TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

GLOSSARY

Coincidental Evolution. A tendency of genes present in the same genome to evolve in a non-independent manner. The presence of coincidental evolution makes homologous genes within a genome more similar to each other than to homologous genes from other genomes. Many authors use the term "concerted evolution" as a synonym for coincidental evolution, which was originally defined by Hood et al. (1975).

Contig. An island consisting of at least two fragments.

Fitness. An organism's ability to propagate.

Functional Genomics. The development and application of experimental approaches to assess gene function by making use of the information and reagents provided by structural genomics (Heiter and Boguski, 1997; McKusick, 1997).

Gap. A region of a target that is not represented in an island. Gaps are sometimes referred to as "oceans."

Gene Product. The product of a gene. Most genes code for proteins, but some code for structural RNAs, and some affect the structure and/or regulation of the genome without being transcribed into a downstream product.

Gene. Information encoded in a segment of genomic DNA that affects the fitness of an organism. Semantically, it is useful to consider some genes as consisting of multiple gene segments, such as the TCR β gene.

Genome. The information encoded in the DNA of a cell. Every individual, with few exceptions, has a distinct genome. The genome can vary slightly from cell to cell within an organism. The term was coined in 1920 by Winkler as an elision of the words "gene" and "chromosome" (McKusick, 1997).

Genomicist. One who studies genomes.

Genomics. The study of genomes. By contrast, the term "genetics" refers to the study of inheritance. The term "genomics" was introduced by Roderick in 1986 (McKusick, 1997).

Homologous Genes. Genes that share a common ancestral gene. Orthology and paralogy are subcategories of homology. Genes may be homologous without necessarily being either paralogous or orthologous (see, for example, Tatusov et al., 1997).

Indel. An insertion or a deletion.

Island. A maximal set of fragments each of which is connected to all other island members by at least one path of overlapping fragments.

Isozymes. Enzymes that have identical (or nearly identical) biochemical properties.

Orphon. A term applied to gene segments separated from a complete functional gene locus. A typical example is a TCR Vβ segment on chromosome 9, unable to recombine with Dβ and Cβ segments to form a functional gene.

Orthologous Genes. Genes in different species that share a common function and evolved from a common ancestor. By definition, the split between such genes was caused by a speciation event. The result of speciation (Fitch, 1970).

Paralogous Genes. Multiple genes resulting from duplication within a particular genome. The result of gene duplication (Fitch, 1970).

Parameterization. A particular choice of parameters for a model (or a project). Parameters are variables that one can control, such as the choice of how many clones to analyze, or what length clones to choose.

Pseudogene. A gene that is no longer functional. The ancestral sequence was a functional gene that acquired one or more mutations that destroyed functionality. Typically, this would entail the acquisition of stop codons in an open reading frame. Point mutations are typically responsible for the creation of pseudogenes.

Relic. A fragment of a gene that was once functional. The semantic boundary between a pseudogene and a relic is fuzzy. Generally in order to be classified a relic, one or more recombinations or major deletions must have operated on the ancestral gene.

Repeat. A region of a genome that is nearly identical to another region of the same genome. It should be emphasized that in genomics terminology the word “repeat” does not imply 100% identity. The percent identity used to define a repeat is somewhat subjective.

Scaffold. An ordered and oriented list of islands. Also referred to as a “gapped island” (Port et al., 1995) or a “supercontig” (Lawrence et al., 1994).

Structural Genomics. The construction of genetic, physical, and transcript maps of genomes (Heiter and Boguski, 1997). Genomic sequence is considered to be the ultimate high resolution physical map of a genome. The definition can be rephrased as, “mapping and sequencing genomes” (McKusick, 1997). This original definition has no been replaced in common usage. The term “structural genomics” now usually refers to the elucidation of all protein structures coded by a genome. This new meaning would be better served by the term “structural proteomics”, but this phrase is unfortunately not in usage.

Subclone. A clone of a fragment of a larger piece of DNA. The fragment has been genetically engineered into a vector that facilitates laboratory manipulation of the fragment.

Target. A genome or a subset of a genome that will be analyzed during the course of a project.

LIST OF ABBREVIATIONS

BAC. Bacterial artificial chromosome.

bp. Base pair.

cDNA. Complementary deoxyribonucleic acid.

DNA. Deoxyribonucleic acid.

ds. Double stranded.

EST. Expressed sequence tag.

HMM. Hidden Markov model.

kb. Kilobase pair.

mRNA. Messenger RNA.

OSS. Ordered shotgun sequencing.

PCR. Polymerase chain reaction.

RNA. Ribonucleic acid.

SMG. Sequence-mapped gap.

ss. Single stranded.

STS. Sequence-tagged connector.

STS. Sequence-tagged site.

TCR. T-cell receptor.

YAC. Yeast artificial chromosome.

LIST OF VARIABLES

C. An arbitrary length (a partial target goal less than *G*).

$\boldsymbol{D_k}$. The length of the k^{th} spacing from the left end of the target.

f. The fraction of the target that is covered.

$\boldsymbol{f_g}$. The effective fractional coverage of the target provided by one fragment.

G. The length of a target (in bases). This abbreviation is most appropriate when the target is a genome, but has come to be used even when the target is something else, such as a subclone. I do not employ the convention of setting *G* equal to unity.

$\boldsymbol{G_e}$. The effective length of the target.

$\boldsymbol{g_m}$. The number of short spacings in the m^{th} island from the left end of the target.

I. The length of an insert.

L. The length of a clone (in bases). I do not employ the convention of setting *L* equal to unity.

$\boldsymbol{l_m}$. The length of the m^{th} island from the left end of the target.

N. The number of permitted residues at a sequence site.

n. The total number of clones analyzed in a project, either by mapping or sequencing.

$\boldsymbol{N_{gaps}}$. The number of gaps in a project.

$\boldsymbol{N_{islands}}$. The number of islands in a project.

$\boldsymbol{N_{long}}$. The number of long spacings in a project.

$\boldsymbol{N_{singletons}}$. The number of single fragment islands in a project.

μ. The probability of a mutation per unit evolutionary time.

$\boldsymbol{p_{gap}}$. The probability of a gap following a spacing.

R. Redundancy.

$\boldsymbol{R_e}$. The effective redundancy.

$\boldsymbol{S_k}$. The k^{th} spacing from the left end of the target.

τ. Evolutionary time.

$\boldsymbol{T}$. The number of base pairs necessary to determine overlap during the assembly phase of a random subcloning project. In many practical cases $T<<L$, and can be approximated as zero.

$\boldsymbol{V}$. The length of the vector sequence.

$\boldsymbol{z_m}$. The number of fragments in the m^{th} island from the left end of the target.

PREFACE

The doctoral thesis has a rich history. The tradition of the dissertation binds modern students back through the ages to the history and culture of the Renaissance Universities (Haskins, 1923). The richness of values governing the content of theses dictates certain compromises with respect to the choice of included material and its style of presentation.

My desire in this thesis is to provide knowledge in a understandable fashion to the widest possible audience: biologists, mathematicians, computer scientists, engineers, and those in allied fields. I hope to have made at least a portion of the text accessible to students at all levels. However, for much of this dissertation, I will assume at least a basic knowledge of molecular biology, such as that obtainable from *The Cartoon Guide to Genetics* (Gonick and Wheelis, 1991) or another introductory molecular biology textbook. The reader is encouraged to ignore or merely skim material that is either too basic or advanced. A more compact presentation of much of the material in this dissertation can be found in my original journal articles which are reproduced in the appendices.

My doctoral research has led me down many disparate paths, of which three have been significant enough to include in this dissertation, each as a separate chapter. Two of these chapters group naturally in the area of strategic genomics, while the third strikes out tangentially into the field of molecular evolution. Rather than attempt an integrated and detailed introduction to these three subjects simultaneously, I have kept the overall introduction, which follows, short and generally oriented. I include more detailed topical introductions at the beginning of each chapter. Also, in each chapter I employ several specialized terms appropriate to the subject matter; I encourage the reader to exploit the glossary when in doubt of the meaning of a particular term. Readers may also find the list of abbreviations and the list of variables useful as occasional references.

This thesis is a living document. The most recent edition of this thesis can be found on the web. My current website is faculty.washington.edu/roach. Errata notifications and suggestions for additions or improvements are appreciated.

ACKNOWLEDGEMENTS

Collaboration is the cornerstone of modern research. One can lift a rock; many can move a mountain. The many include: Lee Hood, Roger Perlmutter, Chris Wilson, Dave Lewis, Ken Walsh, Andy Siegel, Steve Henikoff, Kai Wang, and innumerable technicians, graduate students, post-docs, professors, friends, and family, all who have contributed variously their labor, ideas, and encouragement. I am grateful for a grant from the Life & Health Insurance Medical Research Fund.