Flow cytometry data analysis

Basic concepts and statistics

James V. Watson

Clinical Oncology Unit
Medical Research Council
and
Faculty of Clinical Medicine
The Medical School
University of Cambridge

CAMBRIDGE
UNIVERSITY PRESS

Published by the Press Syndicate of the University of Cambridge
The Pitt Building, Trumpington Street, Cambridge CB2 1RP
40 West 20th Street, New York, NY 10011-4211, USA
10 Stamford Road, Oakleigh, Victoria 3166, Australia

First published 1992

Library of Congress Cataloging-in-Publication Data

Watson, James V.

Flow cytometry data analysis : basic concepts and statistics / James V. Watson.

p. cm.

Includes bibliographical references and index.

ISBN 0-521-41545-4

1. Flow cytometry - Statistical methods. 2. Flow cytometry - Mathematics. 3. Flow cytometry - Data processing. I. Title.
QH324.9.F38W37 1992
574.87'028 - dc20 92-14761
CIP

A catalog record for this book is available from the British Library.

ISBN 0-521-41545-4 hardback

Transferred to digital printing 2002

Contents

1

Introduction

Flow cytometry is now a well established technique in cell biology and is gaining increasing use in clinical medicine. The major applications to date in the latter have been in DNA histogram analysis to determine "ploidy" (DNA index, Hidderman et al. 1984) and S-phase fractions for prognostic purposes in cancer patients and in immunophenotyping (Parker 1988). However, more recent applications in cancer work include determination of tumour cell production rate using bromodeoxyuridine (Begg et al. 1985) and estimations which relate to therapy resistance including glutathione, drug efflux mechanisms and membrane transport (Watson 1991). The power of the technology relates to its capacity to make very rapid multiple simultaneous measurements of fluorescence and light scatter at the individual cell level and hence to analyse heterogeneity in mixed populations.

The early commercial instruments were somewhat fearsome beasts with vast arrays of knobs, switches, dials, oscilloscopes and wires hanging out all over the place. At best, they tended to be regarded as "user non-friendly" and at worst as "non-user friendly". However, the recent generation of machines have been simplified considerably, with the in-house computer taking over many of the tasks which the operator previously had to perform manually. The undoubted "user-friendliness" of these modern instruments, together with the relative reduction in initial capital outlay, is a considerable advantage as it makes the technology available to many more users. In turn, this makes it possible for relatively untrained persons, who may not be fully aware of potential problems and pitfalls, to stain samples and operate the instruments to produce "numbers". There appears to be a prevalent philsophy amongst the instrument manufacturers to produce bench-top devices that require a minimum of operator interaction, so that all that needs to be done is stain up the cells, shove them in the instrument, and out come the numbers.

I'm sure this philosophy is fine from the manufacturers' standpoint, as this approach helps to sell more machines because you purchase an instrument specifically designed to do a particular task. Under test conditions the instrument will perform very well the particular task for which it was designed. However,

there are a number of disadvantages to this philosophy. First, a particular instrument designed specifically for a particular task may not do so well with an apparently similar task using different combinations of fluorochromes for different purposes. Second, the "new" generation of flow cytometry users and operators may not even be aware that such problems could exist. Third, the operator is usually insufficiently aware of deficiencies or potential deficiencies in a particular instrument, as no manufacturer will ever say it's not very good at doing this or that. Finally, many operators are completely at the mercy of the software data-handling package supplied, which may contain deficiencies that the manufacturers do not appreciate.

Flow cytometers produce a vast amount of data, which is one of their many attractions, but this can be a two-edged sword. Data, which are just a series of numbers, must be converted to information. Moreover, the information produced from those numbers not only must have meaning, but also must be shown to have meaning. This is the most important single aspect of flow cytometry, but it has received relatively little attention.

One of the frequently voiced advantages of the technology is that it produces "good statistics" because large numbers of cells have been analysed. However, confidence limits are seldom placed on results, and hence the reader has little or no feel for the inherent variability in the information produced. This variability is important and has three major components. The first is due to the measuring system, and applies not just to flow cytometry but to every measurement system. Manufacturers will tend to downplay or ignore this component. The second component is due to variability in the processes involved in making the measurement possible, and in flow studies this includes variability in fluorescence staining procedures (including the various reagents) as well as the technical competence with which the procedures are carried out. The last, and most important, source of variability is within the biology being studied, and it is from this that we might gain some extra information.

This short monograph was compiled from a series of notes originally intended for users of the custom-built instrument in the MRC Clinical Oncology Unit at Cambridge. All of the procedures described in this book are contained within our computer analysis package, which has been updated continuously over the past decade, and most of the examples are drawn from our data base. The statistics sections are limited to those we have found most useful, but I hope this will provide newcomers to flow cytometry with insight into some of the potential problems to be faced in data handling, data analysis and interpretation. The statistics begin with some very basic concepts including measurement of central tendency, distances between points and hence assessments of distributions, and what these various parameters mean. These basic concepts, of which everyone is well aware, are then developed to show how they can be used to analyse data and help convert them into information. A distinction is made here between data handling – for example, gating and counting the

numbers of cells within that gate (a process commonly regarded as data analysis but which, in reality, is data handling) – and data analysis itself, which is the means by which information is extracted. Gating is not covered as a specific topic.

The book is intended for biologists using flow cytometers who know about the basic anatomy and physiology of these instruments. Data analysis obviously implies that mathematical ideas and concepts will have to be considered. However, as the book has been written for biologists, an attempt has been made to make this aspect as simple as possible; if you can add, subtract, multiply, divide and, most importantly, think logically then you should have few problems. If you are also familiar with power functions, logarithms, transcendental functions, differentiation, integration and basic statistics then you probably need not be reading this. This book is not intended for highly experienced users and developers with backgrounds in physics, mathematics or statistics who also have struggled with the various problems considered within these pages.