Preliminary Project Report

Assignment Date: March 30, 2019
Due Date: Monday, April 13, 2019 @ 11:59pm

Each team should submit a PDF of your preliminary project proposal (2 to 3 pages) to GradeScope by 11:59pm on Monday April 13.

The preliminary report should have at least:

- Title of your project
- List of team members and email addresses
- 1 paragraph abstract summarizing the project
- 1+ paragraph of Introduction
- 1+ paragraph of Methods that you are using
- 1+ paragraph of Results, describing the data evaluated and any preliminary results
- 1+ paragraph of Discussion (what you have seen or expect to see)
- 1+ figure showing a preliminary result
- 5+ References to relevant papers and data

The preliminary report should use the Bioinformatics style template. Word and LaTeX templates are available at https://academic.oup.com/bioinformatics/pages/submission_online. Overleaf is recommended for LaTeX submissions. Google Docs is recommended for non-latex submissions, especially group projects. Paperpile is recommended for citation management.

Later, you will present your project in class starting the week of April 22. You will also submit your final written report (5-7 pages) of your project by May 13.

Please use Piazza if you have any general questions!
Part I: Review
Variation across populations

• Not a single variant 100% unique to a given population
• 17% of low-frequency variants (.5-5% pop. freq) observed in a single ancestry group
• 50% of rare variants (<.5%) observed in a single population
Genes mirror geography within Europe

Part 2:
Inherited Diseases
Huntington’s Disease

A Novel Gene Containing a Trinucleotide Repeat That Is Expanded and Unstable on Huntington’s Disease Chromosomes

The Huntington’s Disease Collaborative Research Group*

Summary

The Huntington’s disease (HD) gene has been mapped in 4p16.3 but has eluded identification. We have used haplotype analysis of linkage disequilibrium to spotlight a small segment of 4p16.3 as the likely location of the defect. A new gene, IT15, isolated using cloned trapped exons from the target area contains a polymorphic trinucleotide repeat that is expanded and unstable on HD chromosomes. A (CAG)n repeat longer than the normal range was observed on HD chromosomes from all 75 disease families examined, comprising a variety of ethnic backgrounds and 4p16.3 haplotypes. The (CAG)n repeat appears to be located within the coding sequence of a predicted ~348 kd protein that is widely expressed but unrelated to any known gene. Thus, the HD mutation involves an unstable DNA segment, similar to those described in fragile X syndrome, spino-bulbar muscular atrophy, and myotonic dystrophy, acting in the context of a novel 4p16.3 gene to produce a dominant phenotype.

Introduction

Huntington’s disease (HD) is a progressive degenerative disorder characterized by motor (stiffness, rigidity, and dystonia), cognitive, and psychiatric manifestations (Berg-Larsen, 1986). It is inherited in an autosomal dominant fashion and affects ~1 in 10,000 individuals. The most characteristic feature of HD is a distinctive chorea (involuntary, writhing movements) that typically has a subtle, insidious onset in the late fourth or fifth decade of life and gradually worsens, taking 10 to 20 years until death. Occasional intermediate phenotypes are expressed in juveniles, typically manifesting as early-onset Huntington’s disease. Juvenile onset of HD is associated with Mendelian transmission of the disease, and the paternally-derived chromosome 4 is the most frequent abnormality in HD patients. The sequence of HD also displays a distinct pattern of selective loss of neurons that is most severe in the caudate nucleus and putamen. The biochemical basis of HD is not yet understood, but recent genetic advances have provided clues to its pathogenesis. The genetic defect causing HD was mapped to human chromosome 4 in 1983 in one of the first successful approaches using polymorphic DNA markers.

Figure 6. PCR Analysis of the (CAG)n Repeat in a Venezuelan HD Sibship with Some Offspring Displaying Juvenile Onset

Results of PCR analysis of a sibship in the Venezuelan HD pedigree are shown. Affected individuals are represented by closed symbols. Progeny are shown as triangles, and the birth order of some individuals has been changed for confidentiality. AN1, AN2, and AN3 mark the positions of the allelic products from normal chromosomes. AE marks the range of PCR products from the HD chromosome. The intensity of background constant bands, which represent a useful reference for comparison of the above PCR products, varies with slight differences in PCR conditions. The PCR products from cosmid L191F1 and GUS72-2130 are loaded in lanes 12 and 13 and have 18 and 48 CAG repeats, respectively.
Human disease genes

Gerardo Jimenez-Sanchez*, Barton Childs* & David Valle†

* Department of Pediatrics, McKusick-Nathans Institute of Genetic Medicine, and † Howard Hughes Medical Institute, Johns Hopkins University School of Medicine, Baltimore, Maryland 21205, USA

The complete human genome sequence will facilitate the identification of all genes that contribute to disease. We propose that the functional classification of disease genes and their products will reveal general principles of human disease. We have determined functional categories for nearly 1,000 documented disease genes, and found striking correlations between the function of the gene product and features of disease, such as age of onset and mode of inheritance. As knowledge of disease genes grows, including those contributing to complex traits, more sophisticated analyses will be possible; their results will yield a deeper understanding of disease and an enhanced integration of medicine with biology.

To test the proposal that classifying disease genes and their products according to function will provide general insight into disease processes, we have compiled and classified a list of disease genes. To assemble the list, we began with 269 genes identified in a survey of the 7th edition of Metabolic and Molecular Bases of Inherited Disease. We then searched the 'morbid map' and allelic variants listed in the Online Mendelian Inheritance in Man (OMIM), an online resource documenting human diseases and their associated genes (www.ncbi.nlm.nih.gov), and increased the total disease gene set to 923. This sample included genes that cause monogenic disease (97% of the sample) and genes that increase susceptibility for complex traits. We excluded genes associated only with somatic genetic disease (such as non-inherited forms of cancer) or the mitochondrial genome.

Functional classification
We categorized each disease gene according to the function of its

**Human disease genes**
Genome Wide Association (GWAS)

<table>
<thead>
<tr>
<th>SNP1 Cases</th>
<th>SNP2 Cases</th>
</tr>
</thead>
<tbody>
<tr>
<td>Count of G: 2104 of 4000</td>
<td>Count of G: 1648 of 4000</td>
</tr>
<tr>
<td>Frequency of G: 52.6%</td>
<td>Frequency of G: 41.2%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>SNP1 Controls</th>
<th>SNP2 Controls</th>
</tr>
</thead>
<tbody>
<tr>
<td>Count of G: 2676 of 6000</td>
<td>Count of G: 2532 of 6000</td>
</tr>
<tr>
<td>Frequency of G: 44.6%</td>
<td>Frequency of G: 42.2%</td>
</tr>
</tbody>
</table>

| SNP1 P-value: $5.0 \cdot 10^{-15}$ | SNP2 P-value: 0.33 |

Chi-squared or similar test
Manhattan Plot

Genome-wide significance: $5 \times 10^{-8}$

Marker significance: $1 \times 10^{-5}$

Significant loci

Four Novel Loci (19q13, 6q24, 12q24, and 5q14) Influence the Microcirculation In Vivo

Regional Association Plot

Why does it look like this? Which is causal?
GWAS Catalog

As of 2020-03-08, the GWAS Catalog contains 4493 publications and 179364 associations.

http://www.ebi.ac.uk/gwas/diagram
GWAS In Crisis

Non-Replication and Inconsistency in the Genome-Wide Association Setting


<table>
<thead>
<tr>
<th>Phenotype</th>
<th>Genome-wide association study characteristics</th>
<th>Identified gene/SNPs</th>
<th>Replication status (January 2007)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age-related macular degeneration</td>
<td>Affymetrix 100k (116204/103611), UCC; then sequencing of region, Genomic control, F-ratio</td>
<td>CFH/Intronic rs380390; then sequencing showing exonic rs106170 (Y420H) 2kb upstream of 41-kb haplotype block</td>
<td>Meta-analysis of 11 studies (n = 8,991): OR 2.49 and 6.15 (heterozygotes and homozygotes respectively), no large between-study inconsistency in effect sizes; also replicated in large Dutch cohort (n = 5,681); several studies on Asian populations claim no association</td>
</tr>
<tr>
<td>Obesity</td>
<td>Affymetrix 100k (116204/86604), Family-based, 2-stage, followed by mapping 100 neighboring SNPs, Family-based design, Family-based, then sequencing up to 923</td>
<td>INSIG2/rs7566605 10kb upstream of the transcription start site</td>
<td>Replication in the same publication in 3 of 4 independent populations of n = 9,881 subjects with modest between-study heterogeneity; 7 more independent populations with over 21,000 subjects total failed to replicate the association; no effect and no heterogeneity across the independent replication teams</td>
</tr>
<tr>
<td>Parkinson disease</td>
<td>Perlegen (248535/198345), Family-based, second stage with matched case-controls, Family-based design; matching at second stage; also genomic control</td>
<td>13 different SNPs identified from analysis of both stages; none with genome-wide significance</td>
<td>Several small replication studies and a large collaborative consortium (n = 12,208) failed to replicate any of the 13 proposed SNPs; null results were consistent across the teams participating in the consortium</td>
</tr>
<tr>
<td>Myocardial infarction</td>
<td>Random gene-based (92788/67671), UCC, None (just Japanese nationality)</td>
<td>LTA/Haplotype of 5 SNPs (2 in LTA and 3 in adjacent genes); the two LTA SNPs had association in larger sample and then Thr26Asn had also functional assay support</td>
<td>Replication in the same publication in additional 1,133 cases and two control groups (n = 1,006 and 872); association not replicated in subsequent ISIS-4 case-control study and meta-analysis (n = 18,325) shows no association (non-significant OR 1.07 without significant between-study heterogeneity vs. 1.77 in originally proposed association for recessive model)</td>
</tr>
<tr>
<td>Age-related macular degeneration</td>
<td>Affymetrix 100k (116204/97824), UCC; then sequencing of region, Genomic control, F-ratio</td>
<td>HTRA1/Intragenic rs10490924; then sequencing showing promoter rs11200838 6kb downstream</td>
<td>Independent study (n = 890) published in the same issue starting from dense mapping of locus showing consistent effects with OR 1.90 and 7.51 for heterozygotes and homozygotes, respectively</td>
</tr>
</tbody>
</table>
“Three groups of researchers scoured the genomes of huge populations (>30,000 people) for genetic variants associated with the height differences. More than 40 turned up. But there was a problem: the variants had tiny effects. Altogether, they accounted for little more than 5% of height’s heritability”

- Rare, moderately penetrant or common, weakly penetrant variants?
- CNVs and SVs?
- Epistasis (multiple genes working together)?
- Epigenetic effects, especially in utero?
Penetrance: The proportion of individuals with a specific genotype who manifest the genotype at the phenotypic level.
A central goal of genetics is to understand the links between genetic variation and disease. Intuitively, one might expect disease-causing variants to cluster into key pathways that drive disease etiology. But for complex traits, association signals tend to be spread across most of the genome—including near many genes without an obvious connection to disease. We propose that gene regulatory networks are sufficiently interconnected such that all genes expressed in disease-relevant cells are liable to affect the functions of core disease-related genes and that most heritability can be explained by effects on genes outside core pathways. We refer to this hypothesis as an “omnigenic” model.
Epigenetic Factors

The Key Role of Epigenetics in Human Disease Prevention and Mitigation

Needles in stacks of needles: finding disease-causal variants in a wealth of genomic data

Methods

Predicting Deleterious Amino Acid Substitutions

Pauline C. Ng\textsuperscript{1,2} and Steven Henikoff\textsuperscript{1,3,4}

\textsuperscript{1}Fred Hutchinson Cancer Research Center, Seattle, Washington 98109, USA; \textsuperscript{2}Department of Bioengineering, University of Washington, Seattle, Washington 98105, USA; \textsuperscript{3}Howard Hughes Medical Institute, Seattle, Washington 98109, USA

Many missense substitutions are identified in single nucleotide polymorphism (SNP) data and large-scale random mutagenesis projects. Each amino acid substitution potentially affects protein function. We have constructed a tool that uses sequence homology to predict whether a substitution affects protein function. SIFT, which sorts intolerant from tolerant substitutions, classifies substitutions as tolerated or deleterious. A higher proportion of substitutions predicted to be deleterious by SIFT gives an affected phenotype than substitutions predicted to be deleterious by substitution scoring matrices in three test cases. Using SIFT before mutagenesis studies could reduce the number of functional assays required and yield a higher proportion of affected phenotypes. SIFT may be used to identify plausible disease candidates among the SNPs that cause missense substitutions.

\textbf{SIFT Key Idea:} Substituting one amino acid for another with another with very similar biochemical properties is probably less significant than a more dissimilar substitution. Learn those similarities by comparing orthologs across species.
A probabilistic disease-gene finder for personal genomes

Mark Yandell,1,3,4 Chad Huff,1,3 Hao Hu,1,3 Marc Singleton,1 Barry Moore,1 Jinchuan Xing,1 Lynn B. Jorde,1 and Martin G. Reese2

1Department of Human Genetics, Eccles Institute of Human Genetics, University of Utah and School of Medicine, Salt Lake City, Utah 84112, USA; 2Omicia, Inc., Emeryville, California 94608, USA

VAAST (the Variant Annotation, Analysis & Search Tool) is a probabilistic search tool for identifying damaged genes and their disease-causing variants in personal genome sequences. VAAST builds on existing amino acid substitution (AAS) and aggregative approaches to variant prioritization, combining elements of both into a single unified likelihood framework that allows users to identify damaged genes and deleterious variants with greater accuracy, and in an easy-to-use fashion. VAAST can score both coding and noncoding variants, evaluating the cumulative impact of both types of variants simultaneously. VAAST can identify rare variants causing rare genetic diseases, and it can also use both rare and common variants to identify genes responsible for common diseases. VAAST thus has a much greater scope of use than any existing methodology. Here we demonstrate its ability to identify damaged genes using small cohorts (n = 3) of unrelated individuals, wherein no two share the same deleterious variants, and for common, multigenic diseases using as few as 150 cases.

[Supplemental material is available for this article.]

VAAST Key Idea: Evaluate amino acid substitutions in evolution AND allele frequencies in 1000 genomes project

Genome Research 2011. doi:10.1101/gr.123158.111
A general framework for estimating the relative pathogenicity of human genetic variants

Martin Kircher1,5, Daniela M Witten2,5, Preti Jain3,4, Brian J O’Roak1,4, Gregory M Cooper3 & Jay Shendure1

Current methods for annotating and interpreting human genetic variation tend to exploit a single information type (for example, conservation) and/or are restricted in scope (for example, to missense changes). Here we describe Combined Annotation–Dependent Depletion (CADD), a method for objectively integrating many diverse annotations into a single measure (C score) for each variant. We implement CADD as a support vector machine trained to differentiate 14.7 million high-frequency human-derived alleles from 14.7 million simulated variants. We precompute C scores for all 8.6 billion possible human single-nucleotide variants and enable scoring of short insertions–deletions. C scores correlate with allelic diversity, annotations of functionality, pathogenicity, disease severity, experimentally measured regulatory effects and complex trait associations, and they highly rank known pathogenic variants within individual genomes. The ability of CADD to prioritize functional, deleterious and pathogenic variants across many functional categories, effect sizes and genetic architectures is unmatched by any current single-annotation method.

CADD Key Idea: Evaluate amino acid substitutions AND allele frequencies in 1000 genomes project AND ENCODE regions AND ... (63 annotations total :)
A method for calculating probabilities of fitness consequences for point mutations across the human genome

Brad Gulko¹, Melissa J Hubisz², Ilan Gronau²–³ & Adam Siepel¹–³

We describe a new computational method for estimating the probability that a point mutation at each position in a genome will influence fitness. These ‘fitness consequence’ (fitCons) scores serve as evolution-based measures of potential genomic function. Our approach is to cluster genomic positions into groups exhibiting distinct ‘fingerprints’ on the basis of high-throughput functional genomic data, then to estimate a probability of fitness consequences for each group from associated patterns of genetic polymorphism and divergence. We have generated fitCons scores for three human cell types on the basis of public data from ENCODE. In comparison with conventional conservation scores, fitCons scores show considerably improved prediction power for cis regulatory elements. In addition, fitCons scores indicate that 4.2–7.5% of nucleotides in the human genome have influenced fitness since the human-chimpanzee divergence, and they suggest that recent evolutionary turnover has had limited impact on the functional content of the genome.

fitCons Key Idea: Evaluate amino acid substitutions AND allele frequencies in 1000 genomes project AND aggregate by ENCODE regions
Using VAAST to Identify an X-Linked Disorder Resulting in Lethality in Male Infants Due to N-Terminal Acetyltransferase Deficiency

Alan F. Rope,1 Kai Wang,2,19 Rune Evjenth,3 Jinchuan Xing,4 Jennifer J. Johnston,5 Jeffrey J. Swensen,6,7 W. Evan Johnson,8 Barry Moore,4 Chad D. Huff,4 Lynne M. Bird,9 John C. Carey,1 John M. Opitz,1,4,6,10,11 Cathy A. Stevens,12 Tao Jiang,13,14 Christa Schank,8 Heidi Deborah Fain,15 Reid Robison,15 Brian Dalley,16 Steven Chin,6 Sarah T. South,1,7 Theodore J. Pysher,6 Lynn B. Jorde,4 Hakon Hakonarson,2 Johan R. Lillegaard,3 Leslie G. Biesecker,5 Mark Yandell,4 Thomas Arnesen,3,17 and Gholson J. Lyon15,18,20,*

We have identified two families with a previously undescribed lethal X-linked disorder of infancy; the disorder comprises a distinct combination of an aged appearance, craniofacial anomalies, hypotonia, global developmental delays, cryptorchidism, and cardiac arrhythmias. Using X chromosome exon sequencing and a recently developed probabilistic algorithm aimed at discovering disease-causing variants, we identified in one family a c.109T>C (p.Ser37Pro) variant in NAA10, a gene encoding the catalytic subunit of the major human N-terminal acetyltransferase (NAT). A parallel effort on a second unrelated family converged on the same variant. The absence of this variant in controls, the amino acid conservation of this region of the protein, the predicted disruptive change, and the co-occurrence in two unrelated families with the same rare disorder suggest that this is the pathogenic mutation. We confirmed this by demonstrating a significantly impaired biochemical activity of the mutant hNaa10p, and from this we conclude that a reduction in acetylation by hNaa10p causes this disease. Here we provide evidence of a human genetic disorder resulting from direct impairment of N-terminal acetylation, one of the most common protein modifications in humans.
Figure 2. Pedigree Drawing and Pictures of Families 1 and 2
(A) Pedigree drawing for family 1. The most recent deceased individual, III-4, is the most well-studied subject in the family and is indicated by an arrow. Genotypes are marked for those in which DNA was available and tested. The following abbreviations are used: SB, stillborn; +, normal variant; mt, rare mutant variant.
(B) Pictures of four affected and deceased boys in this family, showing the aged appearance.
(C) Sanger sequencing results of NAA10 in individual III-4 from family 1.
(D) Pedigree for family 2. Individual III-2 is the most well-studied subject in the family and is indicated by an arrow.
(E) Picture of individuals II-4 and III-2 in family 2 at ~1 year of age.
Part 3:
Cancer Genetics & Genomics
A tumor removed by surgery in 1689.
Benign vs. Malignant Tumors

**Benign** (not cancer) tumor cells grow only locally and cannot spread by invasion or metastasis.

**Malignant** (cancer) cells invade neighboring tissues, enter blood vessels, and metastasize to different sites.
The Six Hallmarks of Cancer

Hallmarks of Cancer
Somatic Mutations In Cancer

Signatures of mutational processes in human cancer
SK-BR-3

Most commonly used Her2-amplified breast cancer cell line

80+ chromosomes,
Many are a patchwork of fragments of other chromosomes

(Davidson et al, 2000)
A firestorm in cancer

Figure 2. Major types of tumor genomic profiles. Segmentation profiles for individual tumors representing each category: (A) simplex; (B) complex type I or sawtooth; (C) complex type II or firestorm. Scored events consist of a minimum of six consecutive probes in the same state. The y-axis displays the geometric mean value of two experiments on a log scale. Note that the scale of the amplifications in C is compressed relative to A and B owing to the high levels of amplification in firestorms. Chromosomes 1–22 plus X and Y are displayed in order from left to right according to probe position.

Novel patterns of genome rearrangement and their association with survival in breast cancer
Aberrations in cancer genomes

Chromothripsis, which literally means 'chromosome shattering', is a phenomenon that has recently been reported to occur in cells harbouring complex genomic rearrangements (CGRs). It has three defining characteristics:

1. Occurrence of remarkable numbers of rearrangements in localized chromosomal regions;
2. Low number of copy number states (generally between one or two) across the rearranged region;
3. Alternation in the chromothriptic areas of regions where heterozygosity is preserved with regions presenting loss of heterozygosity (LOH).

Chromothripsis and cancer: causes and consequences of chromosome shattering
Hypomethylation distinguishes genes of some human cancers from their normal counterparts

Andrew P. Feinberg & Bert Vogelstein
Cell Structure and Function Laboratory, The Oncology Center, Johns Hopkins University School of Medicine, Baltimore, Maryland 21205, USA

It has been suggested that cancer represents an alteration in DNA, heritable by progeny cells, that leads to abnormally regulated expression of normal cellular genes; DNA alterations such as mutations\textsuperscript{1,2}, rearrangements\textsuperscript{3-5} and changes in methylation\textsuperscript{\textsuperscript{6-8}} have been proposed to have such a role. Because of increasing evidence that DNA methylation is important in gene expression (for review see refs 7, 9-11), several investigators have studied DNA methylation in animal tumours, transformed cells and leukaemia cells in culture\textsuperscript{8,12-30}. The results of these studies have varied; depending on the techniques and systems used, an increase\textsuperscript{12-15}, decrease\textsuperscript{20-24}, or no change\textsuperscript{25-29} in the degree of methylation has been reported. To our knowledge, however, primary human tumour tissues have not been used in such studies. We have now examined DNA methylation in human cancer with three considerations in mind: (1) the methylation pattern of specific genes, rather than total levels of methylation, was determined; (2) human cancers and adjacent analogous normal tissues, unconditioned by culture media, were analysed; and (3) the cancers were taken from patients who had received neither radiation nor chemotherapy. In four of five patients studied, representing two histological types of cancer, substantial hypomethylation was found in genes of cancer cells compared with their normal counterparts. This hypomethylation was progressive in a metastasis from one of the patients.

and (3) \textit{HpaII} and \textit{Hhal} cleavage sites should be present in the regions of the genes.

The first cancer studied was a grade D (ref. 43), moderately well differentiated adenocarcinoma of the colon from a 67-yr-old male. Tissue was obtained from the cancer itself and also from colonic mucosa stripped from the colon at a site just outside the histologically proven tumour margin. Figure 1 shows the pattern of methylation of the studied genes. Before digestion with restriction enzymes, all DNA samples used in the study had a size >25,000 base pairs (bp). After \textit{HpaII} cleavage, hybridization with a probe made from a cDNA clone of human growth hormone (HGH) showed that significantly more of the DNA was digested to low-molecular weight fragments in DNA from the cancer (labelled C in Fig. 1) than in DNA from the normal colonic mucosa (labelled N). In the hybridization conditions used, the HGH probe detected the human growth hormone genes as well as the related chorionic somatotropin

<table>
<thead>
<tr>
<th>Table 1</th>
<th>Quantitation of methylation of specific genes in human cancers and adjacent analogous normal tissues</th>
</tr>
</thead>
<tbody>
<tr>
<td>Patient</td>
<td>Carcinoma</td>
</tr>
<tr>
<td>1</td>
<td>Colon</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Colon</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Colon</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Causes of Cancer

Cancer is a Preventable Disease that Requires Major Lifestyle Changes

Variation in cancer risk among tissues can be explained by the number of stem cell divisions.

Tomasetti and Vogelstein (2015) Science. DOI: 10.1126/science.1260825
Stem cell divisions, somatic mutations, cancer etiology, and cancer prevention

Tomasetti, Li, and Vogelstein (2017) Science. DOI: 10.1126/science.aaf9011
Oncogenes

- **HER-2/neu**: encodes for a cell surface receptor that can stimulate cell division. The HER-2/neu gene is amplified in up to 30% of human breast cancers.
- **RAS**: The Ras gene products are involved in kinase signaling pathways that ultimately control transcription of genes, regulating cell growth and differentiation.
- **MYC**: The Myc protein is a transcription factor and controls expression of several genes.
- **SRC**: First oncogene ever discovered. The Src protein is a tyrosine kinase, which regulates cell activity.
- **hTER**: Codes for an enzyme (telomerase) that maintains chromosome ends.


**Tumor Suppressors**

- **TP53**: a transcription factor that regulates cell division and cell death.
- **Rb**: alters the activity of transcription factors and therefore controls cell division.
- **APC**: controls the availability of a transcription factor.
- **PTEN**: acts by opposing the action of PI3K, which is essential for anti-apoptotic, pro-tumorogenic Akt activation.
TP53: The first and most important tumor suppressor

Surfing the p53 network

<table>
<thead>
<tr>
<th>Mechanism of inactivating p53</th>
<th>Typical tumours</th>
<th>Effect of inactivation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Amino-acid-changing mutation in the DNA-binding domain</td>
<td>Colon, breast, lung, bladder, brain, pancreas, stomach, oesophagus and many others</td>
<td>Prevents p53 from binding to specific DNA sequences and activating the adjacent genes</td>
</tr>
<tr>
<td>Deletion of the carboxy-terminal domain</td>
<td>Occasional tumours at many different sites</td>
<td>Prevents the formation of tetramers of p53</td>
</tr>
<tr>
<td>Multiplication of the MDM2 gene in the genome</td>
<td>Sarcomas, brain</td>
<td>Extra MDM2 stimulates the degradation of p53</td>
</tr>
<tr>
<td>Viral infection</td>
<td>Cervix, liver, lymphomas</td>
<td>Products of viral oncogenes bind to and inactivate p53 in the cell, in some cases stimulating p53 degradation</td>
</tr>
<tr>
<td>Deletion of the p14ARF gene</td>
<td>Breast, brain, lung and others, especially when p53 itself is not mutated</td>
<td>Failure to inhibit MDM2 and keep p53 degradation under control</td>
</tr>
<tr>
<td>Mislocalization of p53 to the cytoplasm, outside the nucleus</td>
<td>Breast, neuroblastomas</td>
<td>Lack of p53 function (p53 functions only in the nucleus)</td>
</tr>
</tbody>
</table>

Figure 1 The many ways in which p53 may malfunction in human cancers.

>10,000 known mutations
>17,000 publications
**BRCA1 and BRCA2 (breast cancer type 1/2 susceptibility genes)**

Normally expressed in the cells of breast and other tissue, where they help repair damaged DNA, or destroy cells if DNA cannot be repaired. They are involved in the repair of chromosomal damage with an important role in the error-free repair of DNA double-strand breaks.
The Clonal Evolution of Tumor Cell Populations
Tumor Evolution

Evolution of the cancer genome
Tumor Heterogeneity

The evolution of tumour phylogenetics: principles and practice
The evolution of tumour phylogenetics: principles and practice
Cancer Mutation Analysis

http://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1002824
DNA sequencing of a cytogenetically normal acute myeloid leukaemia genome

Acute myeloid leukaemia is a highly malignant haematopoietic tumour that affects about 13,000 adults in the United States each year. The treatment of this disease has changed little in the past two decades, because most of the genetic events that initiate the disease remain undiscovered. Whole-genome sequencing is now possible at a reasonable cost and timeframe to use this approach for the unbiased discovery of tumour-specific somatic mutations that alter the protein-coding genes. Here we present the results obtained from sequencing a typical acute myeloid leukaemia genome, and its matched normal counterpart obtained from the same patient's skin. We discovered ten genes with acquired mutations; two were previously described mutations that are thought to contribute to tumour progression, and eight were new mutations present in virtually all tumour cells at presentation and relapse, the function of which is not yet known. Our study establishes whole-genome sequencing as an unbiased method for discovering cancer-initiating mutations in previously unidentified genes that may respond to targeted therapies.
Figure 1. Overlap of SNPs detected in 933124 and other genomes
(A) Venn diagram of overlap between SNPs detected in the 933124 tumor genome and the
genomes of Watson and Venter. (B) Venn Diagram of overlap among 933124 tumor
genome, skin genome, and dbSNP (ver. 127). Single nucleotide variants were defined with a
MAQ SNP quality ≥ 15.

Figure 2. Filters used to identify somatic point mutations in the tumor genome
See text for details.
First Melanoma Genome

• Insertions (light-green rectangles);
• Deletions (dark-green rectangles);
• Heterozygous (light-orange bars) and Homozygous (dark-orange bars) Substitutions
• Coding substitutions (coloured squares: silent in grey, missense in purple, nonsense in red and splice site in black);
• Copy number (blue lines); regions of LOH (red lines);
• Intrachromosomal rearrangements (green lines);
• Interchromosomal rearrangements (purple lines).

A comprehensive catalogue of somatic mutations from a human cancer genome
Mutations in Breast Cancer

Comprehensive molecular portraits of human breast tumours
Finding Driving Mutations

Integrative Annotation of Variants from 1092 Humans: Application to Cancer Genomics
Regulatory mutations in PDAC

Coding alterations of PDAC are now fairly well established but non-coding mutations (NCMs) largely unexplored

- Developed GECCO to analyze the thousands of somatic mutations observed from hundreds of tumors to find potential drivers of gene expression and pathogenesis

- NCMs are enriched in known and novel pathways
- NCMs correlate with changes in gene expression
- NCMs can demonstrably modulate gene expression
- NCMs correlate with novel clinical outcomes

NCMs are an important mechanism for tumor genome evolution

Recurrent noncoding regulatory mutations in pancreatic ductal adenocarcinoma
Feigin, M, Garvin, T et al. (2017) Nature Genetics. doi:10.1038/ng.3861
Driving Non-Coding Mutations

Recurrent noncoding regulatory mutations in pancreatic ductal adenocarcinoma
Feigin, M, Garvin, T et al. (2017) Nature Genetics. doi:10.1038/ng.3861
The evolution of tumour phylogenetics: principles and practice
Tumor-Normal Pairs

Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples
Bulk Heterogeneity

THetA: inferring intra-tumor heterogeneity from high-throughput DNA sequencing data
Somatic Variant Detection

Evolution of the cancer genome
The evolution of tumour phylogenetics: principles and practice
Tumour evolution inferred by single-cell sequencing

Nicholas Navin, Jude Kendall, Jennifer Troge, Peter Andrews, Linda Rodgers, Jeanne McIndoo, Kerry Cook, Asya Stepansky, Dan Levy, Diane Esposito, Lakshmi Muthuswamy, Alex Krasnitz, W. Richard McCombie, James Hicks & Michael Wigler
What causes “outlier” families?

Eli Van Allen,
Dana-Farber Cancer Institute