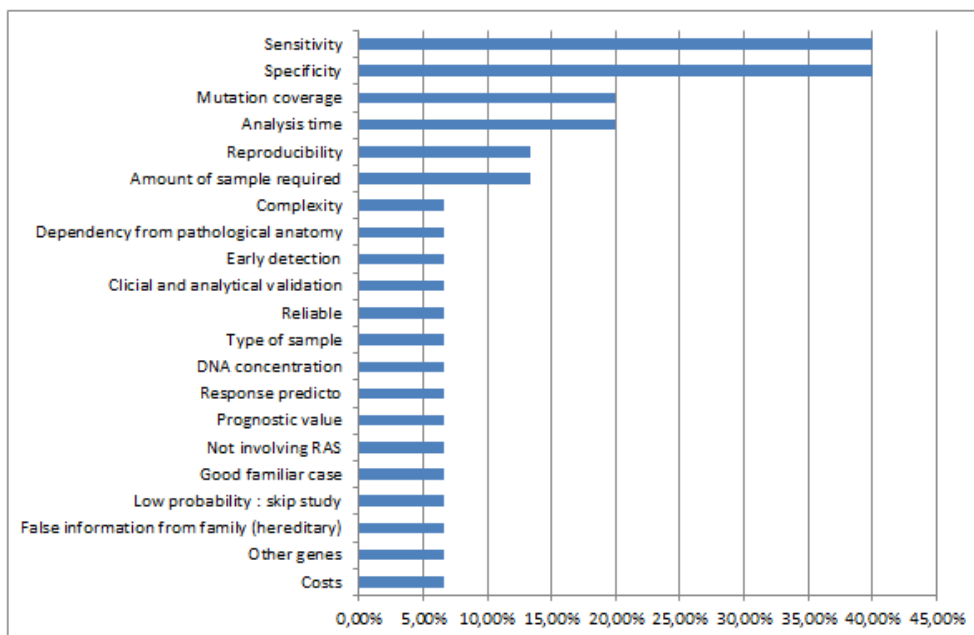


Figure 57: Priorities for a Diagnosis/Prognosis Genetic Analysis Instrument. Answers to open question.

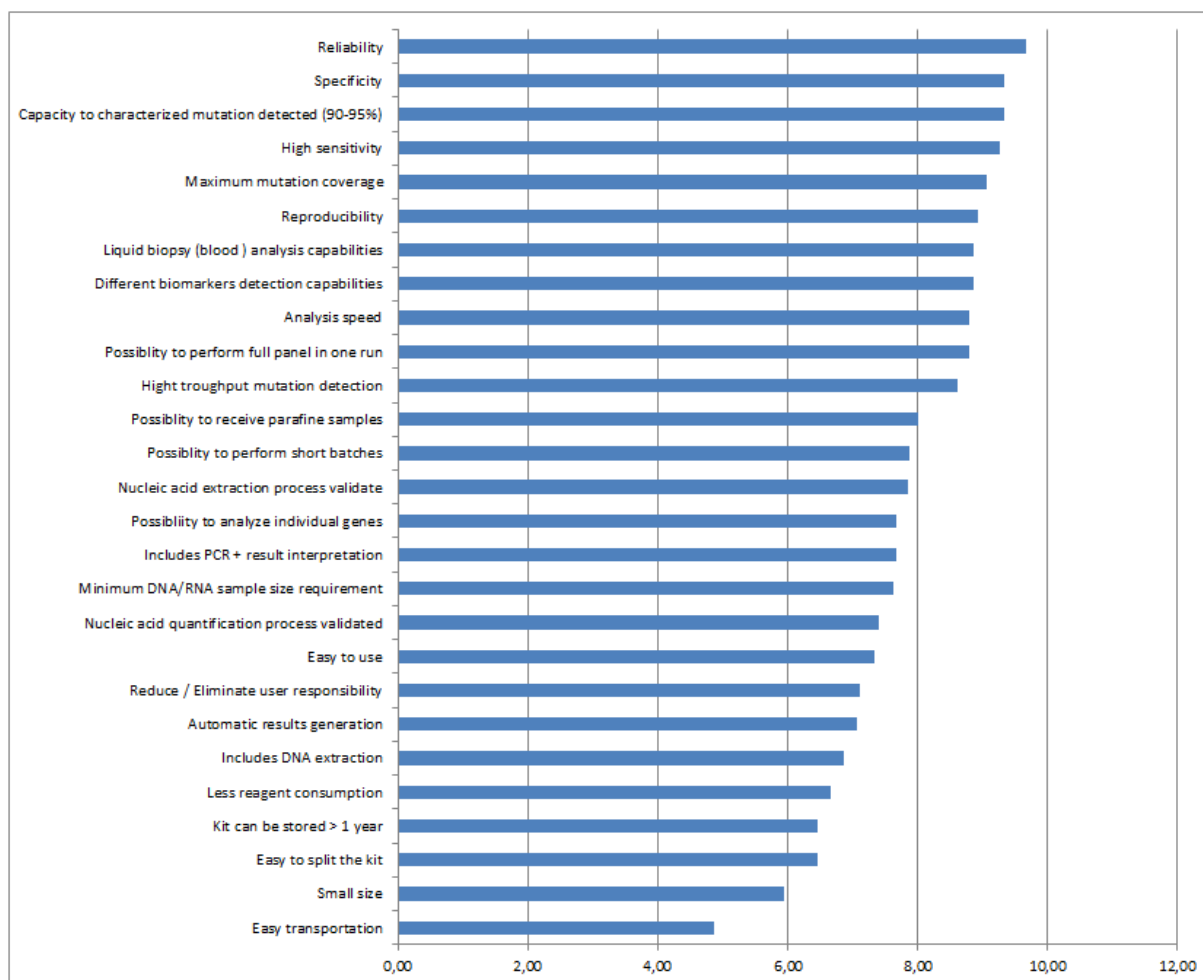


Figure 58: Priorities for a Diagnosis/Prognosis Genetic Analysis Instrument. Scoring list of prefixed priorities.

Most of the consulted experts (73.33%) indicated that having the DNA extracted from the sample in a separated process – and not embedded in the testing system - could have potential advantages, such as 1) the genetic test could be repeated if necessary 2) the sample could be reused in future tests, 3) the specificity could increase, or 4) the same specimen sample could be used for research purposes. (See [Figure 59](#))

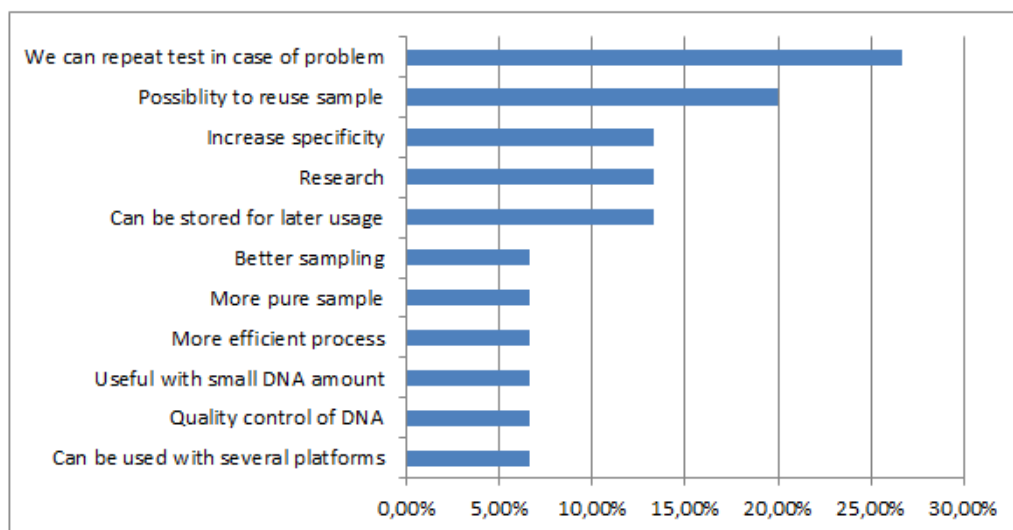


Figure 59: Reasons why having a separated process for DNA extraction can be an advantage in an analysis system.

When experts were asked about desired sensitivity, the average value of all answers was 91.5%, and when asked about specificity 90.24%.

Average number of hours waiting for receiving the test results was 164 hours.¹

Current biomarkers. Biomarkers that are used more often for CRC prognosis are mutations in KRAS and NRAS genes, followed by mutation on gene BRAF and microsatellites instability. The following biomarkers were mentioned in the survey as supporting biomarkers: MLH1, PIK3CA, MH2, MIH, MH6, Toxicity DPD, EGFR, CpG, PMF2, APP, P53, CA19-9, CA, PTEN, PMS2, MSH2, APC, TERT, IDH1/2, MGMT, UGT1A1, PDGFR-alpha, CKIT and MLH2

¹ This value could be biased by the fact that some experts answered the question considering only instrument analysis time, and others answered it considering time elapsed between the test was requested and tests results were sent to the physician.

Future biomarkers. There was no consensus among the surveyed experts about which genes or biomarkers should be used for CRC diagnosis or prognosis in the future. The following genes and biomarkers were mentioned in the survey : PI3K, SRC, EGFR-ECD, HER3, TNRK, AKT, MEK, MET, HER, ALDH1B1, ALDH1A1, LGR5, EPCAM, CD166, CD44, CD29, CD24, CD133, 18QAI, PD1/PDL1, immunity related genes, genetic signatures, angiogenesis related genes, circulating DNA / liquid biopsy.

5.2 Colorectal Cancer DNA Microarray Reader Subsystems Design and Implementation.

Five independent subsystems were designed, prototyped and integrated to build a functional automatic DNA microarray reader. [Table 23](#) summarizes the subsystems implemented.

Table 23. DNA microarray reader subsystem implemented during the process of development.

Subsystem	Degree of successful implementation achieved
Automatic microscopic stage (XY) positioning	100%
Autofocus (Z stage) positioning	100%
Fluorescence sensor based on LED lightning	NOT ACHIEVED
Light Absorbance sensor based on LED lightning	100%
Automatic microarray value reading based on image analysis (AI)	100%

5.2.1 Automatic Microscopic Stage (XY)

An automatic positioning stage for scanning a microarray slide of (75mm x 25mm) on the XY axis was implemented (Figure 60). The slide could hold up to 12 different microarrays. Microarrays had a size of 5 x 5 spots, with a spot size ranging from 400-800 μm . Several prototypes were developed based on a modified version of the commercial microscope stage (Optika B-60 and B-190 microscopes).

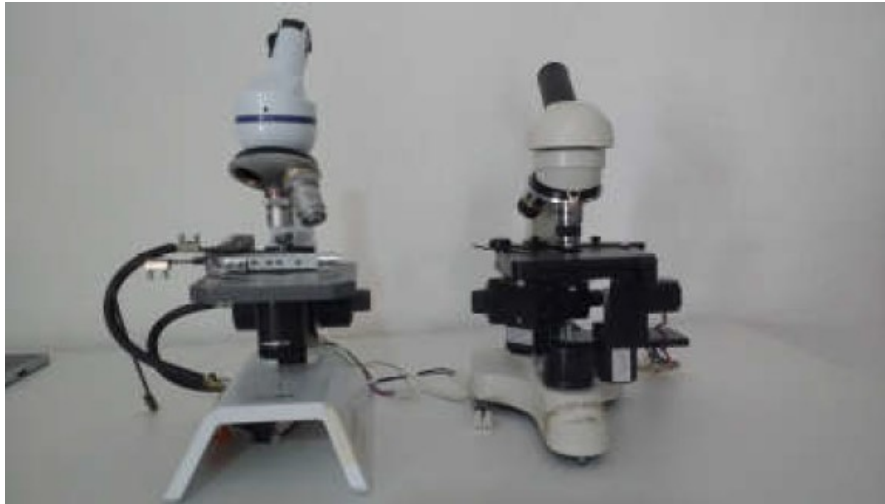


Figure 60: Picture of the two first XY automated stage prototypes with small footprint stage based on Optika B-60 and B-190 microscope stages.

The second prototype configuration included : 1) moving XY stage with microarray plate holder 2) motorization of stage on axis XY (mechanical adaptors and two stepper motors) 3) electronic motorization system (Arduino One board + L6470 Full Stepper Driver board) 4) External PC for stage control. 5) Software for own communication protocol programmed in C (Arduino) and C# (control PC) (See Figure 61)

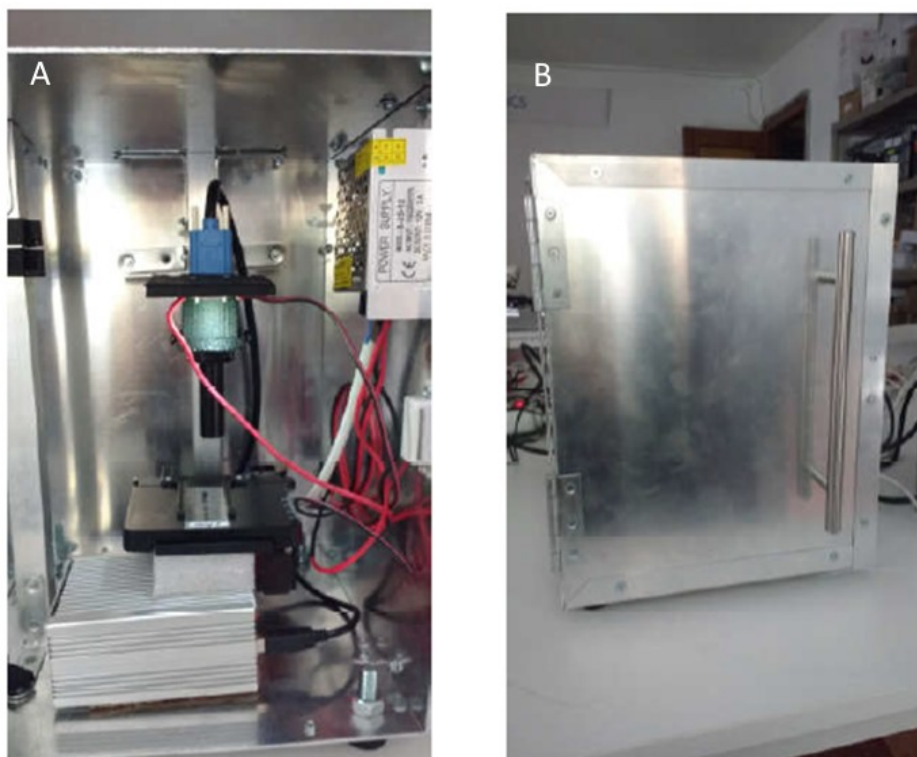


Figure 61: Second XY automated stage prototype, including optical and lighting system and dark chamber for fluorescence microarray sensing (A). External prototype with front door closed (B)

5.2.2 Autofocus Z stage positioning.

Movement on the Z axis was implemented using the microscopic stage (Optika B-190) by plugging a stepper motor to the microscope focus knob. (See [Figure 62](#)). The subsystem allowed for automatic re-focus on the sample between different microarray reads to compensate for the irregularities on the height of the slide. The autofocus was programmed in C# running in a control PC. The image focus calculation was performed using the Matlab® function *estimate_sharpness*. The libraries for image processing were compiled using MCR (Matlab Component Runtime) and integrated into the C# code running on a Windows 7 embedded PC.

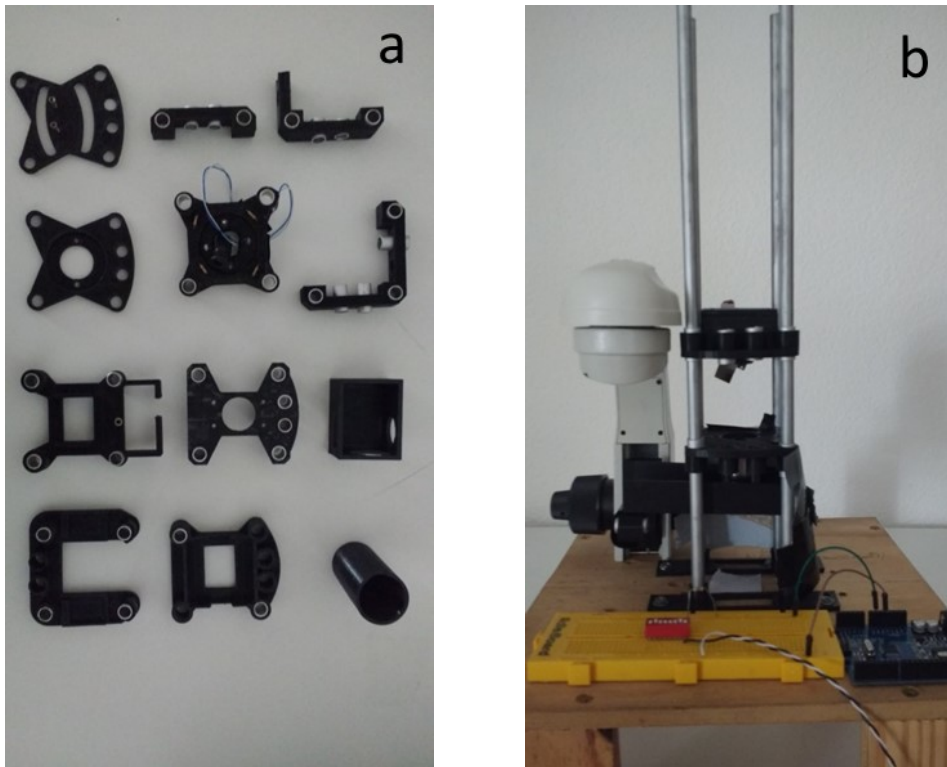


Figure 62 (a) 3D parts designed and PLA printed for motorization of XY and Z stage
(b) First prototype of the Z autofocus stage.

The system initially performed a coarse autofocus at high speed looking for a maximum peak of sharpness and a second pass afterwards to fine focus the lens. The complete autofocus algorithm is described in [Figure 63](#).

5.2.3 Fluorescence sensor based on LED lightning.

It was not possible to achieve a satisfactory level of signal on the microarray spots to guarantee an acceptable signal to noise ratio (SNR) for the application. (See [Figure 64](#)). We continued the development with a light absorbance system instead.

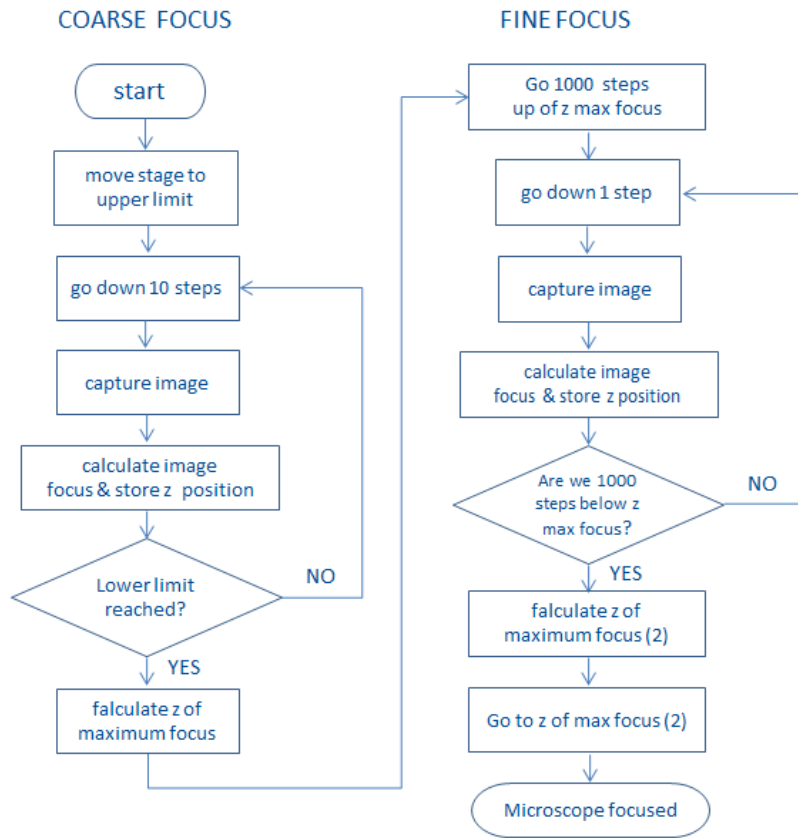


Figure 63: Autofocus algorithm used in the DNA microarray reader.

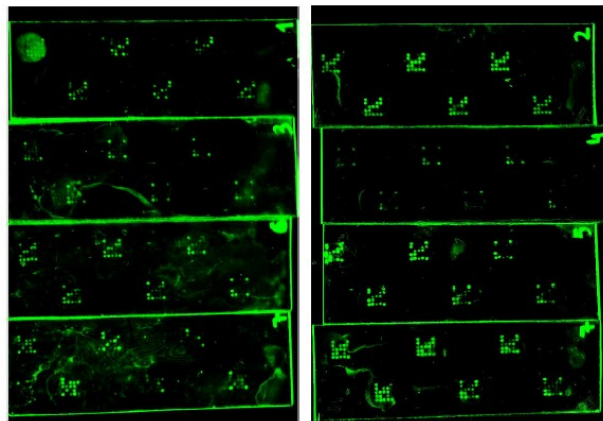


Figure 64: Microarray FITC fluorescence tests. Source: IDM-UPV

5.2.4 Light Absorbance Sensor Based on LED Lighting.

Based on the configuration above described a light absorbance sensor system was implemented. Instead of using fluorescence to detect PCR products, the system used a white lighting system. As shown in [Figure 65](#) the PCR product was measured as dark spots on a similar transparent plastic slide.

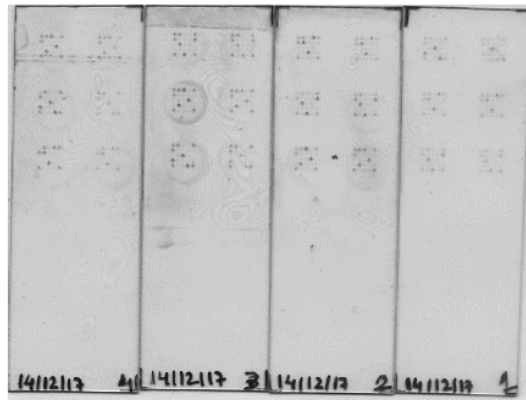


Figure 65: Absorbance microarray prototype. Source: IDM-UPV

5.2.5 Automatic Microarray Spot Value Reading Based on Image Analysis.

This subsystem was implemented programming a C# graphic interface and an IA detection algorithm designed to automatically locate the microarray reference pattern. Reference spots located at the corners of the array were used to identify the array location on the slide (See [Figure 66](#))

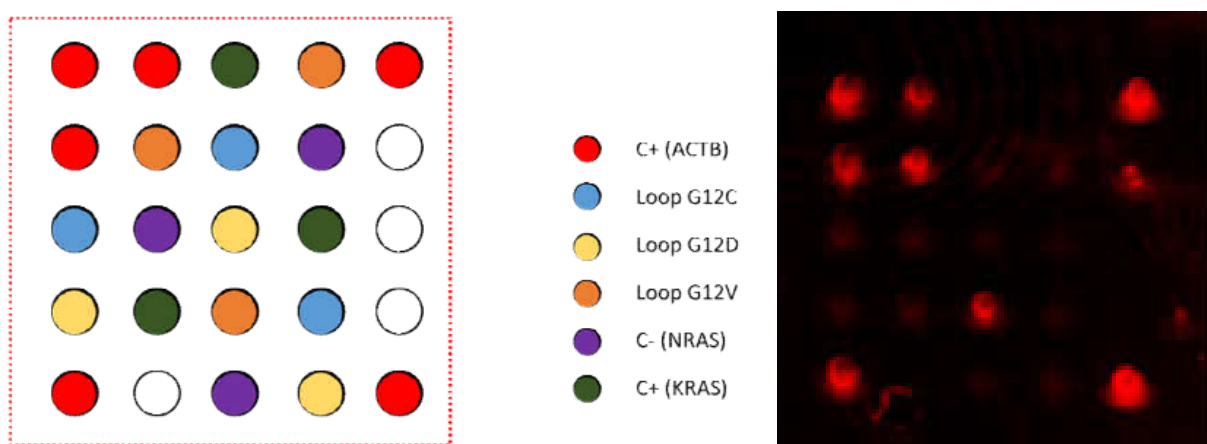


Figure 66 Microarray reference pattern configuration. 6 Spots located at the array corners (in red, C+) are used for microarray location by the image-based AI reader.

The CCD camera captured the image, removed the background, and processed the image for the microarray reference pattern identification. The accuracy of positioning was 100% with the five slides used in the control test. (See [Figure 67](#))

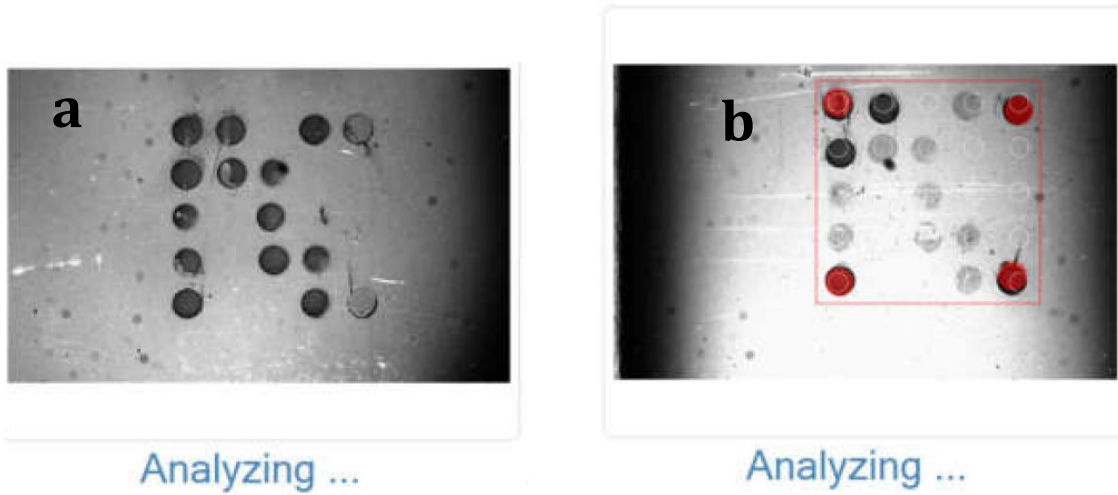


Figure 67: Example of microarray image (a) and reference pattern automatic location with AI-based image analysis (b).

5.3 Improved Methodologies for Variant Interpretation

5.3.1 Experts Panel Survey.

We conducted personal interviews to a group of six geneticists with high expertise in variant interpretation. [Appendix 9.8](#) summarizes the most relevant findings of these surveys. Among the surveyed geneticists the most popular types of analysis were WES (Whole Exome Sequencing) and NGS panels. Most of the respondents used Windows (6) or Linux (5) and received variants to be analyzed in either VCF or CSV format. The average number of patients analyzed by each geneticists per year was 152 (min=2, max=400). Samples were from different pathologies including cancer, rare diseases, retinal dystrophies, hereditary cardiopathies, ataxias, neurology and developmental disorders. The most commonly drawbacks found in the current systems were 1) internet dependency, 2) databases not updated, 3) not fully automatic, 4) deficient results and variant classification, 5) lack of database integration, 6) requirement of human intervention, or 7) need to use different databases and tools. When asked about the most important features of a variant interpretation system, respondents emphasized the following features in the following order: 1) ensure patient genetic data security, 2) reliability, 3) training for system use, 4) reproducibility, 5) telephonic & email technical support, 6) specificity, 7) sensitivity and, 8) connection to own information systems.

5.3.2 BINOME. Semi-assisted Variant Interpretation Methodology.

We developed a new methodology, which we called BINOME system. It automated most of the repetitive tasks that geneticist did manually, such as database access and variant filtering. Therefore, BINOME could save between 50% and 80% of geneticist analysis time.²

In [Figure 68](#) we represent the flow diagram of the main tasks performed by the BINOME system.

² Estimate rovided by Dr. Laia Pedrola, Genomics Unit Laboratory Responsible at Health Research Institute Hospital La Fe. Estimate based on preliminary variant analysis tests performed in the unit.

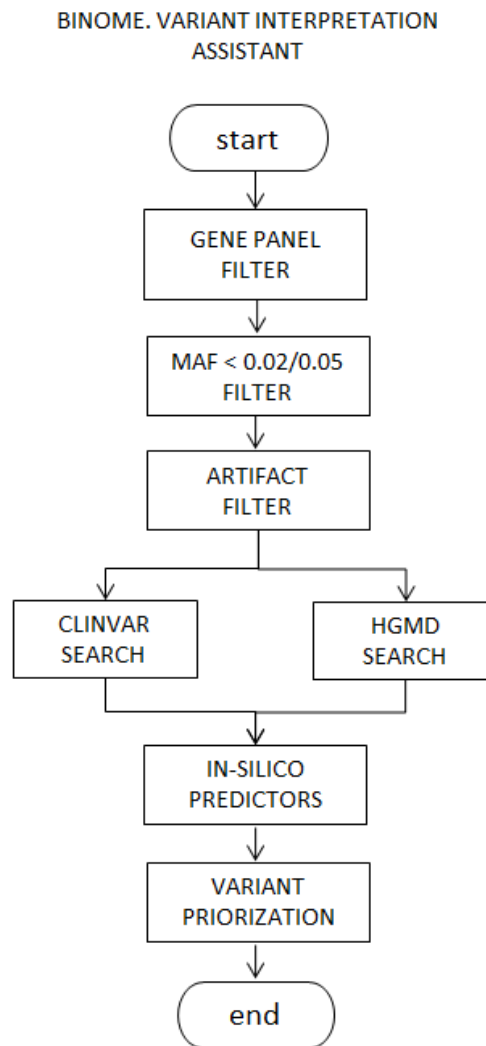


Figure 68: Process flow diagram of the BINOME system. The method is conceived to automate the most tedious parts of the variant interpretation process.

The BINOME system was initially tested with ten patient samples coming from the Health Research Institute La Fe – Genomics Unit mimicking the manual process performed by their geneticists. For each sample patient the geneticist team had selected a single variant to be classified as pathogenic and to be reported to the responsible physician for diagnosis purposes.

As a regular practice, the laboratory evaluated each sample using two separated geneticists that performed the same analysis in parallel, without sharing any information to avoid biases. One of the experts evaluated each sample manually with the only assistance of MS Excel® software, and a second expert performed the same

evaluation with the assistance of the Cartagenia variant analysis software (Agilent). In our evaluation all 10 samples had a single variant identified as pathogenic and both geneticists agreed with the provided finding.

The new method automated several processes that were traditionally performed manually by geneticists. (Table 24)

Table 24: Simplified summary of the tertiary analysis processes performed by geneticists and selection of processes that are automated by the BINOME system.

Variant Step	Interpretation	Automated by BINOME	Technology used
Gene filtering		YES	Python script
MAF filtering		YES	Python script
Artifact probability		YES	Python script, AI for artifact classification.
Clinvar Database search		YES	Database integration
HGMD Database search		YES	Web scrapping.
In-silico predictors		YES	Database integration (dbnsfp), webscrapping (Polyphen2, SIFT, Provean and Mutation Taster)
Gene Structure (splicing)		NO	N.A.
Variant prioritization		YES	AI classifier.
Other databases search		NO	N.A

The VCF files received had an average of 6010 variants (SD 1535). After applying the set of automated process performed by the BINOME system the output was a subset of the input variants. The average number of variants selected by the system was 4 (SD 2.62).

Figure 69 shows a graphic representation of the number of variants that entered the system and were selected for further analysis.

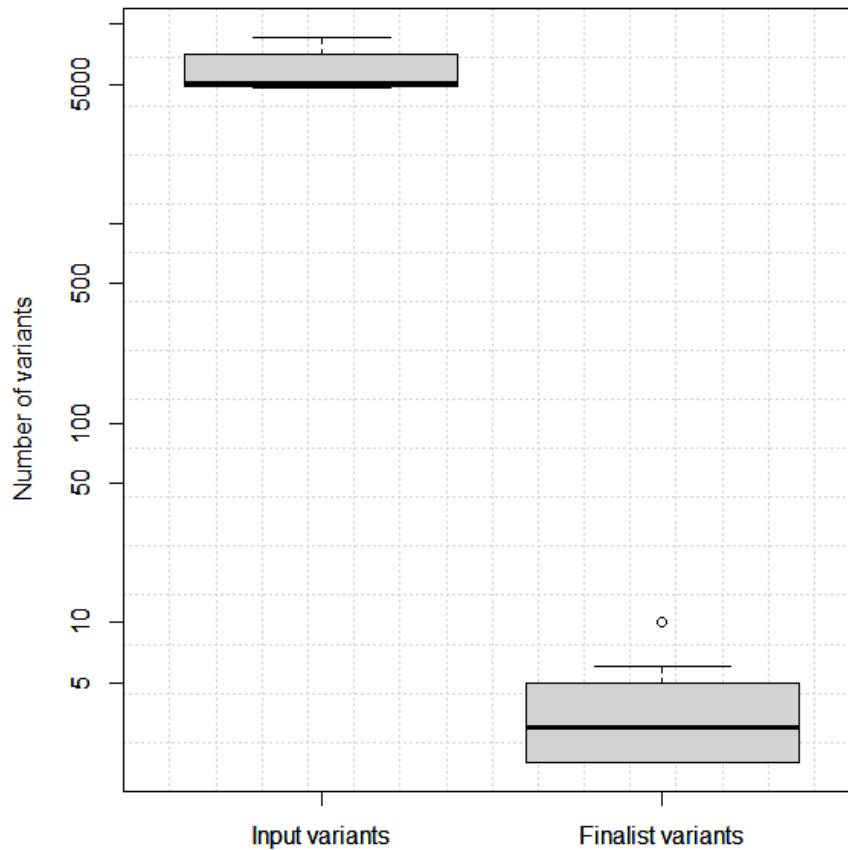


Figure 69: Summary of the variant filtering process performed by BINOME. Note that Y-axis is presented in logarithmic scale.

On average 99.93% of the input variants were filtered out by the automatic BINOME system, leaving 0.07% of variants to be reviewed manually by the geneticists. In the selected sample cohort the system showed a sensitivity of 100% coupled with 99.95% specificity. Table 20 describes the main performance indicators of the BINOME system.

Table 25: Performance showed by the BINOME system compared to manual analysis by expert geneticists.

	Manual & Manual+Cartagenia	BINOME
Sensitivity	100%	100%
Specificity	100%	99.95%
Positive Predictive Value	100%	25%
Negative Predictive Value	100%	100%

5.3.3 New Method for Artifact Classification

The best performing AI algorithm for artifact classification was a neural network with 4 layers and an input dimension of 8. The first hidden layer was composed by hyperbolic tangent functions with dimensionality of 16. The layer takes any real value as input and output a number between -1 and 1. The second hidden layer was composed by rectified linear activation functions (ReLU) with dimensionality of eight. The output layer was composed of a sigmoid function that outputs 1 if the NN considers the sample an artifact or 0 otherwise.

The proposed method reported concordance with human classification in 97.7% of cases in a sample of 45 variants randomly extracted from the original 158 variants sample. Precision reached 94.11% with a recall of 100%. (See [Figure 70](#))

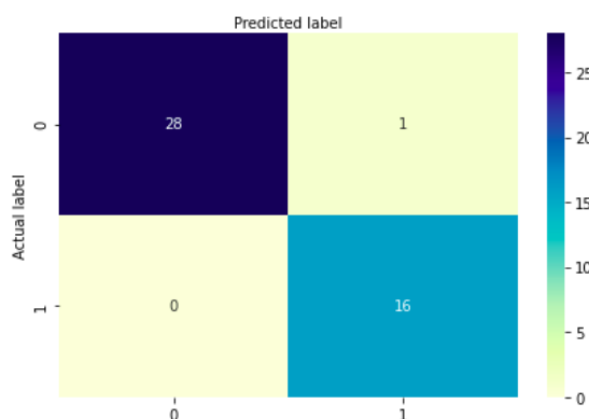


Figure 70: Confusion matrix. Upper-right (true positives), upper-left (false positives), lower-right (false negatives), lower-left (true negatives)

5.3.4 New Method for Sensitivity Increase in Variant Interpretation.

The best performing AI algorithm for variant classification and sensitivity increase was a neural network with 3 layers and an input dimension of 5. The first hidden layer was composed by sigmoid functions with dimensionality of 10. The output layer was composed of a sigmoid function that outputs 1 if the NN considers the sample a potentially pathogenic variant or 0 if potentially benign. Several tested algorithms based on decision trees provided also both acceptable and robust results (and might be considered for future evolutions of the method)

The new methodology was tested in a proof of concept cohort of 13 individuals that were apparently healthy at the time of the genetic test.

Using the proposed methodology the following results were generated:

- One pathogenic variant were identified in one individual the cohort (Thoracic Aortic Aneurism predisposition). The pathogenicity of the variant was validated with past family history: father of the proband.
- Two pathogenic variants were identified for two different recessive disorders (Muscular dystrophy and Rubinstein-Taybi syndrome). One of these two variants was confirmed with past family history: aunt of the proband.
- One High Risk VUS variant was identified (Connective tissue disorder predisposition), with no family health history validation.

The percentage of high risk uncertain variants compared to the total number of uncertain variants was 2.67% on the tested sample, giving an average of 43 extra variants per sample that required further manual interpretation by the geneticist.

6 Discussion

We have interviewed a total of 21 genetics experts in clinical diagnosis and/or variant interpretation regarding their preferences for genetic analysis systems. Experts were geneticists or oncologists that were used to interpret genetic analysis results. They were surveyed to acquire a better understanding of what requirements and features are desired by the users of genetic analysis systems. We did not find any similar survey in the academic bibliography for this matter.

The most requested features were sensitivity and specificity, analysis time, mutations coverage, reproducibility and capability to work with small amounts of sample.

Another relatively surprising finding was that while most oncologists agreed on the usage of mutation on KRAS, NRAS and BRAF genes, there is a lack of consensus about the usage of additional biomarkers in clinical practice. Oncologists surveyed were using 23 different additional genes with a low degree of coincidence among answers. When they were asked about genes that should be taken into account in the future, a set of 25 different genes were mentioned by different respondents showing clear disagreement. Lorans M et al. (2018) also remarked a similar lack of consensus regarding inclusion criteria in CRC gene panels.

6.1 mCRC mutation detection with DNA microarray

We successfully developed a microarray automatic reader based on absorbance measurement, including the development of an automated stage, an autofocus system and automatic image analysis based on AI to read the microarrays spots and generate the clinically relevant information for detection of CRC mutations. This prototype reader could be used in further applications that may require microarray reading, image capturing or automatic positioning.

We tested several configuration of a reader system based on fluorescence LED lightning but we did not achieve satisfactory results. Even after testing the best combination of material found which included highest sensitivity CCD fluorescence cameras, best specifications for fluorescence filters on the market (Semrock) and high power LED of 30W. However, the amount of technical problems found suggested us that incremental improvements in this direction were not achievable with our current setup.

The single similar system that we found in academic bibliography was analyzed by Pierzchalski et al. (2009). They analyzed an instrument with a similar configuration suitable for fluorescence microarray analysis, the Lumisens system from Sensovation AG. The system was equipped with 10-LEDs and a 8.3 MPixel 16-bit CCD camera that achieved detection sensitivity of solutions as low as 0.0004 µg/ml. The higher sensitivity achieved by this device was probably due to the usage of several high power LEDs (> 100W) and light concentration lenses to focus the light in a reduced area of the slide. The lack of concentration lenses and very high power LEDs are considered the main reasons of our limited capacity to detect acceptable fluorescence levels with our setup.

Besides, the optimization of the multiplex PCR designed in the ONCOMARKER project proved to be a difficult endeavor. (Results not available in the present document). As described by Markoulatos P et al. (2002) and Sint D et al. (2012), the design and optimization of a multiplex PCR should take into account multiple factors such as primer specificity, primer efficiency, thermo cycling conditions, assay sensitivity etc. Setting up a system for detection of 13 mCRC mutations during the ONCOMARKER took more time and effort than initially forecasted. Further increase in the number of mutations will certainly increase the difficulty of the multiplex PCR reaction design, reducing the system flexibility for future expansion. Microarrays systems tend to be easier to use than their NGS counterparts. But they are less flexible and more difficult to design and modify in case they need to be adapted to new mutations detection.

We believe that despite some of the problems and limitations NGS technology, it will progressively displace the use of microarrays and other technologies for clinical diagnostic and prognostic applications thanks to its broad analysis genetic range. We consider that system flexibility and evolution capacity is becoming a must for any

genetic diagnosis instrument, given the constant changing nature of the genomics sector that require constant adaptation to new discoveries.

6.2 Interpretation of genetic variants for genetic diseases diagnosis.

We successfully analyzed the current and historic state of the art of NGS variant interpretation recommendations and guidelines followed by most laboratories around the world. According to Amendola et al. (2016) only 34% concordance between variant interpretations across laboratories was found, which in our opinion points out the high amount of uncertainty, human dependency and lack of reproducibility of laboratories processes, and by extension of the current ACMG guidelines. Based on our analysis we identified several point of improvement in current variant diagnosis systems such as automation of repetitive tasks and database information retrieval, automatic artifact identification and sensitivity increase for specific applications.

With our proposed methodologies we automated more than 80% of the repetitive tasks performed by geneticists for some specific genetic analysis clinical applications. Besides we predicted the variant probability of being an artifact, and defined a method that increased the amount of high risk variants reported by 7.7%. We estimate that this new method has the potential to increase the equivalent diagnosis yield by 5-15% by increasing sensitivity when compared to the strict application of current ACMG guidelines. (Richards S et al., 2015). In this sense, we have achieved significant improvements that could be applied to the process of many genetic laboratories workflows for variant interpretation.

For both the artifact classification methodology, it would be desirable to train the system with a higher number of samples to increase accuracy of the AI system. We suggest this method to be used only for variant prioritization, avoiding using it to filter out variants in a standard clinical variant interpretation pipeline since by doing so we could be reducing the system sensitivity. The overall system detecting capabilities would remain unaffected, and substantial time savings would be achieved with the proposed approach for variant prioritization.

Also, care should be taken when applying the proposed method for sensitivity increase. This method comes at the expense of an increased rate of false negatives. ACMG guidelines do not recommend reporting any VUS for clinical diagnosis. Physicians could

wrongly interpret the reported VUS as pathogenic (when it is not), and they could take a wrong medical decision prescribing a suboptimal treatment to the patient.

Our recommendation would be to use this increased sensitivity method only under the following specific situations:

- 1) When the geneticist and the physician with genetic background agree on the pathogenicity probability of the high risk VUS reported.
- 2) When the treatment to be provided to the patient has no negative consequences even if the variant reported as pathogenic happens to be a false positive.
- 3) For preventive medicine applications, where there are usually no harmful prescriptions for consultants.

Additionally, the proof of concept of the improved sensitivity methodology should be completed with a larger population sample. It is planned to increase the number of samples to 200-300 individuals and validate the method by confirming each high risk VUS found with the proband or the proband's family phenotypes.

6.3 The future of variant interpretation.

Currently, there is a relative lack of reproducibility between results from different laboratories, the increasing complexity of the constant revisions of the ACMG guidelines in recent years and several "gaps" that the standard leaves to the interpretation of the geneticist. Therefore, we venture to predict that the ACMG guidelines will undergo further revisions in the near future.

The current ACMG guidelines are already out-dated, at least database and in-silico predictor wise. Some companies such as Invitae (USA) and Varsome (Switzerland) are already applying their own optimizations and improvements ahead of the ACMG.

Other reviews of the ACMG guidelines arise to modify or complete them. Nykamp et al. (2017) proposed a refinement that protected against overcounting conceptually related evidence and replaced the rigid "clinical criteria" style of the guidelines with semi quantitative criteria. Houge G et al. (2021) proposed an ABC system that complements the ACMG guidelines separating the grading based on functional effect, and introducing the penetrance concept on the variant interpretation.

From a strictly technical point of view, our view is that ACMG has tried to adapt the evidence-based diagnostic methodology that works quite well in other clinical sectors in a field of enormous complexity such as the interpretation of genetic variants. We have seen how the complexity of the proposed guidelines has progressively increased over time to the point that at present the application of these guidelines is practically impossible without automated systems.

ACMG guidelines represent a trade-off between simplicity, flexibility and accuracy. From the point of view of an AI systems expert, ACMG is trying to solve a highly complex problem such as the classification of genetic variants with a relatively simple algorithm based on a sort of decision tree. This push for simplicity is explained by the need for the guidelines to be fairly well understood by a human geneticist while keeping inter-laboratory results reproducibility and sensitivity high. The price paid by ACMG in their current guidelines is that they leave several high-uncertainty decisions to the discretion of the geneticist as a way to alleviate the relatively limited and rigid capabilities of the variant classification algorithm suggested.

In our opinion, in order to increase the accuracy and sensitivity of the method in future revisions they would need to sacrifice simplicity. Nevertheless, the current 2015 revision is already excessively complex and impractical for most geneticists to apply manually. Additionally, ACMG guidelines pretend to be a one-size-fits-all solution for all clinical applications, which also limits its adaptive capabilities to disease specific applications.

6.4 Future research based on the present work

NGS technology continues to increase its presence in clinical applications, at the same time that clinical databases multiply in number and size thanks to the constant increase of scientific findings. In our view, NGS will continue to evolve in this direction unless another disruptive technology for DNA sequencing appears in the near future.

We suggest and plan to continue this research through increasing the number of samples analyzed, up to 200-300 samples, and compare them with alternative systems. Afterwards, and based in our experience with this research, we will consider to optimize the ACMG guidelines and presented methodology for specific pathologies or clinical

conditions such as cancer, cardiovascular or rare diseases that may require specific filtering adjustments.

As per the evolution of the AI systems integrated in the variant interpretation method, we wish to continue exploring future improvements with Graph Neural Networks (GNNs) that intuitively may fit the variant interpretation process under study in this document.

In futures lines of research, the current work performed with genetic database and in silico predictors' integration could be adapted to systems that integrate thousands of patients samples coupled with their corresponding health history for clinical research purposes. In case that samples available for training and testing increased to ten/hundreds of thousands we could adapt our systems to Deep Learning, Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) that seem to be well suited to infer knowledge from complex and large amounts of genomic data. (Dias R and Torkamani, 2019)

6.5 The future of variant interpretation and genomics clinical research

We envision a promising line of research to continue optimizing the interpretation of genetic variants with AI algorithms, especially in classification and prioritization as an extension of the present research and the adaptation of the guidelines to specific clinical applications such as diagnosis and prognosis of heart disease, cancer and rare diseases.

We consider that the challenge of genetic variant interpretation is especially well suited to the capabilities of modern AI system. Variant interpretation requires well-trained experts, automation, large amount of data analysis, and physicians knowledgeable in genetics for patients receive the best possible treatment. However, it is a cumbersome process to define clear rules that can be followed unequivocally by humans or machines.

<p>On the clinical research side, I believe that with the right amount of computer power, storage capacity, and a large enough set of human samples (genotype-phenotype), no genomics challenge will resist to artificial intelligence.</p>

At the present there is a shortage of this kind of databases, so that computational biology and bioinformatics researchers still cannot take full advantage of the capabilities of modern AI systems; although some initiatives have been arising recently in that direction (Schatz MC et al., 2022) (Li et al., 2021)

7 Conclusions

1. Following two survey among 21 genetics experts the most requested features for genetic analysis system to be used in clinical setups were 1) sensitivity 2) specificity 3) analysis time 4) coverage 5) reproducibility 6) capacity to work with small amount of sample.
2. In a survey performed to 9 oncologists, the genes that were currently used as biomarkers in clinical practice with strong consensus were KRAS, NRAS and BRAF. Other 25 genes are currently used by different oncologist interviewed independently with no clear consensus among them.
3. We have designed and built up from scratch two innovative genetic analysis systems based on AI for clinical diagnosis and prognosis. One system targeted for detection of the most prevalent mutations of mCRC using a multiplex PCR microarray (ONCOMARKER), and the second one designed to perform the tertiary analysis of a standard NGS genetic analysis workflow (BINOME)
4. The ONCOMARKER microarray reader incorporates an automatic XY stage for microarray positioning, lightning system, and AI-based software for automatic microarray placement and sample analysis.
5. The BINOME system is able to save between 50%-80% of geneticist hands-on analysis time by automating database access and in silico predictors. It incorporates an AI-based artifact-detection algorithm that was tested with 94.11% precision and 100% recall. In includes an improved sensitivity filter to detect high risk VUS that has the potential to increase the diagnosis yield by 5%-15%. This new method is considered suitable for specific diagnosis cases and preventive medicine applications.

8 References

- ACMG recommendations** for standards for interpretation of sequence variations. *Genet Med.* 2000;2(5):302-303. doi:[10.1097/00125817-200009000-00009](https://doi.org/10.1097/00125817-200009000-00009)
- Amendola LM**, Jarvik GP, Leo MC, et al. Performance of ACMG-AMP Variant-Interpretation Guidelines among Nine Laboratories in the Clinical Sequencing Exploratory Research Consortium. *Am J Hum Genet.* 2016;98(6):1067-1076. doi:[10.1016/j.ajhg.2016.03.024](https://doi.org/10.1016/j.ajhg.2016.03.024)
- Behjati S**, Tarpey PS. What is next generation sequencing? *Arch Dis Child Educ Pract Ed.* 2013;98(6):236-238. doi:[10.1136/archdischild-2013-304340](https://doi.org/10.1136/archdischild-2013-304340)
- Dias R**, Torkamani A. Artificial intelligence in clinical and genomic diagnostics. *Genome Medicine.* 2019;11(1):70. doi:[10.1186/s13073-019-0689-8](https://doi.org/10.1186/s13073-019-0689-8)
- Gauthier M**. Simulation of polymer translocation through small channels: A molecular dynamics study and a new Monte Carlo approach. Published online September 1, 2007.
- Gelb BD**, Cavé H, Dillon MW, et al. ClinGen's RASopathy Expert Panel consensus methods for variant interpretation. *Genet Med.* 2018;20(11):1334-1345. doi:[10.1038/gim.2018.3](https://doi.org/10.1038/gim.2018.3)
- Ghosh R**, Harrison SM, Rehm HL, Plon SE, Biesecker LG, ClinGen Sequence Variant Interpretation Working Group. Updated recommendation for the benign stand-alone ACMG/AMP criterion. *Hum Mutat.* 2018;39(11):1525-1530. doi:[10.1002/humu.23642](https://doi.org/10.1002/humu.23642)
- Green RC**, Berg JS, Grody WW, et al. ACMG recommendations for reporting of incidental findings in clinical exome and genome sequencing. *Genet Med.* 2013;15(7):565-574. doi:[10.1038/gim.2013.73](https://doi.org/10.1038/gim.2013.73)
- He MM**, Li Q, Yan M, et al. Variant Interpretation for Cancer (VIC): a computational tool for assessing clinical impacts of somatic variants. *Genome Med.* 2019;11(1):53. doi:[10.1186/s13073-019-0664-4](https://doi.org/10.1186/s13073-019-0664-4)

- Hershberger RE**, Givertz MM, Ho CY, et al. Genetic evaluation of cardiomyopathy: a clinical practice resource of the American College of Medical Genetics and Genomics (ACMG). *Genet Med*. 2018;20(9):899-909. doi:[10.1038/s41436-018-0039-z](https://doi.org/10.1038/s41436-018-0039-z)
- Hershberger RE**, Givertz MM, Ho CY, et al. Genetic Evaluation of Cardiomyopathy-A Heart Failure Society of America Practice Guideline. *J Card Fail*. 2018;24(5):281-302. doi:[10.1016/j.cardfail.2018.03.004](https://doi.org/10.1016/j.cardfail.2018.03.004)
- Hoskinson DC**, Dubuc AM, Mason-Suares H. The current state of clinical interpretation of sequence variants. *Curr Opin Genet Dev*. 2017;42:33-39. doi:[10.1016/j.gde.2017.01.001](https://doi.org/10.1016/j.gde.2017.01.001)
- Houge G**, Laner A, Cirak S, de Leeuw N, Scheffer H, den Dunnen JT. Stepwise ABC system for classification of any type of genetic variant. *Eur J Hum Genet*. Published online May 13, 2021:1-10. doi:[10.1038/s41431-021-00903-z](https://doi.org/10.1038/s41431-021-00903-z)
- Kalia SS**, Adelman K, Bale SJ, et al. Recommendations for reporting of secondary findings in clinical exome and genome sequencing, 2016 update (ACMG SF v2.0): a policy statement of the American College of Medical Genetics and Genomics. *Genet Med*. 2017;19(2):249-255. doi:[10.1038/gim.2016.190](https://doi.org/10.1038/gim.2016.190)
- Kelly MA**, Caleshu C, Morales A, et al. Adaptation and validation of the ACMG/AMP variant classification framework for MYH7-associated inherited cardiomyopathies: recommendations by ClinGen's Inherited Cardiomyopathy Expert Panel. *Genet Med*. 2018;20(3):351-359. doi:[10.1038/gim.2017.218](https://doi.org/10.1038/gim.2017.218)
- Ku CS**, Loy EY, Salim A, Pawitan Y, Chia KS. The discovery of human genetic variations and their use as disease markers: past, present and future. *J Hum Genet*. 2010;55(7):403-415. doi:[10.1038/jhg.2010.55](https://doi.org/10.1038/jhg.2010.55)
- Lander ES**, Linton LM, Birren B, et al. Initial sequencing and analysis of the human genome. *Nature*. 2001;409(6822):860-921. doi:[10.1038/35057062](https://doi.org/10.1038/35057062)
- Li MM**, Datto M, Duncavage EJ, et al. Standards and Guidelines for the Interpretation and Reporting of Sequence Variants in Cancer: A Joint Consensus Recommendation of the Association for Molecular Pathology, American Society of Clinical Oncology, and College of American Pathologists. *J Mol Diagn*. 2017;19(1):4-23. doi:[10.1016/j.jmoldx.2016.10.002](https://doi.org/10.1016/j.jmoldx.2016.10.002)

- Li B**, Wang Z, Chen Q, et al. GPCards: An integrated database of genotype–phenotype correlations in human genetic diseases. *Computational and Structural Biotechnology Journal*. 2021;19:1603-1611. doi:[10.1016/j.csbj.2021.03.011](https://doi.org/10.1016/j.csbj.2021.03.011)
- Lorans M**, Dow E, Macrae FA, Winship IM, Buchanan DD. Update on Hereditary Colorectal Cancer: Improving the Clinical Utility of Multigene Panel Testing. *Clinical Colorectal Cancer*. 2018;17(2):e293-e305. doi:[10.1016/j.clcc.2018.01.001](https://doi.org/10.1016/j.clcc.2018.01.001)
- Markoulatos P**, Siafakas N, Moncany M. Multiplex polymerase chain reaction: a practical approach. *J Clin Lab Anal*. 2002;16(1):47-51. doi:[10.1002/jcla.2058](https://doi.org/10.1002/jcla.2058)
- Miller DT**, Lee K, Chung WK, et al. ACMG SF v3.0 list for reporting of secondary findings in clinical exome and genome sequencing: a policy statement of the American College of Medical Genetics and Genomics (ACMG). *Genet Med*. 2021;23(8):1381-1390. doi:[10.1038/s41436-021-01172-3](https://doi.org/10.1038/s41436-021-01172-3)
- Miller DT**, Lee K, Gordon AS, et al. Recommendations for reporting of secondary findings in clinical exome and genome sequencing, 2021 update: a policy statement of the American College of Medical Genetics and Genomics (ACMG). *Genet Med*. 2021;23(8):1391-1398. doi:[10.1038/s41436-021-01171-4](https://doi.org/10.1038/s41436-021-01171-4)
- Morales A**, Kinnamon DD, Jordan E, et al. Variant Interpretation for Dilated Cardiomyopathy. *Circulation: Genomic and Precision Medicine*. 2020;13(2):e002480. doi:[10.1161/CIRCGEN.119.002480](https://doi.org/10.1161/CIRCGEN.119.002480)
- Nykamp K**, Anderson M, Powers M, et al. Sherlock: a comprehensive refinement of the ACMG–AMP variant classification criteria. *Genet Med*. 2017;19(10):1105-1117. doi:[10.1038/gim.2017.37](https://doi.org/10.1038/gim.2017.37)
- Pierzchalski A**, Marecka M, Müller HW, Bocsi J, Tárnok A. Evaluation of Slide Based Cytometry (SBC) for concentration measurements of fluorescent dyes in solution. *Progress in Biomedical Optics and Imaging - Proceedings of SPIE*. 2009;7182. doi:[10.1117/12.808797](https://doi.org/10.1117/12.808797)
- Richards CS**, Bale S, Bellissimo DB, et al. ACMG recommendations for standards for interpretation and reporting of sequence variations: Revisions 2007. *Genet Med*. 2008;10(4):294-300. doi:[10.1097/GIM.0b013e31816b5cae](https://doi.org/10.1097/GIM.0b013e31816b5cae)

- Richards S**, Aziz N, Bale S, et al. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet Med*. 2015;17(5):405-424. doi:[10.1038/gim.2015.30](https://doi.org/10.1038/gim.2015.30)
- Sanger F**, Coulson AR. A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *J Mol Biol*. 1975;94(3):441-448. doi:[10.1016/0022-2836\(75\)90213-2](https://doi.org/10.1016/0022-2836(75)90213-2)
- Sanger F**, Nicklen S, Coulson AR. DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci U S A*. 1977;74(12):5463-5467. doi:[10.1073/pnas.74.12.54631](https://doi.org/10.1073/pnas.74.12.54631).
- Schatz MC**, Philippakis AA, Afgan E, et al. Inverting the model of genomics data sharing with the NHGRI Genomic Data Science Analysis, Visualization, and Informatics Lab-space. *Cell Genomics*. 2022;2(1). doi:[10.1016/j.xgen.2021.100085](https://doi.org/10.1016/j.xgen.2021.100085)
- Šimundić AM**. Measures of Diagnostic Accuracy: Basic Definitions. *EJIFCC*. 2009;19(4):203-211.
- Sint D**, Raso L, Traugott M. Advances in multiplex PCR: balancing primer efficiencies and improving detection success. *Methods Ecol Evol*. 2012;3(5):898-905. doi:[10.1111/j.2041-210X.2012.00215.x](https://doi.org/10.1111/j.2041-210X.2012.00215.x)
- Venter JC**, Adams MD, Myers EW, et al. The Sequence of the Human Genome. *Science*. Published online February 16, 2001. doi:[10.1126/science.1058040](https://doi.org/10.1126/science.1058040)

9 Appendix

9.1 PCT patent filed : “Particle counting system adaptable to an optical instrument”

ABSTRACT

PARTICLE COUNTING SYSTEM ADAPTABLE TO AN OPTICAL INSTRUMENT

The invention describes a microscopic counting system, designed for counting particles, in particular microorganisms, with a microscope or a magnifying glass. The system is adaptable to any microscope or magnifying glass. It allows analyzing biological samples and/or particles that have previously been prepared for observation and introduced into a counting chamber or other container (Neubauer Chamber, Thoma Chamber, etc.) but also can be used with microorganisms on its culture medium without exposing the culture (Flasks, Petri dishes, bioreactors, etc.). Using our own calibration method, the system allows calculating automatically or in a semi-assisted manner the cell concentration (or of particles) in the sample quickly and efficiently.

Technical field of the invention

The present invention is related to the field of cell and particle counting. In particular, it is related to the cell and particle counting systems with magnifying glass or microscope.

More specifically, is related to the counting systems with microscope or magnifying glass based on image analysis that carry out an automatic, semi-automatic or semi-assisted counting.

State of the Art

Various old techniques of cell count such as the one disclosed by the patent US-5135302 (Flow cytometer) or by the patent US-1918351 (hemacytometer/Neubauer Chamber) are known. These techniques have various disadvantages and problems.

Among which are highlighted the following:

Counting in Neubauer Chamber has the following disadvantages and problems:

- (a) human counting errors;
- (b) statistical errors;
- (c) errors in the calculation of concentration, when applying the corresponding mathematical formula;
- (d) low reproducibility / high variance in the measurements;
- (e) very tedious and monotonous process for the laboratory technicians;
- (f) difficult process for people with visual impairments;
- (g) tiring process for the view of the laboratory technicians.

Counting in flow cytometer has the following disadvantages and problems:

- (a) destruction of the sample when performing the counting;
- (b) system with a high maintenance cost;
- (c) system requires periodic calibrations;
- (d) if the system is not used often, it is damaged;
- (e) the system does not allow the visual observation of the samples by the technicians.

The counting chambers of the type described by US 1918351 (Neubauer Chamber) are chambers adapted with a bright field or a phase contrast microscope. They consist, generally, of a slide with a depression in the centre, at the bottom of which a grid with a given size has been marked with the help of a diamond, with a known separation between two consecutive lines. For counting the cells the reticle is observed with a microscope with the suitable magnification and the cells are counted.

Based on the number of cells counted, knowing the liquid volume that the reticle field holds, the concentration of cells per volume unit of the initial liquid sample is calculated.

One of the problems of this technique lies in the inaccuracy that occurs when the count is carried out, since a statistical formula that introduces certain error is used, and also the human error is present. This inaccuracy results in a lower reliability and reproducibility (reliability= absence of error on the measurement, reproducibility= precision= coherence in the measurements of the same concentration).

US 5135302 describes a Flow cytometer. Flow cytometry is a technique of cellular analysis which involves measuring the characteristics of light scattering and fluorescence that the cells have as they are passed through a beam of light. For their analysis by flow cytometry, the cells should individually be suspended in a fluid. When they pass through the beam of light, the cells interact with this causing light scattering. Based on the diffraction of the light frontally, the size of the cells that pass can be assessed and by measuring the reflection of the light laterally the granularity or complexity of these is assessed. In addition to light scattering, if the cells are placed in the presence of monoclonal antibodies marked with fluorescent molecules prior to their analysis, it can be evaluated which cells have the antigens complementary to the monoclonal antibodies used.

A major problem of this technique lies in the destruction of the sample that will be used to carry out the count, since by exposing the cells to the beam of light and to fluorescence the extracted sample is destroyed.

Various techniques based on other cell counting principles are also known. Examples of this are patents and patent applications such as US 2007/0143033 that discloses systems and methods for counting particles by Beckam Coulter, US 2007/0012784 discusses the authentication of product by Thomas J. Mercolino, US 2008/0050619 "Fuel cell life counter and method of managing remaining life" by Life Technologies, US 3973194 from 1976 by Daniel W. McMorris and William J. Skidmore, US 5159642 from 1992 by Tokihiro Kosaka or US 5741648 from 1998 by George P. Hemstreet which describes a method of cell analysis using fluorescent quantitative image analysis.

Brief description of the invention

The present invention proposes a system adaptable to any microscope or magnifying glass. Currently, the cell automatic counting systems are closed machines that only allow performing cell counts. The claimed system leverages the existing microscopes and converts them in machines for counting cells allowing measurements and counts on a wider range. The mounting on the microscope, also allows saving space on the workbench.

Advantageously, the invention allows in addition the automatic counting in any container of known depth that can be observed in the microscope.

To achieve these objectives, several difficulties have had to be solved:

- Problems with the illumination of some microscopes. According to the microscope, it may be necessary to include a light detector, and to include a brightness parameter associated with each cell profile. If the brightness with which the system was configured is changed when trying to perform the count the profile must be redefined.
- Blur problems. There are microscopes which are defocused in a matter of seconds or by moving the plate. This problem has been mitigated by instantly showing the analysis on screen (if the analysis is wrong, it is usually a problem of defocusing and the user realizes it immediately and corrects it).
- Problems of artifacts, dirt of the microscope and aberrations in the lenses in low quality, used, etc., microscopes. Exclusion areas can be defined to eliminate especially problematic areas of the screen and prevent their use in the count.

Thus the particle counting system is adaptable to an optical instrument and includes:

- Means of image acquisition for acquiring images from a container with a sample of particles through the optical instrument.
- Means of visualization for viewing images acquired by the capturing means associated with the sample.
- Means for processing the acquired images. Said processing means identify edges of possible particles, identify a plurality of regions of the image, at least partially defined by edges, to associate them with the background of the image or to associate them with a region with at least one potential possible particle. They also check if, indeed, said region contains at least one particle. This is done on the basis of the fulfilment of a condition based on at least one of the following geometric parameters: concentration of edges, maximum length, minimum length, perimeter, area or coincidence with a preset contour pattern. The processing means assign a number of particles greater or equal to 1 to said region and can count the particles contained in a plurality of regions.

Optionally, the processing means are configured for assigning the number of particles to the region on the basis of a previous classification of said region. For example an extrapolation can be carried out if the number of particles in an area is known to assign the number of particles of the region.

Optionally, the processing means can convert the acquired image to a scale of shades according to its luminous intensity and wavelength.

Optionally, the means of visualization can also distinctly mark the particles

counted.

Optionally, the means of visualization comprise a user interface for validating a recorded region or for allowing discarding it as recorded.

Optionally, the processing means can also assign a value in the scale of shades to the background of the image.

Optionally, the processing means can associate a particle size according to the number of pixels in the corresponding image.

Optionally, the processing means can calculate the concentration of particles per volume unit or per area unit when the sample is placed in a container of known dimensions.

Optionally, the processing means can exclude from an acquired image. The region of exclusion can be defined by a user through the interface of the means of visualization.

Optionally, the captured image can be converted to an image in greyscale.

The system is particularly applicable when the particles are biological microorganisms. The biological microorganisms can be, among others, cells, fungi, algae or platelets. Also protozoa, virus, bacteria, mites or spores.

The processing means can optionally carry out a selective count in the captured image when filtered when it is illuminated with light of a wavelength associated with a particular feature of the biological microorganisms, if said microorganisms were previously marked with a marker sensitive to such wavelength.

Optionally, a selective count in the acquired image can be carried out when it is illuminated with light of a wavelength associated with a particular feature of the biological microorganisms if said microorganisms were marked with a marker sensitive to said wavelength.

Optionally, the means of visualization can detect the illumination of the sample and for modifying the luminous intensity applied to the cell sample.

Optionally, the image capturing means comprise a digital camera.

Optionally, the means of visualization of images comprise a touch screen.

Optionally, the system may include a mechanism for automatically moving the container of the sample.

Optionally, the image capturing means are calibratable, such that a pixel is associated with a real dimension value. Thus a correspondence is possible between size

on the image and actual size.

Optionally, the processing means can calculate a correspondence between the total area of the screen covered by particles (confluence) and the concentration of particles per area or volume unit.

Optionally, the counting system can also include the optical instrument.

Optionally, this optical instrument may be a magnifying glass or a microscope.

When the instrument is a microscope it may optionally include an auto focus mechanism that focuses automatically.

Figures

Aspects relating to an embodiment of the invention are schematically represented in the following drawings.

Figure 1. Scheme of counting system adapted for a microscope (4).

Figure 2. Example of definition of an area of exclusion (14) for avoiding false detections.

Figure 3. Example of several operations carried out by the counting system with the schematic image produced by each operation: image capture (21), edges detection (22), delimitation of areas of possible particles (23), filtering based on geometric criteria (24) and resulting count (25).

Detailed description of the invention

In the next pages, the invention is illustrated in addition and without limitation by means of its integration in an optical microscope (4) (phase contrast, fluorescence, etc.) with coupling means to the digital camera (2).

However, for counting particles and in particular microorganisms the invention is applicable both to a magnifying glass and a microscope.

For example, said coupling means can be carried out in:

- 1) the trinocular of the microscope if it exists (usually a coupling from the digital camera to the C mount adaptor with a diameter of approximately 25 mm will be used, although also couplings for camera thread, bayonet mount, etc. are available).

- 2) if there is no trinocular, a coupling to one of the binocular lenses would be carried out, which have a diameter usually of 25 mm. This option is more uncomfortable because one of the lenses that allow the visualization with the microscope is disabled. In monocular systems there is no other possibility of visualization rather than the screen of the invention.

Another possibility is the adaptation to the image capturing camera in the microscope (if it exists).

The system may include also the following elements:

- Image capturing camera (2).
- Processing device (1) (PC or equivalent) with storage capacity.
- Visual interface / screen (3).
- Communications cable between the camera and the processing device
- Calibration device (it can be a Neubauer Chamber or a chamber wherein a reference measurement can be taken, a microscope gauge, etc.).
- Sample holding chamber (6) (it can be a Neubauer Chamber, a Thoma Chamber, etc. The sole requirement is that the depth of the chamber must be known). The holding chamber can be washable or disposable.

Also it is necessary to have a microscope (4) for adapting the system. The microscope must be clean, to the possible extent, have light for illumination of the samples and have at least one optics, preferably of at least 10x.

The system can also be optionally coupled to the local data network through Ethernet RJ45, WiFi or similar connection, with the object of:

- 1) performing backups on the server,
- 2) sending images to the data receiving centre for maintenance, quality control, out of calibration / malfunction of the system detection, updates, etc.

The option of use of identification credentials in the same device, to facilitate traceability is envisaged.

Advantages: complete traceability with user identification, entry barriers to external staff, increase of the responsibility in sampling by staff, improves the quality of the measurements as the employees will make greater efforts in the preparation of the sample, dilutions etc. since their measuring operations will be registered.

The system supports connection with other peripherals such as keyboard, mouse

and/or plastic pen for Tablet PC.

The system even allows through a touch screen (3), to be used with the finger. The options are selected by pressing the touch screen and through the use of a virtual keyboard that appears on the screen when it is required.

Neubauer Chamber Counting

<i>Problem</i>	<i>Solution</i>
Human counting errors	The automation of the count allows reproducible results
Statistical errors	The statistical error is reduced by increasing the number of samples, and the area analyzed by the system. (With human count it entails investing considerable time by laboratory staff)
Concentration estimation errors, when applying the formula	The automation eliminates this type of errors
Low reproducibility / high variance	The system carries out a sampling of a larger number of samples, significantly reducing the variance, and the statistical error.
Monotonous and tedious process	The counting part is completely eliminated with the automatic counting.
Difficult process for people with visual impairments	Display screen allows the people with deficiencies to count and observe images with a microscope.
Strenuous process for the eyes	Display screen allows counts with less visual fatigue.
Expenditure on disposable material	It allows the use of washable counting chambers
Expenditure on reagents	The use of the microscope allows to count

	cells without using reagents
--	------------------------------

The advantages of performing the automatic counting with respect to traditional solutions are:

- 1) Reproducibility. The error introduced by the system is restricted, and is systematic. It does not depend on the laboratory staff performing the count.
- 2) Reliability. Human errors are eliminated; taking more images reduces the statistical error.
- 3) Elimination of the tedious manual counting process. The measurement by manual counting with a microscope can take between 1 and 10 minutes of a laboratory technician, depending on the concentration.
- 4) Less maintenance, less setup and cleaning time than a flow cytometer.
- 5) Traceability (images and counts associated with a user and determined in time).
- 6) Recovery of images.
- 7) Visual detection of contamination.

The advantages of performing the semi-assisted count with respect to traditional solutions are:

- 8) Reliability. Certain human errors in the calculations of the concentration are eliminated.
- 9) Usability. Reduction of visual fatigue and it allows staff with certain visual impairments to use the microscope.
- 10) Usability. Reduction of intellectual fatigue. The system automatically saves the counted cells, and by using the marks on the screen of those already counted it allows the staff to "get distracted" without losing count.
- 11) Responsibility. The counts undertaken are recorded with the name of the laboratory employee who carried out the count. The poorly made counts can be related with a certain person, subsequently improving their habits through training, etc.
- 12) Less maintenance, less setup and cleaning time than a flow cytometer.
- 13) Traceability (images and counts associated with a user and determined in time).

14) Subsequent recovery of images, and of the counts together with the information of what has been considered as a cell / particle to be counted.

15) Contamination visual detection.

16) Count of elements with high visual complexity (adherent cells, very dirty samples, etc.) where the system cannot be configured to count automatically.

17) Used in conjunction with automatic counting, semi-assisted counting allows validating the calibration carried out for the automatic operation mode, and verifying that the system is operating within the acceptable operating ranges.

The use of the system for the first time requires the following steps, in this order

- 1) Size calibration.
- 2) Definition of cell profile
- 3) Launch of cell count

In a second use, the cell count can be launched in a direct way, without performing the size calibration and definition of profile, provided that the same microscope is used, and the type of cell to be counted is the same (or that it has been previously defined).

Each of the steps is described below.

CONFIGURATION AND CALIBRATION OF THE SYSTEM PRIOR TO CARRYING OUT THE COUNT.

1) A calibration of the system is carried out, where the following actions are performed

(A) Size calibration. A known distance in the microscope (4) is taken as reference to obtain the real distance to number of pixels on screen ratio (size calibration).

(B) Definition of the biological profile. For this are defined:

- a. Depth of the container of count (6). (distance / depth calibration in the Z axis of the microscope (4))
- b. Maximum and minimum size of cells (eventually)
- c. Maximum and minimum illumination (it is not selected, it is detected automatically)
- d. Sensitivity and contrast of the sample

- e. Other morphological characteristics of the type of cell to be measured (eventually: roundness, form factor, etc.).
- f. Features dependent on the wavelength of the element to be analyzed (for example, in the visible spectrum)

Next these settings are defined in more detail.

(A) SIZE CALIBRATION.

The calibration can be done in different ways, although it must always be done with an object of which we know the exact distance between 2 points with a microscope. Among others, the following elements of calibration can be used.

- 1) Standard microscope calibration plate. It is standard in some commercial microscopes. This is a plate where a pattern with lines is printed, where the distance between the lines is known.
- 2) A Neubauer Chamber, Thoma Chamber, Improved Neubauer Chamber, disposable chamber or any other type of chamber with known depth. In this type of chambers, there is a grille / grid on the microscope in the central part. This grid has been used historically as a reference for hand counts in the microscope. The distances and dimensions of the grid are generally written at the top part of the chamber.
- 3) Other systems. The system calibration can be done with any system, provided that the exact distance between 2 points visible with a microscope is known.

(B) DEFINITION OF THE BIOLOGICAL PROFILE.

The calibration of the biological profile determines the morphological characteristics of size, shape, texture, colour and/or absorbance in the visible spectrum (or invisible depending on the image sensor), and contrast in the sample.

The calibration of the biological profile is carried out always subsequently to the size calibration, since in order to perform this calibration, we must know the distances of the elements that we are visualizing on the screen, to be able to select the maximum and minimum range of geometric parameters of the biological elements that the system will count.

In the biological profile, the user selects the features of the elements of the image that they want to count.

- Sensitivity (value between 1 and 100). It determines the minimum contrast that the body or edge of the biological element (cell or equivalent) must have to be considered

valid for count.

If a very high sensitivity is selected, the system will capture strange elements of the image, such as dirt from the camera or the microscope, artifacts, etc. producing false positives (detection of elements where there should not be any).

If a very low sensitivity is selected, the system will ignore elements of the image that should be taken into account in the count, producing false negatives (no detection of elements that should be detected).

- Geometric parameters (24): They determine the size (maximum length, minimum length, perimeter, area or coincidence with a preset contour pattern) that the biological element must have to be taken into account in the count (25).

- Light colour / absorbance calibration: This filter determines the range of acceptable colours counting the elements. For example, if the colour red is selected for the cells, because one wants to count only the cells that have absorbed a red dye, the system will ignore the cells/elements of colours very different to red, and will count the elements of the selected shade of red as well as similar shades (close in the colour and frequency spectrum).

The profile filters can be activated or deactivated. If the profile filter is deactivated, there will not be discrimination of the elements according to the characteristic of the profile. For example, if the colour filter is deactivated, the system will ignore colour when considering the elements for the count, counting all the elements of the image that meet the rest of the filters, and ignoring the colour.

- Viability: The system allows a specific calibration for the measurement of cell viability (percentage of dead cells on total cells, live and dead). The measurement of viability can be deactivated. The system performs a simple counting and provides only the cell concentration in cells / ml, or activated, in which case the cell concentration will be provided in cells / total ml and the percentage of living cells in the sample, in percentage.

In the case of activation of the viability measurement, the specific colour filters for living (usually white) and dead cells (usually blue when using Trypan blue dye) must be defined.

- Used optics: The user selects the used optic. This is necessary to take into account the distances in the image, and make the relevant adjustments once the initial calibration has been carried out. The most common optics in optical microscopes are 4x, 10x, 40x and 100x.

- Depth of the chamber / Container: As a rule, the commercial chambers (Neubauer, Thoma, etc) have a standard and known depth, and also said depth is written on the surface of the chamber.

To perform this calibration, the exact depth of the measurement container in mm must be introduced. The most common depths are 0.1 mm and 0.2 mm.

Therefore the chamber (image sensor) and the biological samples that you wish to measure are independent from the microscope used.

TYPE OF COUNTING [automatic] vs. [semi-assisted]

In cases in which the nature of the images or the type of cell to be counted do not allow to carry out an automatic count with sufficient reliability, the system allows configuring a profile for semi-assisted count.

In the case of selecting semi-assisted count in a cell profile, the system will ignore all the filters previously described and will not perform the automatic analysis of the images, but the user will be the one that will indicate manually or in a semi-assisted manner what they consider as a cell in each one of the images captured on the screen (by pressing with the finger or with the mouse).

THE CONFIGURATION OF THE BIOLOGICAL PROFILE STEP BY STEP.

While the characteristics of size calibration are common and do not vary, provided that the camera (image sensor) and the microscope are not changed, the characteristics of the biological profile change with each type of particle or cell to be measured. Therefore the user must define a different biological profile for each type of biological element that they want to measure.

The system allows the storage of the features defined for each profile in the memory of the system, for later retrieval.

Example of profile name: hepatocytes -type-a-10x-viability-María.

The operations for carrying out the calibration of the biological profile are the following:

- 1) Preparation of biological sample of the type of cell / microorganism to be measured.
- 2) Introduction of the sample in a Neubauer Chamber or similar. If necessary, a dilution has been previously carried out by introducing a sample into a test tube with an inoculating loop. This step is performed to achieve a proper

concentration that allows the visual analysis on the screen. It is estimated that the system can perform a calibration if we can visualize on the screen or with a microscope between 1 and 2 cells as minimum and 50-100 cells as maximum, which corresponds to a concentration of between 200,000 cell / ml and 10,000,000 cell / ml (approximately).

- 3) Selection of the Configuration section in the device.
- 4) A Counting Profile (a set of parameters that will define what should and what should not be counted in each image) is selected.
- 5) The parameters corresponding to the type of cell / element that we want to count are selected.
- 6) If the microscope has dirt, or a part of the image appears blurred due to imperfections in the lens, the part of the screen having a problem will be eliminated by an Exclusion Area (14) (equivalent to a mask for ignoring problem areas).
- 7) After the adjustment of the parameters (and eventual configuration of the exclusion area) it can be visually checked that the system detects the cells of the image by drawing a coloured circle (10) on top of each cell. The verification that the system is well calibrated consists of manually counting the cells in the image, and checking that the system has drawn a superimposed circle on all of them.
- 8) In the case that all the cells / elements are not detected correctly, steps 5, 6, and 7 are repeated until at least 90%-95% of the cells of the image are detected correctly.
- 9) After the visual check, the microscope is moved and it is checked that the detection of cells is carried out correctly with 2 or 3 additional images. (Correct detection of at least 95% of the cells / elements)
- 10) The profile data are saved, and the system is ready to be able to carry out counts with this particular microscope, and with the type of element for which it has been configured.
- 11) This step is only necessary with adherent, overlapping cells, or cells with high level of agglomeration. In the case of adherent cells that tend to agglomerate against each other, the system must be configured to perform a calculation of extrapolation of the cell concentration from the confluence (or percentage of visual field occupied by cells). For this purpose a CONFLUENCE –

CONCENTRATION internal ratio must be configured, through the following steps:

- a. Measurement of the CONFLUENCE of the sample.
- b. Measurement of the Cell concentration of the sample using an alternative method (e.g. manual counting, flow cytometer, etc), and introduction into the system of the value of said concentration.

From this ratio, the system will be able to calculate the cell concentration by means of the analysis of the confluence of the sample.

It is only required to perform this configuration the first time you work with a cell line.

BIOLOGICAL ELEMENTS COUNTING - AUTOMATIC MODE (WITH PREVIOUSLY CALIBRATED / CONFIGURED SYSTEM).

After completion of the calibration of the system (see previous paragraph) and definition of the biological profile of the element we want to count, the following steps are followed for carrying out a count.

- 1) The biological sample is prepared and introduced in the counting container (Neubauer, Thoma, Howard ch., Slide + coverslip, or a proprietary container and made to measure for the system, a chamber with special calibration marks, a 24 or 96 wells plate, a Petri dish, a culture flask, etc.).
- 2) The PROFILE that has been previously configured in step 1) for this microscope and specific cell type is selected.

The images are taken with the digital camera (the device for moving the microscope tray is used, and the touch screen or the keyboard is used to indicate the system that the image can be captured). A number of images that can vary are captured. Several images are taken to reduce the statistical error (in the same way as in a manual counting with Neubauer Chamber, the custom is to measure 5 quadrants and perform an average of the same). In our case, taking more images entails a minimum effort for the user that translates into a significant reduction of the error.

As a general rule between 5 and 20 images will be taken, although we intend to take only 1 image with high resolution in a next version of the product.

If you want to reduce further the statistical error it is possible to take as many images as you wish, the statistical error being inversely proportional to the number of images taken.

- 3) The images are sent to the data processing system (1).
- 4) In the case that the images do not have the sufficient illumination or excessive illumination, the system will make the appropriate adjustments.
 - a. By means of the adjustment of the level of illumination in the captured image.
 - b. By means of an adjustment loop, where the processing unit sends a signal to the control unit of the intensity of the light source ordering to increase or decrease the intensity of the light source.
 - c. Returning to item 4 (and it is iterated until the illumination falls within the range).
 - d. If it were not possible to perform an automatic adjustment, it will be indicated to the user that the illumination levels are out of range and that the system is out of range for performing the count.
- 5) In the case that the focus level of the images is not appropriate, the system will make the appropriate adjustments.
 - a. By means of an adjustment loop, where the processing unit (1) sends a signal to the focus control unit of the microscope (4) ordering the microscope to get closer or away from the sample.
 - b. Returning to item 4 (and it is iterated until the focus falls within the range).
 - c. If it were not possible to perform an automatic adjustment of the focus (because the microscope does not include focus control), the system systematically indicates on each image analyzed the elements being recognized, so that if the system is not correctly focused, the user can see on the screen that the cells are not being detected correctly.
- 6) The analysis system processes the images, by applying:
 - a. size filters
 - b. filters on the morphology of the object
 - c. filters on the wavelength that passes through the element / or is reflected by the element.
 - d. filters on the eccentricity of the element (similar to the eccentricity of an ellipse)

- e. filters on the length of the contour of the object
- f. filters on the area of the object
- g. filters on the area to be analysed (areas of exclusion (14))
- h. elimination of the “background image (12)” (noise and dirt of the image that remain constant in each image)

7)

- a. It calculates the number of cells in each image.
- b. It performs an averaging of the same.
- c. It subsequently multiplies by the average volume of the image (this average volume is calculated from the calibration in size and the depth of the chamber used - detailed in the previously defined biological profile).

The system displays on screen the results of the element count:

The system provides:

- Cell or particle concentration per volume unit.
- Cell or particle concentration per area unit

It can also provide:

- Total number of cells counted on screen.
- Degree of cell confluence (the cell confluence is the percentage of area occupied by the cells or particles with respect to the total percentage of the screen).
- Percentage of cells of a cell type or profile with respect to the number of total cells.
- Percentage of cells of a cell type or profile with respect to the number of cells of another cell profile.
- Percentage of living cells with respect to total cells.
- Percentage of dead cells with respect to total cells.
- Total area analysed.
- Total volume analysed.
- Statistical error

8) The system stores the samples and images for subsequent consultation, generation of growth charts, etc. The display of results, images and graphics is done through the graphical interface (screen).

BIOLOGICAL ELEMENTS COUNTING - SEMI-ASSISTED MODE (WITH PREVIOUSLY CONFIGURED / CALIBRATED SYSTEM).

The system allows the semi-assisted counting of elements with a microscope.

In this mode of operation the system does not apply any filter defined in the biological profile nor performs any automatic analysis of the image (is the own user the one that does the counting manually, and their own intelligence is used to select the cells on the screen).

The steps to be followed in this case are:

- 1) size calibration. (Which is performed only once for each microscope)
- 2) definition of the biological profile (in this case only the depth of the chamber used and the optics have to be defined).
- 3) item 2) of the automatic method is carried out.
- 4) item 3) of the automatic method is carried out.
- 5) the images are taken with the digital camera (the device for moving the microscope tray is used, and the touch screen or the keyboard is used to indicate to the system that the image can be captured). A number of images that can vary are captured. Several images are taken to reduce the statistical error.

In semi-assisted mode, after the capture of each image the cells have to be marked manually on the screen using the mouse, the finger or the plastic pen. Whenever a cell has been marked, a semi-transparent circle is drawn on the cell to indicate to the user that the cell has already been counted.

- 6) The system displays on the screen the cell concentration (case of simple count) or the cell concentration together with the viability percentage (in the case of count with viability).

The system provides:

- Cell or particle concentration per volume unit.

It can also provide:

- Total number of cells counted on screen.

- 7) The system stores the samples and images for subsequent consultation, generation of growth charts, etc. The display of results, images and graphics is carried out through the graphical interface (screen).

In this case the data processing has been limited to the calculation of the cell concentration or the calculation of the total sum of cells marked by the user on the screen.

The calculation of the cell concentration can be done in this case thanks to the innovative calibration system of the system.

One of the most advantageous innovative components of the system is the coupling to any microscope on the market. This is achieved thanks to the following set of factors:

- a. the mechanical adaptation, which is done through common mechanical adapters, which usually exist on the market.
- b. the size calibration.
- c. the detection of changes in luminosity and the adjustments in brightness
- d. the detection of blur and focus adjustment.
- e. the edge detector that prevents problems of illumination.
- f. the exclusion areas.

The system can also be considered as a whole, attached to a specific and pre-calibrated microscope for the set of lenses of the microscope.

Numeric references

1. Processing unit.
2. Camera.
3. Touch Screen.
4. Microscope.
5. Eyepiece.
6. Sample container.
10. Edge of particle/cell.
11. Particle.
12. Photography background.
13. Inclusion area.
14. Exclusion Area.
21. Image capture.

22. Edge detection.
23. Delimitation of areas of possible particles/cells.
24. Filtering according to geometric criteria.
25. Count.

CLAIMS

1.- Particle counting system adaptable to an optical instrument (4) comprising:

- means of image acquisition (2) configured for acquiring images from a container (6) with a sample of particles through the optical instrument (4),
- means of visualization (3) configured for viewing images acquired by the capturing means (2) associated with the sample,
- means for processing (1) the acquired images,

characterized in that

the processing means are configured for:

identifying edges (10) of possible particles,

identifying a plurality of regions of the image, at least partially defined by edges, to associate them with the background of the image (12) or to associate them with a region with at least one possible particle,

checking if said region contains at least one particle (11) depending on the fulfilment of a condition based on at least one of the following geometric parameters: concentration of edges, maximum length, minimum length, perimeter, area or coincidence with a preset contour pattern;

assigning a number of particles greater or equal to 1 to said region and counting the particles contained in a plurality of regions.

2.- Counting system according to claim 1, characterized in that the processing means are configured assigning the number of particles to the region on the basis of a previous classification of said region.

3.- Counting system according to claim 1 or 2, characterized in that the processing means (1) are configured for converting the acquired image to a scale of shades according to its luminous intensity and wavelength.

4.- Counting system according to any one of the previous claims, characterized in that the means of visualization (3) are also configured for distinctly marking the

particles (11) counted.

5.- Counting system according to claim 4, characterized in that the means of visualization (3) comprise a user interface configured for validating a counted region or for allowing discarding it as counted.

6.- Counting system according to any one of the previous claims, characterized in that the processing means (1) are also configured for assigning a value in the scale of shades to the background of the image (12).

7.- Counting system according to any one of the previous claims 4 to 6, characterized in that the processing means (1) are configured for associating a particle size (11) according to the number of pixels in the corresponding image.

8.- Counting system according to any one of the previous claims, characterized in that the processing means (1) are configured for calculating the concentration of particles (11) per volume unit or per area unit when the sample is placed in a container of known dimensions.

9.- Counting system according to any one of the previous claims, characterized in that the processing means (1) are configured for excluding from an acquired image an exclusion region (14) according to that defined by a user through the interface of the means of visualization (3).

10.- Counting system according to any one of the previous claims, characterized in that the captured image is converted to an image in greyscale.

11.- Counting system according to any one of the previous claims, characterized in that the particles counted are biological microorganisms.

12.- Counting system according to claim 11, characterized in that the biological microorganisms are selected at least from the following:

- cells,
- fungi,
- algae,
- platelets,
- protozoa
- virus
- bacteria
- mites
- spores.

13.- Counting system according to claims 11 or 12, characterized in that the processing means are configured for performing a selective counting in the image acquired when it is illuminated with light of a wavelength associated with a particular feature of the biological microorganisms if said microorganisms were marked with a marker sensitive to said wavelength.

14.- Counting system according to any one of the previous claims, characterized in that the means of visualization (3) are configured for detecting the illumination of the sample and for modifying the luminous intensity applied to the cell sample.

15.- Counting system according to any one of the previous claims, characterized in that the image capturing means (2) comprise a digital camera.

16.- Counting system according to any one of the previous claims, characterized in that the means of visualization (3) of images comprise a touch screen.

17.- Counting system according to any one of claims 8 to 16, characterized in that it comprises a mechanism to automatically move the container of the sample.

18.- Counting system according to any one of the previous claims, characterized

in that the image capturing means (2) are calibratable, such that a pixel is associated with a real dimension value.

19.- Counting system according to any one of the previous claims, characterized in that the processing means are configured for calculating a correspondence between the total area of the screen covered by particles and the concentration of particles per area or volume unit.

20.- Counting system according to any one of claims 1 to 19, characterized in that it comprises the optical instrument (4).

21.- Counting system according to any one of claims 1 to 20, characterized in that the optical instrument (4) is a magnifying glass.

22.- Counting system according to any one of claims 1 to 20, wherein the optical instrument (4) is a microscope.

23.- Counting system according to claim 22, characterized in that it comprises a mechanism that automatically focuses the microscope.

9.2 Most common statistical requirements of life science journals

1	Report confidence intervals
2	Report statistical limits / significance
3	Suggest to consult with professional statistician
4	Report standard deviation
5	Report 'Center Values', such as median or mean
6	Report number of samples
7	Report exact p-value (not < 0.05)
8	Report statistical tests used
9	Report randomization method used
10	Report statistical method and measures in general
11	State what n represents
12	Report number of experiment replicas
13	Report inclusion /exclusion criteria
14	Reviewers will be asked to check statistical methods
15	Report statistical package or program used
16	Report data points if $n < 20$
17	Use plots to report data distribution
18	Report if technical or biological replicas
19	Rationale for the number of samples used (n)
20	Report quartiles of data
21	Report a list of all variables examined
22	Report outliers
23	Report if blind method reporting

24	Report the degrees of freedom
25	Report methods of data normalization
26	Reports if one-side or two-side tests
27	State clearly hypotheses tested
28	Statistical tests results should be included
29	Report all data generated and analyzed
30	Make all data freely available without restriction
31	Reports methods of data transformation
32	Report statistical methods for high dimensional data
33	Report missing value handling
34	Report analyzed data across multiple experiments
35	Experiments with at least 3 biological replicates
36	Report & discuss statistical power
37	Check if data meet the assumptions of the tests.

9.3 Cell counting needs and habits interview. List of researchers and technicians interviewed

Group / Company name	Location	Person interviewed (removed for personal data protection)	Position	Activity
Celartia Europe SL	USA / Valencia	XXXXXX	Founder	Minireactor manufacturer
Synapcell	Grenoble, France	XXXXXX	CEO	Neurology research. Mice
Departamento conservación celular, UV	Valencia, Spain	XXXXXX	PI	Biobank. Bacteria, fungi and yeast.
Agrenvec	Madrid, Spain	XXXXXX	CEO	Biofactory. Enzymes and proteins
Biotools SL	Madrid, Spain	XXXXXX	Scientific Director	Biotech. PCR kits, microarrays, other
EMBL Grenoble	France	XXXXXX	PI	Automation for extraction and crystallization of proteins.
Inmunostep	Salamanca, Spain	XXXXXX	Manager	Materials and reagents for biotech research
Cedivet	Valencia, Spain	XXXXXX	Assistant and R&D	Clinical veterinary analysis
Abba Gaia	Valencia, Spain	XXXXXX	R&D Director, CEO	Plant biology for waste management
Calantia Biotech	Valencia, Spain	XXXXXX	CEO	White Biotech, Energy
Symboro SL	Murcia, Spain	XXXXXX	Founder	Natural fertilizers
Project	Madrid, Spain	XXXXXX	Cytomics Director	Diagnosis and therapy for common and serious diseases
Durviz	Valencia, Spain	XXXXXX	CEO and founder	Equipment and reagents for science distribution
Gregorio Marañón Hospital	Madrid, Spain	XXXXXX	Immunology group responsible	Immunology research
CIB-CSIC	Madrid, Spain	XXXXXX	Researcher (former Scientific Director Pharmamar)	Cell cultures
CIB-CSIC	Madrid, Spain	XXXXXX	Researcher	Cell Cultures

9.4 Cell counting needs and habits interview. Summary of quantitative and qualitative results.

Average number of people in the interviewed company / research group: 8

Location:

Valencia	41%	7 interviewees
Madrid	29%	5 interviewees
Grenoble	13%	2 interviewees
Barcelona	6%	1 interviewee
Salamanca	6%	1 interviewee
Murcia	6%	1 interviewee

80% of the interviewed companies / institutions perform cell counting regularly

Average cell countings per week : 91 (min 3, max 300)

The **perception of accuracy of cell counting (Neubauer) is 82%** on average.
Error \pm 18%

Minimum accuracy that a cell counter should have according to respondents is **89%**, or an **error \pm 11%**.

(In cell cultures accuracy is less important than in disease diagnosis. In this case a rough estimate is enough)

Average counting time for a sample to be prepared and analyzed = 15.5 minutes

Cell types:

Cell culture	45%
Tumors	18%
Stem cells	18%
Spores	9%
Micelium	9%
Blood	9%
Neurons	9%

Degree of satisfaction with current systems : **3 / 5**

Main sources of dissatisfaction:

Slow (2), Monotonous and tedious task (2), Visually tiring (1), Depends on the user (1), Cell processing and preparation (1)

Agreement with statements:

“Cell counting systems save time” **4.7 / 5**

“Cell counting systems save money” **3.6 / 5**

“Cell counting system increase reproducibility” **4.9 / 5**

Cell counting system and brands mentioned:

Hemocytometer (9), Coulter (5), Beckman (3), Invitrogen (2), Stereo Investigator(1), Partek (1), Accuri (1), Uaba Tech (1), Nihon-Cohen (1), Advia (1), Bayer (1), Fisher (1), GE (1)

Most used cell counter systems

Cell counter chamber (Neubauer)	80%	13 interviewees
Flow cytometer	50%	8 interviewees
Image-based cell counter	25%	4 interviewees

Operating systems :

Windows	54%
Mac OS	36%
Linux/Unix	9%

Most important parameters of a cell counter, in order of importance:

Simplicity / easy to use	4.3/5
Exporting data to computer	4.1/5
Counting on medium	4.0/5
Speed	3.9/5
Non-destructive	3.9/5
Distinguish dead cells	3.5/5
Stained cell counting	3.2/5
Morphology measurement	2.9/5

9.5 Test run document for automated cell counter validation with alternative methods.



TEST RUN - Celeromics Technologies, S.L. - CONFIDENTIAL

(This document is valid for internal pilot tests, pilot tests with clients and tests at research centers).

CELEROMICS STAFF PRESENT AT THE TEST

Responsible for the test: <small>(Institution / Company)</small>			
Qualification / title of the responsible institution.			
Responsible for the test: <small>(Celeromics Staff)</small>			
Names and surnames of the persons present at the test.			
Date and time of the start of the test		End time:	
Location / Customer			
Count ref. number# <small>(format: year-month-day-hour-minute)</small>			
Version of the tested product:			
Algorithm of the tested version :			

DEVICE COUNTING REFERENCE

Reference counting devices Name, model and version <small>(other than SimpleCounter)</small>	
	<input type="checkbox"/> The calibration certification of the device is enclosed. <input type="checkbox"/> Uncertainty in the reference of the counting system
Microscope make and model <small>(where SimpleCounter is installed)</small>	

ANALYSIS AND REPORTING OF BIOLOGICAL SAMPLES

Cell / organism name <small>(yeast, liver cells, etc.)</small>	
Cell recipient used <small>(Neubauer chamber, etc.)</small>	
Method of sample preparation	



SHEET 1 Counting ref. number #: _____

INFORMATION ABOUT THE SYSTEMS

MANUAL METHOD WITH MICROSCOPE

Concentration calculation formula used:

(F1)

SIMPLE COUNTER METHODS: MANUAL BY DISPLAY / AUTOMATIC

Multiplication factor (necessary to obtain cells/ml): _____(a)

SEMI-ASSISTED METHOD WITH SIMPLE COUNTER LITE

Living cells concentration = $\frac{(v1+v2+v3+v4+v5)}{5} \times (a)$

Dead cells concentration = $\frac{(m1+m2+m3+m4+m5)}{5} \times (a)$

Signatures (of all persons present at the test):

For Celeromics

For the lab responsible

For the institution
responsible (with seal)



SHEET 2 Counting ref. number #: _____

EXECUTION TESTS - SIMPLE COUNTER

(All measurements (x20) must be from the same original sample and of the same concentration).

Measurement 1

Measurements / Samples	1	2	3	4	5	Concentration (cel / ml)	Time
1 - Manual	V	V	V	V	V	Living cells	
	M	M	M	M	M	Dead cells	
2 - Manual on Screen	V	V	V	V	V	Living cells	
	M	M	M	M	M	Dead cells	
3 - Alternative method	Vivas		Muertas				

Concentration: _____ % Living: _____ % Dead: _____

Measurement 2

Measurements / Samples	1	2	3	4	5	Concentration (cel / ml)	Time
1 - Manual	V	V	V	V	V	Living cells	
	M	M	M	M	M	Dead cells	
2 - Manual on Screen	V	V	V	V	V	Living cells	
	M	M	M	M	M	Dead cells	
3 - Alternative method	Vivas		Muertas				

Concentration: _____ % Living: _____ % Dead: _____

Measurement 3

Measurements / Samples	1	2	3	4	5	Concentration (cel / ml)	Time
1 - Manual	V	V	V	V	V	Living cells	
	M	M	M	M	M	Dead cells	
2 - Manual on Screen	V	V	V	V	V	Living cells	
	M	M	M	M	M	Dead cells	
3 - Alternative method	Vivas		Muertas				

Concentration: _____ % Living: _____ % Dead: _____

[PAGES 4 to 14 ARE A COPY OF THIS ONE, AND WERE REMOVED]

9.6 Clinical Genetic Analysis System User Expert Panel Survey



Name of Respondent										
Cargo										
Institution / Company		Size								
Date / Time										
Involved in :		Diagnosis <input type="checkbox"/> Prognosis <input type="checkbox"/> Monitoring <input type="checkbox"/> Monitoring [Treatment optimisation <input type="checkbox"/> Clinical trials <input type="checkbox"/> Other : _____								
Detailed description :										
Do you work with human samples? [Yes <input type="checkbox"/> No <input type="checkbox"/>		Blood <input type="checkbox"/> Solid tumour <input type="checkbox"/> Tumour in paraffin <input type="checkbox"/> Tumour in paraffin [] Other : _____								
Do you think precision medicine is necessary for the management of CRC patients?		<input type="checkbox"/> Yes <input type="checkbox"/> No								
Do you work with biomarkers? <input type="checkbox"/> Yes <input type="checkbox"/> No		Types :	Interested in markers for CRC? <input type="checkbox"/> Yes <input type="checkbox"/> No							
Which markers meet the expectations of patient management at your centre?		Do you think other biomarkers will be needed in the future? show table(*) (Which ones?)	<input type="checkbox"/> Yes <input type="checkbox"/> No							
Pathologies analysed:										
Systems currently used for analysis		Current methods, limitations and problems?								
Critical parameters for this analysis:										
CRITICAL FACTORS (1=not at all important, 10=very important)	1	2	3	4	5	6	7	8	9	10
Automatic results report										
Reliability										
High sensitivity : specify range :										
Specificity : specify range :										
Reproducibility										
Quick analysis : Specify time :										
Reduce / Eliminate user responsibility (e.g. Cartridge)										
Automatic results report										
Easy to use										
Small size										
Less reagent consumption										
Easy to transport										
Flexibility in handling: kit fractionation										
Kit shelf life > 1 year										
PCR analysis included / Real-time PCR										
DNA extraction included										
Validated nucleic acid extraction process										
Minimal baseline tumour DNA/RNA requirements										
Maximum mutation coverage										
High-throughput screening for molecular alterations (P/N)										
Ability to analyse different biomarkers										
Ability to characterise the detected disturbance (90% - 95%)										
Ability to introduce the sample as it arrives (in paraffin)										
Ability to analyse liquid biopsy (blood sample)										
Possibility to realise a complete gene panel in one run										
Possibility to make individual genes										
Possibility of short runs (few patients)										

ECONOMIC PARAMETERS

Do you think that manual DNA extraction can have advantages? <input type="checkbox"/> Yes <input type="checkbox"/> No/Develop :		What would be the optimal format kit size? <input type="checkbox"/> 12 rxs <input type="checkbox"/> 24 rxs. <input type="checkbox"/> Other : _____	
Nb. Analysis per week? (approx)	Average cost per analysis? Time: Money:	Would you prefer a loaned system with minimum annual consumables consumption? <input type="checkbox"/> Yes <input type="checkbox"/> No	
How much more would you be willing to pay for a system with automatic DNA extraction?	<input type="checkbox"/> +50% <input type="checkbox"/> +25% <input type="checkbox"/> +15% <input type="checkbox"/> +10%	Would you pay 35,000 EUR for the device? <input type="checkbox"/> Yes <input type="checkbox"/> No	Would you pay 200 EUR per analysis (includes panel of 14 genes KRAS, NRAS, BRAF with the most common mutations)? <input type="checkbox"/> Yes <input type="checkbox"/> No
Would you be interested in being a beta tester of the system? <input type="checkbox"/> Yes <input type="checkbox"/> No Would you like to receive the results of the study? <input type="checkbox"/> Yes <input type="checkbox"/> No		Do you know anyone who could help us with this study?	

Interviewer

Respondent

Date / Signature

Date / Signature

Gen	Codón	Mutación (DNA)	Mutación (aa)	Frec (%)	Frec/gen
KRAS	12	c.35G>A	p.Gly12Asp	34,54	95,98%
	12	c.35G>T	p.Gly12Val	23,01	
	13	c.38G>A	p.Gly13Asp	12,86	
	12	c.34G>T	p.Gly12Cys	11,96	
	12	c.35G>C	p.Gly12Ala	5,46	
	12	c.34G>A	p.Gly12Ser	4,86	
	12	c.34G>C	p.Gly12Arg	3,30	
NRAS	61	c.182A>G	p.Gln61Arg	31,30	95,07%
	61	c.181C>A	p.Gln61Lys	22,22	
	12	c.35G>A	p.Gly12Asp	12,26	
	61	c.182A>T	p.Gln61Leu	7,03	
	13	c.38G>A	p.Gly13Asp	5,68	
	12	c.34G>A	p.Gly12Ser	3,70	
	12	c.34G>T	p.Gly12Cys	3,13	
	13	c.37G>C	p.Gly13Arg	2,63	
	12	c.35G>T	p.Gly12Val	2,03	
	61	c.183A>T	p.Gln61His	1,93	
	13	c.38G>T	p.Gly13Val	1,65	
	12	c.35G>C	p.Gly12Ala	1,53	
BRAF	600	c.1799T>A	p.Val600Glu	97,70	97,70%

9.7 Variant Analysis Geneticists Expert Panel Survey

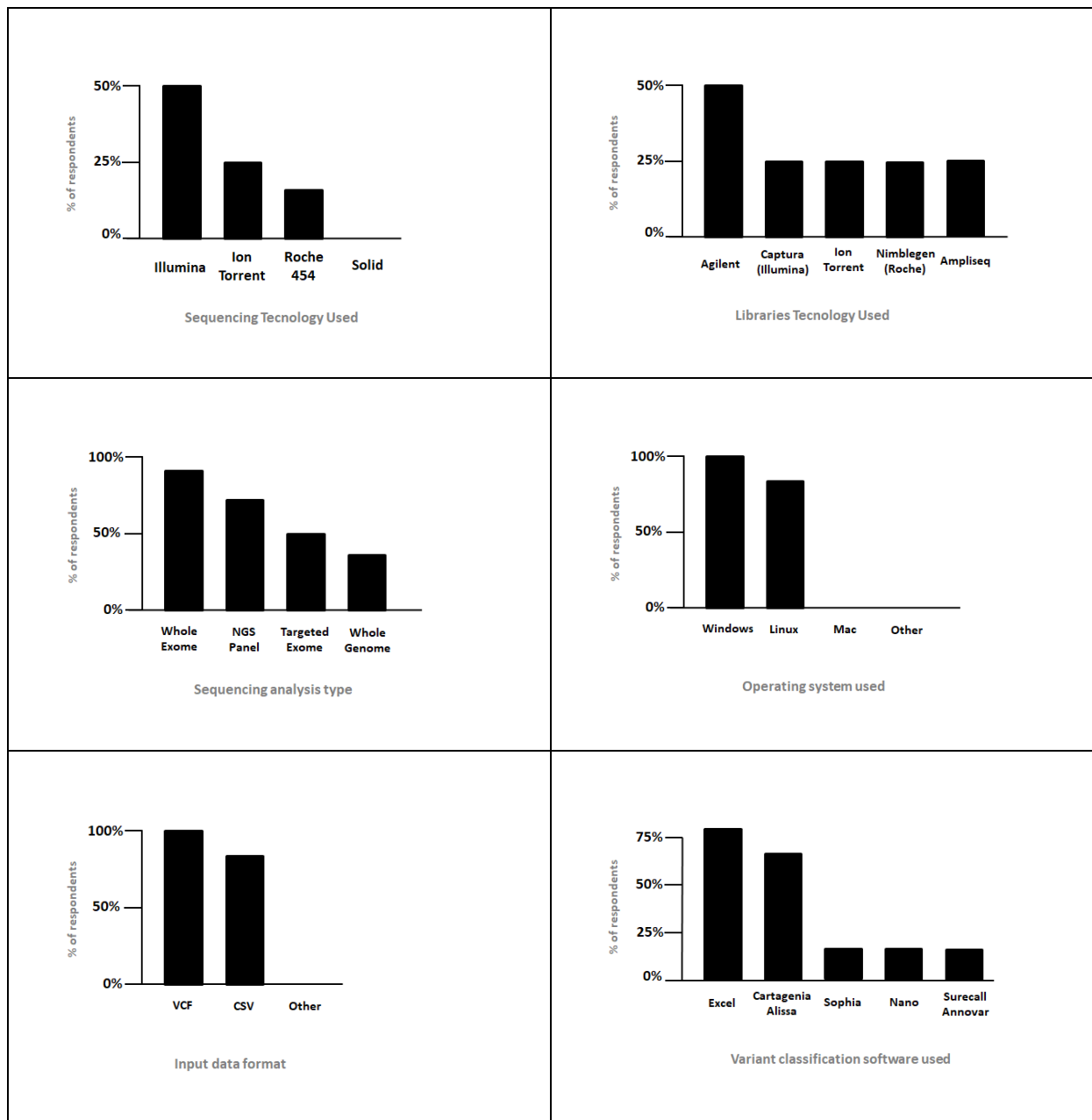
EXPERT SURVEY ON GENETIC ANALYSIS

BINOME SYSTEM : ASSISTANT FOR CLASSIFICATION OF GENETIC VARIANTS.

Name of Respondent			
Cargo			
Institution / Company		Size	
Date / Time			
Involved in :	Diagnosis <input type="checkbox"/> Prognosis <input type="checkbox"/> Monitoring <input type="checkbox"/> Monitoring [Treatment optimisation <input type="checkbox"/> Clinical trials <input type="checkbox"/> Other : _____		
Detailed description :			
Type of samples :	Blood <input type="checkbox"/> Tumour Solid <input type="checkbox"/> Tumour in paraffin <input type="checkbox"/> Tumour in paraffin [] Other : _____		
Do you work with NGS technology?	[] Illumina [] Roche 454 [] Ion Torrent [] SOLiD		
What technology did you use to build the library?			
You perform classification of genetic variants on a recurring basis.		Number of patients analysed / month :	<input type="checkbox"/> Yes <input type="checkbox"/> No
Most common pathologies for which you carry out the analyses	#1 : _____ #2 : _____ #3 : _____		
Type of analysis		Percentage of total analyses	
	<input type="checkbox"/> complete genome		
	<input type="checkbox"/> complete exome		
	<input type="checkbox"/> targeted exome		
	[Own panel : _____		
	[Own panel : _____		
	[Own panel : _____		
Do you use any software to support the classification of genetic variants?	<input type="checkbox"/> No. <input type="checkbox"/> Yes. [] Excel [] Carthage / Alisa [] Sophia []	Other systems for filtering variants?	Operating System <input type="checkbox"/> Windows <input type="checkbox"/> Linux <input type="checkbox"/> Mac [Other : _____
Input data format	<input type="checkbox"/> CSV <input type="checkbox"/> VCF <input type="checkbox"/> OTHERS	Output data format:	
Limitations and problems of current systems?			
Critical parameters for these types of analysis:			

CRITICAL FACTORS (1=not at all important, 10=very important)	1	2	3	4	5	6	7	8	9	10
Reliability										
High sensitivity : specify range :										
Specificity : specify range :										
Reproducibility										
Quick analysis : Specify time :										
Reduce / Eliminate user liability.										
Easy to use										
Flexibility in configuration										
Automatic report generation										
F : filter IGNORE, ignore variants										
F: filter based on GENES / GENE PANELS.										
F: PRT-based filter										

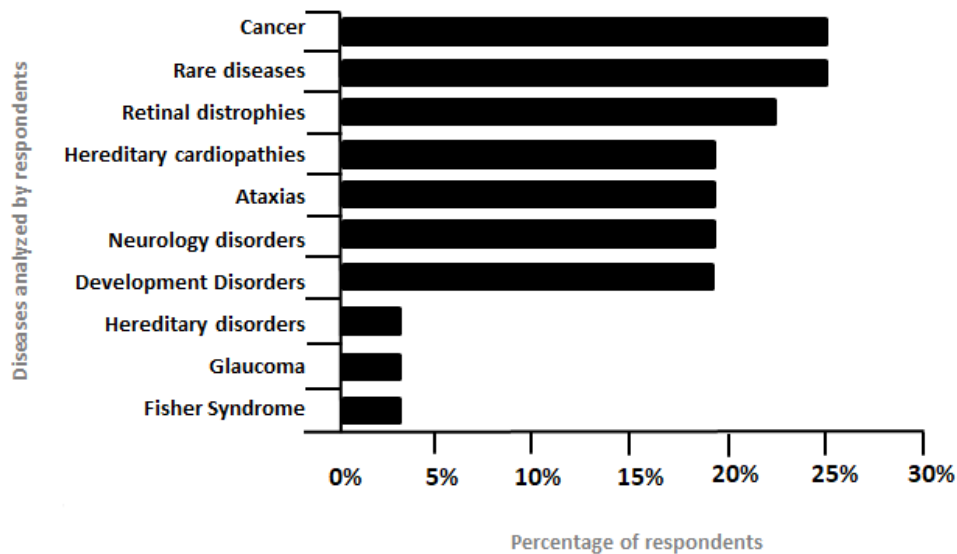
9.8 Variant analysis geneticists' survey detailed results extract.



Nature of biological samples : Blood (6), solid tumour (4), paraffin tumour (4), saliva (1), feces (1), marrow (1) and tears (1).

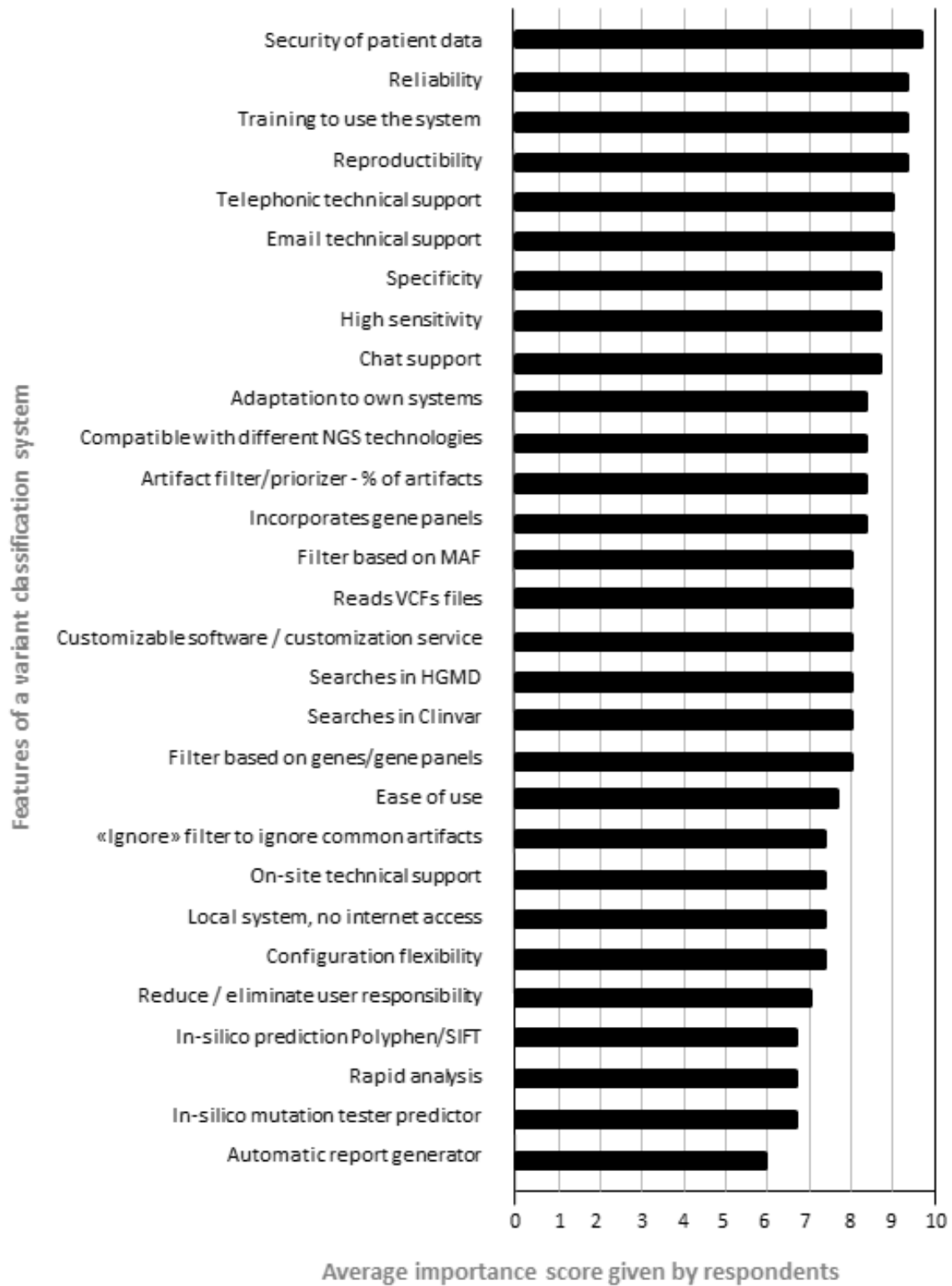
Average number of patients analysed per year by each respondent:

152 (min=2, max=400)



Current analysis systems problems and limitations enumerated by respondents: System functioning depends on the internet, not updated databases, not fully automatic, deficient results analysis, deficient variant classification, not integrated with databases, requires human intervention, lacks population databases, pathogenic classification.

Critical parameters for variant interpretation enumerated by respondents: Need to use several tools scattered in different places (websites), need to use a set of different tools/webs/sites, updated databases, reliability, low false negatives rate, access to databases and customizable systems.



9.9 Artifact Detection with Neural Networks Python Algorithm.

```
import pandas as pd
import seaborn as sns
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LogisticRegression
from sklearn import metrics
from sklearn import preprocessing
from keras.models import Sequential
from keras.layers import Dense
from tensorflow.keras.optimizers import SGD
from matplotlib import pyplot

# load dataset
art = pd.read_csv("A250_Res.csv")
art.head()

#split dataset in features and target variable
feat_cols = ['REF-F', 'REF-R', 'REF', 'ALT-R', 'ALT-F', 'ALT', 'TOTAL', 'REPETITIVE']
X = art[feat_cols]
y = art.ARTIFACT
print(X)
X_scaled = preprocessing.scale(X)
print(X_scaled)

#Split 75% train, 25% test
X_train,X_test,y_train,y_test=train_test_split(X_scaled,y,test_size=0.3,random_state=2)

# define model
model = Sequential()
model.add(Dense(8, input_dim=8, activation='relu', kernel_initializer='he_uniform'))
model.add(Dense(1, activation='tanh'))
opt = SGD(learning_rate=0.01, momentum=0.9)
model.compile(loss='hinge', optimizer=opt, metrics=['accuracy'])

# fit model
history = model.fit(X_train, y_train, validation_data=(X_test, y_test), epochs=200, verbose=0)

# evaluate the model
_, train_acc = model.evaluate(X_train, y_train, verbose=0)
_, test_acc = model.evaluate(X_test, y_test, verbose=0)
print('Train: %.3f, Test: %.3f' % (train_acc, test_acc))
```

```
# plot loss during training
pyplot.subplot(211)
pyplot.title('Loss')
pyplot.plot(history.history['loss'], label='train')
pyplot.plot(history.history['val_loss'], label='test')
pyplot.legend()

# plot accuracy during training
pyplot.subplot(212)
pyplot.title('Accuracy')
pyplot.plot(history.history['accuracy'], label='train')
pyplot.plot(history.history['val_accuracy'], label='test')
pyplot.legend()
pyplot.show()
```