Contents

Chapter 1

Molecular, Cellular & Developmental Biology

1.1 Principles and Background

1.1.1 Structure-Function Paradigm

Paradigm: explain the function/behavior of a system in terms of its structure, aka., its components and their interactions. For a biological system, the central problems are:

- What is the function of the system, or why does the behave in certain ways? This may not be obvious especially for complex control systems. Ex. the TRN controlling the global transcriptional response to some signal (it is not clear what should be the right response, thus the function of the TRN is not precisely known).
- How is the function/behavior of a system achieved? What are the important biological features that make it possible? Operationally, an important feature is one without which the function cannot be properly implemented.
- A more general view of the function/behavior of a system is: the evolutionary benefits of the system, or the reason it is selected by evolution. Important considerations are: (1) there may be many ways of implementing a function; (2) the evolution is constrained by where it starts from.
- Knowing the function/adaptation of a system would allow one to make predictions of the structure/design of the system. Ex. antioxidants protect against UV radiations, thus we can predict the amount of antioxidant would be higher in plants exposed to more sunlights.

Problems: to understand the structure-function relationship of a system, divide into multiple problems of different levels/details or different angles:

- What is the function of the system being studied?
- What are the components of the system? Identification of relevant genes or cells, and understanding the function of components.
- How do the components interact to achieve the function? Mechanistic understanding of the system and identify important features. Even if we successfully map all components and interactions in a system, it does not necessarily mean we understand it. In these cases, the analysis of the system would be more important (related to the next point of design principles).

- Why is the system designed/behaved in this way? Rationalize the features of the system. This may need evolutionary thinking, especially when the function is quantitative.

Structural and functional pespectives:

- Perspectives: focus on the change of the structure or function of a system.
- Example: detecting cancer genes in tumor cells: (1) structure pespective: they are often located in regions with mutations (not match the normal cells); (2) functional perspective: putting these genes in normal cells would transform the normal cells.

Techniques of exploring a biological system:

- Observation: of internal states of the system (e.g. gene expression) could allow one to dissect the system structure. This could be combined with other types of analysis.
- Functional analysis: how the function of a system varies with environments/etc. or changes upon manipulations.
- Manipulation: change one or multiple parts of a system, observe the effect(s) on the behavior or internal states of the system. (Internal states of a system can be viewed as a special type of phenotype of the system).
- Association: (when cannot manipulate system under controlled conditions) associate changes with effects (behavior or internal states).
- Engineering: create new system and understand its behavior (in this sense, pertubation analysis is one special kind of engineering).

Remark:

- Mechanism-driven research: the goal is to understand mechanisms, at different level of details. Often when one hypothesis is proven, new problems at a higher-level of details will emerge. Ex. once a candidate gene of a trait is found, the next question is naturally how does this gene work.
- In biology, normally any single method (for testing a hypothesis) is imperfect, thus it is important to pursue multiple methods. This is often because of the difficulty of establishing causality.

1.1.2 Methods for Studying Biological Systems

Understanding the mechanism of a biological process (environmental response, cellular phenotypes, etc.):

- Identifying players involved:
  - Perturbations: genetic deletions, or knockdowns, or overexpression experiments, and test how the responses/phenotypes are affected. Natural variations (e.g. genotypes) can also be used.
  - Associations: correlation of expression patterns with phenotypes or environmental stimuli. Ex. genes whose expression change significantly in response to perturbations, genes whose expression correlates with some quantitative traits, etc.
- Interactions among molecular players:
  - Observing the process at molecular level: e.g. imaging of translocation and association of molecules.
  - Mechanistic studies of molecular interactions: e.g. manipulation of the level of one protein, and examine the level of other genes.

  - Network analysis: use PPI networks, TRN networks, etc., to identify the partners, targets of the genes, and other candidate genes, etc.

- Linking genes and phenotypes: generally, for complex phenotypes, link with lower-level events that may be explained by molecular mechanisms (intermediate phenotypes). Examples:
  - Expression patterns of genes: spatial-temporal aspects. Example: to link FOXP2 with linguistic ability, regions of expression suggest its role in higher cognitive function (Broca's region), as well as sensory-motor integration. The temporal pattern suggests the specific roles in brain development.
  - The tissue and organ-specific phenotypes of the mutants. In FOXP2 example: phenotype of FOXP2 mutant patients includes the changes in brain tissues, neural circuits (synpatic connections), etc. Another example: how cellular developmental programs differ in mutants of C. elegans (thus to identify genes involved in cell fate determination).

- Examples:
  - UPR in yeast [Jonikas & Schuldiner, Science, 2009]: phenotype is measured through the activity of Hac1 promoter (Hac1 is the main regulator of UPR); genetic screen to identify genes that affect UPR; genetic interactions to identify molecular pathways.

Dissecting mechanisms of complex processes:

- Difficulty: association is not equal to causation. In a complex process, many simultaneous changes may have occured, e.g. a cell adapting to an environmental stress, and it is difficult to establish causal relationship.

- Association analysis: the changes that are associated/correlated with the process of interest may play a role in the process. Ex. genes that change expression level during aging may play some role in aging.

- Control through manipulation/perturbation: when everything else is fixed except X, and changing X does not have an effect on the process, then X is part of the mechanism. Examples: (1) to establish the role of ROS in aging, manipulating the level of ROS in cells and observe how it affects aging (and other cellular phenotypes). (2) To study the effect of Foxp2 on human evolution, put human FoxP2 in the background of chimp.

- Spatial-temporal changes: this could reflect the causality (with limitations). In practice, tracking the progress of events (e.g. through visualization) is important.

- Analysis of observational data: it may be possible to extract causal relationship from the observational data (structural theory).

Understanding the dysfunction of a complex system: this leads to diseases in human

- Disease heterogeneity: the same symptom may be produced by dysfunction of a number of parts of the system. Specifically, the source of the symptom could be: the subsystem directly responsible for the symptom (the subsystem A), the subsytem that determines the input of the subsystem A (the subsystem B), the subsystem that determines the input to B, and so on.

- Formal analysis: represent the state variables of the system as x_i, and x_i is determined by a differential equation f_i. Then the disruption of some x_i can be analyzed using f_i, say x_j may affect x_i; then we could go on to analyze the equations f_j, and so on.

- Example: autoimmune diseases (AID). AID results from the abnormal number of self-active T/B cells. This number depends on a number of subsystems/processes, and any genetic variations affecting these subsystems may lead to AID:
  - Peripheral tolerance: antigen presentation, T cell regulation, T cell activation and apoptosis

- Tissue response to self-reactive T/B cells: some tissues may secrete cytokines that suppress the self-reactive cells; tissue repair.
- Central tolerance: also antigen presentation, positive and negative selection.
- Environmental trigger: the innate immunity determines whether pathogens are detected and the degree of response.

Principles of understanding a complex phenotype:

- Intermediate changes and complex phenotypes: find the relevant changes at the molecular and cellular levels. Ex. for diabetes, the change of glucose mechanism and insulin.
- Differentiate causality from correlations: changes of intermediates may be correlated to the complex phenotype, but may not represent causal relations.
- Animal models and in vitro systems: critical for testing the mechanisms, especially to establish causality. Ex. animal model of gill withdrawl reflex for studying memory formation and learning.
- Guilt by association: if a genetic change affects a related phenotype, then the change may also affect this phenotype; or similarly, if one gene affects some phenotype, then a related gene may also affect the same phenotype.

Understanding a complex disease - lessons from Alzheimer's Disease (AD): [Insights Give Hope for New Attack on Alzheimer's disease, NY Times, 2010]

- The sympotoms of a disease: for AD, this involves the finding that AD patients usually have plaques (β-amyloids) in their brains from their brain dissections.
- The connection between tissue/organ-level changes and phenotypes: it is not clear if the β-amyloids are causes or consequences of AD, so needs to study the connection and establish causality, e.g.:
  - In vitro, normal β-amyloid is part of the neuron feedback loop, and helps reduce signaling between neurons; and too much β-amyloids will disrupt cell communication and cause cell death.
  - Genetic mutants carrying excessive β-amyloids show AD symptoms.
  - β-amyloids mainly target the default neuron network: e.g. glucose consumption in the default network is unusally low in AD patients.
- The causes (molecular mechanisms) of the disease-related changes:
  - Failure of disposal instead of excessive synthesis leads to accumulation of β-amyloids: test by moninoring the level of newly synthesized β-amyloid in the brain (using a carbon marker).
  - Since β-amyloids are synthesized faster when people are awake (when the default network is most active), sleeping disorder may lead to accumulation of β-amyloids.
  - The effect of β-amyloid on the default network depends on the protein tau.
- Remark: the molecular and tissue/organ level changes are closely related, and if one knowns the possible molecular mechanisms, it may be helpful to establish the causal effect of higher level changes. Formally, to establish $X \to Y$, where X is some physiological change and Y is the trait, if we know $G \to X$, where G may represent genetic mechanism of X, then we could test if $\Delta G \to \Delta Y$. Note the limitation of this approach, as ΔG may change things other than X.

Important perspectives: the fundamental strategy is to test hypothesis through the predictions/implications (evidence) of this hypothesis, thus it is important to consider all available evidence in different types in testing a hypothesis.

- System/Network perspective: the different aspects of a system are related (sampled from different methods): one aspect may be explained by another (e.g. expression pattern explained by the functional requirements, which are defined by metabolic networks, PPI networks, etc.). In general, to analyze data of one aspect, explain it in the framework of our knowledge of the system.
- Guilt-by-association: if some object has some property, then a related object may have the same or related property. Ex. if a protein X is involved in a disease, then a protein interacting with X may be involved in a related disease.
- Design perspective: what will be needed to design a system that implements this function. Important to understand the function of a system, when it is not clearly. Ex. to understand expression profiles: from design perspective, we need to have interacting proteins co-regulated.
- Pattern perspective: if something happens repeatedly, then it is likely to be important. Ex. convergent evolution at independent origins; overrepresentation of some features in a group of genes; network motifs; etc. In general, formuate hypothesis testing as searching for patterns: if some hypothesis is correct, then there would be some relationship across entities. Ex. hypothesis: a set of genes work together; pattern: often expressed together, evolved together.
- Modularity perspective: often groups of genes behave in similar/related ways.
- Statistical perspective: the underling system (which is being reconstructed) should produce behavior consistent with the data. This often takes the form of explaining the variations of data (e.g. variations of expression of different promoters). A formalism: find a best system, S, to maximizes $P(D|S)$, where D is the data about the system; or S miminizes some errors. When applying this perspective, it is important to understand the sources of variations, as many may be noises or irrelevant to the true source of interest. Designing proper controls may solve the problem.
- Evolutionary perspective: evolutionary patterns (conservation and divergence/polymorphism) are shaped by the underlying structure-function of the system.

Control for confounding variables in observational studies:

- Importance: in observational or even some interventional studies, need to recognize the problems caused by confounding variables. Examples:
  - Comparison of two sequence groups (e.g. bound by one TF vs not-bound) for motif finding: the GC content, or repeat are confounding variables.
  - Evolutionary rates of sequence regions: local mutation rate, which depends on GC content, the repeat content, is a major confounding variable.
  - GWAS: the population structure in the case and control groups is a confounding variable.
- Strategy: the most important thing is to recognize possible confounding variables, then the possible strategies are:
  - Comparison of groups: select only groups where the values of confounding variables are similar. This may not always be feasible as the groups chosen in this way may be too small.
  - Regression: a more general way of performing group comparison. However, not always applicable, especially with linear model.
  - Hierarchical model: explicitly encode the group structure. Ex. in motif finding case, GC content is confouding, thus define response variable (e.g. binding intensity) as a linear model of motif content, and assume the coefficient is a function of GC content.

Testing quantiative hypothesis:

- Problem: in some cases, we may not have a precise hypothesis, instead, we have questions such as, how Z is influenced by X and Y. If quantitative/analytic theory is impossible, then we would rely on experimentation, including simulation.
- Formualting the hypothesis: to make the hypothesis testable, we focus on the trend of the quantitative relationship, including:
  - Influence: which factors have influence, and the relative importance among different factors
  - Non-linearity: monotonicity or peak, phase transition, etc.
  - Long-term behavior: temporal or at the extreme values of factors. Saturation/equilibrium, fluctuation/periodic behavior, etc.
  - Synergistic effects of multiple factors

Systems biology of single cells: high-throughput methods to take snap-shots of cells [Synder & Gallagher, FEBS Letters, 2009]

- Gene expression profiling.
- TF-DNA interactions by ChIP-chip and ChIP-seq.
- Protein activitity profiling.
- Protein-protein interactions.
- Protein localization.
- Protein modifications: phosphorylation, acetylation, etc.
- Genetic mutant profiling.
- Double mutants/E-MAP.

Characterizing function of genes, motifs, or other gene-related features (e.g. the presence of a sequence feature or chromatin modification in the promoter of a gene):

- Functions of associated genes (through GO-enrichment or similar analysis): associated genes can be defined via different ways - interacting genes, regulated genes (if a TF), co-expressed genes, etc.
- Association with phenotypes: RNAi experiments, differential expression (or correlation) under different phenotypes, variation of expression and phenotypes in a population, etc.

Gene groups: many hypothesis can be formulated/tested in terms of gene groups (and the sub-processes/pathways that they represent); and the activity of gene groups can be signatures of a (complex) process. Advantages:

- Biological interpretation: easier than individual genes.
- Statistical support: if multiple genes have the same property, stronger evidence.
- Characterization of gene function: genes in the same group are associated, thus the function of one member can be inferred.

Network analysis: provide a framework to interpret the functional (including evolutionary) data.

- Networks could encode the units of systems and how they relate to each other.
  - Connections among biological molecules, or pathways - paths;
  - A functional unit (which may contain different types of molecules, e.g. kinases, phosphatases, TFs) or a set of related pathways - components / subnetworks;

  - Cross-talk/linking among pathways - component connections;
  - Information/material flow between and within pathways - network propagations/flows.
  - Degree of nodes: may be related to the importance of the gene. Ex. the degree of proteins in the networks may be correlated to the evolutionary rates of the proteins.

- A major advantage of using networks is that different genes could be connected through intermediate nodes, thus revealing connections not obvious in the original data. To apply network analysis, formulate the hypothesis to be tested in terms of paths, components, flow, etc. and translate to a network problem. Specific network analysis methods:
  - The mechanism how one gene could influence (say, expression) of another gene: finding paths/flows.
  - Testing if a set of genes form a pathway or functionally related: detecting tightly linked components.
  - The basic functional units of networks (e.g. feedforward loops, kinase-kinase-TF-target pathway): network motifs.
  - Relationship at the level of gene groups/modules: e.g. groups all genes with the same GO, and see how GO-represented processes are related to each other (two GO terms are linked if their members are highly linked).

- Integrative analysis: link genotype (perturbations) to phenotype through molecular networks. [Markowetz, PLCB, 2010; Przytycka & Slonim, BriefBioinfo, 2010].
  - Search for molecular processes (underlying the phenotypes): sub-networks enriched with genetic hits (or simply enrichment analysis); sub-network where genes tend to have physical or functional interactions. Could be used to annotate gene functions.
  - Discover connections among molecular processes/subnetworks: coordination, or cross-talk among pathways.
  - Predict regulator-target maps: e.g. reconstruct TRNs from expression data.
  - Map pathways linking related genes: e.g. perturbation of one gene changes expression of another, the two genes can be connected in the physical network of interactions [Yeang, JCB, 2004], [SPINE], [Tu & Sun, Bioinfo, 2006], [ResponseNet], [PCST].
  - Infer the hidden structure in molecular networks: e.g. perturbation of signaling molecules lead to change of the transcriptional profiles, which can be used to infer the connections of signaling molecules [Nested effect model].

- Mapping molecular networks:
  - PPI: from Y2H, and other proteomic methods
  - Genetic interaction: the phenotype of the double mutants - synthetic lethality, synergistic relationship, synthetic rescue, etc.
  - Co-expression across different conditions.
  - Same relationship with other genes: e.g. sharing the same target enhancers (for TFs).

Classical experiments:

- Evolution: mutation exists before the change of conditions.
- Mutation: can be induced by external conditions.
- Virus as pathogen of some infectious diseases.
- Short-term memory and long-term memory
- Proto-oncogene theory of cancer
- Identification of oncogene (ras) and tumor suppressor gene

1.1.3 Understanding Biological Systems

How to understand the design of a biological system?

- Biological design: some problems that biological systems are designed to solve can be viewed as design or algorithmic problems. Ex. neuron network: communication among nodes; gene regulatory network: response and memory of stimuli; cell differentiation: switching of stable steady states; etc.
- The functional requirements of the system: we could call it the functional model, and it may not be obvious. Ex. for the TRN controlling the global transcriptional profiles. Formally, this is the question of how the fitness depends on the behavior (phenotype) of the system. Ways of specifying the functional requiments:
  - Understanding of how the system works: and relate this to the fitness of the system. Ex. stress signal should lead to expression of protective proteins.
  - Experimental observations: of the product/behavior of the system being studied. The actual cells tend to behave in some optimal/functional way. Ex. the actual expression patterns of genes (the end product of the TRN).
  - Evolutionary conservation: the behavior of the system should be conserved if it satisfies the functional requiments. Ex. the conservated gene expression patterns are those that should be implemented by the TRN.
  - Information perspective: many systems (gene regulatory networks, neural systems) are information-processing systems, so the analysis of the performance in terms of information transmission and processing may provide important insights. This perspective leads to the problems such as: how noises in the environments are overcome.
- Understanding the structure-function of the systems:
  - Cost-benefit analysis of the system, and optimization of performance (by comparing different designs).
  - Features crucial for the function: e.g. importance of negative feedbacks for removing noises.
  - Evolutionary comparison of system designs: conservation, adaptation or convergent evolution.
  - Recognize design problems: these are problems that may not be obvious at first sight, e.g. how specificity of a process is achieved, stability - how a process can be maintained in the presence of constant protein turnover, etc.

How do we design a biological system?

- Maintaining homeostasis: e.g. we need to maintain level of ATPs. When ATP is insufficient, we need to generate more ATPs. However, lack of ATP may be caused by insufficient amount of metabolic enzymes, or of glucose transporters. How do cells know? In general, cells should implement local, negative feedback s.t. the product of a reaction (or several reactions) regulates the production of related genes.

Design considerations: the design of a biological system at a cellular level should meet requirements in multiple aspects:

- Functional specificity and modularity: a component of a system needs to carry out a specific function, and avoid interference with other components.
- Communication: different parts of a system need to communicate with each other. Distinguishing signal and noise, whether the communication can be achieved in a specific way (only certain targets will see the signal), and the information carried in the signals (so that the receiver could take the appropriate actions) are important. Ex. false activation of unnecessary proteins in a transcriptional response may be both costly and dangerous (these proteins may have other effects, especially if they are regulatory).

- Control: the function of a system needs to be controlled - the function could change with many different conditions/environments. Ex. a cell should turn on/off genes based on a combination of inputs/variables.

- Stability: the system needs to carry out its function under unstable environments. Thus need homeostatis of internal environment, and certain robustness of the system.

- Memory and prediction: a biological system needs to make predictions of the future, thus need some form of memory and capability of prediction. This is important for adaptations. Ex. metabolic adaptation to long term starvation (make strategic changes of the system and have other consequences, e.g. lifespan extension).

- Cost and efficiency: the cost of carrying out a function may be important, and the design of a system may particularly improve the efficiency, e.g. allowing recycling of wastes. Ex. of protein production or energy use in general, of dealing with toxic wastes, etc.

- Competing needs and trade-offs: two different needs may compete for the same system (or part of it). Ex. two different nutrients, but only one transporter, need to choose among the two.

Trade-offs among different requirements:

- Trade-off between performance (growth rate for unicellular organisms) and robustness to perturbations, alternatively, trade-off between performance under different conditions. A quantitative framework of analyzing optimal trade-off: similar to the efficient frontier in the portfolio theory, representing the optimal yield-risk trade-off [Kitano, MSB, 2010].

- Trade-off between reliability and sensitivity of responses to signals: similar to precision-recall tradeoff.

- Trade-off between competing needs of the same resources: the allocation strategy.

Principles of biological design: features/tricks used by cells/organisms to meet the design requirements:

- Molecules as computing devices: a single protein or CRE coudl serve as a computing device that integrates multiple signals. For proteins, this is implemented via allosteric regulation, including chemical modifications, of multiple sites; for CREs, via interactions among TFs and arrangement of TFBSs.

- Multi-level control: the behavior of a system needs to adapt to different conditions, and often requires control at multiple levels. Ex. short-term conditions and long-term conditions often require signaling and transcriptional control, respectively (e.g. memory, starvation).

- Linking different processes/modules/genes: to coordinate two activities (e.g. transcriptional responses of two modules), could use (1) shared regulator(s), or (2) a regulator for one module, A, (or some molecules in A that are indicative of the state of A) that could influence module B, through cross-talking with the regulators of the module B (see below).

- Cross-talk of signaling/regulatory pathways: consider a situation where a signal leads to large cellular response. This can be achieved via: the signal turns on/off the direct response pathway, as well as the other pathways (through cross-talk among regulators of different pathways), and the effectors of all the affected pathways implement the total response. Ex. response to nutrient or stress of free-living cells: nutrient $\rightarrow$ PKA and TOR pathways (which turn on the growth program), and PKA/TOR pathways turn off stress response through regulating Msn2/4.

- Reuse/multi-functionality of components: this is important for economic design, e.g. the olfactory system uses compressed sensing to distinguish roughly 10,000 different odors. Instead of assigning each odor a unique receptor, the system uses a small collection of combinatorial testing sensors, as each smell consists of a limited number of basic odors.

- Negative feedbacks: enhance stability of a system, when the input or intermediates fluctuates. And shut down a response when it is no longer necessary.
- Positive feedbacks: faster dynamics of response, binary switches. A related mechanism is cooperativity among multiple molecules or subunits.
- Signal amplification: the signal specificity is limited in a biological environment, e.g. many proteins may bind the same compound, or the stochastic variation of a reguatlor is relatively high. To improve signal-noise ratio, amplify the signals before taking actions through signaling pathways.
- Learning, memory and anticipation: if a system can learn from the past, and better predict/anticipate the future, it would be advantageous. E.g. if an event A is always followed by event B, then an organism may learn to link the response module of A and that of B.
- Cooperation/interaction with other organisms (of the same or different species).

Picture of a unicellular organism: inter-related and coordinated functional modules that direct cell survival, growth and proliferation. Analogy: a city that maintains itself and produces.

- Functional units: energy production (power plant) and delievery (power cables), metabolism (manifacturing factories), cytoskeleton and cargo transport (roads and vehicles), protein folding and turnover (waste management/reuse), etc.
- Control units: sensors and signaling networks (communication system), transcriptional networks (switches of functional units).
- Coordination within units: many mechanisms may be involved, such as specific recognition between proteins or small molecules, protein complexes (assembly lines), signal integration, etc.

Understanding celluar responses to external signals:

- High-level behavior of the cells: the large change of internal state of cells, the growth/differentiation decision, the specialization function of cells, etc. Examples:
  - Glucose sensing of yeast: cells adapt to the use of glucose as carbon source and initiate growth.
  - Stress response of yeast: cells stop growth and cell cycle, prepare for the stress such as protein misfolding.
  - Neurons responding to electric signals: make the decision of whether fire action potential or not.
- Regulatory, especially transcriptional, programs as cellular responses: the behavior change of cells is accomplished through regulatory programs, in particular, the change of gene expression level.
  - Global response: the transcriptional changes basically serve the cellular behavior changes. It is important to recognize that not just the part directly related to the celluar behavior, but also many other parts of the cell, need to make certain changes to accommodate.
  - Example: transcriptional program of glucose sensing involve: (1) increase glucose utilization: transporters and glycolysis; (2) decrease utilization of other carbon sources; (3) decrease alternative use of glucose: respiration; (4) change of nitrogen metabolism: accommodate growth, e.g. RP gene expression; (5) change of stress response genes: may be less necessary.
  - Anticipatory response: sometimes, the signal may suggest more environmental information, thus part of transcriptional program is anticipatory.
  - Example: stress response: some genes are up-regulated to protect cells against future stress (cross-protection).
- Molecular mechanisms that implement cellular/transcriptional responses: the signaling pathways/networks that implement the responses

  - TFs are often the effectors of the signaling pathways: posttranslational modifications, nuclear translocations, etc.
  - Combinatorial interactions of TFs: typically require multiple TFs, however, in many cases, the change of a single TF may trigger the response if other TFs have alreay been pre-occupied to the promoters/enhancers.
  - Signaling cross-talk: necessary for a signal to influence the effectors of another pathway, e.g. glucose signal to influence enzymes involved in nitrogen metabolism. The equivalent way of understanding cross-talk is: the decision to turn on/off a gene (or any other response) depends on a combination of multiple signals, thus requiring cross-talk. Note that the whole regulatory system of a cell (signaling network and TFs) is designed for handling many different types of signals, not just the being studied.

- Reference: [Roberts & Friend, Science, 2000].

Example systems:

- Autophagy system:
  - Logic of the system: when nutrient is abundant, more efficient to express proteins such as transporters, and use the energy to grow and divide. During long-term starvation, need recycling of cellular components, thus induce autophagy system.
  - Transcriptional vs. translational control: when starvation is long, the autophagy and lysosome system is consumed, thus need replenishment, i.e. transcriptional induction.
  - Selectivity of the system: only certain cargos will be recycled, thus need specific mechanism for cargo recognition. Also during starvation, the energy efficiency of the targets may be another determinant of selectivity.
  - Coordination of lysosome biogenesis and autophagy: accomplished through TFEB.
- The olfactory system uses compressed sensing to distinguish roughly 10,000 different odors.
- Some cells could have multiple distinct stable phenotypes from the same genotype: e.g. C albicans white-opaque switching.

1.2 Metabolism

1. Major classes of molecules [MBOC, 5th Ed, 2.1; Vander, Chapter 2]

   Terminology:

   - Hydroxyl group: -OH
   - Carbonyl group: -CO-
   - Aldehyde group: -CO-H
   - Carboxyl group: -CO-OH
   - Ammonia: NH_3 or NH_4^+
   - Nitrate: NO_3^- or NO_2^-
   - Amine (or amino) group: -NH2, -NH-, -N-(containing a basic nitrogen atom with a lone pair), derived from ammonia, NH_3.
   - Amide group: -CO-N-

   Lipids: a loosely defined collection of biological molecules that are insoluble in water, while being soluble in fat and organic solvents such as benzene. Typically contain either long hydrocarbon chains, as in the fatty acids and isoprenes, or multiple linked rings, as in the steroids.

- Fatty acids: a long hydrocarbon chain, which is hydrophobic and not very reactive chemically, and a carboxyl group. Saturated or unsaturated (mono if a single double bond; poly if multiple double bonds). Derivatives of fatty acids may play important physiological functions.
- Triacylglycerol (also called triglyceride): three fatty acid chains joined to a glycerol molecule. The degree of saturation of the fatty acid chains may vary.
- Phospholipids: the glycerol is joined to two fatty acid chains and a phosphate group (plus a small molecule containing nitrogen), the phosphate group is hydrophilic, mkaing the molecular amphipathic.
- Steriods: four interconnected rings of carbon atoms form the skeleton of all steroids. Examples of steroids are cholesterol, cortisol from the adrenal glands, and female (estrogen) and male (testosterone) sex hormones secreted by the gonads.

2. Physiochemical foundations [Lehninger, Principles of Biochemistry, 4th Ed, 1.3; MBOC, 5th Ed, 2.2]

   Chemical reactions:

   - $\Delta G = \Delta H - T\Delta S$. ΔH is contributed from the change of chemical bonds and weak interactions (van der Walls, etc.); ΔS contributed from the loss of randomness (e.g. more small molecules are created). Spontaneous if $\Delta G < 0$.
   - Equilibrium: the previous analysis is based on completion of reaction (e.g. 1 mol substrates becomes 1 mol products). For ongoing reactions, ΔG is related to ΔG^0, the standard free energy, and the concentrations of substrates and products. At equilibrium, $\Delta G = 0$, this gives the relation: $\Delta G^0 = -RT \ln K_{eq}$.

   Enzymes: a reaction is often very slow even if its ΔG is very negative. The activation energy ΔG^a can be reduced by enzymes through mechanisms, e.g. stablize the transient states (complementary structure), provide an interface s.t. the productive collisions between reaction molecules are more common.

   Oxidation-reduction reactions:

   - One biologically important source of ΔG is e^- transfer from donor to acceptors. e^- transfers along the electrochemical potential.
   - Electron transfer may involve partial charges (in polar chemical bond): e.g. $CH_4 \longrightarrow CO_2$, e^- is shifted from C atom to O atoms in CO_2.
   - In aqueous solution: often hydrogenation (reduction), $A + e^- + H^+ \longrightarrow AH$ and the reverse reaction is dehydrogenation (oxidation). Thus if the number of C-H bonds are increased: reduction; decreased: oxidation.
   - Strong e^- donors (reducing agents): e.g. H_2, $R-CH_3$, NADH, metals; Strong e^- acceptors (oxidizing agents): e.g. O_2.

   How cells overcome unfavorable reactions:

   - Couple an unfavorable reaction with one with large negative ΔG. The mechanica analogy where a heavier object is used to life another object. The most common strategy is to use ATP $\longrightarrow$ ADP + Pi.
   - Energy carriers: often coupled to energetically unfavorable reactions, then the release of energy from activated carrier (e.g. ATP hydrolysis) can be coupled to energetically unfavorable reaction.
   - Common energy carriers: ATP, NADH and NADPH.

   Energy carriers:

- NADH and NADPH: similar to ATP. NAD^+ can receive an electron, and becomes an activated carrier, NADH. The NADH molecules is not stable, thus it will transfer electron to another molecule (reduce that molecule). E.g. for NADPH:
  $NADP^+ + R_1R_2{-}CH{-}OH \longrightarrow NADPH + H^+ + R_1R_2{-}C{=}O$
  $NADPH + H^+ + R_1{-}C{=}C{-}R_2 \longrightarrow NADP^+ + R_1{-}CH{-}CH{-}R_2$.
  Both NADH and NADPH work similarly, but different structure, thus work with different targets: NADH on catabolism (energy generation), thus NAD^+ is kept low; and NADPH on reductive biosynthesis, thus kept high.
- Acetyl-CoA: acetyl-S high energy linkage. The energy can be used to drive biosynthesis, often adding two carbon atoms.

3. Metabolism overview [Britannica, metabolism]

   Catabolism: provide ATP and raw materials from food and current bio-molecules. Generally three phases ([catabolism.gif]):

   - Digestion of food
   - Incomplete oxidization and fermentation (for microorganisms)
   - Complete oxidization: combustion of food materials.

   Central metabolism and for other molecules (such as AAs), break down to other intermediates in the central metabolism.

   Anabolism: construct bio-molecules from raw materials and ATP. To synthesize, say an AA, generally use intermediate compounds as precursors, different from the compounds that this AA is broken down into in catabolism.

   Six types of common reactions: [Stryer, 5th Ed, 14.3]

   - Oxidation-reduction: -CH- bond is oxidated to become -CO- or -C=C-. Can often derive energy and generate activated energy carriers, NADH, $FADH_2$.
   - Ligation: combine small molecules to form large ones, e.g. 3-carbon molecule and CO_2 combines to become 4-carbon molecules. Often driven by ATP.
   - Hydrolysis: common reaction to break down large molecules with H_2O.
   - Group transfer reactions: e.g. glucose $+$ATP $\longrightarrow$ glucose$-_6-$Pi $+$ ADP.
   - Isomerization: rearrange particular atoms within the molecule, e.g. citrate $\longrightarrow$ isocitrate in TCA cycle.
   - The addition of functional groups to double bonds or the removal of groups to form double bonds: e.g. in glycolysis, fructose-1,6-BP is broken down to two 3-carbon molecules; and dehydration to form -C=C- double bonds.
   - Condensation and hydrolysis: condensation is usually energetically unfavorable, thus often need ATP.

4. Carbon metabolism [MBOC, 5th Ed, 2.3]

   Glycolysis: Glucose becomes pyruvate (cytosol): (1) energy investment (2 ATPs); (2) breakdown into two small molecules; (3) energy generation: 2 NADH and 2 ATP. The main reactions are: oxidiation, $R{-}CHO \longrightarrow R{-}COO{-}Pi$ and substrate-level phosphoylation, $R{-}COO{-}Pi + ADP \longrightarrow R{-}COO^- +$ ATP.

   Fermentation: needed to regenerate NAD^+. The product is lactate or ethanol and CO_2.

   - pyruvate $\longrightarrow$ acetaldehyde $+ CO_2$ through pyruvate decarboxylase (PDC).
   - acetaldehyde $+$ NADH $\longrightarrow$ ethanol $+ NAD^+$ through alchohol dehydrogenase (ADH).

Pentose phosphate cycle: alternative pathway of glycolysis: glucose-6-P $+NADP^+ \longrightarrow$ pentose$+CO_2+$ $NADPH + H^+$. The results are:

- NADPH, used in reductive biosynthesis reactions within cells;
- Production of ribose-5-phosphate (R5P), used in the synthesis of nucleotides;
- Production of erythrose-4-phosphate (E4P), used in the synthesis of aromatic amino acids.

Energy storage: (1) long-term: fat in specialize cells (adipocytes); (2) short-term: glycogen in cells. Glycogen breakdown: glucose-1-Pi, to glucose-6-Pi, then glycolysis.

Oxidization of pyruvate: decarboxylation and forms acetyl-coenzyme A through pyruvate dehydrogenase (PDH).

$CH_3{-}CO{-}COO^- + CoA{-}SH + NAD^+ \longrightarrow CH_3{-}CO{-}S{-}CoA + CO_2 + NADH + H^+$

TCA cycle:

- Oxidation of acetyl group (no O_2 needed) and generates CO_2 and NADH, also FADH2 and GTP. NADH then combine with O_2 in electron-transport chain to regenerate NAD^+ and energy in oxidative phosphorylation.
- At the first step, acetyl-CoA reacts with oxaloacetate to generate 6-carbon citrate. At the end of cycle, oxaloacetate is regenerated.
  $acetyl{-}CoA + 3\,H_2O \longrightarrow 2\,CO_2 + CoA{-}SH + 4\,[2\,H]$

and are used as precursors for biosynthesis. Figure 2-84.

Oxidative phosphorylation: NADH, FADH2 along electron transport chain to pump H^+ across Mt. memberane. O_2 is used in the last step to form H_2O (with H^+). The H^+ gradient can then drive ATP synthesis.

Glyoxylate cycle: [Lehninger, 4th Ed, 16-4]

- In plants or in some microorganisms: convert fatty acid or acetate (derived from fatty acid) to glucose. Fatty acid is first converted to acetyl-CoA through β-oxidation, then acetyl-CoA is used to generate oxaloacetate, which can be converted to phosphoenolpyruvate.
- The net reaction: $2\,acetyl{-}CoA + NAD^+ + 2\,H_2O \longrightarrow succinate + 2\,CoA + NADH + H^+$.
- The cycle: [glyoxylate-cycle.pdf] the first two steps are identical to part of TCA cycle:
  $acetyl{-}CoA \longrightarrow citrate \longrightarrow iso{-}citrate$
  Then iso-citrate is cleaved to generate succinate and glyoxylate. Glyoxylate then completes the cycle through malate and oxaloacetate, while succinate can enter TCA cycle again to generate malate and oxaloacetate. Once oxaloacetate is available, it can generate phosphoenolpyruvate for gluconeogenesis.

5. Nitrogen metabolism [MBOC, 5th Ed, 2.3; Lehninger, 4th Ed, 22.1-22.2]

   Overview of nitrogen metabolism:

   - Animals obtain essential AA from food, and other AAs from raw materials, including intermediates of the citric acid cycle.
   - All of the nitrogens in the purine and pyrimidine bases (as well as some of the carbons) are derived from the plentiful amino acids glutamine, aspartic acid, and glycine, whereas the ribose and deoxyribose sugars are derived from glucose. Certain amino acids or parts of amino acids are incorporated into the structure of purines and pyrimidines, and in one case part of a purine ring is incorporated into an amino acid (histidine).
   - Because each amino acid and each nucleotide is required in relatively small amounts, the metabolic flow through most of these pathways is not nearly as great as the biosynthetic flow leading to carbohydrate or fat in animal tissues.

- Because soluble, biologically useful nitrogen compounds are generally scarce in natural environments, most organisms maintain strict economy in their use of ammonia, amino acids, and nucleotides. Indeed, free amino acids, purines, and pyrimidines formed during metabolic turnover of proteins and nucleic acids are often salvaged and reused: amino acids not used in biosynthesis can be oxidized to generate metabolic energy (pass to mitochondrial, converted to acetyl-CoA, and TCA cycle). Their nitrogen atoms are shuttled through various forms and eventually appear as urea.
- Control: independent of catabolism, physical separation (in cytoplasm), and different pacemaker enzymes.

Nitrogen cycle and incorporation into AA biosynthesis:

- Nitrogen fixation and nitrification: N_2 to NH_3 or NH_4^+. Although ammonia can be used by most living organisms, soil bacteria that derive their energy by oxidizing ammonia to nitrite (NO_2^-) and ultimately nitrate (NO_3^-) are so abundant and active that nearly all ammonia reaching the soil is oxidized to nitrate.
- Plants and many bacteria can take up and readily reduce nitrate and nitrite through the action of nitrate and nitrite reductases. Animals then use plants as a source of amino acids, both nonessential and essential, to build their proteins.
- Glutamate is the source of amino groups for most other amino acids, through transamination reactions. The amide nitrogen of glutamine is a source of amino groups in a wide range of biosynthetic processes. In most types of cells, and in extracellular fluids in higher organisms, one or both of these amino acids are present at higher concentrations, sometimes an order of magnitude or more higher than other amino acids.
- Nitrogen assimilation: (1) Glutamine synthetase: Eq. 22-1, incorporate NH_4^+ into glutamate, and generate glutamine. (2) Glutamate synthetase: Eq. 22-2, NH_4^+ into glutamine and creates two glutamate. Thus effectively one NH_4^+ is used to create glutamate. (3) Glutamate can also be formed by a reaction catalyzed by glumate dehydrogenase: $\alpha-$Ketoglutarate + NH_4 + NADPH $\longrightarrow$ L$-$glutamate + NADP + H_2O
- Regulation of the activity of glutamine synthetase: in E. coli, alanine, glycine, and at least six end products of glutamine metabolism are allosteric inhibitors of the enzyme. Superimposed on the allosteric regulation is inhibition by adenylylation of (addition of AMP to). The net result of this elaborate system of controls is a decrease in glutamine synthetase activity when glutamine levels are high, and an increase in activity when glutamine levels are low and alpha-ketoglutarate and ATP (substrates for the synthetase reaction) are available.

Classes of reactions in Nitrogen metabolism:

- Pyridoxal phosphate is required for transamination reactions involving glutamate (glutamate provides an amine group, and becomes alpha-ketoglutarate) and for other amino acid transformations.
- One-carbon transfers require S-adenosylmethionine and tetrahydrofolate.
- More than a dozen known biosynthetic reactions use glutamine as the major physiological source of amino groups. As a class, the enzymes catalyzing these reactions are called glutamine amidotransferases: amino group transfer involving the amide nitrogen of glutamine. After reaction, glutamine becomes glutamate.

AA biosynthesis: [AA-biosynthesis.gif]. All amino acids are derived from intermediates in glycolysis, the citric acid cycle, or the pentose phosphate pathway. Figure 22-9. The main features:

- Ammonia is incorporated into the intermediates of pathways mainly via the glutamate dehydrogenase: Glu $\longrightarrow \alpha-$ketoglutarate.

- A group of several amino acids may be synthesized from one amino acid, which acts as a "parent" of an amino-acid "family".

Specific families of AA biosynthesis:

- Glu family: (1) α-ketoglutarate (TCA) $\longrightarrow$ Glu; (2) Glu $\longrightarrow$ Gln: glutamine synthetase; (3) Glu $\longrightarrow$ Pro: reduction of -COO-, then cyclization; (4) Glu $\longrightarrow$ Arg: reduction of -COO-, then urea cycle to add the functional group of ammonia, also need Glu to provide one -NH2 group.
- Ser family: (1) 3-phosphoglycerate (glycolysis) $\longrightarrow$ Ser: obtain -NH2 from Glu, then remove -Pi group (hydrolysis); (2) Ser $\longrightarrow$ Gly: remove -CH2-OH; (3) Ser $\longrightarrow$ Cys: effectively -OH -¿ -SH through an intermdiate.
- Asp family: (1) Oxaloacetate (TCA) $\longrightarrow$ Asp: transamination; (2) Asp $\longrightarrow$ Asn: amidation, -COOH $\longrightarrow$ -CO-NH2; (3) Asp $\longrightarrow$ ThrMetLys, through multiple branch points.
- Pyruvate family: (1) Pyruvate (glycolysis) $\longrightarrow$ Aln: transamination; (2) Pyruvate $\longrightarrow$ Iso (using Thr), $\longrightarrow$ ValLeu.
- Aromatic AA family: Erythrose 4-phosphate (pentose phosphate) + Phosphoenolpyruvate (glycolysis) $\longrightarrow$ chorismate (branch point), Branch 1: $\longrightarrow$ Trp; Branch 2: TyrandPhe.
- Histine: Ribose-5-phosphate (pentose phosphate) $\longrightarrow$ His

Allosteric regulation of AA biosynthesis: Figure 22-22, also [Asp-family.gif]. Several common patterns:

- End-product inhibition: the end product inhibits the first enzyme in the branch leading to it.
- Concerted inhibition: an enzyme is inhibited by multiple molecules, e.g. glutamine synthetase.
- Enzyme multiplicity: a step is catalyzed by multiple isozymes, each independently controlled by different modulators. This enzyme multiplicity prevents one biosynthetic end product from shutting down key steps in a pathway when other products of the same pathway are required. E.g. the enzyme of the first step of the aromatic AA family has three isozymes: one is allosterically inhibited (feedback inhibition) by phenylalanine, another by tyrosine, and the third by tryptophan.
- Sequential feedback inhibition: one AA inhibited multiple points of its synthesis. E.g. threonine inhibits its own formation at three points: from homoserine, from aspartate-β-semialdehyde, and from aspartate.

Question: if different products inhibit different isozymes, then suppose one product is enough, it will shut down its corresponding isozyme, but other isozymes are still functional?

6. Lipid metabolism

Lipid biosynthesis: [Wiki, cholesterol; Wiki, triacylglycerol]

- Acetyl-CoA: the main precursor of fatty acid and cholesterol biosynthesis. Generated from pyruvate, a produce of glycolysis.
- Cholesterol biosynthesis: from acetyl-CoA, with the rate-limiting step catalyzed by the enzyme HMGCR.
- Triacylglycerol biosynthesis: (1) glycerol from an intermediate of glycolysis (DHAP); (2) fatty acid: from acetyl-CoA.

Fatty acid oxidation: [MBOC, Chapter 2] generate acetyl-CoA, which can be used by TCA cycle. Usually take place at mitochondria, could also be in peroxisome.

- Fatty acid activation: fatty acid becomes fatty acyl-CoA (hydrocarbon chain attached to acetyl-CoA).
- β-oxidation: at each step of the cycle, two carbon atoms are cleaved and form acetyl-CoA.

7. Regulation of metabolism [Britannica, metabolism; Stryer, 5th Ed, 14.3; Lehninger, 4th Ed, 15.2-15.5]

   Requirements of regulation: generally need to: (1) maintain homeostatis in time of perturbation; (2) change metabolic activities in response to signals. To achieve these two goals, need to design the regulatory system so that (maximize efficiency and avoid waste):

   - Partition metabolites between alternative pathways.
   - Draw on the fuel best suited for the immediate needs of the organism (glucose, fatty acid, etc.)
   - Avoid waste by preventing simultaneous operation of pathways in opposite directions.
   - Shut down biosynthetic pathways when their products accumulate.

   Example: the requirements of regulation of glycolysis/glycogenesis in liver and muscle are different in response to glucose change.

   - Muscle: primarily consumption of glucose, thus as glucose increases, need to increase the use/flux of glucose via glycolysis.
   - Liver: primarily maintain homeostatis of glucose in the blood, thus as glucose increase, need to increase the storage of glucose; and as glucose level falls, need to reduce glcolysis to avoid competition with other tissues for glucose.

   Integration/control:

   - Coarse control (transcriptional regulation): determined by pacemaker enzymes, which are induced from stimuli (e.g. nutrients) for catabolic reactions, and regulated by pathway products for anabolic reactions.
   - Fine control: catabolism by ATP, ADP and AMP; anabolism also by ATP, ADP and AMP (stress signals), but mainly by end product inhibition.
   - Compartmentalization: segregates opposed reactions (i.e. biosynthetic and degradative pathways - almost always distinct, further enhanced by compartmentalization).

   Fine control: example in [Asp-family.gif]

   - End product inhibition: linear pathways, the end product inhibits the first enzyme in that pathway or branch.
   - Cross-talk among pathways: e.g. aspartate carbamoyltransferase, the first enzyme of pyrimidine biosynthesis, is inhibited by product UTP. However, the inhibition is relieved by high level of ATP (signals that DNA synthesis is needed).
   - Multivalent repression (control): an enzyme is controlled by a combination of signals. Ex. One form of aspartokinase, the first enzyme in the biosynthesis of aspartate family AAs, is inhibited by threonine, but only when isoleucine is present (because threonine is both a product of this enzyme, and also a substrate for isoleucine [Asp-family.gif].
   - Positive modulation: if a reaction proceeds only when certain substrates are available, then the enzyme leading to this reaction may be positively stimulated by this substrate. Ex. acetyl-coA acts as a positive allosteric effector of pyruvate carboxylation (which eventually need acetyl-coA).

   Principles of metabolic regulation:

   - Substrate-limited reactions vs enzyme-limited reactions: some reactions are close to equilibrium, thus its rates rise and fall with the change of substrates; other reactions are far from equilibrium, and the enzymes are the main point of control. In fact, cells need to ensure many reactions to be far from equilibrium, e.g. ATP $\longrightarrow$ ADP + Pi, so that its $\Delta G < 0$ to drive other reactions.

- Energy state: ATP, ADP and AMP; and other compounds indicating level of ATP (e.g. reduction state of respiratory chain molecules) may regulate the rates of metabolism: (1) ATP-generating (catabolic) pathways are inhibited by an energy charge (a measure of [ATP] vs [ADP] and [AMP], ranged from 0 to 1), whereas ATP-utilizing (anabolic) pathways are stimulated by a high-energy charge. (2) The energy charge, like the pH of a cell, is buffered. The energy charge of most cells ranges from 0.80 to 0.95 (at 0.90, the catabolic and anabolic pathways have equal rates).
- Isozymes have different regulatory and kinetic properties: adapted to different usages of isozymes. E.g. Hexokinase isozymes in muscle and in liver have different properites, one of them (liver) is not regulated by end-product inhibition.
- Metabolic flux of a pathway: the effect of changing one enzyme (concentration or activity) on the flux through the pathway may not be obvious. E.g. in glycolysis, three enzymes are important to set the rate - hexokinase, PFK-1 and pyruvate kinase, but their effects are very different (Figure 15-33). Adding some enzyme may have little effect on the flux. There may not be a single rate-limiting enzyme, in fact, often need to change levels of every enzyme in the pathways.

1.3 Control of Transcription

Genetic switches in Bacteria [Chap. 7, MBOC V4]

- Activation: activator binding provides contact surface for RNAP or stablize the transition state of RNAP. Example: CAP.
- Repression: operator sites often overlap with promoter, thus repressor binding blocks RNAP access to DNA. Example: Trp repressor.
- Activation and repression of a single gene: e.g. lac operon. On only if glucose level is low (CAP is ON) and lactose level is high (lac repressor is OFF).
- A protein can be both activator and repressor, depending on the position of its BS.

Genetic switches in Eukaryotes [Chap. 7, MBOC V4]

- Basic differences with bacteria:
  - Looping: allow more distant interaction. DNA as tethers: increase probability of interaction between two bound proteins.
  - TC: RNA Pol + General transcription factors (GTFs)
  - Chromatin structure.
- Activation: transcription activator (TA) has both DNA-binding domain and activation domain.
  - Interaction with holoenzyme complex or other GTFs: help assembly of TC.
  - Change chromatin strcuture: (i) chromatin remodeling; (ii) histone acetylation: by recruiting histone acetyltransferase (HAT)
  - Help other activators (synergy): multiplicative effect of different TAs, without needing PPI between 2 TAs.
- Repression:
  - Competitive binding with TA
  - Mask the activation domain of TA
  - Direct interaction with TC (GTFs): make the surface not available to TA, etc.

  - Change chromatin structure: (i) repressive chromatin remodeling; (ii) histone deacetylation.

- Effect of coactivators and corepressors: not bind DNA, but form complex with TA and TR.
- Design of eukaryote CRS:
  - Often intersperse with large control region, e.g. eve CRS within 50k bps.
  - Two-level control: LCR of a gene cluster (chromatin decondensation) and cis-regulatory modules, e.g. β globin.
  - Insulators: (i) buffer genes from repressor effect of heterochromatin; (ii) block activation of enhancers to nearby genes.

Activation model: cooperative DNA binding vs multiple contact with BTM (or multiple steps of transcription), experimentally test:

- DNA binding: if DNA binding is increased from a neighboring site in the absence of BTM (multiple contact model would also predict increased DNA binding, although indirectly).
- Linkage: if sites must be linked, likely due to cooperative DNA binding. However, linkage may also be required under the multiple contact model where the bound activators must create a somewhat continous surface, see [Svaren & Chalkley, JCB, 1993].
- Saturating [TF]: if synergism is observed under saturating [TF], then multiple contact model.
- PPI between interacting TFs: identify PPI domains and test if the synergistic transcriptional activation depends on this domain.

Repression model: activator masking vs chromatin modification, experimentally one can test the following:

- Activator specificity: activator masking model predicts repression of specific activators.
- Activator binding: not reduced in the activator masking model.
- Activator concentration: high activator concentration can overcome the chromatin effect, thus still bind to DNA.
- Repressor concentration: high repressor concentration could relax the need of repressor binding, i.e. even without DNA-binding of repressor, it can still inhibit activator.
- Co-repressor activitiy: whether they physically interact with activator or repressor; or they possess HDAC and other chromatin modification activities.
- Chromatin/nucleosome state of the enhancer: e.g. if other activators bound to the enhancer are repressed (yes, under the chromatin model).

Repressors may employ multiple mechanisms: example, Kr:

- Competitive binding with activator: overlapping sites with Bcd in eve stripe 2
- Quenching by activator masking: activator specificity - repress some activators but not the others [Licht & Hansen, PNAS, 1993; Hanna-Rose & Hanse, MCB, 1997].
- Quenching by chromatin modification: quenching depends on CtBP, which has chromatin modification activity [Nibu & Levine, MCB, 2003]
- Direct repression by interacting with BTM: interaction with TFIIEβ in Schneider cell line [Sauer & Jackle, Nature, 1991; 1995].

- Direct repression by chromatin modification: depends on CtBP, which has chromatin modification activity [Nibu & Levine, MCB, 2003]

Some regulator functions both as an activator and a repressor:

- Concentration dependence: Kr, in monomeric form, activates transcription in Drosophila cell lines at low concentration, but activates at high concentration in dimeric form [Sauer & Jackle, Nature, 1991; 1993].
- Context dependence: e.g. A_1 is an activator, and A_2 is also an activator, but A_2 can also bind with A_1, thus making it less productive. Therefore, in the absence of A_1, A_2 is an activator; but in the presence of it, A_2 will appear as a repressor.

Different logic of gene regulation in eukaryotes and prokaryotes [Struhl, Cell, 1999]

- Prokaryotes:
  - Promoter activities vary with genes: some are strong, some weak. The promoter activities (of naked DNA) both in vitro and in vivo are ON. Thus the ground state is ON.
  - Need repressors to shut down transcription: through occlusion or direct interaction with RNAP.
  - Weak promoters need activators: through direct interaction with RNAP.
- Eukaryotes:
  - Promoters in vitro are active, but silent in vivo. Thus the ground state is OFF.
  - Characteristics of regulation: recruit chromatin modifying activities to control BTM access to DNA; and epigenetic memory - even after the activator or repressor is gone, the chromatin state is still maintained (Analogy: the effect of enzyme-catalzed reaction, after removing the enzyme, the reaction will not automatically reversed).
  - Activation (e.g. HO expression in yeast): follows the steps
    - Activator binding (to chromatin active regions): recruit chromatin modifying activities.
    - Chromatin change could propagate along chromosomes (alternative mechanism of DNA looping).
    - Promoter is changed to active state, then BTM binding.
  - Repression: need special mechanism, e.g. heterochromatin.

Transcriptional Coregulators in Development [Mannervik and Levine, Science, 1999]

- Overview of mechanisms of transcriptional regulation:
  - Principle: each component of the process (generally) has a DNA-binding domain and a communication domain (communicate the signal from enhancer to core promoter). Transcriptional regulation can occur at either level. Furthermore, a transcriptional regulator often recruits coactivator or corepressor to achieve one of the two functions.
  - Transcriptional activation: send the activation signal to the core promoter. Can either decondensate the chromatin at the core promoter to faciliate BTM binding to DNA; or recruit BTM components via protein interactions.
  - Transcriptional repression: disrupt the signal from the activator binding regions to the core promoter. It can target either activator or BTM:
    - Targeting activators: either condensate the enhancer regions to block activator binding to DNA; or interact with the activation domain of the bound activator molecules to mask its communication with BTM; or competitive binding to the same DNA sequence the activator binds.

        * Targeting BTM: either condensate the core promoter to block BTM binding to DNA; or interact with GTFs of the BTM to block its assembly and communication with activators; or competitive binding to the same DNA sequence the BTM binds.

- Coactivators:
  - CBP and p300: essential coactivators of a variety of transcriptional activators. CBP has dual role: (i) histone acetyltransferase activity; and (ii) interaction with subunit of BTM.
  - Relative role of the two mechanisms: double mutant of CBP and Rpd3 (a corepressor that condensates chromatins) shows normal development, suggesting that transcriptional regulation by direct interaction with BTM is sufficient.
- Corepressors:
  - Rpd3: a mammalian corepressor of several repressors. It has histone deacetylase activity.
  - Groucho: corepressor of long-range repression in Drosophila (with Hairy and Dorsal). Either position nucleosomes over the core promoter or directly inhibit BTM.
  - CtBP: corepressor of short-range repression in Drosophila (with Snail, Knirps and Kruppel). It has chromtin modification activity.
- Interaction of TF with multiple coregulators:
  - One TF can interact with different co-regulators and either activate or repressor transcription. Thus regulating the level of co-regulators can regulate the role of this TF, and thus the transcription.
  - Example: Dorsal is an activator if associated with Cut and Dri; but repressor if associated with Groucho.
- Question: how does a co-activator or co-repressor, recruited near the distal enhancer, affect the chromatin state of the core promoter? Or general, how does the chromatin state propagate?

The Enhanceosome and Transcriptional Synergy [Carey, Cell, 1998]; Gene Regulation. A Paradigm for Precision [Struhl, Science, 2001]

- Examples:
  - IFNβ enhancer (about 100bp): activated by viral transfection. Activators include: IRE-1, p60 and p50 subunits of NFκB, ATF/cJun and HMG1 (DNA-bending protein)
  - TCRα enhancer: activators include CRE/ATF, AML-1, Ets-1 and LEF-1 (DNA-bending protein, similar to HMG1).
- Process of IFNβ ativation:
  - Enhancersome assembly: 4 hours after viral inferection.
  - Enhancersome recruits GCN5, histone acetylase (HAC), acetalyating the adjacent nucleosome.
  - Shortly after recruting GCN5, recurits CBP and Pol II complex.
  - 1-2 hours later, Swi/Snf nucleosome remodeling complex is recruited. It disrupts the nucleosomes in the core promoter regions, and allows TFIID to gain acess and starts transcription.
- Model:
  - Enhancersome aseembly: cooperative interactions among activators form the enhancersom, which present a complementary surface for interaction with BTM and co-activators. The cooperative interactions (or the creation of the surface) depend on the arrangement of the sites, as repositioning (change of helical phasing, e.g. 6bp but not 10bp) abolishes the transcriptional synergy.

  - Enhancersome interaction with BTM: only two or three activators are sufficient for synergy, by contacting a limited set of targets (e.g. TFIID, TFIIA, TBP). One co-activator, CBP, can interact with each of the activator in the IFNβ enhancersome.
  - Reciprocity: the enhancersome recruits BTM, but the machinery also reciprocally stablizes the enhancersome.
  - The activators need to bind DNA cooperatively, but multiple interactions with BTM are essential.
  - The role of DNA-bending (architectural) preteins: facilitates cooperative interactions among activators, but can be bypassed if the strength of the interaction is enough.
- Significance: enhancersome is different from modular enhancers (important for diversity and evolutionary flexibility), in that all activators are required to achieve a precise ON/OFF response.

Transcriptional Repression in Eukaryotes: Repressors and Repression Mechanisms [Gaston and Jayaraman, Cell Mol Life Sci, 2003]

- Types of repressors:
  - General repressor: repress a central component of BTM, or sequester the promoter DNA by chromatin modification.
  - Class I gene-specific repressor: sequence-specific DNA binding. Example, Eve, Kr.
  - Class II gene-specific repressor: bind to DNA-binding proteins (co-repressor). Example: Tup1, Groucho, CtBP.
  - Class III gene-specific represso: bind to activators, co-activators or BTM, post-translational modification.
- Repression via inhibiting activator-target interactions:
  - By class II repressors: interaction with the activation domain and prevent it from interacting with its targets. Examples include:
    * Gal4 inhibition by Gal80: the binding site for Gal80 in Gal4 overlaps with the Gal4 activation domain.
    * E2F inhibition by Rb (tumor suppressor): Rb masks the activation domain of E2F.
    * p53 inhibition by MDM2: MDM2 masks the activation domain of p53.
  - By class I repressors: block interactions between a DNA-bound activator and its target, either co-acviator or BTM (quenching). Examples include:
    * AP1 inhibition by YY1: YY1 is mammalian Kr-like TF. YY1 binds to the co-activator CBP and prevents the interaction of AP1 and CBP.
    * Supression of IFNβ expression by IRF-2: IRF-2 may "repel" incoming the co-activator CBP, or destablize the interaction of CBP and activator proteins such as NFκB, IRF-1 at the IFNβ enhancer.

Transcriptional Enhancers: Intelligent Enhanceosomes or Flexible Billboards? [Arnosti and Kulkarni, J of Cell. Biochem., 2005]

- Enhancersome model: assembly of many proteins, high cooperativity, $\rightarrow$ single complex. Examples:
  - Interferon β cis-regulatory element (one of the best studied)
  - dorsal (repressor) + groucho corepressor
  - Immune regulated genes
  - neuroectoderm: 1 dorsal site + 1 or 2 twist sites within 20bp + Dip3 binding site

- Billboard model: independent/additive effects of multiple sub-elements. Examples:
  - eve 2: tolerate different activator placement
  - activator + repressor sites arrangement [Kulkarni & Arnosti, Development, 2003]
- Remark: no clear boundary between the two. The sub-elements within a billboard CRM could still have high cooperativity; and the affiniity, number, arrangement could be important for expression output under billboard model.

Regulation of Transcription: From Lambda to Eukaryotes [Ptashne, TIBS, 2005]

- Conservation of activation mechanism: yeast Gal4 (TF) is expressed in higher organisms and can stimulate the expression of reporter genes bearing Gal4 binding sites.
- Evidence of recruitment: activation domain of Gal4 is deleted, and Gal4 is fused with BTM, strong activation, suggesting that the function of activation domain is recruiting BTM.
- DNA looping: binding between multiple Gal4 sites, favored when the sites on the same side of the DNA helix. Cooperative binding even separated by ¿3,000 bps.
- Transcriptional activation by recruitment: recruited complex modify histones and increase the affinity of other activators and complexes.
- Transcriptional repression by recruitment: repressing machines may "poison" the BTW, and/or recruit histone- and DNA-modifying enzymes that decreases affinity for BTM.

Eukaryotic Transcription Activation: Right on Target [Green, Mol. Cell, 2005]

- Background:
  - PIC: RNA Pol II and GTFs, including TFIIA, TFIIB, TFIID, TFIIE, TFIIF and TFIIH, assemble on the core promoter to form a preinitiation complex (PIC).
  - Mediator complex: evolutionarily conserved, about 25-30 units, a component of PIC.
  - Artificial recruitment experiment: for identifying the target of activators, fuse the putative target with DBD and test if the fusion protein is able to activate transcription.
- Mechanisms of activation:
  - Simulate PIC assembly: direct interaction between activation domain (AD) and component(s) of PIC. Example: (i) yeast Gal4 target: Tra1, a component of SAGA complex; recruits the Meidator complex; (ii) adenovirus E1a target: MED23 (Mediator subunit).
  - Stimulate subsequent steps such as initiation, elongation and reinitiation.
  - Recruiting chromatin modifying activities: (i) ATP-dependent chromatin remodeling complex which non-covalently modify chromatin structure; (ii) histone modifying complexes, which add or remove covalent groups, e.g. acetyl groups, methyl groups, from histone tails.
- Relative importance of PIC and chromatin: modification of chromatin structure may be necessary but not sufficient in yeast:
  - Atificial tethering of chromatin-modifying components activate transcription very poorly.
  - Global perturbation of chromatin structure has a small effect on genome-wide activation of transcription.
  - Well-studied yeast chromatin modifying activities, such as Swi/Snf and HAT Gcn5, are required for expression of only a small percentage of genes.

In higher organisms, the chromatin is more condenses, thus transcriptional activation may be more dependent on chromatin modification.

- Two models of transcriptional synergy:
  - Different activators act at different steps, e.g. PIC assembly, elongation.
  - Multiply bound activators simultaneously interact with different PIC components, synergistically increasing PIC assembly. Evidences: in vitro PPI experiments show that a single activator can interact with multiple components of BTM; likewise, many BTM components can stimulate transcription in artificial recruitment experiments.

1.3.1 Transcriptional Regulation in Prokaryotes and Yeast

E. coli lac operon [lac operon, Wikipedia]

- Model: lac operon promoter has one BS for the protien CAP and another BS (called operator, O) for lacR. CAP binding activates transcription, and lacR binding to O blocks the access of RNAP to DNA. The levels of these two proteins are controlled by:
  - lacR is constitutively expressed, and bind to O when [lactose] is low. But when [lactose] is high, it can inactive DNA binding of lacR.
  - CAP is activated by cAMP, which is induced by low level of glucose.
- In summary, the operon expression is the highest if [lactose] is high and [glucose] is low; it is expressed at a low level if [lactose] is high and [glucose] is high; it is not expressed if [lactose] is low.

Cooperative binding of yeast TF Gal4 [Giniger & Ptashne, PNAS, 1988]

- Problem: does transcriptional synergy of 4 binding sites of Gal4 in the region between Gal1 and Gal10 genes come from cooperative DNA binding or multiple contact with BTM?
- Results:
  - Binding of Gal4 to site 4: no binding if the adjacent site 3 is deleted via DNA footprinting assay; adding a synthetic site restores site 4 binding.
  - Response of single and multiple Gal4 binding site: measured by lacZ activity in yeast cells
    - Weak sites: the effect is synergistic. Site 3 - 16; site 4 - 1.3; site 3 + site 4 - 302.
    - Strong sites: the effect is additive. 17-mer (sythetic consensus site) - 454; 2 * 17-mer - 1062
- Conclusion: the results support cooperative DNA binding as the mechanism of synergy. If multiple contact model, synergy will still be effective for strong sites. Under cooperative binding model, synergy will diminish at saturating conditions (for strong sites, the occupancy is close to 1).

Synergistic activation of a mammalian gene by Gal4[Carey & Ptashne, Nature, 1990]

- Problem: cooperative binding or multiple contact with BTM?
- Results:
  - Transcriptional synergy *in vivo*: mammalian CHO cells transfected with sequences containing Gal4 binding sites (1, 2, 5, 11 copies resprectively; 17-mer, the strong consensuse site). Strong synergy with 1 vs 2, 5, 11; weaker synergy when comparing 2 vs 5, 2 vs 11 or 5 vs 11.
  - Saturated DNA binding: verify that [TF] saturated binding in the experimental conditions.
  - Transcriptional synergy *in vivo*: same results found.

- Conclusion:
  - Multiple bound Gal4 molecules simultaneously contact BTM, somewhere between 5-10 activator molecules.
  - DNA cooperative binding can be observed under this mechanism, but not from direct PPI between bound activators, but from indirect interaction mediated by BTM.

Transcriptional activation by yeast IBF [Svaren & Chalkley, JCB, 1993]

- Aim: the activation mechanism of yIBF.
- Results:
  - Activation by an enhancer with different number of yIBF binding sites: 1 - 12; 2 - 550; 3 - 2,200; 5 - 8750; 6 - 14400; 8 - 14000 (repeats in the opposite orientation show little difference). Therefore, synergism from 1 to 2 sites; reduced with 3 sites or more; and saturated at 6 sites.
  - Occupancy of sites: essentially complete for all sequences (even single site). Thus synergism cannot be attributed to cooperative DNA binding.
  - Multiple site occupancy is important for achieving synergism.
  - Spacing requirement: the synergy depends on the distance between sites. For a two-site distance, increasing spacing from 33bp to 85bp will reduce transcription by about 5 fold (but still much higher than single site).
- Conclusion: synergism depends on the binding of multiple sites, which create a continuous linking surface that make multiple simultaneous contact with the BTM.

Cooperative binding of Gal4 under physiological conditions [Xu & Johnston, PNAS, 1995]

- Problem: cooperative DNA binding or multiple contact with BTM?
- Methods:
  - Plasmid titration assay: directly measure TF binding with some sites. Put the sequences of interest to plasmids and transfect to cells. The sites in plasmid sequence will compete TF with endogenous genes, thus endogenous gene expression level measures the binding of TF with plasmid sequence (since plasmid does not have promoter sequence, its binding must be only due to the introduced sequence).
- Results:
  - Low affinity Gal4 sites interact synergistically at low [Gal4] but not high [Gal4].
  - Strong sites: no synergy at low or high [Gal4].
  - Plasmid titration assay: support cooperative binding.
  - Same results were observed with a different (VP16) activation domain.
- Discussion:
  - Support cooperative binding model. The earlier results supporting multiple contact model may come from incorrect conditions: different cells, high concentration of TF, naked DNA instead of chromatin, etc.

1.3.2 Transcriptional Regulation in Drosophila

Transcriptional Repression in the Drosophila Embryo [Philos Trans R Soc Lond B Biol Sci, 1995]

- Problem:
  - AP patterning: in particular for eve 2, how the narrow stripe is determined?
  - DV patterning: how dl gradient leads to 3 different expression patterns/territories?
- Repression mechanisms: a repressor could act in several ways
  - Competition: between activator and repressor for overlapping sites
  - Direct repression: through interaction/inhibition of transcription complex (TC), e.g. turn off some coactivator of TC
  - Quenching: make contacts with bound activators s.t. they cannot activate BTM
- AP patterning: eve 2. Two mechanisms may be involved
  - Two gt sites, gt1 and gt3, overlap with bcd sites: if mutated will slightly disrupt eve 2 pattern
  - The gt3 site: not overlap, but about 50bp away from bcd site: if mutated, will greatly change the eve 2 pattern $\Rightarrow$ quenching is more important repression mechanism
- DV patterning: translation of dl gradient to different patterns
  - Mesoderm(ventral): e.g. sna. Type I CRMs (low affinity to dl), activated only by high [dl]
  - Neuroectoderm (lateral): e.g. rho. Type II CRMs (high affinity to dl) and sna sites. Type II will respond to lower [dl], thus activated in both ventral and lateral regions, however, sna can shut down the expression in ventral region (which is only expressed in ventral) through quenching. The evidence of quenching include:
    - If sna site is 150 bp away, no repression
    - Synthetic CRM of eve 2 and rho (separated by 175bp) show additive pattern of expression (i.e. the repressor sites in rho cannot affect eve 2 expression)
  - Dorsal ectoderm (dorsal): e.g. zen. Ubiquitious activators and dl sites: the neighboring corepressor sites convert dl into a long-range silencer (thus will not be expressed in ventral and lateral).

Transcriptional Repression in Development [Gray and Levine, COGD, 1996]

- Types of repression: could be divided by
  - Target: activator (quenching) or transcription complex, or BTM (direct repression)
  - Range: short or long
  - Mode: direct (through the repressor protein itself) or indirect (through corepressor)
- Repression mechanisms:
  - Short-range quenching: the simple scenarios is specific PPI between bound repressor and nearby activator or through co-repressors (basic regions within repressor might interact with acidic region within activator [Saha & Ptashne, Nature, 1993]). E.g. gt, Kr in eve 2; sna in rho.
  - Short-range direct repression.
  - Long-range quenching: e.g. E2F recruits Rb, which works over 1kb on specific activators
  - Long-range direct repression: e.g. yeast α_2 protein through corepressor Tup1 to inhibit BTM; hairy through corepressor groucho to inhibit BTM of the target gene achaete.

- Remark:
  - The effect of coactivators and corepressors can be "absorbed" into the effect of activator or repressor molecules, as needed for thermodynamic modeling.
  - A protein can be both activator and repressor, e.g. dl; and a repressor can act through different mechanisms, e.g. gt may work through both competion and short-range quenching.

Active Repression Mechanisms of Eukaryotic Transcription Repressors [Hanna-Rose & Hansen, TIG, 1996]

- Active repression: inhibitory PPI with components of BTM or transcriptional activators.
- Repressor domains: activator protein (usu. large activation domain) must interact with multiple target proteins to effect max. transcriptional activation, whereas the majority of repressors might only require contact with a single target in order to block transcription effectively (a small region defines the entire repression activity).
- Repressor interations: Fig. 2
  - With BTM components: may prevent the assembly of the preinitiation complex or medidate the formation of inactive preinitiation complex. Example: (i) TR interaction with TFIIB and TBP; (ii) Kr interaction with TFIIEβ; (iii) Eve interaction with TBP.
  - With activators or co-activators: more specific, limited to certain enhancers (where certain activators are used). May prevent productive interactions between the activator and its ultimate target. Example: (i) YY1 interaction with cAMP-binding protein, thus repress the transcription of Fos gene; (ii) Kr show activator specificity: can repress Sp1 but not Gal4.
  - With co-repressors: which in turn repress through interaction with BTM or activators/coactiavtors. Example: Ssn6p-Tup1p corepressor complex can interaction with components of the RNA Pol II holoenzyme.

Going the Distance: A Current View of Enhancer Action [Blackwood and Kadonaga, Science, 1998]

- Elements of transcriptional control: Fig. 1
  - Enhancers: 50 bp to 1.5 kbp in size, perform a specific function, such as activation of its cognate gene in a specific cell type or at a particular stage in development
  - Promoters: -50 to -200 bp relative to TSS, including core promoter and promoter-proximal region
  - Boundary elements (insulators): 0.5 to 3 kb, blcok the spreading of the influence of enhancers or silencers
- Enhancer-promoter selectivity: Fig. 2
  - Specific interaction between enhancers and promoters (due to protein-protein interactions), thus an enhancer may be able to activator only one type of promoter, but not another.
  - Boundary elements block interactions with inappropriate promoters
- LCR and transcriptional competence: Fig. 3, LCR is activated and the locus becomes transcriptionally competent (may have multiple genes), then ehancers could act.
- ON-OFF vs progressive models: Fig. 4
  - ON-OFF model: each cell is either ON or OFF, the enhancers only increase the probability of a cell being ON.
  - Progressive model: the expression level is continuous, enhancers increase the level (for each cell).

- Enhancer mechanisms:
  - Protein-protein contact: between enhancer-associated factors and basal transcription machinery.
  - Covalent modification of proteins: many transriptional activators and coactivators possess histone acetyltransferase activity, while corepressors histone deacetylase activity. Histone acetylation can reduce the repressive nature of chromatin
  - Nucleosome modeling: alter chromatin structure and increase the mobility of nucleosomes.
- Facilitated tracking model: Fig. 5, combining DNA looping and scanning mechanisms. An enhancer-bound complex tracks via small steps (perhaps scanning) along the chromatin untile it encounters the cognate promoter.

Transcriptional repression: the long and the short of it [Courey & Jia, GD, 2001]

- Mechanisms of repression by HDAC: Fig. 2
  - Long range corepressor: e.g. Groucho, recruits HDAC to nearby histone tails. Corepressor can bind hypoacetylated histones, and recruit additional HDAC, thus forming corepressor polymers and resulting in the spreading of a large repressed state of chromatin.
  - Short range corepressor: e.g. CtBP, also recruits HDAC to nearby histone tails. However, the short-range corepressors lack the ability of binding hypoacetylated histones, thus will not polymerize and spread along the chromatin.
- Short-range repression:
  - CtBP-dependent repression: CtBP may function, at least in part, by recruiting HDAC. Evidence: Gal4-CtBP fusion protein at least in some cases, is sensitive to TSA, a specific inhibitor of HDAC.
  - Activator quenching:
    - HDAC interactions do not fully account for CtBP-dependent repression (in some cases, such as in 293 cells, Gal4-CtBP is not sensitive to TSA).
    - The recurited corepressor can interact with nearby bound activators to block activation, perhaps by obstructing the interaction between activation domain and BTM.
    - Lack of activator specificity: CtBP-dependent repressor can quench different types of activators, evidence against specific quenching. Possible explanation is: even though the co-repressor activator interaction is promiscuous, the local high concentrations of both co-repressor and activator make their interaction favorable.

Multiple TAFIIs Directing Synergistic Activation of Transcription [Sauer and Tjian, Science, 1995]

- Background:
  - Bcd alone can activate hb transcription in vivo.
  - Most euk. TFs contain multiple ADs, and different ADs (such as glutamine-rich, acidic and isoleucine-rich) contact distinct components of BTM. Example, distinct subunits of TFIID that comprise TBP and at least 8 TAFs.
- Methods:
  - System: in vitro reconstituted Drosophila transcription system containing TFIIA, B, E, F, H and RNA Pol II + the target to be tested (TBP, TFIID or TBF-TAF$_{II}$ complex). The DNA template is hb promoter and 1, 2 or 3 Bcd sites.
- Results:

  - Effect of the number of sites: at saturating condition of Bcd, single Bcd site - 3 fold; two Bcd sites - 10 fold; three Bcd sites - only additional 25%.
  - Determining Bcd targets: $TAF_{II}110$ as target of Bcd glutamine rich domain (Q) and $TAF_{II}60$ as target of Bcd alanine-rich domain (A).

- Conclusion: multiple contact between activator and BTM component may represent a general mechanism for achieving transcriptional synergism.

DNA Template and Activator-Coactivator Requirements for Transcriptional Synergism by Drosophila Bicoid [Sauer and Tjian, Science, 1995]

- Methods:
  - System: similar to the Bcd synergism experiment. The DNA template is hb promoter containing 3 Bcd sites and 1 Hb site. An additional template tested is eve promoter and hb promoter with 1 Bcd site and 1 Hb site. Use only truncated Bcd (Q domain, Bcd-Q) and Hb.
- Results:
  - Bcd-Q target: $TAF_{II}110$; and Hb target: $TAF_{II}60$.
  - No cooperative DNA binding between Bcd and Hb is detected (i.e. the binding of one to its site is enhancer by the other).
- Conclusion: multiple contact between Bcd, Hb and BTM components, instead of cooperative DNA binding, is the mechanism of transcriptional synergism.

Cooperative DNA binding of Bcd [Ma and Ma, Development, 1996]

- Goal: explore if cooperative DNA binding of activator is a plausible mechnaism of achieving sharp boundary of gene expression?
- Background: Bcd activates expression of hb (Hb protein is not necessary for its activation even though hb enhancer has a Hb site) in the anterior, while creating a sharp posterior boundary. In particluar, about 2 to 3 fold [Bcd] change is sufficient to trigger the transcriptional switch.
- Method: in vitro assays. Use hb enhancers, which contain 6 Bcd sites.
- Results:
  - Cooperative DNA binding: Bcd binds cooperatively to DNA, less than 4-fold [Bcd] can trigger the unbound/bound transition. Estimated Hill coefficient 2.56.
  - Binding affinity of two-site sequence is much higher than single-site sequence: the dissociation constant about $5 \cdot 10^{-8}$ M^{-1} vs $5 \cdot 10^{-7}$ M^{-1}.
  - Bcd proteins can interaction with each other, weak in solution, enhanced if bound in DNA.

Interaction between Dorsal and Twist [Shirokawa & Courey, MCB, 1997]

- Aim: molecular mechanism of the transcriptional synergy between Dl and Twi.
- Discussion:
  - PPI between Dl and Twi: they are found to interact in vitro; and the domains of interaction are required for transcriptional synergy.
  - Insufficiency of PPI as explanation of synergy: PPI is not sufficient to create the cooperative DNA binding; furthermore, the [TF] are high enough to saturate DNA binding, thus it is unlikely that cooperative DNA binding alone contributes to synergy.

- Alternative mechanism: multiple contact with BTM (but that does not require direct PPI); or PPI between Dl and Twi changes the conformation of one TF that make it activate transcription efficiently.

dCBP is a co-activator of Dorsal [Akimaru & Ishii, NG, 1997]

- Aim: the mechanism of Dorsal-dependent twi expression.
- Results:
  - dCBP mutant: fails to express twi, and generates twisted embryo.
  - PPI between Dl and dCBP: suggests that dCBP is a co-activator of Dl (CBP has been shown to be a co-activator of many TFs, including mammalian NFκB, which is related to Dl).

Binding site patterns of Bcd [Yuan & Ma, J Biochem., 1999]

- Aim: how binding site arrangement (spacing, orientation) affects the affinity of Bcd?
- Background: it is known that some activators bind with certain preference to BS arrangement, e.g. Prd, Lab, Exd of Drosophila; and α2 of yeast.
- Methods: in vitro selection assay; hb enhancer (6 Bcd sites).
- Results:
  - Bcd prefers tail-tail arrangement, with spacing 7 - 15 bp in in vitro selection experiment using 48 bp random oligonucleotides.
  - In a different experiment (one head plus 13 bp random), Bcd prefers head-head arrangement, with fixed spacing 3bps.
  - Long spacing relaxes the strict requirement of arrangement: with natural spacing and tandem repeat (36bp), inverting one site does not change the affinity.
  - Contribution of sites to hb activation: site X2 and X3 are the most important (mutation of other sites has little effect). There exists two tail-tail or head-head site pairs, only X2 and X3s is at the optimal spacing. Mutating X3s reduces expression by 10 fold; while mutating X3t (disrupt another pair not at optimal spacing) only 5-fold.
- Discussion: the site arrangement (at optimal alignment) may be more important than if a site matches the consensus sequence. Example, X2 and X3s in hb enhancer matches consensus poorly than other two sites in hb, but they contribute most; in kni enhancer, the 6 Bcd sites match the consensus poorly (but satisfy the requirement defined here).

Bcd function without TAF interacting domain in vivo [Schaeffer & Wimmer, PNAS, 1999]

- Aim: the mechanism of transcriptional activation by Bcd in vivo? Depend on the interaction with TAF?
- Results:
  - Transgenic fly with the TAF interaction domains all deleted rescues the Bcd null phenotype.
  - Development of embryo is not affected by the presence of dominant negative mutation of $\text{TAF}_{II}110$ and $\text{TAF}_{II}60$.

CBP as a coactivator of Bcd [Fu & Ma, JBC, 2004]

- Aim: the mechanism of transcriptional activation by Bcd.
- Results:

- CBP is a Bcd coactivator: (i) in Drosophila S2 cells, co-transfection with CBP can increase Bcd activity; (ii) Bcd has interaction with CBP; (iii) CBP mutant shows a reduced, but not abolished, co-activator function of Bcd
- Differential role of dCBP in facilitating Bcd activation in hb and kni enhancers.

Cooperative binding of Bcd [Lebrecht & Hanes, PNAS, 2005]

- Methods: bcd mutants: K57R, S35T (less severe) that lack cooperative binding
- Results:
  - Phenotypic definition of the mutants: head defects and lethality
  - Change of expression of gt, hb, Kr
- Discussion: other possible mechanisms of transcription activation and synergy
  - One factor changes chromatin structure to facilitate binding by other factor(s) [24]
  - Coactivation through other cofactors. E.g. bcd binds with dCBP (histone acetyltransferase activity); dCBP increase bcd occupancy and recuritment of GTFs [20]

1.3.3 Cell Type Determination

Ref: Gene regulation and cell types [MBOC, Chapter 7]

Problem: gene regulatory sequences respond to signals, however, the signals are often transient. How do cells maintain a stable expression pattern, or in multi-cellular organisms, how are different cell types created and maintained?

DNA rearrangement: the change of DNA sequence can create inhertable change of gene expression pattern.

- Phase variation in bacterial: two flagellin genes, H1 and H2. In some case, H2 and a repressor protein is expressed and the repressor stops H1 expression; in the other case, the promoter is inverted by site-specific recombination, and H2 and repressor cannot be expressed. Thus either H1 or H2 is expressed, and it is switched randomly (escape host immue system).
- Yeast cell type determination: either MAT-a or MAT-$\alpha 1/\alpha 2$ is expressed in a or α cells. The two cell types can randomly switch through mating type switching involve site-specific recombination.
- V(D)J recombination in vertebrate immune systems.

Gene regulatory circuit:

- Flip-flop: could create bi-stability. Ex. lambda phage, the two proteins, lambda repressor (cI) and Cro, mutually represse the expression of each other. When cI is expressed, the phage is in the propage state (integrated in the host gene); when Cro is expressed, in the lytic state (multiplication inside the host cells). The switching of the two states is triggered by DNA damage repair mechanism of the host cells (thus if host cells are under favorable conditions, virus does not kill host cells; if not, then multiply and kill the host).
- Positive feedback: self-perputating expression, thus when expression is trigged by some transient signal, it can be sustained by its own expression. This creates memory.
- Negative feedback: the expression level will be maintained at some small range.
- Feed-forward loop (FFL): suppose $X \to Y \to Z$ and $X \to Z$, and expression of Z depends on both X and Y. Then a transient X will not activate Z, as Y is N.A. In contrast, a stable signal of X will trigger accumulation of Y, and lead to Z expression. Thus FFL may be important in distinguish noise from true signals.

- Circidian clock: the central feature is the accumulation and decay of regulatory proteins. In fruit fly, accumulation of Tim and Per → dissociation of Tim-Per complex (delay) → Per negatively regulates expression of Per and Tim (delay) → next cycle of accumulation.

Cell type determination by regulatory proteins:

- Principle: achieve cell type specific expression. Consider two cell types A and B, and two gene groups G_A expressed in A, and G_B expressed in B. Then G_A needs to be ON in A, and OFF in B; and G_B needs to be OFF in A and ON in B. In general, this would be some regulators specific to A or B, and CREs of G_A and G_B that interpret these regulators.
- Yeast mating type determination.
- One regulator may coordinate expression of a large set of genes: e.g. glucocorticoid receptor (GR), activated by cortisol under fasting or intense physical activity, turns on a number of genes in liver cells, including those involved in converting AAs and other metabolites to glucose. GR activates different sets of genes in different cell types.
- One regulator can lead to expression change and a specialized cell types/organs: (1) Skin fibroblast can be converted to muscle by induction of MyoD expression. During development, myoblast cells, upon signal, express several myogeneic factors including MyoD, MyogG, Mrf4, Myf5, and they activate Mef2, all these regulatory proteins activate a large battery of downstream genes: muscle structure genes (myosin, actin, etc.), muscle-specific metabolic genes (such as phosphokinases), Ach receptor (for neural stimulation). (2) Ey (Pax6 in vertebrates) in fly, when expressed in the precursor cell of legs, can lead to eye formation in the leg later in development.
- A small number of regulators can create multiple cell types, by combinatorial control (Figure 7-76) the meaning of a regulator depends on the other regulators in the cell.

1.4 Epigenetics

Chromatin structure: [MBOC, Chap. 4]

- Nucleosomes: 146 bps DNA in 8 core histone proteins (histone octamer), two of each H2A, H2B, H3, H4. Each nucleosome core particle is separated from the next by a region of linker DNA, which can vary in length from a few nucleotide pairs up to about 80. On average, therefore, nucleosomes repeat at intervals of about 200 nucleotide pairs.
- Chromatin fiber: nucleosomes are packed together into 30nm compact chromatin fiber. Packing is achieved by: (i) linker histone H1; (ii) histone tails may help attach one nucleosome to another.

Regulation of chromatin structure: [MBOC, Chap. 4; Felsenfeld and Mark Groudine, Nature, 2003]

- Function: in compact chromatin, DNA is not accesible by other proteins in the cell, particularly those involved in gene expression, DNA replication, and repair. Thus it is important to regulate the nucleosome organization in the chromatin to allow DNA access when the cell need changes.
- Chromosome remodeling: by complexes such as SNF/SWI, use the energy of ATP hydrolysis to change the structure of nucleosomes temporarily so that DNA becomes less tightly bound to the histone core. Chromatin remodeling complexes are carefully controlled by the cell. When genes are turned on and off, these complexes can be brought to specific regions of DNA where they act locally to influence chromatin structure.

- Covalent modification of histones: by complexes such as SAGA. Each histone tail is subject to several types of covalent modifications, including acetylation of lysines, methylation of lysines, and phosphorylation of serines. For example, acetyl groups are added to the histone tails by histone acetyl transferases (HATs) and taken off by histone deacetylases (HDACs). The modifications can attract specific proteins to a stretch of chromatin that has been appropriately modified. Depending on the precise tail modifications, these additional proteins can either cause further compaction of the chromatin or can facilitate access to the DNA.

- Histone variants: may replace the consensus histones, with functional consequences. Example, histone H2AZ, which is associated with reduced nucleosome stability, replaces H2A non-randomly at specific sites in the genome.

Heterochromatin: a tightly packed form of DNA, which comes in different varieties.

- Constitutive heterochromatin is usually repetitive and forms structural functions such as centromeres or telomeres, in addition to acting as an attractor for other gene-expression or repression signals.

- Facultative heterochromatin is not repetitive, and can, under specific developmental or environmental signaling cues, lose its condensed structure and become transcriptionally active. Heterochromatin is often associated with the di and tri-methylation of H3K9. The formation of facultative heterochromatin is regulated, and is often associated with morphogenesis or differentiation, e.g. X-chromosome inactivation in female mammals.

- Heterochromatin functions: from gene regulation to the protection of the integrity of chromosomes; some of these roles can be attributed to the dense packing of DNA, which makes it less accessible to protein factors. For example, naked double-stranded DNA ends would usually be interpreted by the cell as damaged or viral DNA, triggering cell cycle arrest, DNA repair, or destruction of the DNA fragment.

- Heterochromatin is generally clonally inherited; when a cell divides the two daughter cells will typically contain heterochromatin within the same regions of DNA, resulting in epigenetic inheritance.

Ref: [Felsenfeld & Groudine, Nature, 2003; Li & Workman, Cell, 2007]
Chromatin structure:

- Histones: highly conserved proteins. Function as a general repressor of expression, and DNA packaging (as a by-product).

- Chromatin structure: (1) Nucleosome: two subunits of H2, H3, H2A and H2B forming octamer, around which DNA wrapps around (147 bp). (2) Each nucleosome is connected to its neighbors by a short linker DNA (10-80bp) - "beads on string". (3) This polynucleosome is further folded into a 30nm compact fibre, which is stabilized by histone H1 (binding to nucleosomes and ajdacent linkers).

- Once bound, histone-DNA forms very stable protein-DNA complex. The time of histone occupancy is on the order of minutres for H2A/B and hours for H3/H4.

- The nucleosomes generally make DNA less inaccessible: some DNA sequences are outward-facing on the nuclesome surface or in the linkers between nuclesomes, but most are buried inside nucleosomes. The most compact form of chromatins is inaccessible for transcription.

Nucleosome occupancy and regulation: it is basically determined by the DNA sequences, and regulated by three mechanisms:

- Chromatin remodeling:

  - The chromatin remodeling complexes such as SWI/SNF family in yeast and human, RSC in yeast, continually shuffle the positions of individual nucleosomes s.t. sites are randomly exposed for a fraction of time, and some mobilize nucleosomes, causing them to move short distance along DNA. They can also help form DNA loops.
  - Remodeling complexes interact selectively with other regulatory proteins that bind specific DNA sequences. E.x. only certain classes of TFs interact with mammalian SWI/SNF family.

- Histone modificiation: the pattern of histone modifications at different resiudes of different histones, histone code, may encode transcription information.
  - Common modifications: methylation of Arg (R); methylation and acetylation of Lys (K). Euchromatin modifications (active transcription): acetylation of H3 and H4, di- or tri-methylation of H3K4. Heterochromatin modifications: H3K9me and H3K27me.
  - Mechanisms: (1) With the exception of methylation, histone modifications result in a change in the net charge of nucleosomes, which could loosen DNA-histone interactions. Demonstrated with acetylated histones (easier to dissociated from DNA). (2) Histone code may be read by other regulatory proteins, and lead to further effects.
  - H3K4 methylation: monomethylation is enriched towards the 3' end, dimethylation peaks in the middle and trimethylation occurs around TSS. Possible mechanism: Set1 associates with the elongating PolII at the beginning of the ORF and convert monomethyl into di- and tri-methyl group. The H3K4 methylation pattern can be recognized by chromating remodeling and histone-modification complexes containing PD domains.

- Histone variants:
  - Histone variants, e.g. H2A.Z usually have lower nucleosome stability.
  - Substitution of one of the core histones either though ATP-dependent histone exchange reactions, or with the help of a histone chaperon Nap1.

Chromatin states:

- Propagation of transcription states: certain modification (e.g. H3K9Me) recruits a protein HP1, which in turn recruits histone methyltransferase that methylates the adjacent histone, thus the active state will be proprogated. By the same mechanism, the inactive state of chromatin can also be propagated.
- Boundary elements: block the propgation of states, and help estabilish chromatin domains, e.g. CTCF.
- DNA methylation: methylated CpG may recruite HDAC, and silence gene expression.

Special chromatin structures: telemere, centromeres and inactive X chromosomes.

- Inactive X chromosomes: may be enriched for certain histone variants, macroH2A.
- Centromeres: at vertebrates, H3 is replaced by CENP-A.
- Telemeres: in Scer, gene silencing is mediated by RAP1 and SIR proteins.

Chromatin dynamics and transcription:

- Chromatin states in promoters: promoters are often associated with low level of nucleosomes (in yeast, there are 200 bp nucleosome-free regions (NFRs) in promoters) Also acetylation of H3 and H4 peaks sharply at active yeast promoters. The histone variant H2A.Z (Htz1) is also enriched at promoters that are poised for activation.

- Transcription initiation: nucleosomes pose a significant barrier for the formation of pre-initiation complex (PIC), which makes many DNA contacts. Nucleosomes are lost during initiation by transcription activators through: recruting chromatin remodeling complexes (such as Swi/Snf or SAGA), then (1) eviction of H2A.Z by remodeling complex and acetylation and PIC assembly; or (2) displacement of nucleosome by TFIIH and PolII binding.

Epigenetic inheritence:

- During DNA replication, the bound nucleosome may be randomly assigned to one strand, then in the other stand, the adjacent nucleosome (with certain modification) may recruit a histone and do the same modification (according to the mechanims of propagation). Question: how can this explain the epigenetic inheritence in the absence of self-propagation (for some types of modifications)?

- During transcription: with the help of histone chaperons, histones evicts in front of elongating Pol II rapidly deposit onto DNA behind Pol II.

- Efficient mechanism of cellular differentiation: suppose some genes are silenced in one stage through epigenetic mechanism, then in late stages, even without the negative signal (such as TFs), the silencing state can be maintained.

DNA methylation [personal notes]

- Function of DNAm: primarily as a mechanism to lock transcriptional states.

- Hourglass model: specificity of DNAm, sequence > histone marks > DNAm. At sequence level, there is enormous variability, and this is interpreted by TFs, which then creates histone marks/labels. They are in turn interpreted by DNAm machinery and lead to DNAm.

Functions of DNA Methylation: Islands, Start Sites, Gene Bodies and Beyond [Jones, NRG, 2012]

- DNAm writers and erasers: (1) De novo DNA methyltransferase: DNMT3A, DNMT3B: recognize nucleosomal DNAs. Also need maintanence DNA methyltransferase. (2) Erases: TET, AID.

- Three types of CGI promoters: (1) Most are non-methylated: thats why they are maintained during evolution despite high mutation rates in mCpGs. (2) Non-methylated but repressed: via other mechanisms, e.g. Polycomb, e.g. MYOD1 and PAX6. (3) Methylated: long term inhibition, e.g. imprinting and X-inactivation. Could last 100 years.

- Remark: need type (2) to regulate/repress expression of genes, while long-term inhibition is not desirable.

- Function of DNAm in promoters: (1) methylated CGIs cannot initiate transcription. (2) Relationship between gene silencing and DNAm: Likely silencing comes before DNAm (Figure 2): e.g. OCT4 enhancer and NANOG promoter, TF leaving, then nucleosome association, and then DNMT3A/3B establishes DNAm. Remark: consistent with DNAm hourglass model (sequence to histone marks to DNAm).

- Gene body methylation: most gene bodies are CpG poor, and methylated in repetitive sequences and TEs. However, methylation in gene body does not suppress transcription elongation, even they are bound by MeCP2.

- Functions of DNAm in gene body: (1) Possible that elongation (H3K36me3) may recruit DNMTs. (2) Possible that DNAm regulates splicing: DNAm levels are higher in exons, and transition to lower levels in introns. Possible mechanism: recruit CTCF, which pause RNA II, influencing splicing.

- Alternative promoters: very common (most genes do). DNAm in the downstream promoter cannot suppress transcription.

- DNAm at insulators: CTCF binding may not be affected by DNAm, but can lead to DNA demethylation.
- DNAm and TF binding/transcription at CpG poor regulatory regions: possible causal direction is unclear. (1) TF binding can change DNA methylation: e.g. lead to passive DNA demethylation. (2) DNAm can affect TF binding: e.g. block MYC binding, but no effect on SP1 binding. (3) Even when there is a causal effect, the mechanism may be unclear, e.g. DNAm may block Oct4 binding, but this may happen outside motif (100bp).

Function and information content of DNA methylation [Schubeler, Nature, 2015]

- DNAm at gene bodies: DNAm may help maintain chromatin during transcription, since nucleosomes are displaced by RNA Pol 2. H3K36me3 recruits HDAC, which leads to compact chromatin.
- DNAm at promoters: CGI promoters, usually DNAm does not change during normal development, except germline cells. Inactive CGI promoters: repressed by H3K27me3/PolyComb, not DNAm.
- Setting and writing DNAm: de novo: DNMT3A/3B; maintainence DNMT1. TET: could affect DNAm at enhancer regions.
- DNAm readers: CXCC proteins recognize unmethylated DNA, e.g. CFP1 and KDM2A/2B (histone demethylase). MBD family (including MeCP2) recognize methylated DNA.
- DNAm and TF binding (Figure 2): some TFs bind to DNA, and lead to passive DNAm; some TFs bind to methylated DNA; binding of some TFs are blocked by mCpG. Pioneer factor may reduce DNAm (passive), which then facilitate binding of methylation sensitive TFs.
- DNAm in cancer: DNMT and TET2 are both implicated in myeloid cancer.
- Utility of DNAm: easy to measure - frozen samples, low amount. Help identify disease/cancer subtypes.
- Remark: the complex interactions of DNAm and TFs may be explained different roles of TFs: they participate in DNA methylation (writer), demethylation (eraser) and reader functions. Ex. for reader-related role, a TF needs to bind to methylated DNA.

1.5 Post-Transcriptional Control

Types of ncRNAs in cells [MBOC, Ed6, Chapter 6]

- Small nuclear RNAs (snRNAs): function in a variety of nuclear processes, including pre-mRNA splicing.
- Small interferring RNAs (siRNA): degradation of selective mRNAs and establishment of compact chromatin structures.
- Piwi-interacting RNAs (piRNA): protect the germ line from transposable elements.

Transcription processing and termination [MBOC, Ed6, Chapter 6]:

- Bacterial: at the 3'end of the mRNA, form termination hairpin (mRNA folding), which disengages RNA polymerass, leading to termination.
- Euk RNA processing: In Euk, RNA processing is tightly coupled with elongation. The C-terminal domain (CTD) of RNA polymerase is bound by proteins for RNA processing (Figure 6-22). Gradual phosphorylation of CTD facilitate binding of RNA processing proteins: initially 5' capping proteins, then splicing proteins, then 3' end processing proteins.

- Euk Termination (Figure 6-34, 6-35): consensus sequences near 3' end including AAUAAA, 10-30 bases, then CA (cleavage site), ≤ 30 bases and GU or U-rich sequences. The sequence before cleavage, specially AAUAAA, directs binding of proteins, CPSF and CstF. After cleavage, Poly-A polymerase adds poly-A tail (about 200 A's) to mRNA.
- In Euk, 5' capping and 3' polyA protects mRNA. Without 5' capping, 5' to 3' endonuclease in the nuclease can cut and degrade RNA. There are many debris from RNA transcription in the nucleus.
- Q: can sequences outside 3' UTR affect termination?

Pre-mRNA splicing [MBOC, Ed6, Chapter 6]

- Basic process of splicing: Figure 6-25, two phosphoryl-transfer reactions. Intronic branch point adenine (A), about 20-50 bp upstream of 3' end (acceptor site), attacks the donor site at the upstream exon (cleavage) and forms a lariate, then the -OH end of the first exon attacks the second exon and form covalent bond. The lariate is released.
- Splicesome: small nuclear RNAs (U1 to U7) and 200 proteins. The specificity of reactions are controlled by base-pairing between snRNPs and splicing consensus sequences (Figure 6-27), which signals the beginning and end of introns. The consensus sequences have three parts: (1) 5': AG (exon donor site) and GU (intron acceptor site), and several intronic bases. (2) 3': G (exon) and AG (intron), and several intronic bases. (3) branch point: A and some nearby bases.
- Possible mechanisms for accurate splicing: to prevent errors such as exon skipping and cryptic splice sites. (1) Exon definitions (Figure 6-32): SR proteins tend to associate with exons. (2) hnRNPs preferentially associate with introns. (3) Chromatin structure: nucleosome association with exons, and histone modifications in exons that directly recruit splicesome components.
- Mutations on splicing: several different cases (Figure 6-33):
  - A mutation destroy splice site > exon skipping.
  - A mutation destroy a splice site and exposes a cryptic splice site > extended exon.
  - A mutation create a new splice site in the intron > new exon. This could be far from donor/acceptor sites.

Mechanisms of post-transcriptional regulation [MBOC, Ed6, Chapter 7]

- Alternative splicing and its functional significance: a number of forms of AS including: exon skipping, alternative 5' and 3' splice sites, intron retention. Functional significance: e.g. Dscam important for immunity, multiple versions of exons A and B and C (many combinations).
- Regulation of alternative splicing: Figure 7-58. The strength of splice sites can be modified by activators or repressors: e.g. repressor binds to splice site, making it inaccessible; activator enhances the ability to splice out an intron (could be far from the splice site, similar to enhancers). In general, splicing machinery has many potential splice site pairs to choose from (they are all similar without additional information), so it is a matter of relative strength of these splice sites.
- Alternative polyadenation and cleavage: 50% of mRNA species in human have alternative polyeadnation. Happens in pre-mRNA. E.g. Figure 7-59, B-cell antibody production, an alternative cleavage site in the intron that is weak. Usually skipped, but in activated states, induction of CstF allows it to cleave at the weaker, intronic polyadenation site, resulting a different form of antibody (no C-terminal membrane-bound domain, secreted form).
- RNA export and subcellular localization: in cytoplasm, mRNAs species could be localized to different places, to concentrate mRNA on their intended locations. Ex. synaptic genes in neurons. The signal is usually in 3' UTRs.

- Translation initiation: controlled by both 5' and 3' UTRs.
- RNA degradation: most mRNAs are not stable, having half-life less than 30 mins. mRNA degradation process: by exonuclease from 3' to 5', gradual shortening of polyA. Once reach a critical threshold of 25 A's, either decapping of 5' end, or continued 3' to 5' degradation.
- Regulation of RNA stability: the rate of degradation depends on specific mRNA sequences. 3' UTR sequences are particularly important, carrying RBP binding sites that can modify the rate of degradation. RNA stability is also affected by translation: Poly-A shortening and decapping compete directly with the machinery that translates the mRNA (Figure 7-70).
- Additional mechanism for regulating RNA stability: RNA endounclease may cut mRNA from the middle. Ex. transferrin receptor (import iron), usually, the 3' UTR is bound by an enzyme (aconitase). When there is excess of iron, aconitase is released from 3'UTR, then the endounclease cleavage site is exposed, degrading mRNA.
- **Lessons**: 5' and 3' UTRs (especially the 3' one) control these steps, however, the sequences are highly heterogeneous. 3' UTR in different genes bind to different RBPs and miRNAs to regulate localization, translation and stability. The RNA secondary structure at UTRs are probably important in determining and modifying binding of RBPs (perhaps miRNAs as well).

RNA interference [Genetics, Brooker, 15.5; MBOC Chapter 7]

- The discovery of dsRNA in plants: in transgenetic plants, introduction of many copies of a gene can actually silence the expression of this gene. Why?
- Explanation: inserted gene has a promoter. When it is inserted into a plant region, it may be adjacent to another promoter, which transcribes the gene in the antisense strand. The result is a ds-RNA that leads to gene silencing.
- RNA interference by miRNA or siRNA: both are typically 21-23 bp small RNA molecules. The mechanism:
  - Pre-miRNA or pre-siRNA: stem-loop structure. The hairpin is recognized by Dicer, an endonuclease, forming a ds-RNA of 21-23 bp long.
  - The ds-RNA is recognied by RISC: one strand binds to RISC (guide strand) and the other (passenger strand) degraded.
  - RISC recognizes cellular mRNA, due to complementarilty. (1) High complementarily: the mRNA is cut and degraded. (2) Low complementarity: often in human miRNA, the mRNA is unable to be translated.
- In plants, miRNA binding sites are usually found in the coding regions, while in animals, they are often in the 3' untranslated region of mRNA (Scitable, Small Non-coding RNA).
- Regulation of gene expression by miRNA: about 1,000 miRNA in human genome. Each miRNA can regulate expression of hundreds of genes. MiRNAs can work combinatorially to shut down translation.
- RNAi as defense mechanism: against virus or transposable elements. Often dsRNA in a cell indicates foreign invaders (e.g. virus), so the strategy of cells is: use dsRNA as markers, and then try to degrade all RNA (single strand) molecules that match the dsRNA. So RNAi pathway first involves recognition of dsRNA, processing (removing passanger strand), and action (guide strands as probes to detect ss RNA).
  - Defense against virus: some virus may have ds-RNA in life cycle. For bacterial, against bacteriophages.

  - Defense against transposons: transposons may create ds-RNA (similar to plant example) if inserted in many places in the genome.

Regulating gene expression by RNA secondary structure and miRNA [https://www.youtube.com/watch?v=02_SVWsXdg0]

- RNA structure can regulate gene expression
  - RNA stem-loop structure from palindromic sequence.
  - Could be a roadblock for transcription/translation, or serve as a binding site of regulatory molecules.
- Estimate that 70% of genes are regulated by RNA, often short RNAs, which suppress expression of genes with sequence homology.
  - siRNA: generated from ds-RNA precursor molecules.
  - miRNA: from miRNA genes.
  - Piwi-interacting RNA (pi-RNA): expressed in germ-line cells.
- miRNA generation: from longer miRNA gene, called primary miRNA (pri-miRNA), then cleaved by DROSHA. The precursor miRNA (pre-miRNA), of 20-25 base, will be exported to cytosol, and bind by Dicer.
- The relation between miRNA and human diseases: examples of miRNA (up or down-regulation) in cancer. A number of miRNAs can affect expression of oncogenes and tumor suppressor genes.
- Using RNAi to manipulate gene expression: in mammals because of ds-RNA induced interferon response, it's difficult to induce gene silencing by ds-RNA. So the preferred strategy is to mimic miRNA by creating synthetic small RNA (si-RNA): a molecule with stem-loop structure where the stem is 19 bp sequence (complementary).

Systematic identification of miRNA targets [Hafner & Tuschl, Manuscript, 2009]

- Problem: (i) what are the target mRNA sites of miRNAs? (ii) what is the mechanims of target recognition? (iii) what is the function of miRNA binding?
- Background:
  - miRNAs binds to partially complementary mRNAs, between the 5' end (the so-called seed region) of the miRNA, particularly the 7-nt segment from position 2 to 8 of the miRNA, and the target mRNA.
  - miRNA acts as guides of ribonucleoprotein complexes (miRNPs), Argonaute (AGO/EIF2C) proteins, to induce mRNA degradation or inhibit translation.
  - There are several hundred miRNA gene families, many of which are specifically expressed in various cell-types.
- Methods:
  - PURE-CLIP: immunoprecipitation of AGO and TNRC6 proteins in HEK293 cells; crosslinked RNA was recovered after separation of the IPed protein complexes by SDS-PAGE; RNAs then converted to cDNA, and sequenced.
  - miRNA binding sites: defined as clusters of overlapped sequence reads, or crosslink-centred regions (CCRs) of length 41 nt.
  - Studying the effect of miRNAs: inhibit the top miRNAs and compare the mRNA level before and after inhibition (degradation level).

- miRNAs: hundreds of miRNAs are found (via sequencing). The top 25 expressed miRNAs account for 72%, and the top 100 account for 95% of the total of miRNA sequence reads.
- Target identification:
  - AGO experiments identified 17,319 CCRs, mapped to 4,648 transcripts.
  - CCRs correspond to 84% exonic, 14%intronic, and 2% not assigned transcribed regions. Of the exonic CCRs, 50% distributed to the CDS, 46% to the 3' UTR, and 4% to the 5' UTR of mRNAs (CDS percentage is unexpectedly high).
- Target mRNA recognition by miRNAs:
  - The most significantly enriched 7-mers corresponded to the reverse complement of the seed regions 2-8 of the most abundant HEK293 miRNAs
  - A total of 14,809 of the 17,319 CCRs contained at least one 6-mer miRNA seed-complementary region.
- mRNA degradation:
  - The magnitude of the destabilization effects dropped from 9- mer, to 8-mer to 7-mer to 6-mer matches.
  - Transcripts containing more than one CCR were more efficiently destabilized than transcripts containing a single CCR.
  - Transcripts with sites exclusively in the CDS were subject to a statistically significant miRNA-dependent destabilization, albeit less pronounced compared to those caused by sites in the 3'UTR.
  - Highly expressed mRNAs appear to avoid miRNA regulation.
- mRNAs with high-confidence predicted miRNA sites: about 8 to 15% sites of the most abundant miRNAs appear in CCRs. Overprediction may come from: some predicted sites may function in different cellular context. The CCR sites, compared with non-CCR high-confidence sites: have lower free energy (RNA secondary structure), and are more evolutionarily conserved.

1.6 Mutation, Recombination and Transposition

Occurrence and causes of mutations [Brooke, Section 16.2]

- Summary of types of mutations:
  - Spontaneous mutations: abnormal cross-over and chromatin segregation (meiosis), errors in DNA replication, transposition, depurination, deamination, tautomeric shifts, toxic metabolic products.
  - Induced mutations: chemical and physical agents.
- Experiments that support mutations are random events: (1) Luria-Delbruck experiment: fluctuation test, large fluctuation of number of bacterial colonies that are T1 resistant in multiple samples. (2) Replica plating experiment by Lederbergs.
- Depurination: purine bases (A and G) are somewhat less stable, and may spontaneously get lost, creating apurinic site (unpaired base). If not repaired, the empty site will match any base in replication.
- Deamination: (1) C deamination becomes U, if not repaired, will match A during replication (thus CG mutations to TA). (2) mC deamination becomes T, then we have (TG) pairing. It is hard for DNA repair enzyme to know what causes the problem and often leads to mutation (CG to TA). This is why methylated C is often mutated.

- Tautomeric shifts: the four bases, each could exist in multiple forms. Immediately prior to replication, in ss DNA, the rare form might match a wrong base.
- Oxidative stress: G is particularly sensitive to oxidation by ROS, leading to 8-oxoG, which matches to A (thus GC to TA mutation).

1.7 Signal Transduction

1. Principles of signal transduction

   Function and organization of signaling networks:

   - Each cell is programmed to respond to specific combinations of extracellular signal molecules. An individual cell often requires multiple signals to survive, and additional signals to grow and divide or differentiate.
     - Example: Many epithelial cells require survival signals from the basal lamina on which they sit.
     - Implication: Because different types of cells require different combinations of survival signals, each cell type is restricted to a specific set of environments in the body.
   - Different types of cells usually respond differently to the same extracellular signal molecule. Ex. achetylcholine decreases the rate and force of contraction in heart muscle cells, but it stimulates skeletal muscle cells to contract (different receptors), and it stimulates salivary gland cells to secrete (same receptors, but different intracellular signaling).
   - Cross-talk of signaling pathways: because a cell needs to respond to specific combination of signals, the downstream pathways of these signals are often cross-linked and partially shared, s.t. one signal may affect the cell's response to other signals.
   - The importance of inactivation of signaling pathways: the inactivation processes play a crucial part in determining the magnitude, rapidity, and duration of the response.
     - The ability of responding repeatedly: during development, transient extracellular signals often produce lasting effects. In most cases in adult tissues, however, the response fades when a signal ceases, s.t. the cell will be able to response to future signals. The speed with which a cell responds to signal removal depends on the rate of destruction, or turnover, of the intracellular molecules that the signal affects.
     - Prompt responses: the turnover rate of signaling molecules can determine the promptness of the response. Many intracellular proteins have short half-lives, some surviving for less than 10 minutes. In most cases, these are key regulatory proteins whose concentrations are rapidly controlled in the cell by changes in their rates of synthesis. The same principle applies to e.g. kinase-phosphatase regulation of signaling pathways.

   Intracellular signaling molecules: these proteins relay, amplify, spread, modulate or integrate the signal(s).

   - Molecular switches: If a signaling pathway is to recover after transmitting a signal so that it can be ready to transmit another, every activated molecule in the pathway must return to its original, unactivated state. Common switch mechanisms:
     - Phosphorylation: kinase and phosphates.
     - GTP/GDP binding: active if GTP-bound. In GTP-binding proteins (G proteins): inactivated by GTPase-activating proteins (GAPs) and activated by GPCRs. In monomeric GTPases: inactivated by GAPs, and activated by guanine nucleotide exchange factors (GEFs).
     - Binding of another signaling protein or a small intracellular mediator such as cyclic AMP or Ca2+.

  - Covalent modifications other than phosphorylation or dephosphorylation, such as ubiquitylation.
- Signal integration: the molecular switches can integrate signals, implementing logic gates. Ex. a protein needs phosphorylation at two sites (by kinases activated via two different signal pathways) to be active.
- Signaling complexes: they can enhance the speed, efficiency, and specificity of the response. How does an individual cell manage to make specific responses to so many different combinations of extracellular signals? The question is especially puzzling because many of the signals are closely related to one another and bind to closely related types of receptors.
  - One strategy makes use of scaffold proteins, which bind together groups of interacting signaling proteins, often before a signal has been received. Because the scaffold holds the signaling proteins in close proximity, the components can interact at high local concentrations and be sequentially activated speedily, efficiently, and selectively in response to a signal, avoiding unwanted cross-talk with other signaling pathways.
  - In other cases, signaling complexes form only transiently in response to an extracellular signal and rapidly disassemble when the signal is gone. Such transient complexes often assemble around a receptor after an extracellular signal molecule has activated it.

Complex behaviors from signaling circuits:

- Switch-like behavior in response to gradually increasing concentration of extracellular signals:
  - Cooperative responses: they become sharper as the number of required cooperating molecules or phosphate groups increases, Four molecules of the small intracellular mediator cyclic AMP, for example, must be bound simultaneously to each molecule of cyclic-AMP-dependent protein kinase (PKA) to activate the kinase.
  - Responses are also sharpened when an intracellular signaling molecule activates one enzyme and, at the same time, inhibits another enzyme that catalyzes the opposite reaction. Adrenalines binding to a G-protein-coupled cell-surface receptor increases the intracellular concentration of cyclic AMP, which both activates an enzyme that promotes glycogen breakdown and inhibits an enzyme that promotes glycogen synthesis.
  - Positive feedback loops (below).
- Positive feedbacks: e.g. S-kinase activates an E-kinase, and I-phosphatase inactives E. E-kinase auto-phosphorylates, forming a positive feedback. It may produce:
  - Switch-like behavior: a runaway increase in the quantity of product when the signal increases above a critical value.
  - Cellular memory via bistability: once the responding system has switched to the high level of activation, this condition is self-sustaining and can persist even after the signal drops back below its critical value. Through positive feedback, a transient extracellular signal can often induce long-term changes in cells and their progeny that can persist for the lifetime of the organism.
- Negative feedbacks: e.g. S-kinase activates an E-kinase, and I-phosphatase inactives E. E-kinase can activates I, if activated, forming a negative feedback.
  - Counteracts the effect of a stimulus and thereby abbreviates and limits the level of the response, making the system less sensitive to perturbations.
  - A delayed negative feedback with a long enough delay can produce responses that oscillate.
  - A negative feedback with a short delay: the system behaves like a change detector - cells respond to changes in the concentration of a signal molecule (rather than the absolute concentration) over a wide range. The target cells accomplish this through a reversible process of adaptation, or desensitization, whereby a prolonged exposure to a stimulus decreases the

cells' response to that level of stimus. The idea is: a strong response modifies the signaling machinery involved, such that the machinery resets itself to become less responsive to the same level of signal.

- Mechanisms of desensitization: receptor inactivation/down-regulation, inactivation of signaling molecules, production of inhibitory proteins, etc.

Reference: [MBOC, chapter 15]

2. Signaling through enzyme-coupled cell surface receptors

   Receptor tyrosin kinase (RTK) pathway:

   - Some signal proteins that act via RTKs: many are growth factors, EGF, insulin, insulin-like growth factor (IGF1 and IGF2), Nerve growth factor (NGF), PDGF, etc.
   - Downstream signaling: Ras/Rho family
     - Adaptor proteins: those composed almost entirely of SH2 and SH3 domains help to couple activated RTKs to the important signaling protein Ras, a monomeric GTPase that, in turn, can activate various downstream signaling pathways.
     - The Ras superfamily consists of various families of monomeric GTPases, but only the Ras and Rho families relay signals from cell-surface receptors. A single Ras or Rho family member can coordinately spread the signal along several distinct downstream signaling pathways, thereby acting as a signaling hub.
     - Ras is often required, for example, when RTKs signal to the nucleus to stimulate cell proliferation or differentiation. 30% of human tumors have hyperactive mutant forms of Ras, which contribute to the uncontrolled proliferation of the cancer cells.
     - Ras activation (in R7 photoreceptor cells of Drosophila, but also conserved in mammalian cells): the RTK activates Ras-GEF (Sos) gene via an adaptor protein (Grb2), the Ras-GEF then activates Ras.
     - MAPK cascase: Ras signal is often transient. Three-component MAP kinase signaling modules: Ras $\rightarrow$ MAPKKK (Raf) $\rightarrow$ MAPKK (Mek) $\rightarrow$ MAPK (Erk) $\rightarrow$ phosphorylation of many downstream molecules, among which some stimulate cell proliferation, such as the genes encoding G1 cyclins.
     - Mammalian cells also use this scaffold strategy to prevent cross-talk between different MAP kinase modules.
     - Rho family monomeric GTPases regulate both the actin and microtubule cytoskeletons, controlling cell shape, polarity, motility, and adhesion; they also regulate cell-cycle progression, gene transcription, and membrane transport.
   - Downstream signaling: the PI-3-kinase-Akt signaling pathway. It is the major pathway activated by the hormone insulin. It also plays a key part in promoting the survival and growth of many cell types in both invertebrates and vertebrates.
     - Survival: One effect of Akt, for example, is to phosphorylate a cytosolic protein called Bad, which, in its nonphosphorylated state, promotes cell death by apoptosis. The phosphorylation of Bad by Akt creates phosphoserine- binding sites for a scaffold protein called 14-3-3, which sequesters phosphorylated Bad and keeps it out of action, thereby promoting cell survival.
     - Growth: the pathway activates a complex serine-theroine kinase called TOR (mTOR in mammalians). mTOR complex 1 promoters growth by promoting ribosome production and protein synthesis and by inhibiting protein degradation. Complex 1 also promotes both cell growth and cell survival by stimulating nutrient uptake and metabolism. The mTOR in complex 1 integrates inputs from various sources, including extracellular signal proteins referred to as growth factors and nutrients such as amino acids.

Five parallel intracellular signaling pathways activated by GPCRs, RTKs, or both (Figure 15-66):

- GPCR → G protein → adenylyl cyclase → cAMP → PKA
- (GPCR → G protein) or RTK → PLCβ → IP3 → Ca2+ → Calmodulin → CaM kinase; and PLCβ → DAG → PKC
- RTK → Grb2 → Ras-GEF (Sos) → Ras → MAPK cascase
- RTK → PI3K → PIP3 (from PIP2) → PDK1 → Akt kinase

Cytosolic tyrosine kinases:

- These tyrosine-kinase-associated receptors thus function in much the same way as RTKs, except that their kinase domain is encoded by a separate gene and is noncovalently associated with the receptor polypeptide chain
- A variety of receptor classes belong in this category, including the receptors for antigen and interleukins on lymphocytes, integrins, and receptors for various cytokines and some hormones. As with RTKs, many of these receptors are either preformed dimers, or are cross-linked into dimers by ligand binding.
- Src family: the largest family of mammalian cytoplasmic tyrosine kinases. Lyn, Fyn, and Lck, for example, are each associated with different sets of receptors in lymphocytes.
- Cytokine receptors: These receptors are stably associated with cytoplasmic tyrosine kinases called Janus kinases (JAKs), which phosphorylate and activate gene regulatory proteins called STATs. STAT proteins are located in the cytosol and are referred to as latent gene regulatory proteins because they only migrate into the nucleus and regulate gene transcription after they are activated.
- Cytokine binding alters the arrangement so as to bring two JAKs into close proximity so that they transphosphorylate each other. The JAKs then phosphorylate tyrosines on the cytokine receptors, creating phosphotyrosine docking sites for STATs. Phosphorylated STATs (by JAKs) then dissociate from receptor and dimerize, then translocate to the nucleus.
- There are at least six STATs in mammals. Each has an SH2 domain that performs two functions. First, it mediates the binding of the STAT protein to a phosphotyrosine docking site on an activated cytokine receptor. Second, the SH2 domain on the released STAT now mediates its binding to a phosphotyrosine on another STAT molecule.
- Some JAK-STAT pathways: (1) IFN-α: increases cell resistance to viral infection; (2) IFN-γ: activates macrophages. (3) Growth hormone: stimulates growth by inducing IGF1 production.

Recetpor Ser/Thr kinases: transforming growth factor-beta (TGFβ) superfamily

- TGFβ superfamily consists of a large number (30C40 in humans) of structurally related, secreted, dimeric proteins. During development, they regulate pattern formation and influence various cell behaviors, including proliferation, specification and differentiation, extracellular matrix production, and cell death. In adults, they are involved in tissue repair and in immune regulation, as well as in many other processes
- The superfamily consists of the TGFβ/activin family and the larger bone morphogenetic protein (BMP) family.
- All of these proteins act through enzyme-coupled receptors that are singlepass transmembrane proteins with a serine/threonine kinase domain on the cytosolic side of the plasma membrane. There are two classes of these receptor serine/threonine kinases-type I and type II.
- The activated type-I receptor directly binds and phosphorylates a latent gene regulatory protein of the Smad family. Once one of these receptoractivated Smads (R-Smads) has been phosphorylated, it dissociates from the receptor and binds to Smad4. Smad complex then translocates into the nucleus, where it associates with other gene regulatory proteins and regulates the transcription of specific target genes.

Reference: [MBOC, chapter 15]

3. Regulated proteolysis of latent gene regulatory proteins

   Several of these pathways depend on regulated proteolysis to control the activity and location of latent gene regulatory proteins, which enter the nucleus and activate the transcription of specific target genes only after they have been signaled to do so. Most importantly in development. Cases: Notch, Wnt, Hedgeho and NFKB.

   NFκB signaling:

   - The NFκB proteins are latent gene regulatory proteins that are present in most animal cells and are central to many stressful, inflammatory, and innate immune responses. NFκB proteins also have important roles during normal animal development (e.g. Drosophila NFκB family member Dorsal).
   - Toll-like receptors, the receptors for tumor necrosis factor α (TNFα) and interleukin-1 (IL1), activate NFκB pathway.
   - There are five NFκB proteins in mammals (RelA, RelB, c-Rel, NFκB1, and NFκB2), and they form a variety of homodimers and heterodimers, each of which activates its own characteristic set of genes. Inhibitory proteins called IκB bind tightly to the dimers and hold them in an inactive state.
   - NFκB signaling: When receptors are activated, they release NFκB dimers by triggering a signaling pathway that leads to the phosphorylation, ubiquitylation, and consequent degradation of the IκB proteins. The phosphorylation of IκB is mediated by IκB kinase (IKK). The released NFκB translocates to the nucleus and turns on the transcription of hundreds of genes that participate in inflammatory and innate immune responses.

1.8 Cell Cycle and Apoptosis

1. Apoptosis

   Caspases:

   - Apoptosis depends on a family of proteases that have a cysteine at their active site and cleave their target proteins at specific aspartic acids. They are therefore called caspase.
   - Caspases are synthesized in the cell as inactive precursors, or procaspases, which are typically activated by proteolytic cleavage. once activated, caspases cleave, and thereby activate, other procaspases, resulting in an amplifying proteolytic cascade.
   - Among the many target proteins cleaved by executioner caspases are:
     - The nuclear lamins, the cleavage of which causes the irreversible breakdown of the nuclear lamina
     - A protein that normally holds the DNA-degrading enzyme mentioned earlier (an endonuclease) in an inactive form
     - Components of the cltoskeleton and cell-cell adhesion proteins that attach cells to their neighbors
   - The two best understood signaling pathways that can activate a caspase cascade leading to apoptosis in mammalian cells are called the extrinsic pathway and the intrinsic pathway. Each uses its own initiator procaspases and activation complex.

   Extracellular pathway:

   - Death receptors: TNF receptor family. The ligands that activate the death receptors are also homotrimers, and belong to the TNF family.

- Fas ligand on Tc cells bind to Fas on the surface of target cells. the death domains on the cytosolic tails of the Fas death receptors recruit intracellular adaptor proteins (FADD), which in turn recruit initiator procaspases (procaspase-8, procaspase-7), or both, forming a death-inducing signaling complex (DISC). Once activated in the DISC, the initiator caspases activate downstream executioner procaspases to induce apoptosis.
- In some cells, the extrinsic pathway must recruit the intrinsic pathway to amplify the apoptotic signal to kill the cell
- In some circumstances, death receptors activate other intracellular signaling pathways that do not lead to apoptosis. TNF receptors, for example, can also activate the NFκB pathway, which can promote cell survival and activate genes involved in inflammatory responses.

Intrinsic pathways: activate their apoptosis program from inside the cell, usually in response to injury or other stresses, such as DNA damage or lack of oxygen, nutrients, or extracellular survival signals.

- Depends on the release into the cytosol of mitochondrial proteins that normally reside in the intermembrane space of these organelle. A crucial protein released from mitochondria in the intrinsic pathway is cytochrome c. When released into the cytosol, it has an entirely different function: it binds to a procaspase-activating adaptor protein called Apafl. The Apafl proteins in the apoptosome then recruit initiator procaspasep roteins (procaspase-9).
- A major class of intracellular regulators is the Bcl2 family of proteins, which regulate the intrinsic pathway mainly by controlling the release of cltochrome c and other intermembrane mitochondrial proteins into the cytosol. Some Bcl2 proteins are pro-apoptotic whereas others are anti-apoptotic. The two classes of Bcl2 proteins can bind to each other in various combinations, and the two inhibit each other's function. The balance between the activities of these two functional classes largely determines whether a mammalian cell lives or dies.
- Anti-apoptotic Bcl2 proteins: Bcl2 and BCL-XL, share four distinctive Bcl2 homology (BH) domains (BH1-4).
- The pro-apoptotic Bcl2 proteins: BH123 proteins. In mammalian cells, Bax and Bak are the main BH123 proteins, and at least one of them is required for the intrinsic pathway of apoptosis to operate.
- The pro-apoptotic Bcl2 proteins: BH3-only proteins, e.g. Bad, Bim, Bid, Puma and Noxa. The largest subclass, and are thought to promote apoptosis mainly by inhibiting anti-apoptotic Bcl2 proteins.
- BH3-only proteins provide the link between apopotic stimuli and the intrinsic pathway of apoptosis. Ex. in some cells, when deprived of the extracellular survival signals, an intracellular signaling pathway activates the transcription of the BH3 gene, Bim, which triggers the intrinsic pathway. Another example, P53 accumulation caused by DNA damage activates the transcription of BH3-only protens Puma and Noxa.
- BH3-only proteins may also be important to link extrinsic and intrinsic pathways. When death receptors activate the extrinsic pathway, the initiator caspase, caspase-8, cleaves Bid, and the truncated Bid translocates to mitochondria, where it inhibits anti-apoptotic Bcl2.

Survival factors: some extracellular signals inhibit apoptosis. Most animal cells require continuous signaling from other cells to avoid apoptosis.

- Ex. neuron development: the neurons compete for a limited amount of survival factors, secreted by the target cells. As a result, only a small number of cells survive, to ensure that an appropriate number of neural connections are formed.
- Survival factors usually bind to cell-surface receptors, which activates signaling pathways that suppress apoptotic program, often by regulating Bcl2 family proteins.

Reference: [MBOC, chapter 18]

1.9 Animal Development

Positional information [Wolpert, TIG, 1996]

- Problem: cells need positional information to acqure their fate. How do they know that?
- Morphogen gradient: French flag model. Positional fields of possibly multiple morphogens, which carry the positional information and determine the cell fate.
  Example: (Fig. 2) insect wing formation, dpp at the boundary of the compartments of the wing imaginal disc and hh is high at the posterior end of the wing $\rightarrow$ pattern of veins.
- Cellular history/types: French flag + Union Jack model (Fig. 3). Cells interpret positional information according to their history/types.
  Example: tissues in presumptive thigh region of chick, when put into wing bud, will develop like a leg, but not wing.
- Remark: both types of positional information interact to determine pattern formation. Initial morphogen gradients $\rightarrow$ cell types $\rightarrow$ further gradients (because different cells may expression different proteins) and further cell types (because the initial cell types could respond to gradients differently) ...

Quantative models of development: overview [Tomlin & Axelrod, NRG, 2007; Reeves & Shvartsman, Dev Cell, 2006]

- **Principle**: developmental pattern formation is driven by: (i) morphgen gradient; (ii) interpretation of the gradient.
- Identifying goals: explain some (main) pattern of observation (phenotype)
  Example 1: in segment polarity networks, the main observation is the stripes and the maintanence of stripes over developmental times. ‖
  Example 2: bcd gradient is invariant over different species (of very different sizes).‖
- Formulating models: choose what components are involved; analyze the processes; choose the mathematical framework
  Example: to explain the Dpp gradient (important in many stages of development, e.g. wing), analyze the processes that may change it, including secretion at the AP boundary, receptor binding in cell surface, endocytosis, etc. ‖
- Analyzing models: parameter fitting, model testing/selection/refinement
  Techniques/ideas for model analysis: (i) use the experimental constraints (e.g. the expression pattern of mutants) to fit parameters; (ii) dimensionality analysis that group multiples parameters to reduce number of free parameters; (iii) analyze the limiting cases of parameter values, which may make the equations tractable.
  Example: 1D diffusion: the gradient (shape of steady-state) depends on a single dimensionaless parameter: $\phi = L/\sqrt{D/k}$, where L is the egg length, D is the diffusion coefficient and k is the degradation rate. ‖
- Using and validating models: gain intuitive understanding of the model; provide access to properties that are directly measureable; prediction through experiments (e.g. mutants or other forms of manipulations)
  Note: both an application and validation of a model and the starting point for future (more powerful or more intuitive) models.
  Example 1: segment polarity network: the earlier kinetic models could be considerably simplified such as a Boolean model, to yield the same qualitative pattern. The simpler model allows insights to be extracted, e.g. the two positive feedback loops are important for the pattern. ‖
  Example 2: diffusion coefficient is generally not easy to measure, the value fitted could provide an approximate range of this parameter. ‖

Hox genes:

- Spatial colinearity: Hoxd4 in anterior, d9 in middle to posterior and d13 in posterior (tail).
- Spatial and temporal colinearity: expression pattern of Hox. In E8.5 (start activation of Hox), d4 is expressed in anterior and posterior; in E10.5 (Hox activation complete), d4 is anterior and d4-13 expressed in posterior. To summarize: anterior genes (d4) activated earlier than posterior genes (d13). This process leads to changing spatial pattern over time.
- Investigating changes of chromatin structure (reflecting regulatory activation) [Temporal dynamics and developmental memory of 3D chroamtin architecture at Hox gene loci, eLife, 2014].
  - Temporal colinearity: examine activation in tail, from ES to E8.5 to E10.5, chromatin changes from one repressive component (marked by repressive histone marks), to 2 components (both activation and repression), to 1 active component.
  - Spatial colinearity: at E10.5, chromatin structure from anterior to posterior.

1.9.1 Early Drosophila development

Reference: [Chapter 9, Gilbert, Developmental Biology, 2006]
Process:

- Cleavage:
  - Syncytial blastoderm: cell division without cell membrane formation (nulcear division), 13 cycles.
  - Cellular blastoderm: after 13 cycles, cell membrane starts to form. About 6,000 cells within 4 hours.
  - Midblastua transition: starting from cycle 10, cell division is getting slower, at cycle 14, takes 75-175 minutes. Transcription is greatly increased in this period.
- Gastrulation: starts at the time of midblastua transition.
  - Formation/segregration of mesoderm, endoderm and ectoderm from cell movement.
  - Segmentation: head end, tail end, 3 thorax segments ($T1$ to $T3$), and 8 abdomen segments ($A1$ to $A8$).

Overview of Anterior-posteior axis formation: Fig. 9.8

- Before cellularization: materal gene gradidents, gap genes, and pair-rule genes (transit patterns)
- After celluarization: segment polarity genes and homoetic selector genes (initial patterns established from the transit patterns above, then maintained by cell-cell communication, etc.)

Maternal genes: set up the gradients critical for the patterns of gap and pair-rule genes.

- Anterior and posterior organizing center: Fig. 9.11. Initially, bicoid mRNA is deposited at the anterior end, and nanos mRNA at the posterior end, hunchback and caudal mRNA uniform throughout the embryo. The gradients are set up by:
  - bicoid mRNA $\rightarrow$ Bicoid protein through translation
  - nanos mRNA $\rightarrow$ Nanos protein through translation
  - Bicoid $\dashrightarrow$ translation of caudal mRNA, thus Caudal is limited at posterior part
  - Nanos $\dashrightarrow$ translation of hunchback mRNA; in addition, Bicoid can transcriptionally stimulate hunchback mRNA, thus Hunchback is high at the anterior, but low at the posterior

- Evidence of bcd as anteror organizing center:
  - bcd^- phenotype - missing anterior structure
  - Bicoid forms anterior-to-posterior gradient
  - Change Bicoid gradient by alloing bicoid mRNA to diffuse further into posterior changes the phenotype
  - Injection of bicoid mRNA to bcd^- mutant restores the wt. phenotype.
- Terminal gene group: Fig. 9.17. torso is deposited throughout the membrane of the egg, but only activated at the two ends by Torso-like protein. The Torso signal (activated from a MAPK pathway) will inactivate the supression of hkb and tll from the transcriptional repressor Groucho.

Segmentation genes: the outcome is 15 segments - 3 head segments, Ma, Mx, Lb, 3 thorax segments, $T1-3$ and 9 abdominal segments $A1-9$, or 14 parasegments (Fig. 9.18).

- The gap genes: (Fig. 9.21) Kr, knirps, gt, tll, hkb
  - Maternal genes determine the initial pattern: high level Hunchback as activators of giant, but repressors of knirps; Caudal as activators of giant and knirps.
  - Mutual repression between gap genes: e.g. Kruppel controls the posterior boundary of hunchback and giant; while Hunchback and Giant control the anterior boundary of Kruppel.
- The pair-rule genes: (Fig. 9.24) each is expressed in 7 stripes. Divide the embryo into 14 parasegments, and each pair-rule gene is expressed in altenate segments.
  - Primary pair-rule genes: hairy, even-skipped, runt. Directly controlled/initialized by gap genes, then stabilized by interactions among themselves
  - Secondary pair-rule genes: ftz, odd-paired (opa), odd-skipped (odd), sloppy-paired (slp), paired (prd). Activated or repressed by gap, primary pair-rule genes, and themselves.
- The segment-polarity genes: (Fig. 9.25) involved in Wingless and Hedgehog signaling pathway. Expressed in 14 stripes, one in each parasegment.
  - Initial patterning: engrailed $\leftarrow$ high eve or ftz; wingless $\leftarrow$ high slp (and low eve, ftz).
  - Maintenance of patterns: Wingless diffusion to adjacent cells and activate engrailed expression via Wnt signaling pathway; Engrailed $\rightarrow$ hedgehog expression; Hedgehog diffusion $\rightarrow$ wingless expression in neighboring cells. The gradients created by diffusion are important for specifying cell fate (Fig. 9.26).

Homeotic selector genes: (Fig. 9.27) each homeotic gene is expressed in a single or a few adjacent segments and give the identity to the segment(s). Two regions in chromosome 3: Antennapedia complex (lab, Pb, Dfd, Scr, Antp) and bithorax complex (Ubx, abdA, AbdB).

- Initial patterning: controlled by gap and pair-rule genes. E.g. activated or repressed by gap genes, then boundaries defined by Eve and Ftz.
- Maintenance of patterns: through mutual inhibition. E.g. Antp expression is negatively regulated by all homeotic genes expressed posterior to it. Once stabilized, the expression pattern is "locked" into place by alteration of the chromatin conformation in these genes.

Terminal system: [Interactive Fly; Gilbert]

- Torso gradient: Torso (receptor tyrosine kinase, RTK) is expressed around the embryo, but only activated at the poles by Torso-like protein. Thus the active Torso forms a gradient peaked at the both ends of the embryo (Fig. 9.17 of Gilbert). Torso plays a role similar to Bcd, a maternal morphogen.

- The gap genes hkb and tll: activated by Torso signal (through a MAPK cascase by removing the suppression by Groucho). Both of them contains Torso response element (TorRE) in their sequences to response to Torso signal. Their roles are similar to gap genes, Kr, knirps, giant.
- The function of Tll:
  - Anterior: represses fushi tarazu, hunchback and deformed, as well as hedgehog, and helps to establish the borders of expression of head gap genes, like orthodenticle.
  - Posterior: acts on gap genes knirps, Kruppel and giant, setting up the posterior borders of expression.

JAK/STAT pathway: [Luo & Dearolf, BioEssays, 2001]

- JAK/STAT pathway in mammalian: most important in hematopoiesis and immune responses. Steps: (i) cytokine activates the receptor, which by dimerization, activates JAK; (ii) then JAK phosphorates the receptor, provides the docking sites for SH2 domain of STAT; (iii) STAT binds to the receptor and is phosphorated by JAK then forms dimers; (iv) STAT then translocates to the nucleus to activate the transcription of the target genes.
- Drosophila homolog of JAK: hop; homolog of STAT: stat92E or D-stat. Furthermore, the binding specificity of stat92E is the same as its mammalian homolog: TTCNNNGAA.
- Role of JAK/STAT pathway in segmentation: affects the expression of pair-rule genes, the best known one is eve.
  - JAK/STAT mutant: deletion of the A5 segment and partial deletion of the A4 segment. Expression of eve, runt and ftx is reduced in stripe 5, and to a lesser extent in stripe 3.
  - Two putative stat92E binding sites in eve stripe 3/7 enhancer, and if mutated, abolish the expression in stripe 3.
  - Model: JAK/STAT pathway is the activator of eve 3/7 by either (i) JAK/STAT is active everywhere because hop, stat92E mRNA is ubiqutiously expressed; or (ii) the proteins are activated by some unknown mechanism in specific stripes. The boundaries are set by other gap genes.

1.10 Misc. Cellular Processes

1. Protein synthesis

   Ribosome assembly: [Chapter 6, MBOC, V5] Figure 6-47 or [nucleolus.pdf]

   - Ribosomes consist of proteins and rRNAs. Note that a cell often needs multiple copies of rRNA genes (for proteins, one copy of mRNA can generate many protein molecules, but this is not the case for rRNA).
   - Nucleolus: ribosomes are assembled within nuclei at the site of rRNA genes, from rRNA (precursor and mature), ribosomal proteins (RP, re-imported from cytoplasm), RNA processing enzymes. Also tRNAs, in fact, the sites of non-coding RNAs.

   mRNA translation: [Chapter 6, MBOC, V5] Figure 6-66 or [mRNA-translation.pdf]. Note the structure and function of rRNA and RPs are mostly based on bacterial (the cystal structure of ribosome is available in bacterial but not in euk).

   - Ribosomes: small subunit (40S in euk) and large subunit (60S in euk). One site for mRNA, and three sites for tRNA: the A-, P-, and E-sites (short for aminoacyl-tRNA, peptidyl-tRNA, and exit, respectively). The small subunit (euk: 16S rRNA and SSU RPs) provides the framework on which the tRNAs can be accurately, matched to the codons of the mRNA, while the large subunit

(euk: 5S and 23S rRNA and LSU RPs) catalyzes the formation of the peptide bonds that link the amino acids together into a polypeptide chain. When not actively synthesizing proteins, the two subunits of the ribosome are separate.

- Four steps of translation: tRNA loading (base pairing with mRNA), peptide bond formation, move of large subunit, resetting ribosomes.
- Ribosome structure: rRNAs and not proteins are responsible for the ribosome's overall structure, its ability to position tRNAs on the mRNA, and its catalytic activity in forming covalent peptide bonds. In marked contrast to the central positions of the rRNAs, the ribosomal proteins are generally located on the surface and fill in the gaps and crevices of the folded RNA.
- Function of RPs: The main role of the ribosomal proteins seems to be to stabilize the RNA core, while permitting the changes in rRNA conformation that are necessary for this RNA to catalyze efficient protein synthesis. The proteins probably also aid in the initial assembly of the rRNAs that make up the core of the ribosome.
- Function of rRNAs: Not only are the A-, P-, and E-binding sites for tRNAs formed primarily by ribosomal RNAs, but the catalytic site for peptide bond formation is also formed by RNA. It is believed that the 23S rRNA forms a highly structured pocket that, through a network of hydrogen bonds, precisely orients the two reactants (the growing peptide chain and an aminoacyl-tRNA) for peptide bond formation.

Ribosomes in euk. cells: [Dinman, JBC, 2009]

(a) Ribosome composition: about 70 RPs, $>$ 170 non-ribosomal proteins (temporary association) and about 70 rRNAs.

(b) Ribosome biogenesis: (1) conservation from bacterial: rDNA operons - large and small unit rRNAs are co-transcribed. (2) Difference from bacterial: 3 Polymerases (I for 25S pre-RNA, III for 5S and II for RPs); assembly and localization - pre-ribosome in nucleoplasm and maturation in cytoplasm; also a number of remodeling steps are required for biogenesis (thus many proteins).

(c) Special ribosome hypothesis: RPs play distinct functional roles and there may be special ribosomes for specialized circumstances.

- RPs can have additional regulatory function: e.g. human L13a dissociation from ribosome by interferon-γ treatment and silences translation of other mRNAs; L11 is involved to signal transduction of c-Myc; etc.
- RPs and paralogs may have distince function: the mutants may have different phenotypes, localization.
- Post-translational modification of RPs: may be different for different RPs, and correspond to nutriental status.

2. Reactive oxygen species (ROS) and oxidative stress [Reactive oxygen species, Wiki]

ROS and damage:

- ROS include free radicals (unpaired electrons) and peroxides (oxygen-oxygen single bond). Include: $O_2^{\cdot -}$ (superoxide anion), H_2O_2 (hydrogen peroxide), HO_2^- (hydroperoxyl), etc.
- ROS species can cause many damages: to DNA, oxidization of AAs, oxidative degradation of lipids, etc.

Generation of ROS:

- Cellular production: during oxidative phosphorylation, O_2 receives e^- in the last step and form water. However, in about 0.1% cases, the reduction of O_2 is incomplete, and $O_2^{\cdot -}$ is generated.
- During environmental stress (e.g. UV or heat exposure), ROS levels can increase dramatically.

Dealing with oxidative stress:

- Superoxide dismutase (SOD): $2\,O_2^- + 2\,H^+ \longrightarrow H_2O_2 + O_2$. In animals, three forms of superoxide dismutase are present. SOD1 is located in the cytoplasm, SOD2 in the mitochondria and SOD3 is extracellular.
- Catalase: $2\,H_2O_2 \longrightarrow 2\,H_2O + O_2$.
- Glutathione peroxidase: reduces hydrogen peroxide by transferring the energy of the reactive peroxides to a very small sulfur containing protein called glutathione: $2\,GSH + H_2O_2 \longrightarrow GS{-}SG + 2\,H_2O$

Redox regulation: [Ansell & Adler, EMBOJ, 1997]

- Redox balance: cells need to maintain a certain ratio of $NAD^+/NADH$, which determines the rate of catabolic reactions (the generate NADH from NAD^+).
- Maintaining redox balance in aerobic organisms: shuttle system that transport cytosolic NADH to the mitochondrial, where it enters the electron transport chain. Also mitochondrial NADH oxidases also contribute to maintaining the ratio.
- Maintaining redox balance under anaerobic conditions: when O_2 is limiting, cells need an e^- acceptor to dispose of the excess reducing power. In Scer, glycerol plays a role as a redox valve.

Chapter 2

Experimental Techniques

Strategy of experiment design:

- High-level design: express a model (to be studied/tested) as a set of specific, testable hypothesis (or questions). At this level, need to decide what part of the model to be focused on, what general stretegy to apply (e.g. genetic vs. biochemical manipulations), etc.
- Detailed design: need to make specfic choices on the technical aspects of the experiments (below). Note that each choice has its limitations, e.g. a model organism may be easier to study than human, but the cross-species difference may make the results invalid to human.

Issues/considerations of experiment design:

- Systems: organisms (e.g. human vs model organism), tissues/cells (e.g. cell lines vs primary cells), in vitro or in vivo.
- Techniques: measurement of the output (e.g. reporter for transcription level), manipulation (eg. overexpression vs. knockdown), etc. Understand the limitation of experimental techniques. Ex. STARR-seq/reporter assay: sequences are devoid of the natural chromatin context inside cells.
- Confounding variables: use negative controls to remove the effect of confounding variables or use statistical methods to account for them.
- Need quality control at each step: e.g. for transformation, need to genes are successfully inserted or deleted. Positive controls: see if we can find what we expect.
- Off-target effects: experiment often involves some kind of specific treatment or detection, think about whether they work as expected, or lead to some unwanted effects. Ex. antibodies.

Lessons for experimental design [MOD project]:

- Cell population or cell lines: cells from a body are often a mixture of different cell types, while only particular cell types may be of interest.
  - Isolating cells: flow cytometry.
  - Characterize cell population: flow ctyometry, single-cell technologies.

  Some experiments may require large number of cells, e.g. ChIP-seq, so may need to immortalize cells; alternatively, some cells can proliferate in culture.
- Stages of cells/tissues: important consideration, as the stages can change transcriptome, epigenome, etc. When comparing between preterm and term birth: the transcriptomes are inherently different because of different stages.

- Verifying the quality of cells:
  - Transcriptome analysis: expression of cell-type specific markers, clustering of transcriptomes (PCs or hierarchical clustering).
  - Cellular behavior: e.g. response to a stimulant. s

Characterize function of a gene in some process/cellular behavior:

- Perturbation of genes: genetic, biochemical (e.g. chemical inhibitor), overexpression.
- Perturbation of regulators or upstream genes.
- Association of this gene with the process or other components involved in the process.
- Expression pattern: correlation with the process.
- Interaction with other genes with known roles.
- The role of homologous genes, the association of gene content with phenotypes, co-evolution with other proteins, etc.

Common strategies for experiments:

- Isolation: (1) Based on general/physical properties. Ex. column chromatography selecting electric charge, hydrophobicity, etc. (2) Interactions with specific partners. Ex. FACS, use Ab. to recognize specific markers.
- Detection: use specific markers in the targets, or use specific probes to recognize targets. Ex. microarray, sequencing (DNA sequence itself is a marker), western blot with Ab.
- Amplification: use biological replication mechanisms. Ex. DNA polymerase, cell replication.
- Manipulation: Enzymes to achieve specific goals. Vectors for delivery.

Common techniques and ideas:

- Labeling: introduce labels/tags to objects of interest. Could be done through targets or probes. Ex. tag proteins with GST, thus easier for separation. In situ hybridization: tag probes. cDNA or oligonucleotide microarray: tag target mRNAs/cDNAs.
- Use isolation for detection: enriching the targets then followed by precise detection methods. Ex. chromatin interactions by Hi-C: first do promoter capture, then sequencing. DNA methylation by Bi-S sequencing: first enrich CpG islands by digesting with restriction enzyme that recognize C-G, then select short fragments.
- Use amplification for detection: of specific things to be detected. Ex. to detect a small amount of DNA in blood, use PCR first.
- Use manipulation for detection: changing the targets so that they can be distinguished from the rest. Ex. Bi-sufite conversion for detecting DNA methylation.
- Fingerprinting: one feature may not be discriminative enough, use a combination of multiple features. Ex. protein recognization in MS (random cleavage of proteins and then identify each fragments - a fingerprint).
- Indicative products (markers): the idea is similar to finding specific markers. To detect interactions, design experiments s.t. the interaction could lead to indicative products. Ex. yeast two-hybrid for PPI.

- Screening/searching: detection at large-scale may require special techniques. Ex. affinity chromatography for finding protein inhibitors.

Strategy for designing high-throughput experiments:

- Basic strategy: with sequencing technology, one can measure many things at the same time, e.g. RNA-seq or ChIP-seq. And each of this measurement corresponds to some outcome that we are interested, say transcript level.
- Identification issue and barcode: the main challenge is that we can identify the interested variables from simultanesous measurement. In RNA-seq, the identification is provided by the sequence of transcript itself. When the variables are not directly identifed, we can use the idea of barcode or some other markers for identification, e.g.
  - STARR-seq: need to measure activities of many enhancers (targeting the same promoter/gene). To identify these targets, introduce enhancer sequence into the construct (self-transcribed enhancer).
  - Allele-specific expression: use natural polymorphism to identify transcripts.

Quality control: especially for genomics experiments [personal notes]

- Principle 1. Possible experimental errors and artefacts. Common experimental errors may include:
  - Sample quality and quantity issue: not sufficient amount of samples, for instance.
  - Sample mix-up/contamination.
  - Batch effects and other bias: e.g. ChIP-seq experiments, the read mappability is not uniform, and may depend on GC content, chromatin accessibility, etc.
  - Efficiency of treatment: experiments often involve some kind of treatment, e.g. antibody to pull down proteins, enzymatic conversions. Understand how low efficiency might affect the results.

  To deal with these issues, one main approach is: discard/filter the data of low quality. Ex. in NGS data, discard duplicate reads; adapter trimming.
- Principle 2. Use replication to assess the quality of data. Replication can mean exact replication, or mean similar results across related samples.
- Principle 3. Check biological expectation. We often expect some properties of the things we are detecting/measuring. Ex. ChIP-seq experiments: enhancers are often conserved, enriched with motifs, functionally related.
- Ex. QC of DNA-seq data in association studies:
  - Possible errors in reads: remove low-quality reads (error during sequencing process), remove duplicate reads, remove reads unaligned to genome (could be due to contamination of non-human species), adaptor trimming.
  - Possible errors in samples: remove samples with unusual number of variants (sample contamination/insufficient DNA).
  - Batch effect: between cases and controls due to different sequencing platforms.
  - Biological expectations: variant rate per sample should be similar to previous findings and similar between cases and controls. There should be fewer variants in important (highly constrained) genes.

2.1 Experimental Techniques for Cells

Reference [MBOC V5, Chapter 8]

Isolating cells:

1. Preprocessing: preprocess the cells with proteolytic enzymes to digest EM proteins; and/or EDTA or bind Ca2+ on which cell adhesion depends.
2. Methods:
   - Physical properties of cells, e.g. density/size using centrifugation
   - Protein markers: Flueorence activated cell sorted (FACS). (1) Choose a fluroescence-labelled antibody that binds specifically to a membrane proteins of cells of interest; (2) Create droplets containing a single cell; detect with fluorescence, and add different electric charges to labeled and unlabeled cells; (3) Droplets move through electric field s.t. the labeled and unlabeled cells are separated
   - Laser caputre microdissection - use laser beams to obtain a very small piece from a tissue sample under inspection

Growing cells in culture:

1. Problem: need to satisify the cell growth requirements
2. Definition: primary culture (direct from tissues without proliferation), secondary cluture (after proliferation), in vitro (cells in culture), in vivo (in body), clone (cells from a single common ancestor cell)
3. Procedure:
   - Physical requirements: cells often need to attach to solid surface, thus use the culture dish; often need EM proteins, thus add e.g. collagens in the dish
   - Growth factors: cells often need growth factor stimulation to proliferate
4. Cell lines: those that divide indefinitely. However, most vertebrate cells stop dividing after a certain number of cell divisions. To immortalize cells, do one of the two things depending on the source of replicatiue cell senescence:
   - Telomease: often suppressed in somatic cells $\Rightarrow$ telemeres get shorter after each generation. Introduce telomease to cells;
   - Cell-cycle checkpoint mechanisms: introduce oncogenes, e.g. via tumor viruses.
5. Transformed vs. nontransformed cell lines: cell lines can often be generated easily from cancer cells, these are called transformed cell lines.

Embroynic stem cells (ESC)

1. Embroynic stem cells (ESC): derived from inner cell mass of embryo, able to proliferate indifinitely while maintaining the ability to give rise to any part of the body.
2. Medical use of ES cells: a potentially inexhaustible supply of cells that might be used to replace and repair damaged mature human tissue, e.g. neurons for Pakinston's disease patients; insulin-secreting cells for diabetes; etc. The difficulties: immune rejection for ES cells from different persons.
3. Obtaining ES cells by somatic cell nuclear transplanation (therapeutic cloning): Figure 8.6. Donor diploid cell nucleus is injected into an unfertilized egg (whose nucleus has been removed), and allow proliferation. The cells from early embryo can be transferred to to culture dish to form ES cells.

Making monoclonal antibodies by hybrid cells:

- Problem: treatment of animal with a certain antigen X, the serum will contain mulitple types of anbibodies, each recognizing one part of the antigen. Ideally, one wants to obtain a single type of antibodies.
- Idea: (1) use a single B lymphocyte as the ancestor; (2) hybridize with tumor cells s.t. it can divide indifinitely

Separating cell components:

1. Goal: separate the organalles and macromolecules of cells s.t. the proteins involved in a process can be identified, their subcellular distribution be determined, etc.
2. Separating organalles: preprocess by disrupting cell membranes (e.g. vai osmotic shock or ultrasonic vibration)
    - preparative centrifugation (low resolution): the large organalles sediment the fastest under centrifugal force and form a pellet at the bottom of the tube. This can be done repeatedly s.t. each time one type of organelles are collected
    - velocity sedimentation (high resolution): during centrifugation, the organelles of different sizes move at diffferent rates and thus stop at different bands
    - equilibrium sedimentation (high resolution): centrifugation in a solution with density gradient, the organelles of different buoyant density will stop at the position where the two densities are equal.
3. Cell-free systems: purified organelles or macromolecules to study a biological process (reconstruct each biological process in vitro using cell-free systems)

2.2 Experimental Techniques for Proteins

Reference: [MBOC V5, Chapter 8]

Separating and purifying proteins by column chromatography:

1. Methods: proteins with different properties pass through the column with different speed.
    - Ion-exchange chromotography: electric charge
    - Hydrophobic chromotography: hydrophobicity
    - Gel-filtration chromotography: size
    - Affinity chromatography: binding to particular molecules, including substrates (for enzymes), anbibodies, short DNA sequences (for DNA-binding proteins), etc. Particularly useful when proteins can be labeled. If use specific antibodies, called immunoprecipitation (IP).
    - HPLC: in conventional chromotography, inhomogenity of matrix makes the flow rate uneven. In HPLC, tightly packed particles in the column, and require high pressure s.t. the solution can flow through.
2. Protein labeling by genetically engineered tags: facilitate the purification of proteins by affinity chromatography.
    - Epitope tagging: protein contains a recoginition tag, which is an antigenic determinant, or epitope. The epitope can be recognized by specific antibodies.
    - GST tagging: attach a small enzyme GST to the protein of interest (fusion protein). The affinity chromatography contains glutathione, the substrate of GST.

Detecting proteins:

1. Gel electrophoresis:
   - SDS-PAGE: separate proteins by molecular weight. Negatively-charged SDS $\Rightarrow$ proteins are unfolded (remove shape effect) and bound with SDS (the intrinsic charge of proteins is neglectable) s.t. only size of protein determines the amount of SDS bound and thus how fast it moves in the electric field. The proteins can then be deteced with a dye or silver/gold stain.
   - Western blotting (immunoblotting): to detect a specific protein, first run SDS-PAGE on samples containing this protein; then transfer the gel to a nitrocellulose paper; and probe the protein via its antibody labeled by radioactive isotope, or enzyme, or fluorescent dye.
   - 2D gel electrophoresis: one D by isoelectric focusing (pH gradient, protein will stop at where pH = pI, its isoelectric point) and the other D by SDS-PAGE
2. Mass spectrometry:
   - Protein identification: cut proteins into peptides (tryptic digestion), then determine the mass/charge ratio of each fragment. The fingerprints of all proteins can be predicted using the genome sequences. Matching the profile of a protein X with this database can determine the identity of X.
   - Sequencing proteins: each peptide is further fragmented and the masses of the fragments measured by a second MS (MS/MS).
   - Posttranslational modification (such as phosphoylation and methylation): modification will impart a characteristic mass to fragments.
3. Remark: the idea of fingerprinting. Selective cleavage of X by enzymes or chemical at specific peptide bonds, e.g. Lys-N (N is any AA) $\Rightarrow$ mass (and charge sometimes) of each fragment is a profile (fingerprint) of the protein X.

Protein-protein interactions:

1. Biochemical methods:
   - Co-immunoprecipitation (co-IP)
   - Protein affinity chromatography: a target protein is attached to polymer beads that are packed into a column.
   - Protein microarrays: thousands of different proteins or antibodies spotted onto glass slides or immobilized in tiny wells. Incubuate fluorescently labeled proteins with the array, similar to DNA microarray.
2. Yeast two-hybrid: to find binding partners of X, use X as the bait: fused with DNA-binding domain of a transcription activator. All other genes are fishes: fused with activatation domain (prey library). Introduce yeast cells with bait and fishes: if X and Y interact, then the reporter gene will be expressed. Determine the identify of Y then.
3. Optical methods:
   - FRET: X, Y are fused with different fluorescence proteins, say X - blue and Y - green. If no interaction, then stimulation of X will only emit blue; however, if interaction, then stimluation of X will emit both blue and green, because green is activated if Y is close to X.
   - SPR: allows determination of the kinetic parameters of the protein-protein interaction (the resonance angle of the light beam depends on the association/dissociation state of the two proteins)

Controlling protein activities by chemical inhibitors:

1. Problem: find small molecules that regulate/interfere with a biological process, and identify its targets.
2. Methods:
   - Screening of many small molecules in living cells or cell-free systems, e.g. inhibitor of cell cycle.
   - Target identification: use the functional molecule as the probe in affinity chromatography.
3. Applicatoin: the chemical inhibitors can be used to switch on/off protein activity.

Protein structure:

- X-ray: X-ray diffraction protein of protein crystals. The locations of atoms determine where the X-ray diffraction is stronger or weaker.
- NMR: H-atom excitation and relaxation will emit radiofrequency (RF) pulses. The frequency depends on the local environment (aa) of the H-atoms. In particular, if the two H-atoms are close, the frequency will have a small shift.
- Cryo-EM [Single-particle cryo-electron microscopy, NM, 2016]: Procedure:
  - Sample preparation: capturing the protein structure at the moment of freezing.
  - 2D electron micrographs are snapped of individual protein particles on the sample grid. Key:direct-detection cameras.
  - many different 2D views of a protein are needed to reconstruct its 3D structure.

  Conformational heterogeneity can also make high-resolution 3D reconstruction a major challenge

2.3 Experimental Techniques for DNA and RNA

Reference: [MBOC V5, Chapter 8], [Brooker, Chapter 18]
Separation and detection of DNA:

- Separating DNA molecules: similar to SDS-PAGE for protein separation, gel electrophoresis can separate DNA molecules of different size (each nucleotide carries a single negative charge).
  - polyacrylamide gel (small pores): <500bp, can detect single nt. difference
  - agarose gele (large pores): medium-size DNA, a few thousand bp long
  - pulsed-field gel electrophoresis: much larger DNA molecules, more than 1 million bp long
- DNA labeling: with radioactive iostype (32P typically); molecules that can be detected chemically or through fluorescence. This is often used for labeling DNA probes.
- Detecting DNA: create DNA probes that recognize its complementary DNA or RNA via DNA hybridization. Sometimes imperfect hybridization will be used for detecting similar molecules.
  - Northern (for mRNA) and Southern (for DNA) blotting: following gel electrophoresis, transfered to a ntrocellulose paper and then hybridize with labeled DNA probe
  - in situ hybridization: detect the localization of DNA or mRNA in cells/tissues via labeled DNA probes.

  Note: to detect DNA position in the chromosome, exposed cells to high pH to denaturate DNA double strand
- Applications of DNA detection (hybridization):

  - Detect/moninor mRNA molecules in cells
  - Intron/exon boundary: hybridization between DNA (with both introns and exons) and mRNA (only exons) molecules; then digest the single-strand DNA to get multiple exon sequences.
  - In situ hybridization for locating genes in chromosomes
  - In situ hybridization for determining the spatial expression patterns of mRNAs in cells/tissues/bodies, espeically useful for studying development

DNA cleavage and ligation:

- Cutting DNA into fragments: restriction nucleases (enzymes) that recognize particular 4-8 nts (typically paliodrome). Some cut into blunt ends, other cut DNA and leave cohesive ends. The cohesive ends from the same nuclease can be rejoined together later: first form H-bond between complementary DNA, then with DNA ligase, can ligate the DNA.
- Common restriction nucleases: EcoRI, HpaI, HindII, PstI, etc. Ex. EcoRI recognizes GAATTC (reverse complement is still GAATTC).
- Application: both as a tool to break large DNA into small fragments (s.t. they can be analyzed separately) and as a tool to manipulate/engineer DNA molecules.

Cloning DNA and RNA:

- Vectors: plamids (circular DNA) and virus for small DNA and BAC, YAC for large DNA. Vectors often contain: (1) Origin of replication: recognized by different species and different strength (copy number of plasmis), and (2) selectable markers: often antibiotic resistance genes. Classes of vectors:
  - Bacterial vectors: eg. pBluescript plasmid.
  - Mammalian vectors: ex. SV40 virus.
  - Expression vectors: contain promoters. Ex. λgt11 for expression in E. coli.
  - For large DNA molecules (>30k bp): use artificial chromosomes (YAC or BAC) that have their own telemeres, centromeres, etc.
- Suppose we want to clone beta-globin gene from rat in E. coli, the main steps: [Brooker, Figure 18.2]
  - Extract chromosomal DNA and purify from rat.
  - Cut both chromosomal DNA and vector with restriction enzymes. The vector contains amp resistance gene, and a unique restriction site in the middle of lacZ gene.
  - Ligation: mix the DNA and allow time for sticky ends to base-pair, then add DNA ligase. After ligation, we have recircularized vector (empty), vector with gene of interest and with other DNA fragments.
  - Transformation: treat E. coli so that they can take plamids. Then place them on medium containing amp (antibiotic) and X-Gal, IPTG (for lacZ).

  After the experiment, blue colonies contain only empty vectors (because it contain lacZ), and white colonies contain chromsomal DNA pieces. Need selection of gene of interest.
- Cloning mRNA: requires cDNA synthesis. (1) poly-dT primer that binds poly-A tail of mRNA; (2) reverse transcripase to synthesize complementary DNA strand; (3) digest mRNA and RNA primers with RNaseH; (4) DNA polymerase and DNA ligase to synthesize second strand.
- Expressing proteins:
  - Procedure: create expression vector (plasmid + DNA of interest + promoter) $\rightarrow$ transfection to host cells (bacterial or eukaryotic) $\rightarrow$ induce the expression of the target protein.

  - Protein purification: usually the expression vector will have protein tags such as cluster of histidine residues, which can be used to purify proteins with affinity chromatography.
  - Promoters: can be designed with specific properties, e.g. only activated by high temperature.

PCR: in vitro amplification of specific DNA sequences. [Brooker, 18.2]

- Background: DNA replication in bacterial and Euk, how to solve the issue of primer (DNA polymerase needs to bind to some DNA duplex to start the process).
  - Bacterial: replication requires primase, which synthesize RNA primer. Circular DNA, so the primer that is removed will be compensated by replication in the other direction.
  - Euk: linear DNA, so replication will cause DNA to be shorter at each generation. The problem is solved by telemere and telomerase, which synthesize the telemere sequence after each cycle.
- Idea of PCR: use heat to denaturate ds DNA, and polymerase to replicate. To solve the primer problem, put two primers in both ends of DNA. Then the ends of target DNA always get synthesized by base-pairing with primers added. Also need two primers: because in each replicate, one primer is used for starting replication (forward primer) and the other primer gets replicated (reverse primer), which will be used for next round of replication.
- Procedure [Brooker, Figure 18.5]: choose primers that flanks the region of DNA of interest. The primers should be chosen s.t. only specific sequences will be amplified. Three key steps in each PCR cycle: denaturation (by heat) $\rightarrow$ primer annealing: at a lower temperature, allow primers to bind to ss DNA $\rightarrow$ primer extension: DNA replication catalyzed by a DNA polymerase that is stable in high temperature (Taq polymerase).
- Advantages of PCR: (1) Extremely sensitive, can detect single DNA molecule in a sample; for mRNA cloning, use reverse transcriptase first to obtain ds DNA. (2) Can use PCR to purify a DNA: start with a mixture of different molecules, after 20-30 rounds, the DNA of interest will dominant.
- Real-time PCR to quantify the amount of DNA: the key idea is to use a reporter that emit fluorescence only when binding to target DNA.
  - TaqMan detector: oligont. complementary to target DNA, reporter and quencher. Normally, no fluorescence.
  - Detection: in primer annealing step, TaqMan binds to target DNA; in primer extension step, it is removed and cut into single nt., emitting fluorescence.

  Three phases of PCR: exponential, linear and plateau. When it reaches the linear phases depends on the starting materials. So quantification based on cycle threshold (when fluorescence becomes detectable).

DNA libraries:

- Genomic library: random fragmentation of genomes by restriction nucleases and insertion into cloning vectors. Transfoormation of bacterial.
- cDNA library: the set of expressed mRNA molecules $\rightarrow$ cDNA, then cloning into vectors. To clone cDNA, need to first add linker DNA (recognized by restriction enzyme), and cut cDNA and plasmid.
- Colony hybridization [Brooker, Fig. 18.12]: use radiolabeled DNA probes to select the colony containing the desired DNA. First make a replica of the colony in membrane, then treatment to create ss DNA and make membrane permeable. The DNA probe then hybridize with ss DNA and reveal the locations of desired colony.

DNA sequencing:

- Dideoxy method: in vitro DNA synthesis in the presence of chain-terminating dideoxyribonucleoside triphophate (dATP, dGTP, dCTP, dTTP). Suppose dATP is present, then the synthesis of DNA will create all fragments that end at an A position. The band of the DNA fragments in the gel will indicate where are the A nts in the DNA molecul.
- Genome sequencing: shortgun method. Random fragmentation of genome, sequencing of each fragments and assembly.

Site-directed mutagenesis

- Idea: use oligonucleotide with the mutation as primer to synthesize new DNA containing the mutation.
- Procedure: ss DNA, add oligo with mismatch as primer, add DNA polymerase, dNTP and DNA ligase. After replication, add to live cells to allow mismatch repair. The resulting clones can be identified via sequencing.

Introducing DNA into cells. We focus on animal Euk. cells.

- Transfection: non-viral mediated method for introducing nucleic acids into cells. Typically involves opening transient pores or "holes" in the cell membrane to allow the uptake of material. Transfection is usually transient; to obtain stable transfection, a marker gene is co-transfected, which gives the cell some selectable advantage. Then by chance, the foreign material is integrated into the genome and can be selected.
- Transduction: the process by which DNA is transferred to a cell by a virus.
- Remark: the term "transformation" is used to describe bacterial cells taking up foreign genetic materials. For animal cells, it means progression to a cancerous state.

2.4 Studying Gene Function

Reference: [MBOC V5, Chapter 8]

Forward genetics: phenotype to genes

- Mutagenesis: create a large number of random mutants and then find those with interesting behavior.
  - Chemical agents or radiations
  - Insertional mutagenesis: randomly insert a tagged DNA fragment (virus or transposon) to the genome, which may disrupt the function of some gene. The main benefit is: the genes mutated can be easily identified via the tag. Ex. P element for Drosophila mutagenesis.
- Genetic screen: find the mutants of interesting phenotype.
  - Need to have some test for those phenotype not directly observable, e.g. the deficiency in learning and memory.
  - It is often helpful to examine the intermediate-level phenotype, e.g. the change of anatomical structure of an orangism, or the change of gene expression patterns in mutants.
  - Conditional mutants: only behave abnormally under restrictive conditions, e.g. high temperature (temperature-sensitive mutations).
- Type of mutations: loss-of-function and gain-of-function.
- Identification and analysis of genes:
  - Genetic linkage: locate the genes in the mutant via linkage analysis. The markers can be SNPs or other forms.

  - Complementation test: test whether two mutants with similar phenotypes are in the same gene or not. One normal copy will restore the function of a deficient copy, so if two mutant parents are a/a;B/B and A/A;b/b, then the offspring a/A;B/b will be normal if they are in different genes.
  - Epistasis analysis: test how two related genes interact and affect the phenotype. Ex. in a linear pathway, the order of gene action can be determined by epistasis analysis: the phenotype of double mutant is the same of the phenotype of the mutant of the upstream gene.

Transgenic animals and mutants [Brooker, Chapter 19]:

- Gene replacement or insertion in the germ line cells ⇒ organisms whose genotype has been permanently changed, called transgenic organisms.
- Application: (1) Loss-of-function mutation to study gene function. (2) Overexpression of a protein to study its function. (3) Introduce external protein: e.g. GFP with condition-specific promoter as a maker (for instance, detection of some chemicals in the environment).
- Bacterial and yeasts: electroporation (electric shock) or other techniques that renders the membrane temporarily permeable. Then introduce engineered mutant gene into the cell and by homologous recombination, it will replace the normal gene.
- Transgenic flies: insert DNA mutant via P element (transposon) into germ-line cells via microinjection.
- Plant cells: particle bombardment (hit the cell will particles containing mutant genes)
- Gene knockout: The challenge with large mammalian genome is that introduced genes are inserted into the genome by non-homologous recombination 99.9% of time, instead of replacing the endogeneous copy. Replacing an endogenous gene (both copies) through transgenic mice [Transgenic animals, ELS, Figure 8-65], [Brooker, 19.2].
  - Construct: the gene of interest contains a Neomycin-resistance gene in the middle (thus LOF), and a TK gene in the clone.
  - Introduce DNA mutant into ESC through some vector: (1) Gene insertion: cells express TK gene, and will be killed; (2) Gene replacement through homologous recombination: no TK and neomycin-resistance gene is expressed, cells survive. In general, we need positive selection: those where the vector has been incorporated in the genome; and negative selection: those that have not undergone homologous recombination.
  - Inject these ES cells into an early mouse embryo and the resulting "chimera" mouse will have some mutant in its germ cell lines ⇒ heterozygotes in the pool of germ cells.
  - Inbreeding between heterozygote offsprings of the above individual ⇒ 1/4 chance of getting a homozygote mutant.
- Gene knock-in: to insert a gene of interest into ESC at targeted locations. Flank the gene of interest with non-critical sites. Thus the gene will be inserted via homologous recombination at the non-critical sites.
- Questions:
  - In gene knockout, how is the Neo resistance gene expressed?
  - Gene knock-in: how to prevent non-homologous recombination and integration to random genomic locations?

Creating loss-of-function mutations:

- Gene replacement (below).

- Dominant negative mutations: gene replacement is much more difficult than gene addition in higher organisms. Thus introduce a mutant copy via gene addition whose expression will inhibit the widetype protein. For instance, the mutant may have a binding but not functional domain, thus will interfere with the normal function of the wt. protein.
- Site-directed mutagenesis: instead of putting a complete mutant in the loss-of-function mutations, put a mutant with only one or a few mutations.
- Conditional mutants: the mutation will take effect only under controlled conditions (e.g. certain time or tissue type) via inducible promotoers. Ex. Cre/lox system: in transgenic mice, the mutant has a fully functional gene flanked by lox sites. It is mated with transgenic mice that express Cre recombinase gene under an inducible promoter. Then in the offspring, one can switch on/off Cre to selectively excise the functional gene.

RNA interference (RNAi): foreign dsRNA will inhibit the activity of endogenous mRNA via RNA interference mechanism. RNA interference works by: the degradation of foreign dsRNA by RNase; degraded RNA fragments will hybridize with mRNA; and the short region of dsRNA will be degraded by RNase thus destroying the endogenous mRNA.

Gain-of-function mutations:

- Overexpression: introduce either many copies of the gene or a strong promoter s.t. the gene is expressed in much higher levels
- Misexpression: change the regualtory sequences of the gene s.t. its expression pattern is changed

Monitor gene expression pattern:

- Monitor gene expression:
  - Reporter gene: replace the coding protion of a gene with a reporter gene (commonly GFP or beta-galactosidase) and then introduce the recombinant DNA into the cells.
  - In situ hybridization
  - RT-PCR: mRNA to cDNA via reverse transcription. Then PCR on gene of interest, with dyes. The fluroescence level can be used to infer the starting mRNA concenctration (considering the PCR cycles).
  - Single cell measurement: using a fluorescent reporter protein whose expression is under the control of a promoter of interest. Quantify the expression level by fluorescence microscope or flow cytometry.
- DNA microarray:
  - cDNA or oligont. chip: probes for mRNA
  - Preparation of sample: mRNA to cDNA with reverse transcriptase, and the cDNA is labelled. In two-channel microarray, one sample is labeled with one color, and the reference sample is labeled with the other color
  - Hybridization: the expression (difference between two samples in the case of two-channel microarray) of all genes with probes in the chip

Genotyping of yeasts using microarray [Winzeler & Davis, Science, 1998]

- Motivation: determine the alleles of a large number of genetic markers across the genome.
- Background: genetic variation across yeast strains estimated to be up to 1% of the genome. Partial sequencing reveals one instance of allelic variation every 160 bases in two yeast strains. Yeast genome size is about 12M, thus the total number of allele variations is about 75k.

- Idea: the hybridization effiency of different strains on the common probe are different.
- Methods:
  - Microarry: a large number of 25-bp probes (157K?) on oligonucleotide array. Ordered in the chromosome positions.
  - Detection of polymorphism: suppose two strains are A and B, in most cases, the probe hybridizes with equal efficiency; in some cases, with different efficiencies - polymorphism, due to single substitutions near the center of probes, or from deletion.
  - Polymorphism markers: the total polymorphisms revealed: 3714 markers (covering about 4.7% of estimated variation), spaced about every 3.5 kilobases.

Protein-DNA interaction:

- Gel retardation
- DNA footprinting: digest DNA with some chemical $\rightarrow$ DNA fragments of all sizes. If X interacts with some region of DNA, then when X is present, that region of DNA is protected, and thus the corresponding DNA fragments will be missing

Chapter 3

Yeast

3.1 Physiology of Yeast

Common features of Fungi: [Fungus, Wiki]

1. A large kindom, separated from animals, plants, bacteria. Including yeasts (unicellular), mold and mushrooms.
2. Fungi lack chloroplasts, requiring preformed organic compounds as energy sources.
3. Fungi possess a cell wall (containing chitin vs celluose in plant) and vacuoles.
4. Fungi reproduce by both sexual and asexual means, and produce spores.
5. Morphology and growth:
    - The cells of most fungi grow as tubular, elongated, and thread-like (filamentous) structures and are called hyphae, which may contain multiple nuclei and extend at their tips. In most fungi, hyphae are the main mode of vegetative growth, and are collectively called a mycelium; yeasts are unicellular fungi that do not grow as hyphae.
    - Dimorphic fungi can switch between a yeast phase and a hyphal phase in response to environmental conditions. E.g. C. albicans.
6. One Phylum is Ascomycota, which include many yeasts.

Yeast ecology: [Yeasts, ELS]

- Associated with plants: sap exudates of certain trees (heavily infected with bacterial and yeasts); rotting tissues; decaying fruits and berries are rich sources; also in leaves and flowers of higher plants.
- Associated with insects: insects are the most important vectors in the distribution of yeasts. Some insects bore tunnels directly into the sapwood or hardwood of weakened trees and they carry symbiotic fungi and yeasts.
- Intracelluar symbiotes: inside specialized insect cells (e.g. mid gut).
- Soils: the organic matter derived from plant or animal residues.
- Warm-blooded animals: they often have the ability to grow in 37 degree. Some have unusally stringent growth requirements; some nonpathogenic are also common in environment. C albicans: opportunistic pathogen that cause infection in susceptible patients, usually in the skin or mucous membrane. C. glabarate: often with urinary tract infection.

- Within a microbial community, one group may supply nutrients to another group by enzymatic breakdown of host tissue, e.g. celluose that cannot be utilized by yeasts.
- Ecology of Scer: found in the musts (grape juice), but not in soils.

Life cycle: [Mating of yeast, Wiki]

- S. cerevisiae (yeast) can stably exist as either a diploid or a haploid. Both haploid and diploid yeast cells reproduce by mitosis, with daughter cells budding off of mother cells. While most yeasts reproduce by budding, some such as S. pombe reproduces by binary fission.
- Pseudohyphal growth: A pattern of cell growth that occurs in conditions of nutrient limitation (nitrogen limitation, poor carbon sources such as lactate, etc.). Cells become elongated, switch to a unipolar budding pattern, remain physically attached to each other, and invade the growth substrate (filamentous growth). Pseudohyphal invasion probably reflects the search for nutrients, and escape from a potentially harmful environment.
- Complete nutrient depletion (nitrogen limitation and absence of glucose) induces another response, which varies in haploid and diploid yeast. Haploid yeast enter a nonmetabolic and quiescent phase, called G0. Diploid yeast undergo a differentiation pathway called sporulation. They can undergo meiosis to produce four haploid spores: two a spores and two α spores.
- Haploid cells are capable of mating with other haploid cells of the opposite mating type (an a cell can only mate with an α cell, and vice versa) to produce a stable diploid cell.

Cell cycle: [Budding Yeast Cell Cycle Model - Basic Mechanism, John Tyson's page]

1. The drivers of cell cycle: cyclin/Cdk complexes, where cyclins are regulatory units, and Cdk are catalytic units. There are nine cyclins: Cln1-3 (G1 cyclins) and Clb1-6 (B cyclins); and one Cdk: Cdc28.
2. Two main checkpoints:
    - G1 checkpoint: If DNA damage is detected, mating pheromone is present, or the cell has not reached the critical size, the cell arrests in G1 and is unable to undergo the Start transition which commits the cell to a new round of DNA synthesis and mitosis.
    - The spindle assembly checkpoint: If DNA damage is detected, DNA is not replicated completely, or chromosomes are not aligned on the metaphase plate, the cell arrests in metaphase and is unable to undergo the Finish transition, whereby sister chromatids are separated and the cell divides.
3. The main mechanism: positive regulators and negative regulators. They mutually antagonize each other, thus cannot coexist. The system exists in one of thw steady states: one where the negative regulators dominate (G1), and one where the positive regulators dominate (S/G2/M). The cell cycle thus consists of two transitions between the two states: START (G1 to S/G2/M) and FINISH (S/G2/M to G1).
    - Positive regulators: Clb/Cdk, inactivate Cdh1 and destabilize CKIs.
    - Negative regulators: Cdh1/APC which degrades Clbs, Sic1 and Cdc6 (CKIs) inhibit Clb/Cdk complexes.
4. START transition: facilitated by Cln/Cdk. Cell growth (critical size) $\rightarrow$ (G1 checkpoint) Cln/Cdk complexes $\rightarrow$ phosphorylatio and inactivation of Cdh1 and CKI $\rightarrow$ bud emerges, Clb/Cdk increases (because of lower Cdh1 and CKI) $\rightarrow$ turn off Cln synthesis, in preparation for the FINISH transition.

5. FINISH transition: facilitated by Cdc20. Clb/Cdk and lifting of spindle assembly checkpoint (full DNA replication and chromsome alignment) → Cdc20 activation → activation of Cdh1 and CKI → decrease of Clb/Cdk activities, meanwhile, this allows Cdc20 activity to drop off after FINISH (since Clb/Cdk are activiators of Cdc20), preparing for the next START transition.

6. Downstream effectors (TFs) of G1 and B cyclins: [Spellman & Futcher, MBC98]

    - MBF (Swi6/Swi4) and SBF (Swi6/Mbp1): activated postranscriptionally (phosphorylation and nuclear localization) by Cln3/Cdc28, and induce expression of G1/S genes (budding and DNA synthesis). Also SBF is inactivated by Clb2/Cdc28 in late cell cycle (no longer needed).
    - Mcm1/Ste12: induce expression of early G1 genes.
    - Mcm1/SFF (unidentified TFs): positive feedback with Clb2/Cdc28, and induce expression of Clb1/2, Bud4, Swi5 in M phase.
    - Swi5/Ace2: activated by Mcm1/SFF, and induce expression of M and M/G1 genes.

Mating of yeasts: [Mating of yeast, Wiki]

- Mating: a cells produce a-factor, a mating pheromone and respond to α-factor through the member receptor (Ste2). Simlarly, α cells produce α-factor and respond to a-factor through the receptor (Ste3). Two haploid yeast of opposite mating types secrete pheromones, grow projections and mate.

- Determination of mating types: The phenotypic difference is due to the different transcriptional program, caused by one of two alleles at the MAT locus, MATa (MATA) or MATα (MATALPHA). E.g. MATa will activate expression of Ste2 and repress Ste3. The alleles present at the MAT locus are sufficient to program the mating behaviour of the cell.

- Determination of haploid vs diploid cells: Haploid cells of both mating types share a haploid transcriptional pattern which activates haploid-specific genes (such as HO) and represses diploid-specific genes (such as IME1). Similarly, diploid cells activate diploid-specific genes and repress haploid-specific genes. After mating, the combination of the information encoded by the MATa allele (the a1 gene) and the MATα allele triggers the diploid transcriptional program.

- Mating type switching: Wild type haploid yeast are capable of switching mating type between a and α. Each yeast cell contains a silent (not transcribed) copy of both a and α alleles at HMR and HML loci, respectively in chromosome 3. During G1 phase, the MAT locus is removed by the HO gene (DNA endonuclease), then the opposite allele at the silent locus will fill in the gap (i.e. HMR will replace the removed MATα and HML replace MATa).

Pheromone response: [Yeast pheromone response model, Google]

- Receptor activation: α-factor binding of Ste2 (GPCR) causes dissociation of Gpa1 (G protein alpha subunit), and Ste4-Ste18 (beta and gamma subunits, respectively).

- MAPK pathway: Ste4-Ste18 dimer binds to the scaffold protein Ste5, and to the kinase Ste20, causing activation of a MAP kinase cascade (Ste11, Ste7, Fus3)

- Fus3 and Kss1 phosphorylation leads to nuclear localization, and they phosphorylate Dig1 and Dig2. Then Dig1 and Dig2 dissociate from Ste12 (the Dig1-Dig2-Ste12 trimer binds to promoters inactively without Dig1/2 phosphorylation). The TFs Ste12-Tec1 then activates transcription of genes involved in mating, polarization of cell growth, and ultimately cell and nuclear fusion.

White-opaque switching in C albicans: [Lohse & Johnson, COM, 2009]

1. White-opaque switching: two types, white and opaque. Each cell type is heritable over many generations and switching between them is stochastic, but influenced by environmental cues.

2. White-opaque switching and mating:

   - Normally, cells in white state with both a and α alleles (tetraploid). The $a1-\alpha2$ heterodimer represses mating functions as well as white-opaque switching.
   - The white cells may loss heterozygosity (similar to sporulation?) and become whilte a and α cells.
   - The swithing from white to opaque states: enable a and α cells to mate and form a/α cells (in white state).

3. The maintanence of the states and switching: Wor1 is the critial regulator, high Wor1 expression leads to the opaque state. Several positive transcritpional feedback loops may stablize Wor1 expression.

4. Properties of white and opaque cells and the switching:

   - Opaque cells may evade innate immune system: less suspectible to phagocytosis by macrophages. White cells are better suited to internal infections while opaque cells in skin infections.
   - Oxidative stress: induce switching from white to opaque. High temperature (37 degree): white cells more stable; however, this can be overcome by anaerobic condition, which favors opaque cells. CO_2 also stimulates whilte-to-opaque switching.

5. Opaque cells can induce biofilm formation of white cells: which respond to pheremone even though cannot mate. The pheromone response occurs through the same pathway as mating, but fewer genes are induced.

6. Evolution of white-opaque switching: Wor1 homolog in other yeast species play a role in regulating morphological changes.

Comparison of yeasts:

1. Sexuality: [Dujon & Souciet, Nature, 2004]

   - S. cerevisiae: the predominantly diplontic cycle is pseudo-heterothallic owing to mating-type switching.
   - C. glabrata displays no known sexual cycle, similar to most pathogens.
   - K. lactis is a heterothallic species with a predominantly haplontic cycle.
   - Y lipolytica has a haplo-diplontic cycle (that is, it alternates between haploid and diploid phases of similar importance).

2. Crabtree effect: positive in all post-WGD yeasts. Also in Spom, but negative in most yeasts.

3.2 Yeast Metabolism

Carbohydrate metabolism of Scer: [Jones et al, The Molecular and Cellular Biology of the Yeast Saccharomyces, 1992]

1. Growth under fermentable carbon source (glucose, galactose, mannose, etc.): uses fermentation to derive energy, regardless of O_2 condition. The ability to ferment in the presence of O_2 is called the Crabtree effect.

   - Alcoholic fermentation: pyruvate $\longrightarrow$ acetaldehyde by pyruvate decarboxylase (PDC), also require cofactor thiamine pyrophosphate (TPP). aceyaldehyde $\longrightarrow$ ethanol by Alcohol dehydrogenase (ADH1 in Scer). The main function of this pathway is to regenerate NAD^+, which is used in glycolysis.

- Main branch point: pyruvate $\longrightarrow$ oxaloacetate $\longrightarrow$ TCAcycle through pyruvate dehydrogenase (PDH); and pyruvate $\longrightarrow$ aceyaldehyde through PDC.
- Mechanism of Crabtree effect: two hypothesis (1) glucose repression of expression of respiration genes; (2) PDC is more effective in driving the flux from pyruvate to ethanol (thus outcompetes PDH).

2. Glucose is the preferred carbon source: other sugar carbon sources have to enter through some pathways (see below for nonfermentable carbon source). Ex. fructose can enter glycolysis through phosphorylation (fructose-6-Pi); galactose needs to be converted to glucose-6-Pi; and mannose needs to be converted to fructose-6-Pi.

3. Growth under nonfermentable source (glycerol, ethanol, lactate, etc.): uses respiration to derive energy. When glucose is depleted, cells switch to non-fermentable source (e.g. ethanol accumulated during fermentation), this is called diauxic shift.

   - Ethanol: ethanol $\longrightarrow$ acetaldehyde via Adh2, and acetaldehyde $\longrightarrow$ acetate. Note that in yeast, there are four ADH enzymes, Adh1 and Adh2 are cytosolic (with opposite functions), and Adh3/4 mitochondrial. Both Adh1 and Adh3 are induced by glucose, but may supply NAD^+ in different pools, cytoplasm and mitochondrial respectively. Adh2 is repressed by glucose (more than 200 fold).
   - Acetate: acetaldehyde $\longrightarrow$ acetyl$-$CoA $\longrightarrow$ TCAcycle.
   - Glycerol: enter glycolysis.

4. Other related pathways: alternative usage of glucose, and glyconeogenesis

   - Storage carbohydrate: glucose may be converted to glycogen to provide internal energy store in case of: adaptation to respiratory growth, sporulation, etc.
   - Glyoxylate cycle: In cell-wall containing organisms, in the absence of available carbohydrates, the glyoxylate cycle permits the synthesis of glucose from lipids via acetate generated in fatty acid β-oxidization. Specifically, acetate $\longrightarrow$ glyoxylatecycle $\longrightarrow$ glucose $\longrightarrow$ polysacchride, the product can be used for cell wall biosynthesis.

5. Glucose repression:

   - Utilization of alternative carbon source: including fermentable (e.g. galactose) and nonfermentable (ethanol, acetate, etc. such as Adh2). Adr1 is a positive regulator of enzymes in utilization of nonfermentable carbon source (ethanol, glycerol, fatty acid, etc.). In particular, Adh2 expression is induced by Adr1.
   - TCA cycle, oxidative phosphorylation.
   - gluconeogenesis, sporulation, etc.

6. Glucose induction:

   - Glycolysis and fermentation: Pgk1 (main enzyme of glycolysis), Pdc1, Adh1. All of them contain Rap1 binding sites. Also Gcr1 is an activator of glycolysis enzymes.
   - Glucose transporters.

Carbohydrate metabolism of Klac: [Breunig & Steensma, Topics in Current Genetics, 2003, Chapter 6]

1. Growth under glucose: Crabtree-negative. Under high glucose condition, respiration in the presence of O_2, and fermentation with low O_2. Generally, Klac cannot grow in the complete absence of O_2.

- Requirement of fermentative growth: since glucose is used inefficiently, need a high glycolytic flux to sustain growth. This poses a problem for strains incapable of acquring high rate of glucose transport. In particular, some strains are sensitive to antimycin A, the respiration inhibitor (Rag- phenotype), and some strains are resistent (Rag+ phenotype).
- Glucose uptake: only four hexose transporters are found in Klac. Some strains have a single low-affinity transporter, Rag1 and other strains have two tandem copies, Kdh1/2. The Rag1 strains are sensitive to antimycin A, while the Kdh1/2 strains are resistent. The other glucose transporter is Hgt1, a high affinity transporter.
- Pentose phosphate pathway: high activity and the NADPH generated can be reoxidized by the NADPH dehydrogenase (not found in Scer).
- Fermentation: four ADHs (two cytosolic, KlAdh1/2; and two mitochondrial, KlAdh3/4). KlAdh4 is used for ethanol consumption. The multiple copies of ADHs might reflect the necessarity to control the level of highly toxic acetaldehyde.

2. Growth under other fermentable carbon source:

   - Kluyver effect: cannot grow anaerobically on some sugars (which can be used if O_2 is available) such as maltose.
   - Lactose utilization: lactose is the natural carbon source for Klac, lactose $\longrightarrow$ lacticacid, rare among yeasts. Lactose metabolism is induced by lactose or galactose in the medium through KlGal4. The enzymes are LAC4 and LAC12, which confer lactose utilization if transferred to Scer.

3. Growth under non-fermentable carbon source:

   - Ethanol: the main enzyme for ethanol consumption is probably KlAdh4, which is strongly induced by ethanol. The acetaldehyde formed will be oxidized to acetate via acetaldehyde dehydrogenase.
   - Acetate: synthesis of acetyl-CoA from acetate via acetyl-coenzyme A synthetase. Acetyl-CoA can then enter TCA cycle.
   - Additional requirement for growth on ethanol or acetate: replenish oxaloacetate via the glyoxylate cycle; gluconeogenesis for building cell wall and storage carbodydrate.

4. Transcriptional regulation:

   - Glucose induction of sugar transport and glycolysis: Rag4 (combines the function of Snf3 and Rgt2) for induction of Rag1 and Kht1. Similar to Scer, Gcr1 and Rap1 may activate glycolysis (the DNA binding domain of Rap1 has diverged though).
   - Glucose repression of alternative carbon source: the glucose uptake rate has a strong impact on glucose repression. The presence of Kht1/2 correlates with high glucose repression than Rag1 strains. Mig1 may be involved in glucose repression.
   - Fermentation and respiration: (1) fermentation is activated at low O_2, not clear how it is activated. KlPDC is auto-regulated, induced by glucose and low O_2 and repressed by ethanol. (2) Another mechanism is Hap2/3/4/5 complex, which is important in Scer to induce respiration genes in the presence of non-fermentable carbon source, but not in Klac.

Redox regulation in Scer: [Ansell & Adler, EMBOJ, 1997]

- Response to low O_2: the sensor of O_2 is heme. A set of hypoxic genes encoding selected enzymes in heme, sterol and fatty acid biosynthesis is repressed under aerobic conditions and induced when oxygen is limiting (compensate for restrictive oxygen availability). Other genes induced at low oxygen tension encode replacement functions that might increase the respiratory capacity of the cells under oxygen limitation.

- Gpd1 and Gpd2: NAD^+-dependent glycerol 3-phosphate dehydrogenase, catalysing the first step in glycerol production.
- Glycerol plays two important roles: (1) osmotic stress, e.g. from dehydration, as a solute; (2) redox regulation, as e^- acceptor to dispose of excess NADH.
- Gpd1 expression is induced by osmotic stress through HOG pathway, and Gpd2 induced by low O_2 (which will generate NADH from glycolysis). The enzymatic functions of the two genes are similar (as overexpression of either one can compensate for the other), but expression patterns are different. Gpd2 induction does not depend on Rox1 and Rox3, the regulators of hypoxic response.

Fermentative lifestyles of yeasts: Saccharomyces complex (including K lactis) [Merico & Compagno, FFBSJ, 2007]

1. Fermentation:
   - Most species possess good fermentative ability, as most can grow well in the presence of antimycin A (respiration inhibitor) under rich medium.
   - Most pre-WGD species show a reduced Crabtree effect, including Klac.
   - Some species show severely impaired growth in aerobic condition in the presence of antimycin A, but grow well in strict anaerobic condtions. This is probably due to the Pasteur effect: inhibition of fermentation by oxygen.
2. Growth in anaerobic conditions:
   - The major problems are: (1) produce energy through fermentation; (2) redox balance; (3) metabolic pathways that may depend on oxygen such as biosynthesis of heme, NAD and uracil.
   - Most pre-WGD species need some oxygen for growth (aerobic), while most post-WGD can grow well in the absence of O_2. The post-WGD species probably adapt to both fermentation and low oxygen condition (they may compete with bacterial for sugars, in the absence of oxygen).
3. Klac vs S. Kluyveri (Sklu) vs K waltii (Kwal): Klac is a strict aerobic, but Sklu and Kwal (especialy Kwal) grow well in anaerobic condition.

3.3 Gene Regulation in Yeast

Key transcriptional regulators:

- Regulation of growth (in response to nutrients and stresses):
  - Rap1, Sfp1, Crf1
  - Dedicated RP regulators: Fhl1, Ifh1.
- Regulation of AA biosynthesis: Gcn4.
- Protein folding and heat shock response: Hsf1
- Cell cycle:
  - Chk protein kinases: Chk1, Rad53, Dun1
  - Transcription factors: Fkh2, Swi4, Swi5, Swi6, Ndd1, Ace2
- Regulation of lipid metabolism:
  - Ino4, Ino2: derepression of inositol-choline-regulated genes involved in phospholipid synthesis.

- Mating type:
  - MATa, MATα.
  - Pheromone receptors: Ste2, Ste3, Gpa1, Ste18 (G protein beta), Ste4 (G protein gamma)
  - Pheromone response: MAPK pathways - Ste5, Ste20, Ste11, Ste17, Fus3, Kss1; transcritpion - Dig1, Dig2, Ste12, Tec1.
  - Mating type switching: HML, HMR, HO.
- Environmental stress response (ESR): common to different stresses.
  - Msn2/4: TFs, translocation from the cytoplasm to the nucleus in stress conditions.
  - Rpd3 (HDAC)
- Osmotic stress:
  - Sensing stresses: Sln1, Msb2/Sho1 (changes in membrance fludity, osmotic shock).
  - Signaling pathways: Ste20, Ste11 (also in pheromone response), Pbs2
  - Hog1 (MAPK) and Sko1 (TF)
- Oxidative stress response: Yap1
- DNA damage response:
  - Sensing DNA damage: Tel1, Mec1.
  - Rfx1: DNA damage and replication checkpoint pathway. Repressor of RNR complex.
- Other TFs:
  - Mcm1
  - Cad1: stress responses, iron metabolism, and pleiotropic drug resistance

Glucose signaling in yeasts [Santangelo, Microbiology and Mol Biol Reviews, 2006; Sabina & Brown, Euk Cell, 2009]

- Transcriptional responses of glucose signaling:
  - Induction: acquision and utilization of glucose, including glucose transporters, glycolysis and RP genes. About 1000 genes with more than 2-fold change.
  - Reduction: genes not need when glucose is abundant, including gluconeogenesis, alternate carbon source metabolism and respiration. About 1000 genes show great than 2-fold change.
- SRR (sugar-receptor receptor) pathway: induction of glucose transporters
  - Signaling: Snf3/Rgt2 transmember proteins (homolog of HXT hexose transporters, but lose the ability to do transportation). Snf3/Rgt2 activates Yck1/2 (kinases), which phosphorylates and inactivates Std1 and Mth1, repressors of Rgt1.
  - Response: the active Rgt1 turns on HXT genes.
- Glucose repression pathway:
  - Signaling: increase of intracellular level of AMP/ATP (through glycolysis) activates Glc7-Reg1 phosphatase complex, which turns off the repressor Snf1 (The phosphorylated Snf1 is able to form complex with Snf4/Gal83 and inactivates Mig1). It is also possible that Snf1/Gal83 state switch is triggered by glucose via PKA pathway (perhaps through Hxk2, hexose kinase, a fraction of which localizes to nucleus upon glucose signal and interacts with Mig1 and Med8 - transcriptional complex).

  - Response: The activated Mig1 represses genes involved in gluconeogenesis and alternative carbons source metabolism.

- cAMP/PKA pathway: activated by RAS or GPA. Overexpression of Ras2 or Gpa2 achieves almost identical transcriptional response. The two branches converge at adeynl cyclase (Cyr1) and cAMP/PKA (Tpk1/2/3 for catalytic subunits and Bcy1 for regulatory subunit). [glucose_sensing.jpeg]
  - Ras branch: monomeric GTPase (Ras1/2). Its state (GTP-bound, active or GDP-bound, inactive) is controlled by Cdc25/Sdc25 (RasGEF, guanine nucleotide exchange factors) and Ira1/2 (GAP, GTPase-activating proteins). It is not clear how Ras is activated by glucose.
  - GPCR branch: Gpr1 (GPCR), Gpa2 (α-subunit) and Rgs2 (regulator of G-protein signaling). Glucose may directly bind with Gpr1 and states signaling, but it is not clear.
  - Response: PKA phosphorylates a number of targets, including Sfp1 and Rap1 (activating expression of RP and ribosome biogenesis), Msn2/4 (limiting stress response). PKA also promoters psuedohyphal growth. Also possible targets: Gcr1 (and/or Gcr2) (activating glycolysis). Gcr1 has (predicted) transmembrane domain and strong similarity with RNA Pol II subunit of M. inverterns. A model: Gcr1 forms part of complex in the perinuclear region, and glucose activation brings target promoters to the complex.

- Glucose sensing in Calb: the differences
  - Scer has a unique galactose sensing mechanism, while Calb uses the same Hgt4 receptor to sense both glucose and galactose.
  - Calb SRR pathway induces expression of genes involved in alternative respiration (lost in Scer) via Rgt1.
  - Morphogenesis may be different: glucose induces morphogenesis (colonization and virulence), while in Scer, pseudohyphal growth (of diploid cells) is induced by nitrogen limitation. In Calb, PKA pathway activates Efg1 (homolog of Sok2 and Phd1 in Scer), which induces expression of hypha-specific cell wall proteins (HWP1, HWP2), adhesins, etc.
  - Calb Mig1 does not have the Snf1 phosphorylation sites in ScerMig1, required for regulation.

- Cross-talk between pathways: some of them may be shared in Scer and Calb
  - Gcr1 is a co-activator of glycolysis genes, but also repress glucose-repressed genes.
  - Mig1 inhibits expression of Snf3 and Mth1 (part of Snf3 signaling) in the presence of glucose.
  - PKA phosphorylates Rgt1 and Efg1 (in Calb): couple glucose sensing with cellular morphogenesis.
  - TOR pathway: also regulate expression of Hxt1 in Scer, and in Calb, affects cellular filamentation in part through Efg1.

Transcriptional regulation of AA biosynthesis pathways [Jones et al, The Molecular and Cellular Biology of the Yeast Saccharomyces, 1992]

- General AA control:
  - At least 40 AAs are under general control: with starvation of one AA, the expression of all these genes will be up by 2 to 10-fold, dependent on Gcn4. The pathways: Trp, Phe/Tyr, Arg, His, Lys, Ilv, Leu, Gln, Glu, Thr, Met. However, note that in one pathway, not all enzymes are subject to the general control.
  - In the minimum medium, the basal expression of many enzymes are also dependent on Gcn4 control. Thus Gcn4 controls both AA starvation response and basal expression at the minimum medium.

  - Regulation of Gcn4: uncharged tRNA → Gcn2 (dependent on Gcn1/20) → phosphorylation of eIF2 (Gcd genes) → derepression of Gcn4 translation.

- Arg pathway: Arg3 as an example (subject to Gcn4 control)
  - Pathway-specific repression: Arg → Arg80-82 $\dashv$ Arg3 expression → Arg biosynthesis.
  - Regulation of Arg catabolism: Arg → Arg80-82 → Car1/2 expression → Arg catabolism.
  - Different roles of Arg80-82 may be implemented by arrangement of Arg80 binding sites: Gcn4 binding site may interfere with Arg80 function in Arg3 promoter.

- Lys pathway: Figure 13. Gcn4 control, meanwhile two pathway-specific mechanisms.
  - Lys80 repression: Lys → Lys80 $\dashv$ Lys biosynthesis enzymes.
  - Lys14: one intermediate product → Lys14 → Lys biosynthesis enzymes. When Lys is abundant, the intermediate product is reduced (because of end-product inhibition of the enzyme), and thus the enzymes cannot be induced. The requirement of Lys14 positive regulator seems to have stronger effect and can overcome the derepression by Gcn4.

- Leu pathway: Figure 14.
  - Leu3 is the positive regulator of the Leu enzymes. Leu3 is activated by the intermediate product α-IPM.
  - Leu starvation → Leu3 activation → Leu1/2/4. Starvation of other AAs → Gcn4 → Leu4, but not Leu1/2.
  - Leu3 also activates Ilv2/5/3.

- Met/Thr pathway:
  - Met → AdoMet (compound synthesized from methionine) $\dashv$ Met biosynthesis enzymes. Can override Gcn4 control if starvation of other AAs.

- His/Trp pathway:
  - PRPP is the common precursor of synthesis of purine, His and Trp (the intermdiate product leading to histidine biosynthesis).
  - Bas1 is a regulator of purine biosynthesis, and Bas2/Pho2 of phosphate utilization.
  - His4 promoter contains Gcn4 and Bas1/2 sites. Purine may down-regulate His biosynthesis enzymes, depependent on Bas1, Bas2/Pho2.

- Interaction between general and pathway-specific control:
  - General AA control and pathway-specific repression together regulates AA biosynthesis, and in some cases, could achieve pathway-specific response (e.g. some Arg biosynthesis enzymes are not indued if Arg is not limiting).
  - Not all enzymes in a pathway is subject to the same control, e.g. Leu4 may be induced by AAs other than Leu (general control), but not Leu1/2.

Regulation of nitrogen metabolism

- AA sensing system (SPS system):
  - Function: senses external amino acid concentration and transmits intracellular signals that result in regulation of expression of amino acid permease genes.
  - Membrane sensing system: Ssy1p-Ptr3p-Ssy5p; the transcriptional regulators: Stp1/2.

- Nitrogen catabolite repression (NCR): [Cooper, FEMS Microbiology Reviews, 2002]
  - Function: the physiological process by which selective use of available nitrogen sources is achieved. Preferred Nitrogen source: Gln, NH_4^+, Asn; and secondary source: proline, glutamate, allantoin, Arg, urea, etc.
  - Regulation: In the presence of excess nitrogen (a good nitrogen source in adequate supply) transcription of genes encoding the proteins needed to transport and degrade poor nitrogen sources is repressed. When the amount of a good nitrogen source becomes limiting, or only poor nitrogen sources are available, the genes needed for their transport and catabolism are transcribed.
  - Regulators: Gln3 is one master regulator.

Mitochondrial retrograde signaling (RTG) [Liu & Butow, Annual Review of Genetics, 2006]

- Function: one main function of the RTG pathway is to ensure that a sufficient level of α-ketoglutarate for glutamate synthesis is made to meet the demand of nitrogen supply for biosynthetic reactions, especially in respiration-deficient cells (e.g. under low O_2).
- Signaling pathway: [retrograde-signaling1.jpeg]
  - Rtg2 activates TFs Rtg1/3, transloating the heterodimer to the nucleus.
  - Mks1 is the main negative regulator: promoting the phosphorylation of Rtg3p and inhibiting the nuclear translocation of Rtg1/3p.
  - Grr1p functions as a positive regulator of the retrograde pathway by mediating ubiquitination of a negative regulator, Mks1p.
  - Lst8p is an integral component of two TOR (target of rapamycin) kinase complexes, TORC1 and TORC2. Lst8p negatively regulates the RTG pathway, with one site upstream and the other site downstream of Rtg2p.
  - Glutamate and glutamine are potent repressors of the RTG pathway: via SPS system that senses AA signals, or via TOR pathway.
- Target genes: the basic process [retrograde-signaling2.jpeg] is to provide α-ketoglutarate. The metabolic strategy is to provide acetyl-CoA and citrate to the first three steps of TCA cycle (acetyl-CoA is one precursor, and citrate is needed to replenish the cycle). Specifically, this is achieved through reconfigure several metabolic processes.
  - Glyoxylate cycle: provides continuous source of citrate (since TCA cycle is incomplete) using fatty acid β-oxidation. The main enzyme, Cit2, the citrate synthase is up-regulated.
  - Pyruvate to acetl-CoA: through acetate, all enzymes are up-regulatd. Acetate is transported from the cytosol to mitochondrial in the form of acetyl-carnitine.
  - The first three steps of TCA cycle: targeted by RTG signaling. Generate α-KG. Also note that the TCA cycle enzymes are also activated by HAP (heme-dependent Hap1 and heme-independent Hap2/3/4/5), thus under dual control.

TOR signaling pathway in yeast [Schneper & Broach, COM, 2004; Rhode & Cardenas, COM, 2008]

- Function of TOR pathway:
  - Rapamycin treatment leads to: G1 arrest, protein synthesis reduction, glycogen accumulation and autophagy. This suggests that TOR is involved in regulating: cell cycle (or cell state), protein synthesis, energy metabolism, autophagy and polarized cell growth/filamentation.
  - TOR responds to AAs and growth factors in mammals, also rapamycin treatment and Nitrogen starvation have similar responses. These suggest AA signals may act on TOR pathway. However, TOR pathway is also involved in responding to other environmental cues through interaction with other signaling pathways (below).

  - Broadly speaking, TOR pathway represents the availability of Nitrogen source (thus if TOR is blocked, the response similar to nitrogen starvation response is activated).

- Signaling by TOR pathway:
  - Two Tor genes in Scer: Ser/Thr kinases. Tor1 is part of the TORC1 complex, and Tor2 part of TROC2 complex.
  - TORC1 signaling: main function of TOR pathway, but also implicated in actin deposition. Two main branches: (1) Sch9 branch: activation of TFs (typically nuclear localization) and then ribosomal biogenesis; (2) Tap42 branch (phosphorylated by Tor1): repression of stress response (STRE), NCR pathway and retrograde signaling (RTG). Also: intraceullular AA signal (inside vacuoles) activates TOR pathway through Gse/Ego complex.
  - TORC2 signaling: actin polarization.

- Targets of TOR pathway:
  - NCR pathway that controls cellular responses to nitrogen quality: the output is activation of non-preferred nitrogen source utilization. Rapamycin treatment leads to activation of NCR pathway.
  - Retrograde pathway (RTG) that responds to mitochondrial dysfunction: the output is replenising TCA cycle components. Rapamycin treatment leads to activation of RTG pathway.
  - Ribosome biogenesis.

- Cross-talk of TOR with other signaling pathways: cross-talk or covergence on the same targets. [carbon-signaling.jpg, nitrogen-signaling.jpg]
  - cAMP/PKA pathway: Snf1 (a regulator in the PKA pathway) phosphorylates Gln3 (the main regulator of NCR pathway).
  - cAMP/PKA pathway: both PKA and TOR regulate the stress response regulators Msn2/4 - PKA prevents Msn2/4 nuclear localization and TOR exports Msn2/4 out of nucleus.
  - cAMP/PKA pathway: both PKA and TOR activates ribosome biogenesis.
  - General amino acid control (GAAC): rapamycin treatment leads to increase of Gcn4 level.
  - Both TOR and PKA regulate exit from mitotic growth and entry into either stationary phase or meiosis partly through modulation of Ime1 and Rim15.
  - Snf1, TOR and PKA all influence the balance between yeast and pseudohyphal growth in response to nutrients.

Yeast mating type determination

- Expression of cell-type specific genes: [mating-type-regulation.gif] [Messenguy & Dubois, Gene, 2003]
  - MAT locus expresses either MAT-$a1$ or MAT-$\alpha 1$ and MAT-$\alpha 2$. $a1$ is not effective (for mating type), and only $\alpha 1/\alpha 2$ are important for mating type (activation and repression respectively).
  - a-specific genes (asgs): activated by Mcm1 and Ste12 in a-cells; and turned off in α-cells by cooperative binding of $\alpha 2$ and Mcm1 to adjacent binding sites. $\alpha 2$ then recruits Ssn6/Tup1 repressor complex to the site and prevents transcription.
  - α-specific genes (αsgs): in α-cells, Mcm1 combines with the $\alpha 1$ protein to activate transcription of at least four αsgs.

- Expression of genes required for mating:
  - In a cells: Mcm1 interacts with the Ste12 protein to activate genes required for mating and cell fusion, conferring pheromone responsiveness to these promoters.

- In α cells: Ste12 interacts with α1 associated to Mcm1 to activate genes required for mating and cell fusion.

Sporulation: [Chu & Herskowitz, Science, 1998] Figure 1. Overlapping processes of meiosis and spore formation.

- Early: DNA replication, homologous chromatin pairing and recombination during prophase, also spore formation starts (prophase). Known regulator: Ume6/Ime1.
- Middle: homologous chromatin separation (meiosis I), division of two spores. Known regulator: Ndt80.
- Mid-later: sister chromatid separation (meiosis II) and formation of four spores.
- Late: spore maturation.

Transcriptional control of sporulation: the function is repress sporulation under vegetative growth, and start sporulation only under starvation for diploids. [Govin & Berger, Int J Dev Biol, 2009]

- Control of early, middle and late genes: early genes are induced by Ime1; middle genes are induced by Ndt80. Under vegetative growth, early genes are repressed by Ume6/Rpd3/Sin3/Isw2; middle genes repressed by Rfm1/Hst1/Sum1 and late genes repressed by Tup1/Ssn6.
- Control of master regulators: Ime1 is repressed by Rme1 in haploids under vegetative growth (not induced in diploids). Ime1 promoter contains about 10 sites that integrate signals from: nitrogen source (repressed by nitrogen source), carbon source (repressed by glucose and activated by non-fermentable source such as acetate) and cell type (activated by MAT-a1/α2).

Yeast nutrient signaling and cellular decision [Zaman & Broach, Annu Rev Genet, 2008]

- Cell cycle progression: influenced by nutrients in different ways. The main regulators of G1/S transition are G1 cyclins.
  - G1 cyclin level reflects cell size: several hypothesis (i) G1 cyclins are unstable, such that the level of G1 cyclin in the cell is proportional to its synthesis rate, (ii) cyclins are deposited into the nucleus (or other subcellular localization).
  - Size threshold (the threshold by which cell cycle proceeds): higher if more nutrient is available. This may be related to ribosome biogenesis (the future translation efficiency).
  - Nutrient signaling can directly influence START: PKA activity is required for executing START, perhaps through stress response (PKA represses stress response, which negatively affects START).
  - Energy storage in the form of glycogen and trehlose (from glycolysis): thus cells may wait till enough storage is availabe, then the increased glycolysis flux may push through START. Also cells may avoid respiration during DNA replication to avoid DNA damage.
- Nutrient signaling and stress response:
  - Both PKA and TOR negatively regulate Msn2/4 nuclear entry.
  - PKA negative regulates Hsf1, an activator of Hsp12/26 (heat shock response).
  - Both PKA and TOR (through Sch9) phosphorylate Sko1, a downstream TF of osmotic stress response (phosphorylated by Hog1).
- Metabolic effects of nutrient addition or starvation:
  - Effects from both the change of transcriptional regulation and the change of post-translational modification. Ex. PKA regulates the activity of Pfk2 (glycolysis) through phosphorylation.
  - Ammnoia starvation: decrease of Gln and increase of α-KG.

- Glucose addition (from glycerol source): adenine depletion, possibly from increased purine biosynthesis.
- Autophage: the process where cells break down bulk proteins to provide nutrients under starvation (through about 30 Atg proteins). Both TOR and PKA negatives regulate autophage (maybe through stress response).

- Filamentous growth:
  - Induction of filamentous growth: Diploid cells subjected to limiting nitrogen or haploid cells subjected to limiting glucose become elongated, exhibit polar budding, suppress budding in mother cells, undergo cytokinesis, but fail to separate. This cohort of features yields chains of cells, referred to as pseudohyphae in diploids or filaments in haploids, capable of invading the underlying substratum.
  - The signaling pathways responsible for the two programs overlap significantly.
  - Effect of nutrient on filamentous growth:
    * Activation of PKA stimulates diploid pseudohyphal growth and haploid invasive growth. Individual PKA catalytic subunits play distinct roles in pseudohyphal growth with Tpk2 stimulating filamentation and Tpk1 and Tpk3 inhibiting the process, the latter likely through feedback inhibition of PKA activation.
    * To resolve paradox (PKA signals nutrient availability): glucose levels influence filamentation through a mechanism other than PKA; PKA is activated under conditions promoting filamentation by a mechanism distinct from glucose activation.
    * Various excreted alcohol metabolites of yeast stimulate pseudohyphal growth.
    * Snf1 serves as the primary locus coupling nutrient limitation to haploid invasive growth and diploid pseudohyphal growth. Nutrient suppresses Snf1 activity, which suppress its target repressors Nrg1/2.
  - A nutrient-independent pathway:
    * Components of the pheromone-responsive MAP kinase pathway also regulates filamentous growth. Besides hyperosmolarity, the extracellular signals to which the FG MAPK pathway responds are not well defined.
    * FG vs pheromone MAPK pathway: This specificity is achieved in part by pheromone-induced degradation of the FG-specific transcription factor Tec1, by subtle aspects of signal channeling achieved by the Ste5 scaffold protein and perhaps by lateral inhibition between the FG and pheromone MAPKs, Fus3, and Kss1.
  - Transcriptional response of filamentation: more than 800 genes involved. Mga1 and Phd1 serving as the predominant modulators of the filamentous transitions. The patterns of genes regulated by these factors highlight two distinct signaling pathways that together regulate this developmental program-Ras/PKA and the FG-MAPK pathway.

- Quiescence:
  - Induction of quiescence:
    * Haploid or diploid yeast cells starved for carbon, nitrogen, phosphate, or sulfur cease accumulating mass, arrest cell cycle progression prior to START, and enter a poorly defined G0 state.
    * G0 cells do have a number of distinguishing characteristics, including a thickened cell wall, increased storage carbohydrates, enhanced resistance to heat and high osmolarity, substantially reduced translation, a specific transcriptional profile, and, most important, the ability to maintain viability under the starvation condition.
  - Signaling pathways:

    - Induction of stress resistance upon starvation results in part from activation of the Msn2/Msn4 stress-responsive transcription factors.
    - Rim15 kinase. Thus, three distinct nutrient-responsive pathways (TOR, PKA, Sch9) converge on Rim15 through distinct mechanisms. Rim15's effect on quiescence derives in part through changes in the transcriptional spectrum of the cell, mediated through the stress response transcription factors Msn2/Msn4 and the related post diauxic shift transcription factor Gis1.

- Meiosis/sporulation:
  - Induction of sporulation: MATa/MAT-α diploid cells in response to: the absence of one essential growth nutrient such that cells arrest in G1; the absence of glucose; and the presence of a non-fermentable carbon source. Although nitrogen starvation is the normal laboratory condition for sporulation, starvation for phosphate or sulfur can also induce sporulation even in the presence of an adequate nitrogen source.
  - Signaling pathways:
    - Nutrient signals impinge on the expression and function of two key regulators of initiation of meiosis, the transcription factor Ime1 and the S/T kinase Ime2. Glucose represses IME1 expression.
    - The presence of a nonfermentable carbon source is perceived by the cell as a consequence of its metabolism to CO_2 and resultant alkalization of the medium. High external pH activates a highly conserved fungal pH sensing pathway comprising cell surface receptors and TF Rim101. Glucose also likely influences this signaling pathway through repression of respiration, thereby blocking alkalization by blocking metabolism of nonfermentable carbon sources.

Tuning gene expression to changing environments [Lopez-Maury & Bahler, NRG, 2008]

- Environmental transcriptional responses in eukaryotes:
  - For plants and microorganisms, their sessile lifestyle leaves them more exposed to the environment than animals, thus their transcriptional responses tend to be larger.
  - Stress responses in Scer and Spom (fast-growing cells): generally induction of heat-shock and antioxidant functions, as well as for carbohydrate metabolism and energy generation; and repression of growth-related genes. The existence of core stress response, responsible for cross-protection.
  - Stress responses in Calb: only a limited core stress response. May reflect the lifestyle: larged shielded from environmental variations.
  - Stress responses quiescent cells (G0 phase) and differentiated cells in multicellular organisms: ribosome and other growth-related genes are not or weakly down-regulated.

- Control of growth and stress response: cells need to balance the expression of growth- and stress-related genes according to environmental conditions (Figure 1).

- Signaling pathways: in theory, cells could use internal conditions to regulate these genes (feedback), but may rely on external signaling. Two pathways (below). It is possible that some regulatory proteins integrate inputs from both (and other) pathways or cross-talk between the two.
  - Growth signaling: TOR pathway. The targets tend to have no TATA box and have robust expression.
  - Stress signaling: SAPK (stress-activated protein kinase) pathway. The targets tend to have TATA box and noisy expression.

- Transcriptional complex: different co-activators are associated with growth- and stress-related genes. Thus cells may regulate the co-activators to achieve massive reprogramming of transcriptomes in response to environmental stimuli.

  - TATA box (about 20% of all genes): SAGA-dominated, stress-related genes.
  - TATA-lacking promoters: TFIID-dominated, growth-related.

3.3.1 Regulation of Metabolic Processes

Cases of regulation of metabolic pathways:

- Galactose utilization pathway in S. cerevisiae: [Ideker & Hood, Science, 2001]
- Arginine biosynthesis in E. coli. [Caldara & Cunin, JBC, 2008]

Integration of metabolism and regulation [Yeang, Methods in Mol Biol, 2009]

- Effect of metabolic shifts on gene regulation:
  - Repression of alternative pathways: e.g. one carbon source in the medium can inhibit the enzymes involved in the metabolism of alternative carbon sources.
  - Maintaining homostasis
  - Stress response: cell cycle arrest, half of biomass accumulation, down-regulation of mRNA and protein synthesis, over-expression of stress response genes, sporulation for unicellular organisms, and apoptosis for multicellular organisms.
  - Gene deletion may have mild responses: e.g. deletion of genes along central carbon metabolism.
- Mechanisms of metabolic gene regulation:
  - Metabolites can directly interact with transcriptional apparatus.
  - Signal transduction pathways: e.g. glucose induction is mediated by Snf3 or Rgt2 (glucose sensors), G proteins, cAMP pathway (PKA, and kinase TPK, which is translocated to nucleus and regulates TFs).
  - Riboswitches: metabolics may bind to mRNAs and modulate translation of mRNAs.
- Principles/issues of metabolic gene relation:
  - Co-regulation of genes: depend on the proximity in the metabolic network. Not simple relationship, though. For instance, the co-expression of enzymes in convergent reactions and divergent reactions.
  - Metabolic fluxes and gene expression: ex. flexibility of fluxes and the diversity of gene expression are correlated, genes active in optimal metabolic fluxes tend to be conserved [Bilu & Ruppin, PLCB, 2006]
  - Feedback control: e.g. input substrate induces expression of enzymes along the pathway and an intermediate product inhibits transcription of key enzymes [Martinez-Perez & Santero, J Bacteriol, 2007]

Regulation of ribosome biosynthesis in budding yeast [Warner, TIBS, 1999; Lavoie & Whiteway, COM, 2009]

- Importance of ribosome biosynthesis:
  - Yeast genome contains a large number of ribosomal RNA (rRNA) genes, about 10% of the genome; and 78 different RPs (encoded by 137 genes).
  - Cells synthesize 2,000 ribosomes per min. rRNA transcription by Pol I represent nearly 60% transcription of the cell; RP transcription - 50% RNA Pol II mediately transcription.

- Design goal of the RP regulation system: the rate of RP production should match the needs of the cell, which depends on nutrient availability, stress, etc.
- Response to internal and environmental signals: Figure 2 of [Warner99]
  - Nutrient availability (including Carbon source): via TOR pathway, Ras/PKA pathway.
  - Stress (heat shock): stop transcription.
  - Secretory pathway: disruption will stop transcription through Wsc1 and Pkc1
  - Cell cycle
- Transcriptional regulation of RP: Figure 2 of [Lavoie09].
  - Rap1, Hmo1, and Fhl1/Ifh1 (Ifh1 associated with RP promoters through Fhl1).
  - Ifh1 is condition sensitive: bind to Fhl1 when TOR pathway is active, but replaced by CRF1 (phosphorylated) when TOR is inactive (Figure 7 of [Martin & Hall, Cell, 2004]).

Transcriptional program of diauxic shift [DeRisi & Brown, Science, 1997]

- Background: diauxic shift - switch from anaerobic growth to aerobic respiration upon depletion of glucose.
- During exponential growth in glucose rich medium, the global pattern of gene expression was very stable (only 19 genes show more than 2-fold change). After glucose depletion, very large change: 710 increase by at least 2-fold, and 1030 decline by at least 2.
- Expression change of central metabolic genes (Figure 3): increase enzymes in ethanol to acetyl-CoA, TCA cycle, and glucogenesis (glyoxylate cycle and synthesis of glycogen and trehalose from gluc-6-P). Decrease enzymes in glycolysis.
- Other changes: (1) up-regulation of glucose repressed genes; (2) induction of stress response genes (Hsp42, Hsp26, etc); (3) induction of electron transport chain; (4) down-regulation of RP genes.
- Change in TFs: Rap1 reduced by 4.4 fold; Hap4 and Sip4 (interaction with Snf1, the master regulator of glucose repression) induced by a factor of at least two.

3.3.2 Regulation of Stress

Evolutionary stategy of stress responses:

- Scer responds to starvation by: quiescence, filamentation (foraging behavior) or sporulation (sexual differentiation).
- Stress responses by increasing variations: (1) By having sex (sporulation), it is possible to increase genetic variations in the population. (2) The noisy expression of genes increases phenotypic variations, a hedging strategy. (3) Transpositions and even mutation rates may be increased upon stress.
- Stress responses may be primordial processes for evolution of cellular differentiation: use of different stress-resistent states; conserved signaling pathways, etc.

Yeast stress responses [Gasch & Brown, MBC, 2000]

- Problem: how cells change expression in response to stresses, and how are they regulated?
- Methods:
  - Stresses: nutrient availability (AA starvation, nitrogen source depletion), temperature, osmolarity and acidity, oxidative stress.

  - Experiment: cells moved to new environmental condition (with stress), e.g. from 25 to 37 degree, then collect samples at various time points over 2-3 h (e.g. heat shock: 5, 15, 30, 60 min).
  - Finding targets of regulators: if R regulates expression of a gene g under condition C, then the expression of g under C in mutant of R will be different from that in wildtype.

- Overall features of the responses to all stresses: the global changes of transcript abundance was largely transient; quantiative response (stronger stress leads to larger and longer response); specific to stress, not environmental change (e.g. 37 to 25 degree, not ESR).

- Environmental stress response (ESR) genes: about 900 genes, shared among all stresses.
  - Represses genes (600): growth-related, RNA metabolism, nucleotide biosynthesis, other metabolic processes.
  - Induced genes (300): carbohydrate metabolism (glycogen synthesis and degradation), cellular redox reactions and defense against ROS, protein folding (chaperones) and turnover, DNA damage repair and signaling pathways (in particular, PKA pathway, which inhibits Msn2/4).

- Regulation of ESR: controlled by multiple systems, activated at different conditions (e.g. chaperons are induced in many stresses, but super-induced in heat shock). About 180 are targets of Msn2/4. Among these, some (e.g. TRX2 cluster) are activated by Msn2/4 in response to heat shock, but by Yap1 in response to H_2O_2.

- Discussion:
  - Why a large ESR common to all stresses? Under all these conditions, cellular instability will be created, thus need to maintain all critical physiological processes. Cross-resistence to various stresses: cells exposed to a low does of one stress becomes resistant to an otheriwse low does of a second unrelated stress.
  - Composite expression reponses: additive in the presence of multiple stresses.
  - Transient response: likely that protein level will reach new steady-state level in new condition.
  - ESR signaling: PKA pathway to nutrientional signals, PKC pathway when secretion is impaired, Mec1 pathway to DNA damage, HOG pathway to osmotic stress. Each is also implicated in regulating more specialized gene expression responses.

3.3.3 Regulation of Life Cycle

Transcriptional program of yeast cell cycle [Spellman & Futcher, MCB, 1998]

- Methods:
  - Methods of synchronization: the basic idea is to create cell cycle arrest, then remove the arresting factor and induce cell cycle. Three different ways: α factor, cdc15 (temperature-sensitive mutant), and elutriation (small cell size). In addition, cdc28 from Cho et al.
  - Dependence of expression on specific cyclins: cln3 and clb2. In cln and clb mutants, expression of cln3 and clb2 and observe the expression change of other genes.
  - Cell cycle regulated genes: Fourier algorithm.

- About 800 cell-cycle related genes.

- Major cell-cycle related biological processes and their expression patterns (the time of peak expression): Figure 7.

  - Replication of chromatins: prereplication complexes (Mcm2/3/4, etc.) in M/G1 phase, DNA replication proteins (Pol1/2,etc.) and initiation (Cdc45) in G1 phase, DNA repair proteins, also in G1 phase, nucleotide biosynthesis (Rnr1/3) in G1 phase, and 11 histones in S phase.
  - Bud formation: site selection (Bud4, etc.) - different proteins in different phases of cell cycle but more in G1 phase; secretion, needed to transport proteins to the budding sites, more in G1 phase; cell wall synthesis - different phases of cell cycle; fatty acid/lipid biosynthesis - different phases of cell cycle.
  - Mitosis: spindle pole body (SPB) in G1 phase; microtubules and accessory proteins mostly in G1 and S phases.
  - Mating: 19 genes including mating pheromones and MAT gene itself in M phase and G1 phases, most regulated by Mcm1. A deep connection between mating, start, and the cell cycle. For instance, if genes involved in mating are turned off at start by multiple mechanisms, it helps explain how passage through start precludes mating.
  - Methionine biosynthesis: about 20 genes (biosynthsis and transport or related) in S and G2 phase. One possible explanation: methionine is limiting AA, e.g. starvation for sulfur or for methionine effectively causes G1 arrest. There are other explanations.

- Promoter sequences: more than half contain binding sites of known TFs: MBF, SBF, Mcm1-SFF, Swi5/Ace2.

- Remark:

  - Even though other processes are needed, e.g. protein synthesis, their expression may not be cell cycle regulated. In general, only limiting genes need to be regulated.
  - Related genes may be expressed in different phases of cell cycle, e.g. bud formation, may be related to different roles of proteins.

Transcriptional program of sporulation [Chu & Herskowitz, Science, 1998]

- Methods:

  - Induction of sporulation: transfer cells to nitrogen-deplete medium.
  - Seven temporal classess: the profiles defined by averaging profiles of known genes in each class.

- Rapid induction of metabolic genes: adaptation to nitrogen starvation. Also repression of many growth-related genes, e.g. RP genes.

- Early genes: DNA replication and sister chromatid cohesion, homologous chromosome pairing and recombination.

- Exit from prophase: duplicated spindly pole bodies (SPB) undergo separation to establish the spindle for meiosis I.

- Chromosome distribution during mitosis: Cdks, motor proteins, kinetochore proteins.

- Anaphase of meiotic divisions: proteolysis mediated by anaphase promoting complex (APC).

- Mid-late induction and later induction: spore formation and spore maturation (e.g. prototypical gene).

Yeast pheromone response [Roberts & Friend, Science, 2000]

- Goal: the transcriptional response to pheromone, and the signaling pathways leading to response. In particular, identify different signaling pathways that may lead to activation/repression of different sets of genes.

- Background: four MAPK pathways in yeasts [yeast-MAPK.jpg].
- Global transcriptional response depends on the Ste12-activation pathway: no response in Δste2, Δste4, Δste12, etc.
- G1 cycle arrest: repression of genes involved in cell cycle progression, DNA replication, budding and mitosis. The repression of these gene depends on Far1 (Δfar1 response), a target of MAPK Fus3.
- Mating projection: require localized cell surface expansion and reorgnization of cortical actin cytoskeleton. The genes: Mpk1, Mlp1 and Rlm1 (targeted by PKC pathway for cell wall integraty response). The genes are induced in prolonged pheromone treatment (2 hour treatment vs 30 min treatment in other experiments). The induction depends on Bni1 and Mpk1 (PKC pathway genes), and overexpression of PKC pathway leads to induction.
- Mating genes: Fus1, Fus3, etc. Activation depends on derepression of Dig1/2 (Δdig1/2 show high expression in the absence of pheromone).
- Filamentation genes: Pgu1, Kss1, Svs1, etc. Induction of filemantion genes by Δdig1/2, induction by basal or pheromone-induced Kss1 signaling, and Tec1-dependent. Pheromone-induced invasion: may enable elongation and growth of dividing cells toward a diffusive pheromone source and thereby failitate mating among haploid cells within dispersed populations.
- Discussion: the cross-talk between pheromone response pathway and PKC pathway, filamentous growth pathway (e.g. through phosphorylation of Kss1 by Ste7), allows cells to activate genes controlled by these two pathways, mating projection and filamentous growth, respectively.

Chapter 4

Model Organisms

4.1 Fruit Fly

1. Enhancer model in fruit fly development

   Billboard model [Kulkarni & Arnosti, Development, 2003]:

   - Motivation: enhancersome model where large protein assembly is formed to direct a single logic ouput vs billboard model.
   - Results:
     - Synthetic promoter of twi, dl and gt: gt reduces expression
     - Synthetic promoter of gt and Gal4 (5 sites): gt has no repressive effect, but if only 3 Gal4 sites, gt will suppress expression
     - Synthetic promoter of twi, dl, Gal4 and gt: gt selectively represses twi, dl, but not Gal4
     - Additional gt sites in twi, dl, Gal4: reduced expression in all the gt domain
   - Discussion: bill-board model. There exist multiple sub-elements which independently interact with TC, thus the stoichiometry, number of sites, etc. dictate the overall output. E.g. one may have an enhancer where one part stimulates TC, and the other part reduces the TC activity.

   cis-regulatory logic of short-range repression [Kulkarni & Arnosti, MCB, 2005]:

   - Problem: what affects the short-range repression?
   - Results:
     - Ratio of activator BS and repressor BS is important. Exp: 2 gt sites can suppress 3 Gal4 sites, but not 5 Gal4 sites.
     - Spacing between activator and repressor sites is important. Exp: moving Gal4 sites for 37 bps abolishes repression.
     - Placement of repressor sites: flanking or interspersed around activator sites will retain repression
     - Affinities of activator sites are important. Exp: 2 gt sites can repress 5 weak Gal4 sites, but not 5 strong ones; bcd sites are more susceptible than Gal4 sites.
     - Repression does not depend on activation domain, but on DNA-binding domain. Exp: Gal4 fusion protein with different activation domain show similar behavior.
     - Short-range repressors show similar functional limits: Kr, gt and kni; but not hairy (long-range repressor).
   - Conclusion/Discussion:

- Mechanisms of repression: (i) the target: either activator or BTM; (ii) mechanism of repressing target: protein-protein interaction or chromotin structure change.
- Hypothesis of short-range repression (quenching): recruiting corepressors (e.g. CtBP) that changes chromatin structure (e.g. deacetylation) that makes activator binding more difficult. Evidences include:
  * Independence of CRMs: repression of one CRM will not affect other CRMs. Support that the activator is the target.
  * Lack of activator specificity: gt, Kr, kni can block activities of many activiators, including bcd, hb, dl, twi and dstat $\rightarrow$ PPI is unlikely to be the mechanism (which is likely to be specific)
  * Importance of DNA binding: sensitivity of activators toward repression mostly closely linked to DNA binding domain and the affinity of binding sites.
  * Function of corepressors: repressors can interact with some corepressors (genetically and physically), which have chromatin remodeling activities.

CtBP independent repression of short-range repressors [Nibu & Levine, MCB, 2003]:

- Background: CtBP is essential for the quenching activity of three short-range sequence-specific repressors in the early Drosophila embryo: Kr?ppel, Knirps, and Snail.
- Results:
  - dCtBP is dispensable for target enhancers that contain overlapping activator and repressor binding sites.
  - dCtBP is essential when Kruppel and Knirps repressor sites do not overlap activator sites but are instead located adjacent to either activators or the core promoter.
- Conclusion: short-range repressor can regulate via different mechanisms - competition, which does not depend on dCtBP; and quenching or direct repression, both of which depend on dCtBP.
- Discussion: the mechanisms of dCtBP - several different models:
  - Activator quenching: dCtBP could disrupt physical interactions between activators and BTM. Perhaps dCtBP masks or modifies the activation domains of upstream activators.
  - Local chromatin modification: dCtBP can modifie proteins such as histones and helps condense DNA within the limits of a nucleosome.
  - BTM poisoning: dCtBP "poisons" BTM and impedes its binding, assembly, or function at the core promoter. The linkage requirement (they must bind within 100 bp of adjacent activators) might reflect a reliance of the repressors on linked activators in order to loop to the core promoter.

Non-homologous structured CRMs from the Ciona genome [Erives, RECOMBSAT, 2008]:

- Aim: the structural rules of CRMs.
- Results:
  - Ciona muscle enhancer characterization: T-box and E-box are important factors.
  - Unstructured query of the cluster of the two types of binding sites: $< 1\%$ sequences are functional.
  - Structured template (matching the known CRMs): 100% of the targets are functional.

2. Anterior-posterior pattern formation

   hb anterior activator [Driever & Nusslein-Volhard, Nature, 1989]:

   - Aim: how the number and affinity of binding sites control the expression pattern: boundary and intensity

- Results:
  - The distance to TSS has a minor effect on expression (pThb7 vs pThb5 and 6).
  - A copy of the enhancer slightly expands the expression domain and increases the expression level (pThb8 vs pThb5). Thus the expression should be near saturation.
  - 3 or 6 or 9 weak TFBSs: similar expression pattern, only different expression level (Fig. 4).
- Conclusion: gradient threshold model, different enhancers responsd to the morphogen gradient differently, by having different affinities to the morphogen.

Hb organizes the expression of Kr and kni [Hulskamp & Tautz, Nature, 1990]:

- Aim: the role of Hb in organizing the expression pattern of Kr and kni.
- Background: the translation of maternal Hb is repressed in the posterior by Nanos; and the transcription of zygotic Hb is activated by Bcd in the anterior. Thus Hb expression consists of maternal and zygotic parts, both in the anterior.
- Results:
  - Removal of hb_{zyg}: anterior expansion of Kr, but its level does not change.
  - Removal of both hb_{zyg} and hb_{mat}: significantly reduced the expression of Kr.
  - Ectopic expression of hb: posterior expansion of Kr.
  - Removal of hb_{mat}: anterior expansion of knirps.
- Conclusion: (Fig. 4) Hb represses the expression of Kr and kni in the anterior (thus determines the anterior boundary of the two), but activates the expression of Kr in the posterior.
- Remark: the secondary effect of Hb must be analyzed, as Hb is also involved in other gap genes such as giant. Therefore, it is possible that reducing the level of Hb will increase the level of, say giant, which further reduces the level of Kr (the activation of Kr by Hb at the low concentration).

Kr enhancers [Hoch & Jackle, EMBOJ, 1990]:

- Aim: identify the enhancers for various temporal and spatial pattern of Kr expression.
- Background/Results:
  - Kr expression pattern: in central domain (CD) at the early to mid- blastoderm stage; in anterior domain (AD, 80%), and posterior domain (PD, 0%) at the late blastoderm stage. The CD is transient and disappears after gastrulation., but PD continues to accumulate.
  - Two elements control Kr CD expression: CD1 and CD2. The wt. expression is observed with either CD1/CD2 both, or CD1 + authetic Kr promoter (instead of heat shock promoter used for reporter gene).

eve stripe 2 [Stanojevic & Levine, Science, 1991]:

- Problem: how bcd, hb, Kr and gt determine the eve stripe 2 pattern?
- Experimental background: techniques for studying gene regulation in development
  - Change of trans- part: mutant of TF
  - Change of cis- part: site-directed mutagenesis
  - Measurement of the effect of a cis-regulatory element: sequence + reporter gene and measure the expression
- Biological background: in Kr^- and gt^-, anterior or posterior expansion of the stripe; in bcd^- and hb^-: reduced or abolished stripe
- Methods: the 5.2kb eve promoter sequence (not the minimal stripe element) as the template for reporter expression analysis.

- Results:
  - Δbcd1 and Δbcd2: reduced expression of stripe 2
  - Δhb3: similar, but less severe reduction of stripe 2
  - Δ all gt sites: anterior expansion
  - Δ all high-affinity Kr sites: posterior expansion, but weaker than in Kr^{-}.
- Conclusion:
  - bcd, hb act as activators; and Kr, gt act as repressors to determine the anterior and posterior boundaries.
  - Low affinity Kr sites are important.

Minimal stripe element, MSE of eve stripe 2 [Small & Levine, EMBOJ, 1992]:

- Aim: the expression pattern of eve stripe 2.
- Results:
  - MSE: a 480 bp region that drives the wt. expression pattern of eve stripe 2.
  - Anterior border by Gt: disruption of all 3 Gt sites leads to anterior expansion of the stripe.
  - Role of Kr: disruption of all 3 Kr sites (reducing the affinity of bcd-1, but not disrupt it) greatly reduces the expression in stripe 2, and no significant posterior expansion.
  - Activation of stripe 2 expression: point mutations in all 5 Bcd sites completely abolished eve 2 (but enhancement in eve 7). Point mutations in bcd-1, bcd-2 (strong sites) nearly abolished the expression, and in hb-3 severly reduced the expression (except the ventral region); point mutation in other weak bcd sites reduced the expression.
- Discussion:
  - Possible additional mechanism for anterior repression: mutation of 3 gt sites does not lead to complete anterior expansion, two models: (i) Tor modified the activity of Bcd and Hb at the poles; (ii) additional repressor in the anterior pole, one candidate is otd, which binds the same sequence as bcd and thus competes with bcd sites in MSE.
  - Kr may be a repressor for the 5.2kb promoter, but not important for MSE repression. The posterior boundary may be set by the decresing concentration of Bcd and Hb activators in the context of MSE. The reduction of expression in the MSE with 3 Kr sites mutated may be due to the lower affinity of bcd-1, instead of the positive role of Kr.
  - Bcd and Hb sites may work cooperatively in MSE, thus mutating a single site will disrupt the whole expression pattern. This may be the mechanism of repression: the binding of Kr and gt repressors can effectively shut off the expression by interfering with just one or two of the activator sites.

knirps posterior enhancer [Pankratz & Jackle, Science, 1992]:

- Aim: the spatial expression pattern of knirps.
- Methods: deletion contructs; and footprinting for binding analysis
- Results:
  - Identification of kni enhancers: kni KD contruct (and KR) drives the wt. pattern.
  - Anterior and posterior border: anterior is set by Hb as (i) Hb binds to the fragment required for anterior border; (ii) deleting Hb binding sites results in the expansion of anterior border. Posterior border may be set by Tll as Tll binds to the fragment required for posterior border, but no evidence of the posterior expansion from deleting Tll binding sites.

Concentration-dependent effect of Kr [Sauer & Jackle, Nature, 1991]:

- Methods:
  - Experimental system: Drosophila cell line, that is co-transfected with reporter gene plasmid (inserted a Kr binding site) and Kr expression plasmid.
- Results:
  - At low concentration, Kr activates expression; and at high concentration, Kr represses expression.
  - The opposite effects are mediated by distinct parts of Kr protein.

Selective repression by Kr [Licht & Hansen, PNAS, 1993]:

- Results:
  - Kr represses the acvitation by Sp1 but not Gal4 in mammalian cells (co-transfected with Kr and activator).
- Discussion:
  - Queching interaction of Kr: the data suggests that Kr may block Sp1 by PPI, with its repression domain queching the activation domain of Sp1. Because it is activator specific, it is unlikely to interact with BTM proteins, unlike Eve, which represses basal transcription.
  - Selective quenching and module independence: if queching is specific to activator, then short-range repression is not a prerequisite for module independence. For example, eve stripe 3 may use different domain of Hb for activation comparing with eve stripe 2, thus the Kr molecule bound to stripe 2 enhancer may not be able to repress stripe 3.

Interaction of Kr with BTM [Sauer & Jackle, Nature, 1995]:

- Results:
  - At low concentrations, Kr monomer interacts with TFIIB to activate transcription; at high concentrations, Kr forms dimer and the interaction between Kr dimer and TFIIEβ results in transcriptional repression.
- Discussion:
  - Different from the embryo or the mammalian system, in Drosophila Schneider cells, Kr mainly interacts directly with BTM to regulate transcription. This interaction may apply to the late stage of development.

Distinct repression regions of Kr [Hanna-Rose & Hansen, MCB, 1997]:

- Results:
  - Evolutionary conservation of Kr amino acid sequence: 96% identical between Dmel and Dvir (60-80 Myr).
  - N-terminal repression region can inhibit transcription of multiple activators (in mammalian cells); while C-terminal repression region can only inhibit transcription of a subset of activators.
- Discussion:
  - Model: N-terminal region directly inhibits basal transcription activity; and C-terminal region interferes with activators when bound nearby within an enhancer.
  - Implications: activator quenching allows enhancer autonomy; while the direct inhibition of BTM could allow Kr to completely repress a gene if necessary.

CtBP dependent and indepdent repression of Knirps [Keller & Arnosti, MCB, 2000]:

- Results:

  - One C-terminal region of Knirps depends on CtBP for its repression, and contains a dCtBP binding motif.
  - One N-terminal region of Knirps does not bind to CtBP and represses even in the CtMP mutant.
  - The even-skipped stripe 3 enhancer, is not derepressed in a CtBP mutant.
- Conclusion: Knirps have two repression pathways, one dependent on CtBP but the other not.
- Discussion:
  - Knirps activity: does not have a phasing effect, but there is a measurable interval (from 100 to 130 bp) over which Knirps activity is attenuated but not entirely abolished. The exquisite distance dependence may contribute to the differential response of endogenous target genes to repressor gradients
  - Quantatitive effect: quantitatively, dual activities may increase the overall level of repression, much as transcriptional activators have been suggested to employ multiple paths to achieve synergistic activation.

Logical analysis of gap network [Sanchez & Thieffry, JTB, 2001]:

- Gap gene network: (Fig. 2)
  - Bcd is the anterior activator: $\rightarrow$ gt, hb_{zyg}, Kr, kni
  - Cad is the posterior activator: $\rightarrow$ gt, kni
  - Hb $\rightarrow$ Kr when at low level, $\nrightarrow$ Kr when at high level
  - Hb $\nrightarrow$ gt and kni
  - Mutual suppression between gt and Kr
  - Additional repressions: Gt $\nrightarrow$ kni, Kni $\nrightarrow$ Kr, Kr $\nrightarrow$ hb
- Analysis of expression domains of gap genes: for each expression domain, analyze (i) activator; (ii) anterior boundary determinant; (iii) posterior boundary determint.
  - The anterior gt domain: (i) Bcd; (ii) terminal system; (iii) Kr.
    Note: Hb acts a repressor of gt, but it plays no role in normal conditions. The repression effect is overcome by Bcd.
  - The posterior gt domain: (i) Cad; (ii) Kr; (iii) posterior Hb domain and terminal system.
  - The hb domain (anterior): (i) Bcd and Hb (synergistic interaction); (ii) n.a.; (iii) lower Bcd gradient.
  - The kni domain (posterior): (i) Cad; (ii) Hb; (iii) Gt.
    Note: activation of Bcd is override by Hb and Gt.
  - The Kr domain (central): (i) Bcd and low level Hb; (ii) Gt (anterior domain) and high level Hb; (iii) Kni, Gt (posterior domain).

CtBP dependent and independent repression of Knirps [Struffi & Arnosti, Development, 2004]:

- Results:
  - CtBP-independent activity, when provided at higher than normal levels, can repress an eve regulatory element that normally requires CtBP.
  - The activity of Knirps containing both CtBP-dependent and -independent repression activities is higher than that of the CtBP-independent domain alone.
- Discussion:
  - Multiple repression activities: quantitative contributions to reaching repression thresholds (Fig. 7). The presence of CtBP only reduces the repression threshold needed to shut down a stripe. When CtBP is low, scine eve stripe 3/7 requires lower threshold to repress than stripe 4/6, stripe 3/7 can still be repressed, while stripe 4/6 not.

  - Mechanism of CtBP-independent repression: unknown, but has very similar characteristics with respect to activator specificity, distance dependence and overall potency.

Activation of posterior kni expression by Cad [Rivera-Pomar & Jackle, Science, 1995]:

- Aim: the expression pattern of knirps.
- Results:
  - Two kni enhancers: kni64 (64bp) and kni223(223bp). The composition: kni64 - 6 Bcd sites and 1 Cad site; kni223 - 5 Cad sites and 2 Hb sites.
  - The expression patterns (Fig. 2): kni64 drives expression at the anterior (only 1 Cad site and no Hb site); kni223 drives expression at the posterior (no Bcd site and Hb sites); combined kni64 and kni223 - similar to kni223; deletion of Hb sites in kni64-223 - ubiquitious expression.
- Conclusion: Bcd is an anterior activator and Cad is a posterior activator; Hb represses the expression at the anterior.
- Question: the results in Fig 2 (in situ hybridization) are contradictory to Fig. 1 and 3 (reporter), where kni 64 alone drives the expression pattern that is similar to wt. Why?

eve stripe 2 CRM [Arnosti & Small, Development, 1996]:

- Problem: how the sharp anterior boundary of eve 2 is achieved?
- Results:
  - Bcd sites: B1, B2 are high affinity sites and B3, B4, B5 are low affinity sites. ΔB1: disrupt the expression pattern. Could be restored by either adding new Bcd sites (at some, but not all positions) or increasing affinities of B3, B4, B5 sites.
  - Hb sites: Δ H1: disrupt the pattern. Could be restored by adding Bcd sites or increasing affinities of B3, B4, B5 sites.
  - Gt sites:
    - ΔG1, ΔG3: minor effects
    - ΔG2: significant anterior expansion (G2 is about 40bp to the nearest Bcd site)
    - ΔG1, ΔG2, ΔG3 and ΔB1: expression in broad anterior domain.
- Discussion:
  - Bcd site arrangement: limited flexibility
  - Hb: synergy with Bcd (because Hb itself has little or no activity): could be replaced by other Bcd sites nearby or transcriptional complementation (a CCN4 or Sp1 domain not directly bound to DNA, but interact with Bcd). Support the promiscuous synergy: as long as a critical number of activators are present that can independently contact BTM.
  - Gt: quenching is important. May disrupt cooperative Bcd binding; and/or the interaction of Bcd activation domain and BTM.

Anterior repression of eve stripes [Andrioli & Small, Development, 2002]:

- Problem: Bcd is expressed in the anterior, and Gt is expressed only in part, but not all, of the anterior domain. What stops the stripe 2 enhancer from expressing in the more anterior region?
- Conclusion: three mechanisms (including one known mechanism)
  - Gt: genetic removal of gt or deletion of Gt binding site leads to anterior expansion, but only in the anterior subregion that lies adjacent to the stripe borde.
  - Sloppy-paired 1 (Slp1): a well-conserved sequence repeat $(CTTT)_4$, required for repression in a more anterior subregion.
  - Torso: downregulation of Bcd activity by Torso, prevents activation near the anterior tip.

Common mechanisms for other stripes: cctopic Slp1 also represses eve stripes 1 and 3, and the eve 1 and eve 3+7 enhancers each contain GTTT repeats.

eve stripe $3+7$, and eve stripe $4+6$ [Clyde & Small, Nature, 2003]:

- Problem: what determines the stripes eve 3+7 and 4+6?
- Background: kni, hb act as repressors for eve, involved in creating eve 3+7 and eve 4+6. The expression of kni and hb: kni high at region corresponding to stripes 4-6; and hb high at regions $\leq$ stripe 3 and $\geq$ stripe 7.
- Model/hypothesis:
  - eve 3+7 enhancer: high kni affinity $\Rightarrow$ OFF at 4-6; and low hb affinity $\Rightarrow$ ON at 3,7
  - eve 4+6 enhancer: low kni affinity $\Rightarrow$ OFF at 5 but ON at 4-6; and high hb affinity $\Rightarrow$ OFF at 3,7
  - Mutual repression between kni and hb expression patterns.
- Results:
  - Bioinformatic analysis of PWM scores of kni and hb in the two enhancers.
  - ectopic expression of kni, hb $\Rightarrow$ disruption of 3+7, 4+6 stripes.
- Conclusion: the combination of repressors can determine complex expression patterns by using different affinities for different enhancers.

CRMs of segmentation gene network [Schroeder & Gaul, PLoS Biol, 2004]:

- Methods:
  - CRM prediction: Ahab
  - TFBS prediction: each putative site has a profile value (between 0 and 1), measuring the fractional occupancy of a site by its factor
- Results:
  - Verification of 16 predicted CRMs: 13 produce patterns of the endogenous genes
  - Correlation between TF expression and the TFBS composition along AP axis: maternal factors (bcd, cad, torRE) show strong positive correlation; gap factors (gt, Kr) show negative correlation; hb effect is context-dependent; kni and tll: no correlation presumbably because of low specificity of their PWMs
- Note: TFBS composition at a location is defined as: the integrated profile values of all CRMs that drive expression at that location
- Analysis: check if the expected TFBSs are found in the CRMs. The mapping between genes to CRMs:
  - gt: gt_(-10) (anterior) and gt_(-3) (posterior)
  - hb: hb_anterior_activator
  - kni: kni_kd
  - Kr: Kr_CD1

  In general, a CRM should have binding sites for 3 types of TFs: (i) activator; (ii) repressor that determines the anterior boundary; (iii) repressor that determines the posterior boundary.
  - gt_(-10): (i) bcd (Yes); (ii) terminal gene (Yes); (iii) Kr (No)
  - gt_(-3): (i) cad (Yes); (ii) Kr (Yes); (iii) terminal (Yes)
  - hb_anterior_activator: (i) bcd and hb (Yes); (ii) NA; (iii) lower bcd (Yes)
  - kni_kd: (i) cad (No); (ii) hb (Yes); (iii) gt (No)
  - Kr_CD1: (i) bcd, low hb (Yes); (ii) gt (anterior), high hb (Yes); (iii) kni, gt (posterior) (Yes)

Bicoid site cluster and bcd-dependent patterning [Ochoa-Espinosa & Small, PNAS, 2005]:

- Motivation: test the differential sensitivity hypothesis (or gradient threshold model) - different CMRs have different sensitivities to bcd, leading to different AP pattern (measured by posterior boundary position, PBP).
- Results:
  - Prediction and identification of bcd clusters (homotypic clusters): by a certain number of sites with a certain PWM score.
  - PBP vs bcd cluster strength: no coorelation with all measures of strength (average score, number of sites, max. score, etc)
  - Combinations of activator and repressor better explain PBP: four classes of bcd clusters
    - bcd only: anterior
    - bcd + hb: extend to posterior (synergistic interactions between bcd and hb increases sensitivity to bcd gradient)
    - bcd + Kr: limit to anterior (Kr is a repressor expressed at posterior)
    - bcd + hb + Kr: patterns vary considerably
- Conclusion:
  - bcd + maternal hb (anterior) / cad (posterior) $\rightarrow$ expression at middel body region
  - Broad pattern of a CRM is refined by repressors (kni, gt, and zygotic hb)

ChIP-chip analysis of Drosophila blastoderm transcription [Li & Biggin, PLoS Biology, 2008]:

- Aim: the properties of regions bound by TFs involved in fly blastoderm development.
- Data processing:
  - Smoothing: obtain the score (log ratio) of each position (or probe) in the genome, then the scores are smoothed using a sliding window of trimmed means 675 bp in length (the score of any position is the mean score of the 675 bp window covering the position).
  - Null distribution of window scores: (i) the symmetric null method assumes that the background window score distribution is symmetric about its mean - reflect the scores of the left of the mode to get its distribution in the right; (ii) the distribution of the negative control. The FDR threshold is then determined using the Q value (p values of all windows).
  - Defining bound regions: first filtering out all windows with scores above the given FDR threshold, then collecting these into contiguous stretches of windows containing a minimum of ten windows, with a maximum allowable gap of 200 bp between any two adjacent windows. Peaks are defined as the oligo or window with the highest scores in the bound regions.
- Methods:
  - Factors: bcd, cad, gt, hb, Kr, kni. PWMs of all except kni are from SELEX experiments (unpublished) and PWM of kni is from FlyReg.
  - TFBS enrichment: use PATSER to predict TFBSs; and the control set consists of random non-coding sequences that do not overlap with 1% FDR regions.
  - Define genes expressed in blastoderm: use RNA Pol II ChIP-chip data in blastoderm.
  - Measuring the sequence constraints of putative TFBSs: (i) PhastCons score; (ii) pairwise substitution rate with D. simulans.
- Results:
  - Comparison with known CRMs: 43 known CRMs are in 1% or 25% FDR regions. And they tend to have higher binding affinities.
  - Co-binding of factors: extensive co-binding, 82% of 1% FDR regions overlap a 25% FDR region by at least 500 bp for at least one other factor.

- Expression pattern and GO of target sequences: top 100 sequences of Kr, 49% are within 10 kb of some gene expressed in blastoderm, with all 1% FDR, about 22%. The GO enriched in the most highly bound regions are associated with: A-P patterning, developmental control and the regulatory of RNAP transcription.
- Many of the weakly bound regions (between 1% and 25% FDR) are probably non-functional: often in coding regions, less associated with the expected GO terms or the genes.
- Novel targets: in addition to sequences close to A-P genes (known to be regulated by A-P factors, or transcribed in blastoderm), there are (i) miRNA genes; (ii) D-V genes, including rho, twi, zen and sna.
- Enrichment of TFBSs: (i) enrichment of TFBSs in bound regions: specific, not found if using random PWMs or using false bound regions; (ii) high score TFBSs are highly enriched in bound sequences (e.g. the 8 highest affinity bcd sites are 8-fold enriched, 184 medium affinity bcd sites only 1.3 times 1.5 times enriched).
- Conservation of TFBSs: (i) strongly-bound > weakly-bound > unbound (integenic) > short introns; (ii) for bcd, cad and gt (not hb, kni and Kr), the conservation is higher than permutated PWMs, but only for bcd, the pattern is consistent with binding data (stronger binding more constraint).

Dual regulation by Hunchback [Papatsenko & Levine, PNAS, 2008]:

- Problem: the mechanism of expression of Kr in early embryo?
- Background: regulation of kni and gt expression
  - kni: by Bcd activator and Hb repressor;
  - gt: Bcd and Cad activators, and anterior boundary by Kr and posterior boundary by Hb.
- Background: three hypothesis of Kr regulation
  - Bcd activator and Hb repressor, similar to *kni* pattern. Quantitative simulations are inconsistent with this model [Jaeger, Mech Dev, 2007].
  - Bcd and Cad work in synergistic fashion to activate *Kr* in central regions. However, neither *bcd* nor *cad* mutations eliminate *Kr* expression.
  - Solely by Hb gradient, acting as concentration-dependent activator and repressor.
- Model: Hb as both activator and repressor. And the model assumes the enhancer is activated if some activator site is bound and no repressor site is bound [Zinzen, Curr Biol, 2006].
  - Dual-B model: some Hb sites are activator sites; some are repressor sites because they overlap with other activator sites, etc.
  - Dual-P model: normally Hb is an activator, but binding of multiple Hb molecules to neighboring sites or dimerization might physically block the activating domain. Ex. in the case of two linked Hb sites, activation if only one site is bound; no activation if both are bound.
- Results:
  - Both dual-B and dual-P model agree well with the observed Kr expression (fitting the parameters). But dual-B requires considerably higher binding constants for Hb activator sites than repressor sites; The dual-P model achieves optimal performance with comparable binding constants for different Hb sites.

3. Dorsal-ventral pattern formation

   Short-range repression of Kr and sna [Gray & Levine, GD, 1996]:

   - Results:
     - Kr, sna: if close to an activator site - quenching, reduce enhancer effect

  - Kr, sna: if close to TSS - direct repression, reduce target gene expression. Dominant effect: affect all enhancers of the same gene

- Discussion:
  - Two modes of repression: short-range (¡100bp), e.g. Kr, gt, kni, sna; and long-range (¿1kb), e.g. dl, hairy
  - Kr, sna: can be either quenching or direct repression, depending on the position of sites. The mechanisms can be: (i) short-range: by queching - PPI with activators. To quench a nearby BTM, may block activator as they loop to bind BTM or the repressor may recognize the common theme in both activator and BTM component. (ii) long-range: PPI with BTM (e.g. Kr with TFIIE-β). (iii) corepressors.
  - Non-specificity of repression: Kr could reduce dl activity in rhoNEE

Long-range repression of dl [Cai & Levine, PNAS, 1996]:

- Background: zen is expressed only in dorsal (reduced expression at ventral and lateral region). Its CRM (VRE, ventral response element) contains 3 dl sites and 3 corepressor sites (AT1-3).
- Results:
  - zen VRE has long-range repression activity: repress eve 2 CRM even if they are a few kb bps away
  - AT2 is important, but not AT1 and AT3
  - Spacing requirement: AT2 and dl2 separated by 10bps. Insertion of 5bp between AT2 and dl2 disrupts the expression, but 10bp insertion is tolerated.
- Discussion:
  - Long-range repression: probably via direct interaction with TC
  - Phasing/spacing requirement: suggests PPI. Not found for Kr, kni and sna, but found for dl2-AT2.

Gradient threshold reponse to Dorsal [Jiang & Levine, Cell, 1993]:

- Aim: how different target enhancers (with different number of Dl sites with different affinities) reponse to Dl graidents. In particular, twi enhancer contains weak Dl binding sites, but activated at neuroectoder where [Dl] is low, how?
- Background: twi enhancer has a distal element (DE) and proximal element (PE). The wt. twi is expressed in the neuroectoderm (laternal) where [Dl] is low, while the expression of PE is considerably narrower and weaker.
- Methods: synthetic enhancers with different Dl, bHLH and sna sites, inspect their expression pattern. The basic sequence template is twi proximal enhancer (PE).
- Results:
  - Number of Dl sites: in PE, there are two low-affinity Dl sites, TD4 and TF5. Increasing the number of sites (2 x PE or 4 x PE) will increase the level of expression, but still narrowed than the wt.. pattern.
  - Affinity of Dl sites: Convert TD4 and TD5 to consensus (PEe) will have a broad expression that is similar to wt.
  - Intearction with bHLL BS: insert Ebox (bHLH BS, twi is one bHLH) close to TD4 and TD5 will have a pattern similar to wt. Further more, use enhanced TD4 and TD5 along with Ebox will have even broader and stronger expression. The interaction depends on close linkage: if placed 260bp away, no effect.
  - Cooperative DNA binding of Dl and bHLH: using gel retardation assay.

  - The function of sna sites: use Ebox which is also recognized by sna will abolish the expression in the mesoderm (ventral).
- Discussion: Fig. 8
  - Number and affinity of Dl sites: both affect the expression pattern, the number of sites seems to affect more the level, while the affinity affects the breadth of expression.
  - Synergistic effect of Dl and bHLH: relies on close linkage, and cooperative DNA binding is involved (the experiment does not exclude the model of multiple contact with BTM).
  - Repression by sna: competitive binding may be involved, but overlapping sites is not necessary, thus short-range quenching seems to be important.

Multiple modes of Dorsal-bHLH synergism [Szymanski & Levine, EMBOJ, 1995]:

- Aim: organizations of different enhancers expressed in neuroectoderm (lateral) vs mesoderm (ventral).
- Discussion: Fig. 7
  - Neuroectoderm enhancers: closely linked Dl and Twi sites, probably involve cooperative DNA binding. The arrangement of sites are flexible (6, 11, 20, 25 bp does not affect expression), however, requires distance about 50-75bp.
  - Mesoderm enhancers: separately placed Dl and Twi sites, thus support the multiple contact model for synergism. Note that the synergism only happens between Dl and Twi sites, not multiple Dl sites. For example, 6 tandem Dl sites only drives weak expression, in comparison with 2 Twi and 2 Dl sites.
  - Minimal stripe unit (MSU): 57 bp of linked Dl and Twi sites, drive weak expression if single copy or separately spaced; if linked, drive strong expression.

Fly neurogenic genes [Markstein & Levine, Development, 2004]:

- Problem: enhancer model of dl controlled neuroectoderm gene expression
- Results:
  - Identify putative targets of dl by finding dl site clusters. Verified targets: rho, vnd, brk, which share gene expression patterns.
  - Search for additional enriched motifs in rho, vnd, brk enhancers: twi, Su(H), CTGWCCY
  - Additional targets by searching both dl and the extra motifs: vein, sim enhancers.
  - Conservation of dl, twi, Su(H), CTGWCCY reguatory in the mosquito genome: enahncer of orthologous sim gene.
- Remark:
  - Bootstrapping strategy: start with dl, find its targets, learn new motifs and find additional targets.
  - Conservation of regulatory relations (dl $\rightarrow$ sim) across divergent species (unalignable).
  - Tight linkage between dl and twi sites: may be an important feature (not captured by the method, which only requires clustering).

Drosal targets [Papatsenko & Levine, PNAS, 2005]:

- Motivation: dl gradient threshold model: 3 types of enhancers with different affinities for dl (affinity determines the expression pattern)?
- Methods:
  - Data: 18 dl target enhancers, times 4 (4 species).
- Results:

- Evoutionary analysis of dl and twi sites: rapid turnover, however optimal sites are more conserved.
- Comparison of 3 groups of enhancers: test if the mean cluster strength is different in differen groups. Find that using average score and best match measure: statistical difference between group 1 and group 2, 3 in both dl and twi.
- dl and twi: in group 1, they are additivie as they are compensatory, i.e. there exists negative correlation between the scores of the two motifs; in group 2, there exists synergy between the two, because certain arrangements (14, 20, 53bps) are favored.

- Discussion: why bcd and dl follow different models, where dl cluster strength is correlated with DV pattern while bcd cluster strength not with AP pattern? Answer may lie in: bcd is cooperative.

Shadow enhancers [Hong & Levine, Science, 2008]:

- Results:
  - ChIP-chip results of Dorsal: many of Dorsal targets contain two separate enhancers (as many as one third to one-half).
  - Some secondary enhancers confer similar expression patterns as the primary ones: (1) brk: the secondary enhancer located at the intron of a neighboring gene, 13kb downstream of brk TSS; (2) sog: 20kb upstream of sog TSS. In earlier studies, the secondary enhancers of vnd and mir-1 were confirmed, located within 5kb of TSS.
  - Conservation of secondary enhancers are lower than primary ones, but still constrained in all 12 species.
- Discussion:
  - What are the functions of the secondary enhancers? Why are they constrained? Possible answers: (1) dosage effect; (2) slighly different function, e.g. at a different stage of development.
  - Origin of shadow enhancers: if resulted from duplication, then the homology may be tested.
  - Shadow enhancers provide opportunity for evolving novel functions.

4. Fly expression atlas

   BDGP in situ data: Release 1 [Tomancak & Rubin, GB, 2002]:

   - Aim: atlas of spatial-temporal expression data in Drosophila embryogenesis
   - Methods:
     - Data generation: (i) RNA probe preparation and determination of probe strength; (ii) in situ hybridization in 96-well plate; (iii) low-magnification image of groups of embryos; (iv) high resolution images and annotation.
     - Temporal stages: 16 stages grouped into 1-3, 4-6, 7-8, 9-10, 11-12, 13-16. A stage ranges from 15 mins to more than 2 hours.
     - Annotation of gene expression patterns: manual because of the variation in morphology and incomplete knowledge of the shape and position of various embryonic structures. Use FlyBase controoled vocabulary, furthermore, 4 categories of developmental structures:
       * Anlage in statu nascendi: the larger domain from which a specific anlage orignates.
       * Anlage: a morphologically indistinct group of contiguous cells, established by lineage tracing, that gives rise to an individual organ. Anlagen for most organs can be distinguished at the late cellular blastoderm.
       * Primordium: can be recognized on the basis of its distinct morphology and will give rise to one or more differentiated organs or gastrula stage.
       * Organ.

For example, Mesectoderm anlage in statu nascendi, Mesectoderm anlage, Midline primorium, Midline glia. In addition, many genes are not expressed during embryogenesis, or not tissue specific (ubiquitous).

  - Hierarchical clustering of both annotation terms (99) and genes (1,257: expressed in at least one of the embryonic structures).

- Results:
  - Validation of indepdent time-course microarray data (Fig. 5). However, microarray data hides spatial information, many genes may share similar expression profiles, but distint spatial patterns.
  - The clustering of annotation terms and genes reveal genes with similar patterns, and tissues with similar expressed genes.
  - Interesting findings: (Fig. 9) 5 metabolic genes are expressed at cellular blastoderm in domains suggestive of roles in embryonic patterning.
- Remark:
  - The technique of in situ hybridization is qualitative, not quantitative. The procedure may be affected by probe sensitivity, thus need independent validation with microarray data.

BDGP in situ data: Release 2 [Tomancak & Rubin, GB, 2007]:

- Aim: atlas of spatial-temporal expression data in Drosophila embryogenesis
- Methods:
  - Annotation of expression patterns: 314 anatomical terms selected from FlyBase controlled vocabulary (CV). Group developmental structures into 16 color-coded organ systems, and reduce the full 314 terms into 145 terms.
  - Anatogram: anatomical signiture of a group of genes (which CV terms are overrepresented, underrepresented, etc.).
  - Hybrid clustering: combine microarray expression data and in situ data to cluster genes. Fuzz c-means clustering so that each gene is assigned to one or more clusters.
- Results:
  - Patterns of 6,003 (44%) of 13,659 protein coding genes. 4,496 (79%) genes have detectable expression in embryo, while the rest 1,244 (21%) are annotated with "No staining" CV term.
  - Clustering: 39 clusters. 2,549 (57%) fall into 10 broad clusters (if a significant fraction annotated as "ubiquitous" or unrestricted maternal only expression); and the remaining 1,947 (43%) fall into 29 clusters represnting restricted patterns.
  - Broad expression patterns: all have maternal expression followed by ubiquitous or broad expression (Fig. 4). The clusters, for example: late midgut and mesoderm; late midgut; late CNS, gonad, midgut; strong ubiquitous; etc. Enriched for genes involved in core cellular processes, such as translation, protein degradation, cell division, energy metabolism and RNA binding proteins.
  - Restricted expression patterns: (Fig. 6) tissue or domain specific, for example: Fat body; Trachea; Brain; Trunk somatic muscle; Optic lobe, SNS; anterior & posterior endoderm primordium. 795 (41%) genes are assigned to more than one cluster. Enriched in genes with sequence-specific DNA-binding domains and signaling molecules.

FlyAtlas [Chintapalli & Dow, NG, 2007]:

- Aim: the expression profile of each adult tissue in Drosophila melanogaster
- Background: most fly biology is focused on development. However, many genes may be involved in adult tissues, e.g. only 20% genes have been identified before the release of Drosophila genome.

- Methods:
  - Tissues: 9 aduilt tissues (brain, head, crop, midgut, Malpighian tubules, hindgut, testis, ovary and male accessory glands) and two larval tissues (tubule and fat body).
  - Platform: Affymetrix array measuring expression of 18,500 transcripts.
- Results:
  - About 25% are housekeeping genes, another 25% are specifically expressed in one tissue.
  - Most (90%) genes expressed in in situ experiment are also expressed in adult, suggesting the mature transcriptional profile of many tissues is substantially established by the late embryo.
  - Unexpected expression of many genes: e.g the olfaction gene, obp56d, is expressed in aduilt hindgut.

VirtualEmbryo of Drosophila (BDTNP) [Fowlkes & Malik, Cell, 2008]:

- Aim: provide a spatial-temporal atlas of gene expression in Drosophila blastoderm.
- Background: from images of multiple individuals to construct the atalas. The difficulties are:
  - Not possible to label the expression of more than a few more gene products in a given animal or tissue.
  - Invididual variability of the shape, cell positions, etc. Need registration techniques to establish correspondences by factoring out individual variations.
  - Absolute fluorescrence levels are not comparable across different embryos (especially at different batches).

  In the blastoderm stage, no cell dividision, and no cell membrane, about 6000 nuclei (small individual variation).
- Methods:
  - Data acquisition: in each embryo, stain expression of one marker gene (eve or ftz) and one gene of interest. 95 genes in 1822 embryos. 50 mins prior to gastraluation, divided into 6 stages, according to the percent of membrane invagination.
  - Spatial registration: the goal is to construct a morphological template with a fixed number of nuclei to match the average embryo shape and nuclei distribution, and the mean locations of the expression boundaries of marker genes. The template is created in an iterative fashion to maximize the correspondence between experimental embryos and the template.
  - Temporal registration: nuclear movement (but no cell division) at different time points. Use a numerical model that predict the direction and distance each nucleus need to move.
  - Compositing expression levels: normalize mRNA expression across multiple batches.
  - Inferring regulatory relations:
    - Segment the expression of each gene into discrete expression domains (modules). A total of 238 modules from 95 genes.
    - Regulators: 17 known regulators, use protein expression level or map mRNA to protein by a 16 min delay.
    - Regression of gene expression over regulator expression: suppose $x_1, x_2, \ldots, x_K$ are expression level of regulators and y be the target gene, then $y = 1/(1+\exp(a_1x_1+\ldots+a_kx_K+b))$.
    - Model fitting: least square minimization over all nuclei at all time points.
- Results:
  - Out of 238 modules, 202 were fit with R^2 of 0.5 or greater.
  - Failuers: e.g. Bcd does not appear as activators in many cases, including eve stripe 2. The analysis favor regulators with sharp boundaries.

BDGP image analysis [Frise & Celniker, MSB, 2010]:

- Motivation: computational processing of expression images (s.t. they can be grouped/compared, etc.) and extract functional information of gene expression (e.g. assign putative functions).
- Methods:
  - Image representation: a mesh of 311 equilateral triangles in the shape of an ellipse, with 16 anchor points (called triangulated image, or TI). Thus each image will have a simple vector representation. About 91% of images can be represented in this way (the rest may be out of focus, overstained, etc.).
  - Expression pattern clustering: use affinity propagation algorithm to cluster expression patterns. The similiarty metric: correlation coefficient (which is insensitive to the total staining intensity).
  - Boundary extraction: find the boundary that minimizes certain energy function under Markov random field (MRF), the function is defined s.t. the neighboring points across the boundary have different values than those within the same region.
  - Recognize complementary patterns: based on extracted boundaries of two images.
- Results:
  - Processed data: the total data has about 6000 genes (on average 10 images per gene). Functional analysis is performed on the stage 4-6 data: (1) 2693 TIs with 1881 genes; (2) 553 distinct TIs (365 genes) .
  - Patterns in stage 4-6: dataset (2) generated 39 distinct clusters. Most common ones are ubiquitous expression and posterior expression.
  - Mining for complementary patterns: may lead to functionally interacting genes, e.g. sna and hkb.

modENCODE: Molecular biology: A fly in the face of genomics [Nature, 2011]:

- Transcriptome data: high-resolution expression data in multiple development stages (whole-animal), which are complemented by an analysis of 25 Drosophila cell lines.
- Chromatin marks: two of the modENCODE studies involved mapping such chromatin marks in Drosophila cell lines and at 11 stages of its life cycle. Kharchenko et al. (page 480) identified nine prominent chromatin signatures, which complement those defined previously.
- Regulatory elements and TF binding: a systematic effort to identify all cis-regulatory elements by examining the occupancy of 38 transcription factors and other chromatin-regulatory proteins at different stages of development. The result is a collection of around 20,000 putative regulatory elements that include insulators, enhancers and promoters.
- Co-binding: tntegrating the binding patterns of all TFs leads to hypotheses of TF partnerships, involving co-binding to regulatory elements. But overlays of transcription-factor binding should be interpreted cautiously, particularly for factors with non-tissue-specific or partially overlapping expression: regions that are co-targeted by multiple factors are not necessarily co-bound in the same cells.
- Remark:
  - Understanding the regulation of enhancer activity requires knowledge of which transcription factors are binding to them, in which cell types, and when. Scaling this up to the roughly 700 predicted Drosophila transcription factors is a monumental undertaking, but feasible given current tagging technologies
  - A major drawback of the data sets is their lack of temporal and spatial resolution. The general absence of functional information is perhaps the most serious limitation of the current work and a major challenge for all genomics projects. Such information is essential to understand the relevance of regulatory connections.

- Many of the transcription factors examined are expressed across a broad range of tissues, which has the advantage of covering a wide range of cis-regulatory elements. But merged transcription-factor occupancy signals from multiple tissues make it very difficult to disentangle regulatory connections and thus to build reliable regulatory networks.
- Moving forward, there is a clear need to integrate diverse types of functional data in order to make the transition from correlations to regulatory function. The thousands of Drosophila mutants available should provide a useful resource for this.

4.2 Mammalian Systems

1. Expression atlas

   GNF gene atlas [Su & Hogenesch, PNAS, 2004]:

   - Methods:
     - Data: 79 human and 61 mouse tissues. Human tissues include: (1) organ level: liver, heart, whole brain, lung, testis, skin, smooth muscle, skeletal muscle, etc.; (2) tissue level: prefrontal cortex, adipocyte, beta cell islets in pancreas, salivary gland salivary gland; (3) cell level: CD4+ T cells, lymphocytes, etc. Customized array coverring 44,775 human and 36,182 mouse transcripts.
     - Region of correlated transcription (RCT): defined as a window of 3-10 genes in chromosomes, where more than 50% of pairwire expression patterns show a correlation coefficient > 0.6.
   - Results:
     - Overall expression statistics: more than 50% genes are expressed in at least one tissue. The average number of transcripts expressed in one tissue is $\approx 8,200$. About 1-3% of transripts are expressed ubiquitously (housekeeping genes): typically about 30-fold higher expression than other genes.
     - RCTs: detection heavily influenced by normalization procedure and data. Found 156 and 108 RCTs for human and mouse. About 60-70% of RCTs have genes with low sequence similarity, thus likely controlled by a multi-gene locus-control region (LCR), instead of tandem gene duplication. The majority of RCTs are not conserved between human and mouse: e.g. a mouse-specific RCT expressed in olfactory bulb (likely from mouse-specific physiology).

   More than a decade of developmental gene expression atlases: where are we now? [deBoer & Moorman, NAR, 2009]:

   - Databases:
     - GXD: mouse, TS 1-26 (most of the embroyic development) and adult. More than 8,000 genes.
     - ZFIN: zebrafish, all 44 stages. More than 10,000 genes.
     - GEISHA: chicken, HH 2-27. 1,000 genes.
     - MEPD: medaka fish (close to zebrafish), Iwamatus stages 15-44. More than 1,000 genes.
     - 4DXpress and COMPARE: combine the expression information of more than one species, mouse, zebrafish and fly.
   - Framework:
     - Two basic ways of annotating expression patterns: (1) manual annotation with anatomical structures; (2) 3D reference models.
     - Biological variation in embryo morphology. Especially, the temporal variation in organ development hampers the mapping of sections to 3D reference models of whole embryos. Mapping gene expression data from mutant embryos will pose additional challenges because their morphology can be very different from that of wild-type embryos.

- The problem of focusing on an individual organ: the area studied can be too narrow to cover the full development of that organ. For example, the main source for the growth of the early embryonic heart tube is the addition of cells from the pericardial mesoderm, which is not part of the heart.
- Controversies often surround anatomical nomenclature. These controversies can be, at least partially, solved by using annotated reference models.
- Gene expression patterns do not respect borders of anatomical structures. This problem can be addressed by combining the anatomical annotation with an annotation based on the expression profiles of a limited set of genes with known expression domains.

Human Expression Map [Lukk & Brazma, NBT, 2010]:

- Methods:
  - Data: 5,372 human samples representing 369 different cell and tissue types, disease states and cell lines. All based on Affymetrix U133A. 96 groups contained at least ten biological replicates.
  - Processing: The raw data were normalized jointly, producing a gene expression matrix of $\sim$ 22,000 probe sets (mapping to $\sim$ 14,000 genes) times 5,372 samples.
- Results:
  - PCA on gene expression matrix: the first three principal components have biological interpretations; we named them the hematopoietic, malignancy and neurological axes.
  - Clustering of samples: (i) cell lines derived from solid tissues, (ii) incompletely differentiated cell types and connective tissues, (iii) solid normal and neoplastic tissues, (iv) hematopoietic system, (v) brain, and (vi) muscle and heart. Note that solid-tissue cell lines form a distinct group, clustering with each other rather than with their respective tissues of origin.

2. ENCODE

Review of ENCODE data: A User's Guide to the Encyclopedia of DNA Elements (ENCODE) [PLoS Biol, 2011]:

- Cell types: (1) The highest priority set (Tier 1) includes B-lymphoblastoid line and ESC. (2) Tier 2 includes HeLa-S3 cervical carcinoma cells, HepG2 hepatoblastoma cells, and primary (non-transformed) human umbilical vein endothelial cells. (3) Tier 3: currently comprises more than 100 cell types that are being analyzed in selected assays.
- ENCODE data analysis: (1) defining promoter and enhancer regions by combining transcript mapping and biochemical marks. (2) assign TF binding to genes: GREAT is the tool used. (2) Defining transcription factor co-associations and regulatory networks.

ENCODE ChIP-chip analysis [Zhang & Gerstein, GR, 2007]:

- Aim: analysis of comprehensive binding data of a large number of factors in different cells/conditions.
- Background: a region bound by TF (promoter, TFBS) is called a TRE.
- Methods:
  - Data: 29 factors, in multiple cells/conditions, a total of 105 profiles; in 30Mb genomic regions.
  - Non-random spatial distribution of TREs: divide into genomic subregions, and count the number of TREs in each region. Test if the number is uniformly distributed: the null distribution comes from random permutation of TREs in non-repetitive regions, then test the observed distribution with random ones using χ^2 test.
  - Correlation of spatial profiles:

        * Build spatial profile: divide into subregions, then either number of TREs in each region (to compare TRE distribution with gene distribution), or the number of nucleotides occupied (relationship between factors/conditions).
        * Pearson correlation of spatila profiles.
        * Significance testing: null distribution from random permutation of the TREs or random shuffling of the subregions (or only the bound sequences). Note: this is necessary because the data is not normally distributed (counts).
    - Enrichment in spatial regions: count the number of TREs in a certain type of region and compare with the null distribution, again from random permutation of TREs.
    - Relationship among factors/conditions: the data is represented by the 2D binary (1 or 0 per nucleotide) or count matrix (number of 1s in a subregion, use 5kb), where rows are regions and columns are factors.
        * Biplot (PCA): (i) standardization of variables (columns); (ii) correlation matrix of the variables; (iii) PCA of the correlation matrix; (iv) each variable is now represented by a line in the low dimension space (choose 2 principal components), and the cosine of the angle between two lines indicate how close the two variables are. Furthermore each observation (row) is now represented as a point in the new space, and its projection in the variable lines represent the observed values.
        * Hierarchical clustering: of the factors, using correlation matrix. The bootstrap value of each branch can be obtained by: sampling with replacement of the count matrix and construct the tree for each new random matrix.
- Results:
    - Spatial distribution of TREs: (i) non-randomly distributed; (ii) correlated with gene density; (iii) enriched in regions close to TSS or TES (transcription end site).
    - Relationship of TFs: (i) two broad classes: sequence non-specific factors (Pol II, histone modification, etc.); and sequence specific ones (Suz12, STAT1, etc.); (ii) some unexpected associations: e.g. cMyc was thought to be sequence specific, but is shown non-specific pattern, has close association with H3ac and H4ac.
    - Reproducibility of labs and microarray platforms: generally high correlation, but relatively low for specific factors; and the labs seem to have larger impact than the platforms.

3. Transcription factors

   Human transcription factors [Vaquerizas & Luscombe, NRG, 2009]:

- TF repertoire in human:
    - Procedure: 347 InterPro DNA-binding domains followed by manual curation (e.g. genes with known function other than transcription, and more TFs from GO).
    - 1,391 high-confidence TFs and 216 possible TFs. Most human TFs are unannotated, and most observations are inferred from other organisms.
    - Famillies: the biggest ones are C_2H_2-Zn-finger (675), homeodomain (257) and helix-loop-helix (87). Note that TF family can suggest function: e.g. homeodomain often associated with developmental process; interferon family often with immue responses against viral infections.
- TF expression patterns:
    - Data: GNF data, 32 healthy tissues and organs. Choosing appropriate threshold to define TF presence or not in each tissue.
    - TF expression level: TFs tend to have lower expression than other genes; of the 873 TFs represented in the array, 510 are expressed in at least one tissue.

  - TF expression pattern: in each tissue, about 150 - 300 TFs are expressed (this result is affect by tissue homogenenity). 161 TFs are expressed in most tissues (ubiquitous) and 349 TFs are selectively expressed (or significantly higher expression) in a few tissues. In general, there is a substantial overlap in TF expression between the developing and the adult stages of the same tissues.
  - Combinatorial usage of TFs: e.g. serum response factor (SNF) is a ubiquitous TF involved in cell proliferation and differentation (MCM1 homolog). In smooth muscle, SNF interaction with NFAT and HERP1; and in prostate, interaction with NKX3-1.
- Evolutionary history of TFs:
  - Expansion: 13% of human TFs appeared in primates (vs 2% for metabolic enzymes). Homeodomain TFs first appear in metazon, then expanded rapidly in vertebrates; helix-loop-helix TFs origniated in metazon but have not expanded significantly since; C_2H_2-Zn-finger family grew at several evolutionary stages. The ability to mutate AAs that directly interact with DNA and the capacity to extend the length of binding sites by linking multiple domains in a sequential manner may explain the expansion of C_2H_2-Zn-finger family.
  - Chromsome distribution and paralogs: 23 high-density clusters of 284 TFs: all are C_2H_2-Zn-finger except one Hox cluster. 15 clusters from tandem duplications. Example: one C_2H_2-Zn-finger cluster in chromosome 19: KRAB-ZNF, which combine C-termianl DNA-binding zinc-fingers with an N-terminal Kruppel-associated box domain. This is the biggest family of TFs in huamn (400 genes). 8 clusters in centromeres and telemeres, probabaly from recombination.
  - Evolution of new regulatory functions: TFs tend to be under positive selection, and there expression level also show accelerated change between human and chimp.

Combinatorial regulation in mouse and human [Ravasi & Hayashizaki, Cell, 2010]:

- Hypothesis: tissue specificity is determined by combinatorial interactions among TFs.
- Methods:
  - TF interaction data: 1222 human TFs and 1112 mouse TFs, interaction in mammailan 2 hybrid (M2H) system using CHO-K1 cells.
  - TF expression data: qRT-PCR in 34 human and 20 mouse tissues. Human tissues (Figure 4A): adipose, liver, heart, kidney, fetal brain, whole brain, frontal cortex, skeletal muscle, etc.
  - Tissue separation: to test what may predict tissue types (expression level of genes, of TFs or TF interactions), use these test variables to cluster tissues, and see if the tissue clusters correspond to their embryonic origins. The clustering methods: PCA and clustering using projection in one PC.
- Results:
  - TF-TF interactions: 762 and 877 interactions in human and mouse. Specificity: 53% (by testing with another technique) and sensitivity: 25% from literature-curated interactions.
  - TF expression landscape: bimodal distribution of TFs (facilitators and specifiers) - a large fraction of TFs are expressed in a broad range of tissue types. TFs with few interactions tend to be expressed in a tissue-specific pattern (strong negative correlation, -0.79, between the number of interactions and the breadth of expression).
  - Tissue specificity by combinatorial TF interactions: TF interactions are more predictive of tissue types than expression levels of genes or of TFs. A network of 15 TFs, and importantly six interactions, are sufficient to classify tissue types. This subnetwork was enriched for homeobox TFs and 10 of 15 were facilitators.
  - Conservation of TF complexes: 80 (305 plus literature) interactions are conserved between human and mouse. 68 conserved TF complexes are identified, most of which were enriched for GO biological processes.

- Discussion: tissue identity is determined not by tissue-restricted TFs, but on tissue-restricted interactions among TFs. Tissue-restricted TFs tend to interact with TFs that are broadly-expressed.

4. Tissue-specific gene regulation

   Predicting human and mouse tissue-specific expression [Smith & Zhang, PNAS, 2006]:

   - Problem: discrimate the genes that are up-regulated vs down-regulated in a certain tissue, via their regulatory sequence features
   - Methods:
     - Data: 28 tissues of human and mouse each from Gene Atalas, and 2 groups of genes that are most strongly up- and down- regulated, respectively in each tissue
     - Sequence features: candidate motifs include the known ones from TRANSFAC and the motifs found via a "discriminative motif finding" algorithm; in addition, pairs and triplets of motifs are also considered
     - Building discrimative model: the predictors are scores based on sequence features (e.g. max binding score of a motif to a promoter sequence). The predictors are used to build a regression model, MARS, where response variable is 1 or -1 (up vs down)

   Predicting tissue-specific enhancers in the human genome [Pennacchio & Ovcharenko, GR, 2007]:

   - Problem: given a set of genes with expression in certain conditions (tissues), and a set of motifs, find enhancers in each gene that leads to the expression profile
   - Methods: regression of TFBS contents in a candidate enhancer over expression (tissue specificity in this case)
     - Candidate enhancers: for each gene, find 3 most evolutionarily conserved regions (ECR), defined by at least 100 bp and 70% PID in human-mouse
     - TFBS mapping: scan each candidate enhancer with Transfac PWMs, find conserved TFBS with rVista, choose p value = 0.0005.
     - Scoring of candidate enhancer: (interpreted as tissue expression level) score of candidate enahncer k of tissue t, S_k^t, is determined by the TFBS content:

       $$S_k^t = \sum_i w_i^t N_k^i \tag{4.1}$$

       where w_i^t is the weight of i-th TF at tissue t and N_k^i is the number of i-th TFBS in candidate enhancer k
     - Regression: each candidate enhancer is associated with the expression of the target gene (up or down in the tissue), optimize parameters by maximize the matches (candidate enhancers whose scores are consistent with expression). No feature selection or Lasso-type regression is done.
     - Selecting tissue-specific factors: each factor is ranked by its relevance to a tissue: the product of its fraction (occurrence in all genes in that tissue) and its weight learned from the regression.
   - Results: 79 human tissues, 554 TransFac PWMs. For each tissue, all genes are divided into up and down sets and apply the methods
   - Remark/Criticism:
     - Multiple ECRs per gene, generally only one of them may contribute, but in the regression, all ECRs contribute. The CRM searching step is missing.
     - Each tissue is analyzed separately; in practice, one may have the constraint that there exists a single CRM leading to the correct expression profile (over multiple conditions).

- No higher level feature is modeled: combinatorial regulation by multiple motifs. Ex. neither of two motifs is predictive of expression, but requiring the presence of both motifs may be very suggestive of an expression pattern.

Tissue-specific regulatory elements in mammalian promoters [Smith & Zhang, MSB, 2007]:

- Problem: find motifs and modules (defined as a set of motifs) that generate tissue-specific gene expression.
- Methods:
  - Identifying tissue-specific genes: combine the microarray data of GeneAtals and Hughes Toronto data, and other functional annotation, a gene is tissue specific is in multiple sources, it has a restricted or unusually high expression in that tissue.
  - Identifying TFs expressed in each tissues.
  - Promoter sequences for tissue-specific genes: -1000 to +100 bps.
  - Identifying tissue-specific motifs: for each tissue, construct a foreground (FG) set of promoters which are promoters of genes specific to this tissue; and a background (BG) set of promoters. Given a motif, each promoter is scored by the best matching to this motif. The motifs that give a good classification of FG and BG sets are tissue-specific motifs, defined as balanced error rates at the optimal threshold.
  - Identifying tissue-specific modules: similar to motifs, find the modules that classify FG and BG sets. Search the module in a greedy fashion with the program MODULATOR: add a motif only when it significantly improve enrichment.
- Results:
  - Correlation between human and mouse CREs: motif enrichment ranks (any motif has an enrichment score in a certain tissue, rank all motifs in human and in mouse) are highly correlated for all but CD4 T cells. However the site order is rarely conserved - fewer than 10% of ortholgous pairs show significant conservation of site order.

5. Case studies

Estrogen receptor binding [Carroll & Brown, NG, 2006]:

- Methods:
  - Data: ChIP identified regions bound by estrogen receptor (ER)
  - Finding co-localized motifs: search for motifs in the ChIP bound regions vs promoters of putative non-targets (expression not changed by estrogen) by either de novo motif finding with MDScan or candidate scanning using Transfac and Jaspar motifs (number of sites by using a threshold, hypergeometric test).

  Results:
  - Colocalized motifs: (in addition to ERE) Forkhead, Ap-1, Oct, C/EBP.
  - Co-binding analysis: test binding of the colocalized motifs to the ER bound regions. Binding of the colocalized motifs is enriched in the ER-bound regions.
  - Correlation of motifs: for each sequence, score its relevance to all motifs; then test the pair-wise correlation of motifs. Some have positive correlations; some have negative ones (AP-1 and ERE).
  - Positional distribution of motif sites: draw position (relative to the center of binding) of all sites of a relevant motif. ERE, AP-1 have peaks at the center, while other motifs have flatter distribution (or multimodal for Oct).

ATF3 in macrophages [Gilchrist & Aderem, Nature, 2006]:

- Problem: from LPS-induced transcription profiles of macrophages, infer the regulatory programs that control important cytokines?
- Methods:
  - Find motifs in a set of co-regulated genes: MotifMogul, scan -3K to +500bp region for known motifs and test the statistical significance.
  - Given the expression profile of a gene, a set of motifs (known to be important to this gene) and the binding profiles of these motifs: find the expression model, i.e. the effect (activation or repression) of the TFs. Do this via kinetic regression: the kinetics (time derivative) of mRNA of a gene is a linear combination of binding of different TFs.
- Results:
  - LPS-indued profile: 11 clusters. Early cluster (cluster 6): enriched with motif of ATF3, and ATF3 is a member of cluster 6 $\Rightarrow$ ATF3 is a candidate for controling the LPS response.
  - Analysis of cluster (important cytokines): ATF3 and NK-κB, or Rel, (known to have PPI) motifs cooccur in IL2b and IL6 within 500bps of TSS.
  - Kinetic regression analysis of expression vs binding: ATF3 as negative regulator
  - Testing and mechanism of ATF3 on IL2b and IL6:
    - ATF3-/- mutant: increased IL2b and IL6 expression
    - HDAC1 (deactylation): recruited by ATF3.
- Discussion: LPS $\rightarrow$ Rel: histone acetylation $\rightarrow$ increased IL2b and IL6 expression; then ATF3 reduces their expression by recruiting HDAC1 (histone deactylation).
- Remark: how to use expression, sequence and ChIP data to come up with specific predictions/hypothesis?
  - Consider the meaning of expression pattern (clusters): e.g. the kinetics of expression provides the interpretation of clusters. Ex. early-activation clusters determine the late response.
  - Motif interactions/coordinations are often important (could use known TF-TF interaction data): better predict target genes or CRMs
  - Determine the role/effect of TFs in expression: integrate expression and sequence/binding data (and expression of TFs themselves).

Glucocorticoid response elements [So & Yamamoto, PLoS Genetics, 2007]:

- Problem: the sequence features (additional motifs), location distribution and conservation of glucocorticoid response elements (GREs, regions bound by glucocorticoid receptor, GR)?
- Results:
  - GRE occurrence strongly correlates (88%) with glucocorticoid response genes in A549 cells. Exp: find 73 GBR (glucocorticoid bound region), 88% of the genes are known to respond to glucocorticoid.
  - Glucocorticoid receptor (GR) binding motif: vary greatly across sites, but individual sites are very conserved across human, mouse, rat and dog genomes.
  - GREs are evenly distributed upstream and downstream, no preference for promoter-proximal regions. Most GREs are remote from TSS (63%: >10kb from TSS). The distribution is consistent with that of conserved non-genic (CNG) sequences.
  - GREs are composite elements: contains multiple TFBSs.

CTCF binding in human [Kim & Ren, Cell, 2007]:

- Aim: chacterization of CTCF binding in human genome.
- Background:

  - CTCF has a zinc-finger DNA binding domain.
  - CTCF is an insulator protein that binds insulator elments to prevent the spread of heterochromatin and restrict transcriptional enhancers from activation of unrelated promoters.
  - CTCF is ubiquitously expressed.
- Methods:
  - CTCF binding: ChIP-chip by whole-genome tiling array in primary human fibroblasts.
  - CTCF consensus motif: learned from discriminant motif finding program [Smith & Zhang, Bioinfo, 2005] where all CTCF binding regions are positive and negative sequences are flanking regions.
  - CTCF binding site identification: matches to the learned 20-mer matches of CTCF motif across genome.
- Results:
  - 13,804 CTCF binding regions are identified, largely invariant of cell types.
  - Distribution of CTCF binding sites: majority are far from TSS, consistent with the role of CTCF as insulator.
  - CTCF motif is highly predictive of in vivo binding.
  - Conservation of CTCF binding sites (motif matches) (i) within the CTCF bound region: high PhastCons scores vs matches to random motifs; (ii) Predicted CTCF binding sites (genome-wide scan): 31,905 potential CTCF-binding sites, out of which 6,553 are conserved (matches in the aligned region) in mouse genome. Only average 149 conserved occurrences if using random matrix.
  - Evolution of CTCF binding sites: using 12,799 conserved sites (conserved in at least one vertebrate genome), an unusually high C-T substitution at position 16.

4.3 Misc Cases

1. EEL validation [Hallikas, Cell, 2006]

   Problem: an algorithm that predicts CRMs and TFBS from given TF specifities, how to validate the predictions?

   Methods:

   (a) If a TF is involved in some response/process, then the target genes of this TF will be enriched in the set of genes that are involved, or equivalently, the number of TFBSs of this TF will be enriched in all genes involved. Use this to verify that the known TF is indeed involved in a certain process.

   (b) Validation of one CRM: the expression pattern of the endogeneous target gene; and compare that with the expression pattern of the reporter gene driven by this CRM. Expectation: reproduce the endogenous pattern or part of the endogenous pattern. Weaker results: produce a specific pattern that is display by some other related genes (e.g. the targets of the same TF)

2. Organizational flexibility of CRM in Ciona [Brown & Sidow, Science, 2007]

   Problem: for genes with similar expression pattern, do they share a certain pattern of CRM organization?

   Methods:

   (a) Data: 19 muscle genes that are tightly coregulated (17 belong to the same macromolecular complex), 3 TF motifs with 77 known sites.

   (b) Determine the activity of each binding site wrt. each target gene:

- Fine-scales deletion or site-directed mutations that target the TFBS;
- For each CRM mutant, measure the activity of CRM through expression of the reporter gene in muscle cells of the transfected embryo;
- The activity of each TFBS is estimated through regression of site presence/absence in the CRM mutants and activities of the mutants

Results:

(a) No apparent rules/features of CRM organization for 19 coregulated genes: little epistasis, spacing, order and orientation

(b) Cnce a CRM formed, little change in motif activity, order or composition between orthologous C intestinalis and C savignyi (divergence mammals and birds). Functional sites are very conserved, PID = 79% vs. the genomewide background PID, 20%. And exists sharp boundary, the PID markedly drops to the backgroud level outside 12bp of the functional sites.

(c) TFBS conservation is strongly correlated with its activity

Question. can the motif activity be predicted? Correlation with the TFBS binding energy? The spacing/placement with other TFBS?

Chapter 5

Physiology & Medicine

Approaches to studying the etiology of complex diseases [personal notes]

- To link a causal factor with a disease: e.g. microbiome and depression. Three aspects: mechanisms of factor to disease, Observational evidence and Experimental/manipulative evidence.
- Understand better the mechanism of the factor and the disease. Ex. microbiome and depression: link through inflammation and chemicals (GABA, BDNF).
- Observational evidence: association and correlation between factor and disease.
  - Association of factor, including its marker, with disease: e.g. depressin patients often have GI tract problem.
  - Association of things that perturb the factor with disease: e.g. autoimmune disease is a risk factor of depression; early life experience that disrupts microbiome (such as C-section) is a risk factor of mental disorders.
- Experimental/manipulative evidence: changing factor would change disease risk.
  - Importance of a model system: e.g. microbe-free mice show behavior changes that mimic depression.
  - Experimental studies with human: e.g. use probiotics, fecal transplanation, etc. to manipulate microbiome and observe their effects on depression risk.
- The same reasoning above applies to the study of downstream effects of factors (or intermediate factors) with disease.
  - Association of intermediate factors with disease: e.g. lower levels of GABA in depression patients; increased inflammation in depression.
  - Experimental evidence of intermediate factors: e.g. incuding inflammation level via TNF-alpha (in hepatitis C patients) increases the risk of depression.

5.1 The Endocrine System and Metabolism

Reference: [Vander, Human Physiology: the mechanisms of body function, 2001 ed]

1. Homeostatic mechanisms [Vander, Chapter 7]

   Design goals of the endocrine system:

- Need of endocrine system: the body needs to take appropriate actions under environmental (and internal) changes. These actions typical involve multiple components of the body, and thus require a messenge system for communication and coordination of different components. Many of these actions may involve long-term changes (metabolic, growth-related, etc.), and cannot be accompolished by nervous systems.
- Design goals: the system basically ensures appropriate actions are taken, thus several aspects of its function are important:
  - Signal recognition: be able to distinguish signal and noise, and recognize specific signals.
  - Appropriate level of response: it is important to have the right level of response, and this often involve integrating signals from multiple sources, especially feedbacks.
  - Anticipation: it is often advantageous to anticipate changes, thus the system may take anticipatory actions. This is particlualry important for long-term responses.
  - Coordination of multiple components: important for complex changes.
- Homeostasis: a particular type of action is homeostasis, which ensures the stability of certain parameters when the environment changes.

General principles of homeostasis control:

- Balancing input and output: generally the internal environment depends on both input and output. Ex. temperature of a body depends both on heat generation and heat loss.
- Set point: the operating value of the internal environmental variable. It is generally maintained in a certain narrow range.
- Negative feedback: the basic mechanism of homeostasis when the environment changes. Ex. when environment temperature reduces, the body constricts blood flow to reduce heat loss, and shiver to increase heat generation, these processes compensate for environmental change.
- Physiological adaptation: the set point may be reset by other conditions. Ex. during fever, the set point of the temperature system is increased to boost immune responses. This also implies that a homeostatic control system takes input from other systems for its optimal function.
- Clashing demands: to main homeostasis of one internal variable, the body may need to make changes of other systems, thus shifting the homeostasis of these other systems.

Implementation of homeostatic control:

- Reflexes: consisting of stimulus, receptor, afferent pathway, integrating center, efferent pathway, effector and response. The response typically serves as negative feedbacks of the stimulus. Earlier used in nervous systems, but also applied to endocrine systems, where the integrating center may be endocrine gland and the effector may be hormones.
- Intercellular messangers: hormones, neurotransmitters and paracrine/autocrine agents. The later consistes of a diverse set of molecules (e.g. growth factors, nitric acid) and play many different roles, including local negative feedback of hormone signals, and mediator of the hormone signals.
- Eicosanoids: an important class of paracrine agents, they are derivatives of polyunsaturated fatty acids. Typically, some stimulus (e.g. hormone) activates phospholipase A_2, which split phospholipids in the membrane. The substrate is metabolized then in two pathways: (1) by cyclooxygenase (COX) and leads to prostaglandins, endoperioxides, etc. (2) by lipoxygenase and leads to leukotrienes. Many common drugs target eicosanoid pathways, e.g. aspirin inhibits cyclooxygenase and is a member of nonsteroidal anti-inflammatory drugs (NSAIDs).

2. Principles of hormone control system [Vander, Chapter 10]

   Hormone structure and synthesis:

- Amine hormones: derivatives of the amino acid tyrosine
  - Thyroid hormones (TH): two iodine-containing amine hormones, thyroxine (T4) and triiodothyronine (T3). T4 is secreted in much larger amounts than is T3, but enzymes in many cell types convert T4 to T3, and T3 is much more active than T4. TH affects almost all cell types: including regulation of oxygen consumption, growth, and brain development and function.
  - Adrenal Medullary (inner part, as opposed to adrenal cortex) hormones: epinephrine (E) and norepinephrine (NE). They exert actions similar to those of the sympathetic nerves (stress response).
  - Dopamine: secreted by certain cells in the hypothalamus.
- Peptide hormones: the great majority of hormones, secreted by cells through secretory pathway. Note that many peptides serve as both neurotransmitters (or neuromodulators) and as hormones.
- Steroid hormones: Cholesterol is the precursor of all steroid hormones. The steriod hormones are highly lipid-soluable, and can easily diffuse across the cell membrane.
  - Adrenal cortex: Aldosterone for salt (mineral) balance; cortisol (glucocorticoids) for regulation of metabolism, stress response and immune system; androgens.
  - Gonad hormones: androgens are precursors of testosterone and estradiol.

Control of hormone secretion:

- Changes in the plasma concentrations of mineral ions or organic nutrients: e.g. insulin.
- Control by neurons: (1) autonomous nervous system: controls endocrine glands such as adrenal medulla; (2) neurons in brain, in particular, hypothalamus and its extension, the posterior pituitary.
- Control by other hormones.

Control system involving hypothalamus and pituitary gland:

- Hypothalamus (neurons) controls anterior (through hormone) and posterior (the hormones synthesized in hypothalamus enter here) pituitary glands.
- Posterior pituitary gland: oxytocin and vasopressin. Vasopressin participates in the control of water excretion by the kidneys and of blood pressure and Oxytocin acts on smooth muscle cells in the breasts and uterus.
- Anterior pituitary gland: (1) folliclestimulating hormone (FSH) and luteinizing hormone (LH): germ cell development and stimulate gonad cells to secrete sex hormones; (2) growth hormone: stimulate liver and other cells to secrete IGF-1 and other function in metabolism; (3) thyroid stimulating hormone (TSH): stimulate thyroid (4) prolactin: breast development and milk production; (5) ACTH: stimulate adrenal cortex to secrete cortisol.
- Negative feedbacks: for hormones controlled by anterior pituitary gland, they often exert negative feedbacks on hypothalamus or anterior pituitary gland, e.g. the increased level of cortisol feeds back to inhibit the hypothalamus and anterior pituitary.
- Control by other hormones: e.g. estrogen enhances the secretion of prolactin by the anterior pituitary.

3. Regulation of metabolism [Vander, Chapter 18]

   Principles of metabolic regulation:

   - The main design goal is to meet energy and material (mainly protein biosynthesis) demands under uncertain conditions, specifically: uncertin food availablility, uncertain energy consumption, and long-term change of energy demand (growth, and conditions such as pregnancy).

- Storage systems: essential for addressing the uncertainty in supply and demand. Multi-level storage systems: both local (e.g. in muscle) where energy is needed and central (adipose tissues) storage are needed, as local storage is quick but has limited capacity.
- Food intake: convert extra food into materials that can be saved into the storage system. Need to regulate to: (1) store the amount of extra food appropriately, and (2) when the storage runs low, signal the body to take more food.
- Consumption: need to regulate to: (1) release the storage to match the consumption and (2) when the storage is low, reduce unnecessary consumption.

Metabolism of absorptive state: from food intake to storage or utilization

- Absorbed carbohydrate/glucose: beyond utilization, storage as glycogen in liver and skeletal muscle, and storage as fat in adipose tissue. The latter involves: (1) glucose to fat conversion in liver; (2) transportation to adipose tissue through VLDL particles; (3) absorption by adipose cells: lipoprotein lipase (LPL) break down fat and absorb fatty acids; (4) within adipose cells, fatty acid and α-glycerol phosphate (synthesized from glucose) reforms TG.
- Absorbed triacylglycerols (TG): forming the lipoprotein, chylomicron and absorbed by adipose tissues through lipoprotein lipase (similar to VLDL transport).
- Absorbed AAs: mainly taken up by tissues for protein synthesis, the excess AAs are converted to carbohydrate or fat in liver (generating urea in the process).

Metabolism of post-absorptive sate: utilization of storage to meet energy demands

- Blood glucose: from three sources (1) the hydrolysis of glycogen stores in the liver and skeletal muscle (only to pyruvate and lactate in muscle and then enter liver to be converted to glucose); (2) lipolysis: the catabolism of triacylglycerols yields glycerol and fatty acids, and glycerol is converted to glucose in liver; (3) protein becomes the major source of blood glucose a few hours after entering post-absorptive state: AAs are converted to glucose in liver.
- Glucose sparing (fat utilization): (1) The circulating fatty acids (bound with albumin, called free fatty acid, FFA) are picked up and metabolized by almost all tissues, excluding the nervous system through β-oxidation; (2) FA converted to ketones at liver, and can be used for all tissues.

Endocrine and neural control of the absorptive and post-absorptive states:

- Question: how would the major transitions (storage during absorptive state, and release during post-absorptive state) are controlled, in different cells (mainly liver, adipose and skeletal muscle)?
- Insulin: increase storage, i.e. glucose uptake (through glucose transporters, e.g. GLUT-4), glucogen synthesis in liver and muscle, TG synthesis in adipose, AA uptake. The secretion of insulin is controlled by:
  - Feedback control: lower plasma glucose level increases insulin secretion.
  - Feedforward control: hormone in GI tract, autonomous nervous system (parasympathetic neurons stimulate while sympathetic neurons inhibit insulin secretion).
- Glucagon: from α pancretic cells, the effect is to increase release from liver cells: increase glucogen breakdown, glyconeogenesis and ketone synthesis. The control of glucagon is similar to insulin: plasma glucose, autonomous nervous system, etc.
- Epinephrine and sympathetic nerves: influence insulin secretion, in addition, epinephrine directly controls metabolism: (1) glycogenolysis in both the liver and skeletal muscle, (2) gluconeogenesis in the liver, and (3) lipolysis in adipocytes.
- Cortisol: increase of cortisol during stress has the opposition effects of insulin: increase gluconeogeneisis and lipolysis, and decrease glucose uptake by muscle and adipose cells.

- Fuel homeostasis during exercise: increased glucagon secretion and decreased insulin secretion during exercise. One main signal for these actions is increased circulating epinephrine and enhanced activity of the sympathetic neurons. During prolonged excerise, decrease of plasma glucose also contributes.

Diabetes Mellitus:

- Causes of diseases: (1) Type 1 diabetes: insulin deficiency from destruction of pancretic cells, this leads to high plasma glucose concentration and high ketone concentration (from liver); (2) type 2 diabetes: insulin resistence, particularly in adipose tissue.
- Consequences of T1D and T2D:
  - High plasma glucose: leads to osmotic diuresis - high osmotic pressue in the small tubes of the kidney, and increase of water and sodium excretion, and this further leads to reduction in blood volume and pressure, and possibly reduced blood flow to brain.
  - High ketones: leads to high proton concentration (some ketones are acids), and can cause brain dysfunction.
  - Chronic abnormalities from high glucose: AGEs, toxic metabolites from glucose.

Regulation of plasma cholesterol:

- Cholesterol balance: (1) Food: ingestion from food, some is excreted in feces; (2) most cells uptake cholesterol for membrane synthesis and synthesis of steroid hormones; (3) liver: production of cholesterol and secretion to blood or in bile.
- Cholesterol delivery: (1) LDL: delievery cholesterol to most cells; (2) HDL: recycle excessive cholesterol to liver and some endocrine cells (which have specific HDL receptors).
- The LDL/HDL ratio correlates with the incidence of coronary heart disease.

Regulation of metabolic rate and energy balance:

- Metabolic rate: the amount of total energy expenditure per unit time. Often measured by basal metabolic rate (BMR). The single most determinant of BMR is thyroid hormone. Other determinants include epinephrine, food-induced thermogenesis and muscle activity.
- Regulation of energy stores: by both food intake and energy expenditure. Long term regulation of food intake is regulated by leptin (secreted by adipose cells, inhibiting the release of neuropeptide Y in the hypothylamus). Short-term regulation, i.e. satiety signals may include insulin, metabolic rate increase, etc.
- Obesity: has a genetic basis. Difficult to treat by diet, as the body tries to maintain the set point of the fat store.

4. Lipoprotein metabolism and regulation [ELS, Lipoprotein Metabolism; Lipoprotein: Genetic Disorders]

   Overview of lipoprotein metabolism:

   - Function: transport system of fat (TG) and cholesterol, from the site of production (food and liver) to the site of consumption (adipose tissue, and other cells).
   - Structure of lipoproteins: phospholipid monolayer, TG and cholesterol ester (CE) in the core, and apolipoproteins and unesterized cholesterol in the surface. Different lipoproteins are different in the amount of TG and CE content: chylomicron, VLDL, LDL and HDL.
   - Functions of apolipoproteins: (1) structural protein: e.g. ApoB; (2) ligand of cell-surface receptors: e.g. ApoE, recognized by liver; (3) co-factor of enzymes: e.g. ApoC is a cofactor of lipoprotein lipase (LPL).

- Two separate systems for transport: exogenous (from intestine) and endogenous (from liver) pathways. The first pathway transports only TG (even though it contains cholesterol) and the second pathway needs to delieve both TG and cholesterol. This may be because the amount and composition of cholesterols from food may not match the body need, thus better recycle them first to liver.

Lipoproteins: [Wiki]

- A lipoprotein is a biochemical assembly that contains both proteins and lipids. Example: chylomicron consists of phospholipid memberane and proteins (ApoA/B/C/E) in the outside, and lipids (choloesterol and triacylglyceride) inside.
- Function: The function of lipoprotein particles is to transport water-insoluble lipids (fats) and cholesterol around the body in the blood. All cells use and rely on fats and, for all animal cells, cholesterol as building blocks to create the multiple membranes.
- Classes of lipoproteins:
  - Chylomicrons carry triglycerides (fat) from the intestines to the liver, skeletal muscle, and to adipose tissue.
  - Very low density lipoproteins (VLDL) carry (newly synthesised) triacylglycerol from the liver to adipose tissue.
  - Low density lipoproteins (LDL) carry cholesterol from the liver to cells of the body. LDLs are sometimes referred to as the "bad cholesterol" lipoprotein.
  - High density lipoproteins (HDL) collect cholesterol from the body's tissues, and bring it back to the liver. HDLs are sometimes referred to as the "good cholesterol" lipoprotein.

Fat and cholesterol delievery pathways: [lipoprotein-APOB-particles.GIF]

- Exogeneous pathway: through chylomicron, whose main structural protein is APOB48. (1) Formation of chylomicron in the intestine; (2) maturation of chylomicron in the blood by getting APOCII and APOE from HDL; (3) release TG in the muscle or adipose cells: APOCII-dependent lipolysis (using LPL in the endothelical cells); (4) chylomicron remnants, or CMR, (including cholesterol content) are absorbed in the liver endocytosis that is dependent on CMR and CMR-receptor interaction (also APOE dependent).
- Endogenous pathway: formation in the liver, containing TG, cholesterol and APOB100 (the same gene as APOB48, but APOB48 from RNA editing). This pathway contains additional step of cholesterol transport.
  - Maturation and TG transport: VLDL taking APOCII and APOEE from HDL, and release TG through lipolysis.
  - Processing in liver: VLDL remnant (or IDL) is processed by Hepatic lipase (HL), and become LDL particles (half TG, half CE).
  - Delievery of cholesterol: in peripheral cells, interactions between LDL and LDL-R and APOE, APOB100, degrade LDL through endocytosis (releasing cholesterol).
- Why separate TG and cholesterol steps in the endogeneous pathway, e.g. why need liver processing to form LDL? This may be related to different transport requirements for TG and cholesterol: TG mainly to skeletal and adipose tissues, while cholesterol mainly for different cells (e.g. endocrine glands).

Reverse cholesterol transport: [lipoprotein-HDL.GIF]

- Function of this system: is probably related to recycling excess cholesterol to improve efficiency (or excess cholesterol may be harmful to the body).
- Pathway: nascent HDL is secreted by liver and intestine, containing APOAI and phospholipid.

  - Collecting cholesterol: from redundant cholesterol in the surface of APOB containing particles (VLDL and chylomicrons), and in the peripheral tissues particularly macrophages and enterocytes, by interacting with a specific cell-surface protein ABCA1. The collected cholesterol is then esterized by LCAT to form CE and enter the core.
  - Exchange with VLDL and LDL: Cholesteryl esters in HDL can be exchanged with triacylglycerols in VLDL and LDL, a process catalysed by CETP.
  - Recycling: SR-B1 in liver catalyses the removal and cellular uptake of cholesteryl esters from the lipoprotein core without uptake and degradation of the HDL particle itself.
- HDL functions: both cholesterol recycling, and also provide APOCII and APOE for chylomicrons and VLDL particles.

Relation to cardivascular diseases (CVD): [ELS - Atherosclerosis: Pathogenesis, Clinical Features and Treatment; Plasma lipoproteins: genetic influences and clinical implications, Hegele, NRG, 2009]

- The most common disease is corotary heart disease, often equated with corotary artery disease, or arteriosclerotic heart disease.
- Only a small amount of circulating cholesterol originates from the diet; 80% is derived from endogenous synthesis, for which HMG-coenzyme A reductase (HMGCR) catalyses the rate-limiting step.
- Relation to atherosclerosis: (1) LDL accumulation in the blood wall and oxidation, triggering the endothelial cells to release cytokines, adhesion molecules, etc. that attract macrophages; (2) modified or oxidized LDL particles are taken up by macrophages (through scavanger receptors), forming foam cells; (3) over time, cholesterol builds up to toxic levels within the macrophage (which cannot remove cholesterol), and trigger macrophage death; (4) this results in the accumulation of large amounts of free cholesterol and cholesteryl esters within the arterial wall; (5) eventually, the cholesterol plaque causes the muscle cells to enlarge and form a hard cover over the affected area.
- Association of lipoproteins and CVD: CVD risk is associated with high LDL and low HDL. The ratio of apolipoprotein B to apolipoprotein A-I has been advocated as the strongest predictor of CHD risk, although the ratio of total cholesterol to HDL cholesterol seems to be equally predictive.
- High LDL cholesterol (LDL-C) and CVD: several lines of evidence of the causal role of LDL-C. (1) The uniform presence of CVD in genetic disorders that cause high LDL levels; (2) the most widely marketed drugs for lowering LDL-C, statins, have been demonstrated in numerous clinical trials to reduce risk of CAD [Teslovich, Nature, 2010].
- Low HDL cholesterol (HDL-C) and CVD: the causal role of HDL-C is uncertain. (1) Not all genetic disorders causing very low HDL levels are associated with CVD; (2) a drug that raised HDL-C via cholesteryl ester transfer protein (CETP) inhibition failed to reduce the risk of CVD [Teslovich, Nature, 2010].
- Plasma triacylglycerol (TG) and CVD: causal role uncertain. In particular, the relationship has been confounded by the association of elevated TG with depressed HDL cholesterol. Nonetheless, elevated TG and familial hypertriglyceridaemias are independently associated with CHD risk.

Regulation of cholesterol metabolism: [Wiki, cholesterol]

- SREBP (sterol response element binding protein) pathway: sensing of intracellular cholesterol level in ER by SREBP. When cholesterol level is low, SREBP is cleaved and migrated to nucleus (acting as a TF) and activate expression of genes such as LDLR and HMGCR.
- Energy control: HMGCR activity is inhibited by AMP kinase, thus when ATP level is low, AMPK is activated and reduce cholesterol biosynthesis.

5.2 The Immune System

Reference: [MBOC, 5ed, Chapter 24, 25]

Overview of the innate immune system:

- Distinguish from the adaptive immune system mainly be the recognition of the pathogens: rely on general, conserved features instead of specific antigenes.
- Importance: the first response to infections, as adaptive immune responses may take days to develop.
- Recognition: of pathogen-associated molecular patterns (PAMP), including cell wall, bacterial flagella, surface molecules (such as LPS), bacterial or viral DNA (CpG motif), double-strand RNA (ds-RNA), bacterial translation. These are often recognized by pattern recognition receptors.
- Response: phagocytosis, membrane disruption (lysis), inflammatation.
- Subsystems: (depending on the carriers) physical and chemical barriers in the surface, specialized proteins, specialized cells, general intracellular systems.
- Q: inflammatory reactions are common in many diseases, esp. chronic inflammation. What triggers them? Hypothesis: innate immunity plays a role in tissue homeostasis, so the tissue debris may trigger inflammation.

Surface barriers:

- Locations: skin and other epithelial surfaces, including those lining the respiratory, intestinal, and urinary tracts. The mucus layer in these surfaces.
- Physical features: tight junction, cilia (help clearance of pathogenes), etc.
- Chemical features: antimicrobial peptides, called defensins. Generally short (12-50 AAs), positively charged, have hydrophobic or amphipathic domains. They may act by using their hydrophobic or amphipathic domains to insert into the surface membrane of their victims, thereby disrupting the integrity of the membrane.

Complement system:

- 20 interacting soluble proteins that are made mainly by the liver and circulate in the blood and extracellular fluid.
- Early complement components: (1) classical pathway: activated by antibodies (conserved chain); (2) lectin pathway: activated by mannan (in fungal cell wall); (3) alternative pathway: molecules in the surface of pathogens.
- Activation: proteolytic cascade, initiated by cleavage of C3.
- Responses: tagging pathogens for phagocytosis, pore formation and lysis, stimulation/attraction of other immune cells.

Specialized cells:

- Macrophages and neutrophils: macrophages are present in both blood, normal tissues (e.g. liver) and places where infections are likely to happen, e.g. GI tract. Neutorphils are mostly in blood, and usually do not survive after fighting pathogens.
- Toll-like receptors (TLRs) and NODs: TLRs are activated by different ligands, e.g. TLR4 by LPS, TLR9 by CpG DNA, and induce transcription of pro-inflammatory genes and interferon-inducible genes. NODs are intracellular proteins that recognize pathogen ligands.

- Phagocytosis: fusion of phagosomes with lysosomes. The lysosomes contain digestive enzymes, defensins, NADH oxidase complex (generate reactive oxygen species).
- Inflammatory response: pain, redness, heat, and swelling at the site of infection. NK-kappa B transcriptional response and activation of TLRs, which induce cytokines, lipid signaling molecules such as prostaglandins. The effects of these molecules: attraction of macrophages and neutrophils, fever, increase of blood permeability, cutting off blood supply of the infection sites, increased activity of antigen presentation (through increasing DC activities and MHC expression, etc.). Pro-inflammatory cytokines: TNF-alpha, IFN-gamma, chemokines and ILs.
- Natural killer (NK) cells: kill cells expressing low or no MHC class I proteins, often virus-affected cells, through induction of apoptosis.
- Dendritic cells: present antigens to T cells, using MHC proteins.

Intracellular systems that prevent viral replication:

- RNAi: ds RNA degraded by the enzymer Dicer, and the ds RNA fragments bind to ss RNA, leading to the destruction of ssRNAs.
- Interferon response by virus-infected cells: triggered by viral components, including dsRNA. Interferon-α and interferon-β (IFN-α and IFN-β) are expressed upon the trigger, and activate expression of hundreds of genes through the JNK-STAT signaling pathway.
- Remark: virus evolved mechanisms to escape from immune system, some of which may be similarly deployed by cancer cells. Ex. virus might suppress the expression of apoptotic genes s.t. they are resistant to killing by CD8+ T cells or NK cells. However, virus have few proteins, how would they manipulate host transcriptome?

Adaptive immune system: overview

- Why need adaptive immunity? Specificity and memory are two main benefits.
- Main problems of adaptive immune system: activation of B and T cells, tolerance of self-antigens or harmless foreign antigens, regulation (turning off) immune responses.
- Lymphocytes: (1) B cells: antibodies inactivation virus and toxin, and mark pathogens for destruction by phagocytotic cells and the complement system; (2) T cells: helper T cells create co-stimulatory signals; cytotoxic T cells, regulatory T cells.
- Lymphocyte development: (1) Primary lymphoid organs: B cells in bone marrow and T cells in thymus. Both start out at bone marrow from HSC. (2) Secondary lymphoid organs: where B and T cells mature and activate, including lymph nodes, spleen.
- Interaction with innate immunity: adaptive immune response generally depends on innate immune response. The dendritic cells in the site of infection migrate to a nearby lymph node, active the T cells, which then either activates B cells, or migrate to the site of infection.
- Migration of lymphocytes: normally circulate in the blood and lymph nodes. They enter an infection site via chemokines: which allow them to pass through blood endothelial cells. When encountering the antigens, express cell adhesion molecules and stay at the infection sites.

Clonal selection theory:

- Immunological memory: when a lymphocyte encounters an antigen in peripheral lymphoid organ (with co-stimulatory signal), it will undergo clonal expansion as well as differentiation into effector and memory cells.

- Specificity: through clonal expansion when specific antigens bind to the receptors, also need (1) co-simulatory proteins (MHC), (2) certain cytokines.
- Memory: after being activated, naive lymphocytes become effector cells (both proliferation and differentiation), or memory cells. Memory cells have much higher antigen affinity, and can mediate a much stronger immune response when exposed to antigens later.

Immunological tolerance: overview

- Natural and acquired immunological tolerance (e.g. through exposure to antigens in newborns): note that self-tolerance persists only for as long as the molecule remains present in the body.
- Immunological self-tolerance: (1) Central tolerance: when lymhocytes encounter self-antigens, they will undergo receptor editing or clonal deletion (apoptosis) (2) Peripheral tolerance: when encounter self-antigens, clonal inactivation, deletion or suppression (regulatory T cells). Foreign antigens lead to different responses because they come with co-stimulatory signals, provided by T helper cells for B cell activation and by DC for T cell activation.
- Auto-immune diseases: It is thought that activation of the innate immune system by infection or tissue injury may help trigger anti-self responses in individuals with defects in their self-tolerance mechanisms.

B cells and antibody molecules:

- B cell activation: each B cell has BCR in cell member with unique binding affinity. Activation requires: antigen recognition by BCR, and T helper cells. Once activated, B cells become antibody-secreting plasma B cells and memory B cells (the same affinities as the original BCRs).
- Antibody structure: 2 identical antigen binding sites (allow cross-linking), 2 heavy and 2 light chains, hinge region and tails (carry function). Both light and heavy chains have: N-terminal variable region (with three hypervariable regions) and a constant region.
- Five classes of antibodies: IgA, IgD, IgE, IgG, IgM, according to the heavy chain. Some has a number of subclassess.
- IgM and IgD: immature naive B cell - IgM, mature naive B cell - IgM and IgD. IgM is always secreted in the early response, IgD mainly serves as surface antigen receptor.
- IgG: the main immunoglobin in the blood (released by B cells after antigen stimulation). Activate complement system, also the tail region (Fc region) binds to the Fc receptors in the phagocytotic cells. Provide protection in babies from mother.
- IgA: the main class of antibodies in secretion, including saliva, tears, milk, and respiratory and intestinal secretions. Transported through epithelial cells via transcytosis into secretion.
- IgE: bind to mast cell in tissues and basophils in blood through the Fc region. Antigen binding triggers mast cells to releast cytokines, in particular histamine. Histamine causes blood vessels to dilate and become leaky, which in turn attracts white blood cells, antibodies, and complement components.
- Primary antibody repertoire: IgM and IgD from naive B cells, low affinity. Second antibody erpertoire: IgG from B cells upon antigen stimulation (and helper T cells), high affinity.

Antibody diversity:

- V(D)J recombination: occurs during B cell maturation in the bone marrow to make the variable regions of light and heavy chains. In human, 40 V segments, 5 J segments, and 25 D segments. VJ recombination in light chain $\Rightarrow$ 200 or 120 for different light chains, and VDJ in heavy chains $\Rightarrow$ 6000. The total is $320 \cdot 6000$. This is called combinatorial diversity.

- Junctional diversity: random gain and loss of nucleotides in the joining sites.

B cell maturation:

- B cell maturation leads to plasma cells, which are very effective at secreting antibodies (5k molecules per second).
- Somatic hypermutation: B cells mutate at the rate of about one mutation per V-region coding sequence per cell generation, made possible by activation-induced deaminase (AID).
- Affinity maturation: a progressive increase in the affinity of the antibodies produced against the immunizing antigen. As a result of repeated cycles of somatic hypermutation, followed by antigen-driven proliferation of selected clones of effector and memory B cells, antibodies of increasingly higher affinity become abundant during an immune response.

Overview of T-cell response:

- Why need T cell response? For virus or parasite-infected cells, antigen may not be visible to antibodies or innate immune system, need a strategy that greatly expands the space of antigens that are recognizable.
- The main features of T-cell response: (1) recognition of the peptide (protein fragments) bound to MHC proteins; (2) the expanded capability of antigen recognition is coupled with a system to kill infected host cells (cytotoxic T cells), or other response systems - innate immunity and B cells (increase the activity of these response systems).
- T cell subtypes: (1) Killing of infected host cells: cytotoxic T (Tc) cells; (2) Warning/activation of other response systems (phagocytotic, B cells, cytotoxic T cells): helper T (Th) cells; (3) Stop the response: regulatory T cells.

Antigen recognition of T cells:

- Basic components:
  - T-cell receptors (TCRs): TCRs are similar to antibodies, except that they are membrane-bound and not secreted. TCR consists of alpha and beta chains (similar to light and heavy chains).
  - MHC: the carriers of the antigens, synthesized in the ER.
- Recognition by Tc cells: [MBOC, Figure 25-59]
  - Signals: typically viruses. The virus infection of host cells may not leave any signals in the cell surface, and may not be recognizable by professional APCs. Distinguished from Th cell signals by: the antigens are cytosolic.
  - Antigen presentation: the viral proteins synthesized in the cytosol of host cells (any cell type) are degraded in proteosomes (ubiquintin-dependent proteolysis, degradation is common for all cytosolic proteins). The resulting peptides are transported into ER lumen by ABC transporters, where they are loaded on class I MHC molecules.
- Recognition by Th cells: [MBOC, Figure 25-61]
  - Signals: typically bacterial and parasites, that can be engulfed by APCs through endocytosis or phagocytosis. Distingusihed from Tc signals by: the antigens are externally-derived (endosomes are topologically equivalent to external enrivonment).
  - Specialized antigen presenting cells (APCs): most importantly dendritic cells (DCs), but also include macrophages, B cells, etc. DCs are located in tissues throughout the body, including the central and peripheral lymphoid organs. The encounter with pathogen (through pattern recognition receptor) activates dendritic cells, which can activate T cells. Remark: APCs play a role of signal amplifier.

  - Antigen presentation: the endocytosized proteins in APCs are degraded in endosomes and loaded on class II MHC molecules. Note: MHC II molecules are synthesized in ER, where they are loaded by invariant chains. The binding of invariant chains direct MHC II molecules to late endosomes, where the invairants chains are cleaved and the binding grooves of MHC II proteins are exposed.

- Cross-presentation of antigens: DCs need to activate Tc cells even when DCs are not infected themselves. DCs phagocytose fragments of virus-infected cells. They then actively transport viral proteins out of the phagosome into the cytosol, where they are degraded in proteasomes; resulting fragments are then transported into the ER lumen, where they load onto assembling class I MHC proteins.

- Positive feedback in antigen presentation in viral-infected cells: Tc cells or Th cells secrete IFN-γ, that acts on the target cells to induce more MHC I proteins, the proteosome units and ER peptide transporters (all related to presenting viral antigens).

MHC proteins: bind a peptide (antigen) and interacts with a TCR.

- Functional requirements of MHC: (1) antigen-TCR interaction needs a carrier protein: e.g. cytosolic peptides need to be moved to the cell surface to be detected by T cells. (2) The carrier proteins should be able to load a huge number of potential antigens, thus within a body, there should be a number of different carrier proteins.

- Function of MHC proteins: needed to distinguish Th and Tc cells: otherwise, APC may activate Tc cells, and get killed. CD4 and CD8 are co-receptors that bind to invariant part of MHC proteins: CD4 in the membrane of Th and regulatory T cells, binding to class II MHC; CD8 in the membrane of Tc cells, binding to class I MHC.

- MHC subclasses: (1) MHC I: expressed by all cells. HLA-A, HLA-B, HLA-C loci. (2) MHC II: expressed only by APCs, including DCs, B cells and macrophages. HLA-DP, HLA-DQ and HLA-DR loci.

- MHC structure: MHC is encoded by the HLA loci in human. (1) class I MHC protein: the transmembrane α chain and an extracellular β_2 microglobin (encoded by a different locus); the α chain contain an immunoglobin(Ig)-like domain. (2) class II MHC protein: heterodimer of two transmembrane chains α and β, both encoded in HLA and contain Ig-like domains.

- MHC diversity within a person: needed because of the diversity of antigens. In the HLA loci, there are 3 genes encoding MHC I: HLA-A, HLA-B, HLA-C, and three clusters encoding MHC II: HLA-DP, HLA-DQ and HLA-DR (each has at least one gene for α and one for β chain). Thus one individual contains 6 types of MHC I proteins, and more than 6 of MHC II proteins.

- MHC polymorphism: some MHC genes have very large number of alleles (some has 400). It is very rare that two individuals have the same set of MHC proteins.

- Causes of MHC polymrophism: an evolutionary arms race between pathogens and hosts - pathogens will try to escape the association with MHC. Thus a new MHC protein will allow presentation of more antigens, potentially more pathogens. On the other hand, more MHC proteins mean a larger fraction of T cells are eliminated during development.

- HLA locus: human HLA locus (about 24M to 35M in chr. 6 of human genome release 37) contains all MHC genes, as well as a number of other genes, many involved in immune function.

T cell activation:

- Behavior of T cells: activation by antigens from pathogens (foreign antigens), and tolerance by self-antigens. The distinction is generally made by: antigens in the absence of infection are self-antigens (peripheral tolerance).

- DC activation: DCs are located in tissues throughout the body, including the central and peripheral lymphoid organs. They are activated by encounter with pathogens, tissue injury or by effector helper T cells.
- Activated DCs can activate a T cell to become an effector cell or a memory cell. Specifically, the two types of cells form three types of interactions: (1) TCR and antigen, presented by MHC. (2) Co-receptors bind to invariant part of MHC: CD4 for class II and CD for class I respectively. (3) Co-stimulatry proteins (on DCs) and receptors (on T cells): among them are CD28 on T cells, bound with B7 on DC.
- T cell tolerance: non-activated DCs help induce self-reactive T cells to become tolerant, both in the thymus and in other organs. Such cells present self antigens in the absence of the co-stimulatory molecules. They induce tolerance in at least two ways: (1) stimulate abortive responses in the T cell that lead to either inactivation or apoptosis; (2) activate regulatory T cells.
- TCR signaling:
  - If activated, TCR brings the Src-like cytoplasmic tyrosine kinase Lck into the signaling complex and activates it. Lck initiates a tyrosine phosphorylation cascade by phosphorylating tyrosines on CD3, which then serve as docking sites for yet another cytoplasmic tyrosine kinase called ZAP70. ZAP70 then helps activate the inositol phospholipid and MAP kinase signaling pathways.
  - Downstream signaling: induce T cells to secrete interleukin-2 (IL2) and simultaneously to synthesize IL2R. The binding of IL2 to the IL2 receptors activates intracellular signaling pathways that turn on genes that help the T cells to proliferate and differentiate into effector cells.
  - Positive feedback: Once activated, a Th cell itself expresses a co-stimulatory protein called CD40 ligand, which acts back on CD40 receptors on the DC surface to increase and sustain the activation of the DC, creating a positive feedback loop.
  - Negative feedback: during activation, T cells express CTLA4, which binds B7 with much higher affinity than does CD28 and, in doing so, it blocks the activating activity of CD28.
- T cell development: in thymus. Positive selection: only TCRs recognizing self MHC can survive. Negative selection: TCRs with high affinity to self-MHC bound by self-antigens die.
- T cell death: Most of the T (and B) effector cells produced during an immune response must be eliminated after they have done their job. This is due to both the survival signal provided by antigen stimulation, and also the death signals. In the case of effector cytotoxic T cells, for example, IFNγ plays an important part in inducing the cell death; as effector cytotoxic T cells make IFNγ, this is another form of negative feedback.

B cell activation:

- Two signals are required: antigen and the signal from Th cells. If a B cell receives the first signal only, it may be eliminated or functionally inactivated, which is one way in which B cells become tolerant to self antigens.
- BCR activation: similar to TCR activation. The antigens cross-link adjacent BCRs, then activates Src-like cytosolic tyrosine kinase, which can be Fyn, Blk, or Lyn. Another Src-like tyrosine kinase called Syk (homologous to ZAP70) is recruited and activated. In addition, the complement-binding co-receptor complexes increases the strength of signaling by activating PI 3-kinase.
- Negative feedback of BCR signaling: later in immune response, Fc receptors in B cells bind the tails of the IgG antibodies, and decrease the strength of signaling.
- Th cell signal:

  - B cells can serve as antigen-presenting cells: through endocytosis, they degrade their bound protein antigen, and the peptide fragments are returned to the B cell surface bound to class II MHC proteins, which are recognized by antigen-specific Th cells.
  - Linkage recognition of antigens: the effector helper T cell activates only those B cells with BCRs that specifically recognize the antigen that initially activated the T cell, even though the TCRs and BCRs usually recognize distinct antigenic determinants on the antigen. This is important for the self-tolerance of B cells.
  - In the Th-B cell contact, the interaction between CD40 ligand in Th cell surface and Cd40 in B cell surface is required for helper T cells to activate B cells to proliferate and differentiate into memory or effector cells.
  - Helper T cells also secrete cytokines to help B cells proliferate and differentiate and, in some cases, to switch the class of antibody they produce. The cytokines include the interleukins IL2 and IL4.

- Both B and T cells require multiple signals for activation: BCR and TCRs bind with additional proteins (CD3 complex for TCR), which relay signal. Require co-stimulatory signals: (1) Activated DCs: B7 to CD28, to activate Th cells. (2) Activated Th cells: CD40 ligand to CD40 receptor, to activate B cells.

T cell responses:

- Tc cells: they form aggregates with the target cells (immunological synpase) through cell adhesion molecules, co-receptors, etc, and kill target cells by inducing caspase cascade (apoptosis), similar to NK cells, through either (1) perforin-dependent killing: activation of a pro-apoptotic Bcl2 protein called Bid (an enzyme enters the target cell through the pore in the membrane fromed by perforin) (2) Fas ligand (target cell) and Fas receptor (Tc cells).

- Th cells: crucial for defense against both extracellular and intracellular pathogens. Proliferation by releasing IL2 in an autocrine fashion. Their effects: help activate B cells, macrophage, cytotoxic T cells, and even dendritic cells by releasing various cytokines. Naive helper T cells differentiate into 4 subtypes of Th cells and T-reg cells upon activation: activated DCs send different cytokines and this determines the fate of naive Th cells:
  - Th1 cells, from IL12, response mainly for intracellular pathogens, through IFN-γ and TNF-α, activation of macrophages and Tc cells.
  - Th2 cells, from IL4: mainly for extracellular pathogens, through IL4 and IL10, activation of B cells.
  - Whether Th1 or Th2 responses are produced depends on the signals in the DC cells. In addition, Th cells have a mechanism to reinforce the initial choice: IFN-γ and TNF-α will inhibit Th2 development, and similarly, IL4 and IL10 inhibit Th1 development.
  - Follicle T helper cells, from IL6 + IL21: secrete IL4 and IL21, important for B cells.
  - Th17 cells, from TGF-β and IL6: secrete IL17, attracting neutrophils and help with wound healing.

- Regulatory T cells: play a crucial part in immunological self tolerance by suppressing the activity of self-reactive effector Th and Tc cells, and help prevent excessive T cell responses to microbial antigens in chronic infections. Foxp3 is a master regulator of regulatory T cell developmet (deletion of Foxp3 causes fatal autoimmune disease). The action of regulatory T cells is thought to be mediated by inhibitory cytokines, including TGF-β and IL10.

- INFγ: effector Tc cells or Th1 cells secrete IFNγ, which greatly enhances anti-viral responses. The IFNg acts on virus-infected host cells in two ways:
  - Blocks viral replication.

- Increases the expression of many genes within the MHC chromosomal region, including class I MHC proteins, the two specialized proteasome subunits, and the two subunits of the peptide transporter located in the ER.

5.2.1 Self-Tolerance of Immune System

Reference: Cellular and genetic mechanisms of self tolerance and autoimmunity [Goodnow, Nature, 2005]
Overview of self-tolerance:

- Central tolerance of T cells: select for T cells capable of binding with MHC (self-MHC), but not self-peptide. This is achieved through:
  - Positive selection: T cells with TCR matching self-MHC survive, and the rest die.
  - Negative selection: T cells with TCR binding strongly with self peptide-self MHC complex die.

  Note that the space of self-antigens in thymus is expanded by self shadow: projection of organ specific genes in thymus, dependent on the TF, autoimmune regulator (AIRE).
- Peripheral tolerance of T cells: important for organ-specific proteins that may not be present in thymus:
  - Deletion of inactivation of the cells recognize self peptides bound to MHC proteins on the surface of dendritic cells that have not been activated by pathogens and therefore do not provide appropriate activating signals.
  - Regulatory T cells in the periphery that suppress the activity of some self-reactive effector T cells.
- Self-tolerance of B cells: in bone marrow, the self-reactive B cells with undergo receptor editing or apoptosis; in peripheral lympoid organs, self-tolerance is induced by Th cells.

Four mechanisms of dealing with self-reactive cells:

- Clonal deletion: the cell displaying the "forbidden" or self-reactive, receptor can be triggered to die. Apoptosis induced by inhibiting BCL2 survival pathway (for example, BIM induction) or by activating death receptors, e.g. FAS.
- Receptor editing: a cell bearing a forbidden receptor can "edit" the offending receptor by further V(D)J recombination or somatic hypermutation
- Clonal anergy or tuning: intrinsic biochemical and gene-expression changes can reduce the ability of the cell to be triggered by self-reactive receptors. Including: BCR/TCR downregulation, induction of inhibitory receptors (CD5, CTLA4), phosphatases (SHP1, SHIP), ubiquitin ligases (CBL, GRAIL, ITCH, ROQUIN).
- Extrinsic controls can limit the danger of self-reactive receptors. These extrinsic controls limit the supply of essential growth factors, costimuli, pro-inflammatory mediators and other factors, and also include active suppression by regulatory T cells.

B cell self-tolerance in bone marrow:

- Self-reactive B cells will internalize the offending BCRs and stop the maturation program: expression of RAG1/2 for receptor editing, receptors for B-cell-activating factor (BAFF) are poorly induced, homing receptors are not expressed. If a B cell with a forbidden receptor fails to edit to a less self-reactive receptor, cell death occurs within 1-2 days, through increasing the levels of BIM (BCL-2-interacting mediator of cell death).
- Signaling pathways:
  - Non self-reactive B cells: BCR $\rightarrow$ NFKB1 $\rightarrow$ BCL2; BAFF $\rightarrow$ NF-KB1 $\rightarrow$ BCL2

  - Weakly self-reactive B cells: BCR → NFKB1 → BCL2; BCR → BIM, death dominates unless increased BAFF is supplied.
  - Strongly self-reactive B cells: BCR internalization and maturation arrest cripples the BCR and BAFFR survival pathways, while BIM induction promotes death.

T cell tolerance in thymus:

- Both receptor editing and clonal deletion happen in TCR tolerance, though deletion appears to be the predominant process.
- TCR may induce NFKB-BCL2 (survival) pathway, BIM pathway (apoptosis) and FAS pathway (apoptosis). In T cells with with strong self reactivity of TCRs, TCR-induced BIM and FAS death pathways predominate.
- Signaling in negative selection: poorly understood. Requires ZAP70, GRB2, activation of ERK, p38 and JNK. At the distal end of the pathway, requires induction of BIM and Nur77 family of orphan nuclear receptors. Strong TCR engagement also induces expression of an extracellular protein, FAS ligand (FASL, also known as CD95L), and triggers T-cell death independently from the BIM/BCL-2 mechanism through FASL-FAS interaction and the caspase-8 proteolytic cascade.
- The combination of strong stimulatory signals through TCR and CD28 is, paradoxically, a potent trigger of nuclear factor-KB activation, the pro-survival pathway that induces expression of BCL-2 proteins in mature peripheral T cells.
- The well established association between particular MHC molecules and susceptibility to specific autoimmune diseases may stem from inefficient presentation of particular self peptides during this phase of TCR deletion.

Clonal anergy:

- B cells: decreased display of self-reactive BCRs on the cell surface owing to accelerated endocytosis and blocked transport of new BCRs out of the endoplasmic reticulum.
- The changes in gene expression that occur selectively in cells bearing self-reactive BCRs. An example is expression of the CD5 cell-surface protein induced selectively by self-reactive BCRs, which provides an additional inhibitory receptor to recruit SHP1 and inhibit BCR signalling and activation.
- T cells: expression of the inhibitory receptor CD5 is induced to 10-50 fold higher levels in self-reactive T cells than in B1 cells or anergic B cells. Expression of another inhibitory receptor, cytotoxic T-lymphocyte antigen 4 (CTLA4), is induced at a high threshold of TCR self reactivity and inhibits T-cell activation by competing with CD28 for ligation with B7.
- Increased expression of the ubiquitin ligases CBL-B, GRAIL and ITCH can also accompany chronic TCR signalling in vitro. These proteins interfere with TCR, CD28 and cytokine receptor signalling by tagging the TCR-CD28 or cytokine receptor signalling molecules with ubiquitin

Extrinsic regulation:

- Survival signal of B cells:
  - BAFF: the survival of peripheral B cells depends on BAFF, which is produced in limiting quantities primarily by radioresistant lymphoid stromal cells. BAFF through receptor binding, triggers an increase in the activity of NFκB2, which maintains survival through induction of BCL-2. BAFF also induces expression of the serine-threonine kinase, PIM2, which has potent pro-survival effects by inhibiting the pro-apoptotic BAD.

  - Constant engagement of self-reactive BCRs, below the threshold required to trigger maturation-arrest in the bone marrow, is still sufficient to increase BIM expression and consequently elevate the survival requirement for BAFF. With large numbers of circulating B cells, the self-reactive cells fail to receive enough BAFF and are competitively deleted.

- Survival signal of T cells: In common with B cells, the survival of mature T cells depends upon continuous signalling in the peripheral lymphoid tissues. This requires TCR signalling through contact with ubiquitous MHC ligands as well as exposure to interleukin-7 (IL-7). In the case of T lymphocytopenia, however, IL-7 levels rise and amplify TCR signalling, causing naive T cells to proliferate. Such homeostatic proliferation may activate T cells reactive to tissue-specific antigens.

- Antibody responses of B cells: depend upon antigen binding to the BCR, and the signal from T helper cells: CD40 ligand (CD40L) and secreted cytokines IL-2, IL-4, IL-5 and IL-21. These signals from T helper cells responding to foreign antigen can be misdirected to self-reactive BCRs that cross-react with a component of a microorganism. The signal may also be misdirected to bystander B cells.

- Self-reactive BCRs are also generated in a second wave of receptor diversification through somatic hypermutation. Remarkably little is known about the mechanism dealing with this problem.

- There is clearly considerable scope for tolerating self-reactive receptors even at the final effector phase. In RA, even when sufficient sautoantibody is present in the circulation, its capacity to localize in joints to produce joint pathology depends on inflammatory cascades involving Fc receptors, mast cells, neutrophils and complement.

Treg cells - the next frontier of cell therapy [Science, 2018]

- Interaction of T-reg cells with inflammation and immune response: T-reg can suppress inflammation, and autoimmune reactions. T-reg cells are defined by the lineage-specific factor FOXP3.

- Manipulating polyclonal T-reg cells: e.g. T-reg expansion by IL2, T-reg stabilization by rapamycin.

- Manipulating antigen-specific T-reg cells: polyclonal T-reg cell therapy can lead to suppression of anti-tumor immunity. To develop antigen-specific T-reg cell therapy, can use APC to present specific antigens.

- Remark: interaction of innate and adaptive immune systems is two way: innate activates adaptive immunity, also is regulated by T-reg.

5.2.2 Immunogenomics

B cell genomics [Aly Khan, 2018]

- Part I. Background: AIDs more common in women than men. MS: T-cell therapy not work. B cell therapy (anti-CD20), stopping maturation of B-cells (but not memory/plasma cells)

- B cell function: cytokine production (IL6, IL10, TNF-alpha), antigen trafficking.

- BCR: VDJ recombination. Constant regions: different classes of antibodies. Ex. IgA dimer. IgG can cross placenta. Each cell expresses only one class of Ig molecules, but class switching can happen.

- Q: VDJ recombination: one per cell? Genome or epigenetic control? Probably genome editing. Q: What protein? AID? Can we modify the enzyme s.t. it can do more editing? What controls the specificity?

- BCR assembly: heavy chain, light chain (two mRNAs). Need to assembly H and L chains (easy for single cells). Hypermutation in variable regions makes the problem harder (CDR1-3 regions are much more mutable). Algorithm: start with anchors (not in CDR regions), choose a read, then find the next one with the highest overlap.

- Idea: use HMM to reconstruct the BCR.
- Ref: BASIC: BCR assembly from single cells. Error rate: 3%.
- Part II. Experiment: Treat patients with flu vaccine. Sort IgG and IgA B cells. BCR clones: defined by BCR (no hypermutation after some point).
- Transcriptome similar between cells of the same clone. Explanation: earlier events define transcriptome; or stimulation (by influenza vaccine) - the first may be more important.
- Q: driven by VDJ or hypermutations? Can we define continuous relationship (tree) of BCR?
- IgG and IgA: have highly similar transcriptional profiles.
- Influenze-specific antibodies have shifted glycan profiles (Ab. decoration by glycan).
- Part III. DNA MMR gene silencing: can predict tumor immunotherapy response. Could be epigenetic level! To measure MMR deficiency: microsatallite instability (change of number of repeat units). Exp: PCR on microsatellite.
- DNN: predict MSI. Training data, 500 patches per slide, 1000 slides. Q: What is useful information in the images? T cell infiltration!
- Learn invariant representation: adversial training, the objective is to: not predict well in cancer type objective (objective in MSS status)
- Guided backpropagation: find which pixels that activate a class with backprop.

5.3 Microbiome

Chapter 12: Human Microbiome Analysis [Morgan and Huttenhower, PLCB, 2012]

- Part I. Techniques of profiling microbiomes. Historical studies: (1) culture-dependent: morphology of colony, medium, etc. (2) Targeted sequencing. (3) FISH: low throughput.
- 16S rRNA: group 16S rRNA genes by sequence similarity, cutoff typically 95, 99%. Results are OTUs. The assignment of sequences to OTUs is called binning.
- Microbial diversity: the Couple Collectors problem (number of OTUs in a population, when we only have a finite sample). Alpha-diversity: number and evenness of species within a community, e.g. Shannons entropy. Beta diversity: difference across samples, e.g. sum of the number of unique species. The diversity can also be defined using phylogenetic breath.
- Shotgun metagenomic sequencing: options for assembly are (1) Direct analysis of reads (part of genes). (2) ORFome: assembly of open reading frames. (3) Maximum unambigous scaffolds: advantage of inferring operons. (4) Whole genome: quite unrealistic. In assembly and later analysis, one can rely on homologous sequences (e.g. using reference database to assign start and end of ORFs), or de novo or a hybrid.
- Part II. Characterization of microbiomes and association with human health. Functional metagenomics: infer functions of a community. Obtain profiles of genes or pathways (KEGG, GO, etc.). Then correlate these profiles with phenotypes of interest.
- Metabolic pathways: multi-organism FBA is difficult but promising. Applications: metabolic engineering, understand the ecological networks/community dynamics (mututalisms, parasitism, competition, etc.)

- Diversity of microbiota: skin < nasal < oral < gut. Gut especially dynamic: change with time, diet, development, etc.
- Part III. Approaches to study host-microbiota interactions: difficulty with model organisms - often micriobiome divergent. Use germ-free model organisms: replace with human microbiome and study its phenotype.
- Example of host-micriobe interactions: establishment of oral bioflim, each time after brushing (1) Streptococci is early colonizer, using its surface adhesions and receptors. (2) Recruit other species that provide nutritive and structural environment. (3) Further taxa. The process involves physical interactions, recognition of surface molecules, intercellular signaling and metabolic dependency. An open question is how such community evolves.
- The mechanisms of how microbiome may influence host health: suppose we have probiotics that benefit human health, questions are: (1) The impact on community: remove pathogenic organisms or overall shift of community? (2) Targets: microbiome or host epithelia cells, immune cells or distant cells? (3) Mediated via increase of beneficial compounds or reduction of detrimental ones?
- Remark: the problem has some similarity with genetics: given the association of microbiome species with trait, test if the effects are mediated by compounds (derived from bacterial genomic annotations).

Where next for microbiome research? [Waldor, PLoS Bio, 2014]

- Summary: understand the interactions among bacterial species and how they form community; how microbiome interacts with the host and affect the health of host. To achieve these goals, need development of technologies.
- Genome-centric view: need not only identification of microbioal species through 16s rRNA, but the whole genome sequences. Single-cell sequencing.
- Systems biology of human microbiome [Borenstein E]: from descriptive reserach to systems biology approaches: e.g. network based anaysis of interactions, predictive model of microbiomes function and dynamics, rational design. Require integration of multiple omics datasets.
- Microbiome evolution: answer the question of why microbiome is as it is. Adaptive evolution of microbiome to host environment; how microbiome affects host evolution (social behavior).
- Understand the role of metabolites derived from microbiome in human health [Brett Finlay]: application of metabolomics. Ex. short-chain fatty acid (SCFA) in T-reg cell maturation through interaction with GPCR.
- Techniques to manipulate specific microbial species.
- Human nutrition and gut microbiome [Jeffrey Gordon]: the effects of foods/nutrition on microbiome, and how microbiome can influence the effects of foods.
- Synthetic biology and manipulation of microbiome for human health: microbiome as diagostic. Fecal transplant. Probiotics. Engineered strains of microbial species.

Drugging the gut microbiome [NBT, 2015]

- Summary: metabolites from microbiome can affect the health of host, and this motivates several strategies for drug development:
  - Metabolite survey to identify active metabolites that can serve as drugs.
  - Identify host targets of these metabolites: new drug targets.

  - Target genes in microbial species to change the microbiome composition to serve therapeutic purpose.

- Cardiovascular disease: unbiased metabolomic survey of heart disease plasma found TAMO was strongly associated with heart disease risk. TAMO: biosynthesis from chemicals in red-meat. Drug development: inhibit the microbial enzymes that synthesize TAMO.

- Autism: metabolomic survey found 4EPS in ASD patients.

- Immune disease and inflammation: short chain fatty acid (SCFA) from dietary fiber, interacts with GPR43 and leads to Treg maturation/production. Lack of SCFA leads to inflammation and autoimmune diseases. Development strategies: delievery of SCFA-like compounds, of microbial species, and targeting GPR43.

- Refactoring synthetic biology: search for biosynthetic gene clustesr (BCGs), and characterizing their functions using E. coli. In 95% of times, such experiments fail.

Sequencing and beyond: integrating molecular 'omics' for microbial community profiling. [Franzosa and Huttenhower, NRM, 2015]

- Meta-proteomics: comparison of proteome data between conditions, e.g. in Crohns disease, proteins involved in SCFA production, and in epithelial integrity are disrupted. Also application in environmental samples to find genes important for adaptation.

- Metabolomics: e.g. SCFA product by bifidobacteria has anti-tumorgenic effects. TAMO has been linked to CVD.

- Integrating multi-omics data: (1) Normalization: comparison of meta-RNA, protein or metabolite data with functional potential, can provide insights on functions of community. (2)Strengtheninghypothesis: e.g. SCFA disruption in IBD confirmed by both proteome and metabolomic data. (3) Mechanistic gaps: e.g. bacterial resistance to antibiotic, metabolite data suggests that the drug is converted to an inactive form, and protein data show the upregulation of the cgr operion.

- Predictive model: constraint-based model such as FBA, enhanced with additional data (e.g. enzyme expression levels)

- To understand how microbiome may influence human health and disease: integration of host and microbiome multi-omics data.

Bridging the knowledge gap: from microbiome composition to function [MSB, 2015]

- Goal: characterizing functions of microbiome genes.

- Temporaral Functional Metagenomics sequencing (TFUMseq): library of DNA fragments containing the genome of a bacterial species. The plasmid library was then transformed into E. coli. Inoculate mice with the plasmid library and follow through time using sequencing: if a strain increases the fitness (colonization), its abundance will increase over time.

- **Lesson**: create library of sequences/constructs/strains, each containing some sequence to be studies, then use natural selection to choose the ones that carrying a desired/defined function.

Gut metabolites and bacterial community networks during a pilot intervention study with axseeds in healthy adult men [Lagkouvardos,Mol. Nutr. Food Res., 2015]

- Background: lignan from plant foods is converted to enterolignan (ED or EL) in the body, which is structurally similar to estrogen (phytoestrogens). Flaxseeds and lignans are beneficial for treating metabolic diseases, CVD and cancer.

- Background:enterolignan producing bacterial are not always the dominant one in gut, and can different between individuals. Thus individuals may respond differently to lignans.
- Experiment: flaxseed intervention in 10 patients, then measure 16S rRNA and metabolites (SCFA, ED, EL, lipids and metabolic profiling)
- Dominant fecal bacterial communities are not changed much by flaxseed: shown in the PCA plot.
- Specific bacterial groups correlated with specific metabolites. Ex. one OTU associated strongly with EL.
- Nontargeted fecal metabolome analysis: similar pattern with OTU profiles (relationship among individuals). Find patients with unusual metabolite changes.
- Remark: the common challenge is that many specific species are altered, how do we find the one most interesting? Association of metabolite (or other trait of interest) with bacterial groups.

Microbiome and Longevity: Gut Microbes Send Signals to Host Mitochondria [Cell, 2017]

- Approach: E. coli mutant library, screening for C elegans life span. Found 29 E coli mutants that extend host life span.
- Mechanisms: two genes target Colanic Acid (CA), a polysacchride. Deletions of two genes lead to higher levels of CA. Found that the effect depends on host mitochondrial ETC/homostasis. So the model is: CA enters mitchondrial and improves Mito. function, particularly increase mito. fission.
- Possible co-evolution: bacterial release CA during stress, and the host may respond to the signal by changing mito.
- Lesson: metabolites from microbiome can influence mito. functions of cells, which can potentially have broad physiological impacts.

Probing the metabolism of microorganisms [Louca, Science, 2017]

- Hypothesis:environmental conditions prescribe the overall biochemical fluxes catalyzed by gene groups, regardless of the precise species involved.
- One microbial system: very high variability of specific species. In contrast, striking similarity in terms of the abundance of genes involved in various pathways such as fermentation, oxygen respiration, and carbon fixation. Note: pathway abundance can be quantified by the fraction of pathway genes over all genes.
- Ocean microbiome: similar finding. Puzzling because ocean current can carry bacterial very far. Possible explanation: complex interactions between organisms can cause fluctuation of species composition.
- **Lesson**: in a microbiome system, focus on metabolic/biochemical functions they perform, instead of specific species.

Do hosts and their microbes evolve as a unit? [PNAS, 2019]

- Evidence of hologenome theory: (1) one species of coral grew resistant to the infection of a bacterial (bleaching), by picking up new microbial members. (2) Cold-adapted tilapia: changes in gut microbiomes as well.
- Role of microbiomes in reproductive isolation: gut microbiomes of Nasonia (wasp) species prevented interspecific breeding. Treatment with antibiotics, hybrids survive.
- Justification of theory: "It really doesnt matter as long as hologenome is reconstituted in every generation.

- Questioning the hologenome theory: coral-microbiome relationship, skeletal microbiome was largely governed by a hosts genes, but the microbes in a corals mucus layer were determined by its environment.
- Can we view host-microbes as a single unit of selection? Need to persist over generations.

5.4 Nervous System and Psychiatric Diseases

Neurobiological functions of transcriptional enhancers [Nord and West, NN, 2019]

- What have we learned from cis-regulatory genomics? (1) Determinants of TF binding: chromatin accessibility and DNAm are good predictors. (2) Higher order structure: super-enhancers may have many physical interactions and form phase separation. (3) It is often hard to predict enhancer effects on gene expression (e.g. in MPRA). P300 is the only good predictor known.
- Neuron life cycle: Cell cycle exit, then migration, axon targeting, and synaptic integration and functional maturation (sensory-dependent development. Challenge of enhancer functions: need to be both long-term (neurons last for a life), and dynamic.
- Model of enhancer changes during neuron development: Figure 3. (1) Pluriopotent TFs: activity go down, and their target enhancers become decommissioned and then silenced. (2) Neuronal enhancers: first pioneer TFs open up chromatin, then lineage/identity TFs. (3) Also constitutive TFs may play a role: recruiting lineage-specific or stimulation-responsive TFs.
- Function of TFs in regulating enhancers: some TFs, e.g. Nkx2.1 can be both activator and repressor.
- Mechanism of enhancer decommissioning: actively regulated, by histone deacetylation (e.g. LSD1) and then DNAm to make enhancers silenced (locked in).
- Maintaining neuronal states: some TFs may work life-long, some only transiently expressed. Ex. spinal motor neurons: the TFs and enhancers during development are different from TFs and enhancer that maintain cell identify.
- Activity dependent transcription: key feature of neuronal enhancers. How is it achieved? Permissive enhancers. Change transcriptional elongation and activation of transcriptionally poised genes (by histone acetylation). Induce p300/CBP and activate many enhancers. Enhancers most activated by electric signals are enriched with AP-1 and Fos-Jun targets.
- Enhancers in diseases: SNP in CACNA1C (Ca channel), associated with SCZ adn BP has allele-specific enhancer activity. FKBP5 (GR regulation): stress associated disorders, change stress hormone axis.
- Remark: the problems of dynamic regulation: (1) What is the primary function of activity dependent transcription? And how it varies between cell types? (2) Mechanism of functional maturation: how enhancer states become locked? Or are they reversible? (3) Important also in adult?

Spatio-temporal transcriptome of the human brain [Kang & Sestan, Nature, 2011]:

- Data: the transcriptomes (exon array) of 16 regions comprising the cerebellar cortex, mediodorsal nucleus of the thalamus, striatum, amygdala, hippocampus and 11 areas of the neocortex. 1,340 tissue samples from 57 developing and adult post-mortem brains of clinically unremarkable donors. We also genotyped donor DNA (2.5 million SNPs).
- Spatial-temporal gene expression patterns:
  - Number of expressed genes: 15,132 (86.1%) of 17,565 genes surveyed were expressed in at least one brain region during at least one period, and that 14,375 (81.8%) were expressed in at least one NCX area.

  - Spatial variation: 70.9% of expressed genes were spatially DEX between any two regions within at least one period, and that 24.1% were spatially DEX between any two NCX areas.
  - Temporal variation: 89.9% of expressed genes were temporally DEX between any two periods across regions, and 85.3% were temporally DEX between any two periods across NCX areas
  - The bulk of spatio-temporal regulation occurred during prenatal development.

- Correlation among brain regions: At the level of NCX areas, clustering formed the following groups during fetal periods: OFC, DFC and MFC; VFC and primary somatomotor cortex (S1C and M1C); and parietal-temporal perisylvian areas (IPC, A1C and STC). V1C had the most distinctive transcriptional profile of NCX areas throughout development and adulthood.

- Region-specific expression: CBC showed the greatest number of region-restricted or region-enriched DEX genes, with 516 (4.8%) of 10,729 genes spatially DEX. By contrast, the numbers of genes highly enriched in the other regions were lower: NCX, 46 (0.43%); HIP, 48 (0.45%); AMY, 4 (0.04%); STR, 137 (1.28%); MD, 216 (2.01%).

- Spatial-temporal expression pattern of ASD-related genes (Suppl. Figure 25): show distinct temporal dynamics. Examples:
  - CNTNAP2, enriched in the areas of the fetal OFC and DFC. It suddenly increased in other cortical areas during early infancy, then remain expressed in all NCX areas.
  - MET was highly enriched in the early midfetal ITC and then increased in the surrounding temporal-occipital areas and the OFC, where it remained enriched throughout development.

  Top 50 genes most correlated with known ASD genes in Suppl. Table 15 (50 related genes per ASD gene).

Overview of FOXP2 [Fisher & Scharff, TIG, 2009; Dominguez & Rakic, Nature, 2009; Spiteri & Geschwind, AJHG, 2007]:

- FOXP2 conservation across species: Orthologs exist in highly similar forms in many vertebrates, but human accelated evolution. Comparable neural expression patterns.

- FOXP2 expression/functional clues: development (brain and lung) and sensory-motor integration.
  - Patients with FOXP2 mutations: developmental abnormalities of basal ganglia (BG) and inferior frontal cortex (IFC).
  - FOXP2 expression: including Broca's region, but also extend more broadly.

- FOXP2 phenotypes:
  - Human: FOXP2 mutant shows a monogenic syndrome characterized by impaired speech development and linguistic deficits.
  - Mouse: reduced FOXP2 dosage yields abnormal synaptic plasticity and impaired motor-skill learning.
  - Songbirds: reduced FOXP2 dosage disrupts vocal learning.

  The picture is: FOXP2 mutation affects motor sequencing actions and synaptic plasticity (two corroborated with each other), and these effects are manifested as disorders of speech and language.

- FOXP2 regulatory networks: Figure 2.
  - Regulation of FOXP2 expression: Lef1 (via Wnt signaling).
  - Heterodimization with FOXP1 and FOXP4, but FOXP1 and FOXP2 expression patterns are not identica. Also homodimerization (may involve different isoforms).

- FOXP2 has a transcriptional repression domain (Zn finger), interacting with CtBP1. Also interaction with Nkx2.1 TF.
- Downstream targets: cell adhesion, ion transport, neurite outgrowth and axongenesis, signaling pathways (e.g. Wnt/Notch), synaptic plasticity.

Molecular evolution of FOXP2 [Enard & Pabbo, Nature, 2002]:

- FOXP2 sequences show accelerated changes in human lineage (since human-chimp split): 2 AA substitutions vs extremely high conservation in other lineages, e.g. 1 substition in mouse lineage over 70 Myr.
- Evidence of positive selection: signature of selective sweep in human polymorphism data (fixation in human lineage of protein sequences).
- Possible functional consequence of two AA substituions: via structural modeling, the human-specific change at position 325 creates a potential target site for phosphorylation by protein kinase C together with a minor change in predicted secondary structure.

FOXP2 targets in human [Spiteri & Geschwind, Vernes & Fisher, AJHG, 2007]:

- Data: two regions of human fetal brain (peak period of neuronal migration and differentiation) - BG and IFC, human fetal lung [Spiteri]; human neuron-like cells, with stable transfection of FOXP2 (little endogenous FOXP2) [Vernes]. ChIP-chip: on about 6,000 promoters (1k).
- FOXP2 binding targets: about 200-300 targets (depending on the tissues). Functional categories: nervous system development, organ development/morphogeneis, Wnt/Notch singaling pathway, synnatptic transmission, axongenesis. Thus supports the role of FOXP2 in development and modeling of neural connections.
- Binding site characterization:

Spiteri : Among 323 targets, 106 contain FOXP2 consensus site, CAAATT, and 82 FOXP1 site, TATTT[A/G]T.

Vernes : Only top 100 promoters are analyzed. 70 out of 100 contain FOXP2, or FOXP, or FOX sites, with 40/100 FOXP2 sites. Possible co-factors: (1) 21 different factors from TRANSFAC are enriched, $P < 0.01$ (no multiple hypothesis correction); (2) by MEME search, the top motif (in addition to FOXP2) resembles UBP1 motif. Also found 29 compound sites (two FOXP sites within 100bp) of the same category, and 37 of different categories.

- Effects of FOXP2 overexpression on ChIP-targets: quantify the expression of ChIP-targets by RT-PCR. Mostly repression, however for some targets, show activation (about 5 to 1).
- Differential expression of ChIP-targets in human and chimp brain: 47 ChIP-targets - TF in neural development, CRS patterning and guidance molecules, neurotransmitter receptors.

Human-specific transcriptional regulation of FOXP2 [Konopka & Geschwind, Nature, 2009]:

- Problem: to understand the adaptive values of FOXP2, map humna-specific targets (including downstream ones) of FOXP2, as these are candidate genes that confer the adapative effects of FOXP2.
- In vitro human-specific FOXP2 targets: in human neuronal cells without endogenous FOXP2, expression of FOXP2 (human) and FOXP2$^{\text{chimp}}$. 61 up-regulated genes (by FOXP2) and 55 downregulated ones. Five FOXP2 direct targets (in ChIP-chip) in a list of 25 differentially expressed genes chosen for RT-PCR. Up-regulated genes: transcriptional regulation of gene expression and cell-cell signaling; down: protein and cell regulation. Enriched with known involvement in cerebellar motor function, craniofacial formation, and cartilage and connective tissue formation.

- Mechanism of differential regulation: (1) differential interaction with FOXP1 and FOXP4 gith MS, no evidence. (2) Differential activation of promoters (upstream 1000 bp, with at least one forkhead BS): 6 out of 8 tested ones.
- In vivo differential expression: the gene list overlap significantly with in vitro results.
- Evolution of the differential targets (from in vitro results): (1) many are associated with HaCNS; (2) five genes are under positive selection.
- Remark: The causes of differential expression: the function of two AAs is unknown. (1) May affect DNA binding of TFs. (2) May affect interaction with co-factors, however, FOXP1 and FOXP4 are excluded from this study. (3) Same TF binding, but differential activation, caused by change in activation domain of FOXP2 (e.g. interaction with BTM). If this is the case, them all the FOXP2 targets will be affected. In addition, note that even though differential expression is only detected in human background, this cannot exclude the co-evolution of CREs (possible as many of differential targets are associated with HaCNS).

Transcriptome comparison between autism and controls [Voineagu & Geschwind, Transcriptomic analysis of autistic brain reveals convergent molecular pathology, Nature, 2011]:

- Data: post-mortem brain tissue samples from 19 autism cases and 17 controls from the Autism Tissue Project and the Harvard brain bank using Illumina microarrays. profiled three regions previously implicated in autism: superior temporal gyrus (STG, also known as Brodmann's area, BA 41/42), prefrontal cortex (BA9) and cerebellar vermis.
- Genes diffientially expressed in autims vs. controls: 444 genes showing significant expression changes in autism cortex samples, and only 2 in cerebellum. For 209 genes downregulated in autistic cortex: GO related to synaptic function, whereas the upregulated genes ($N = 235$) showed enrichment for immune and inflammatory response.
- Comparison of global co-expression network in cases vs controls: the majority (87%) of the autism modules showed significant overlap with the previously described human brain modules.
- Change of relative expression pattern in autisms: whereas 174 genes were differentially expressed between control BA9 and BA41, none of the genes were differentially expressed in the same regional comparison among the ASD cases. These data suggest that typical regional differences are attenuated in frontal and temporal lobe in autism brain.
- Groups of co-expressed genes showing transcriptional differences between autism and controls: a co-expression network using the entire data set (both autism and control). The comparison between autism and control groups revealed two network modules whose eigengenes were highly correlated with disease status:
  - The top module (M12) showed highly significant enrichment for neuronal markers. The M12 eigengene was under-expressed in autism cases. Unlike differentially expressed genes, M12 showed overrepresentation of known autism genes (GWAS).
  - The second module (M16) was enriched for astrocyte markers and markers of activated microglia, as well as for genes belonging to immune and inflammatory GO categories. This module was upregulated in ASD brain. One of the hubs of the M12 module was A2BP1, a neural- and muscle-specific alternative splicing regulator. M16 does not show enrichment of GWAS genes, suggesting these genes may not be causal to ASD.

Expression of ASD genes in Human Brain Transcriptome data [Expression Profiling of Autism Candidate Genes during Human Brain Development Implicates Central Immune Signaling Pathways, PLoS ONE, 2011]:

- Data: NIMH Transcriptional Atlas of Human Brain Development for all genes implicated in ASD (the database AutDB). The NIMH Atlas contains next-generation RNA sequencing data from 16 normal human brain regions, and spans 21 weeks gestation through 40 years of age.
  - We narrowed our focus to those 11 that were most relevant to autism: Dorsolateral Prefrontal Cortex (DLPC), Ventrolateral Prefrontal Cortex (VLPC), Medial Prefrontal Cortex (MPC), Orbital Prefrontal Cortex (OPC), Posterior Superior Temporal Cortex (PSTC), Inferior Lateral Temporal Cortex (ILTC), Hippocampus (Hipp), Amygdala (Amyg), Striatum (Stri), Cerebellum (Cere), and Primary Motor Cortex (PMC).
- Data quality control: Housekeeping gene expression: should be measurable and remain constant across time and brain region. Brain-specific markers: e.g. expect the absence of epithelial and muscle-specific markers; e.g. high expression of neuron-specific markers.
- Expression profiles of Autism, Epilepsy and Schizophrenia candidate genes:
  - Expression level: greater than 70% were not expressed highly in each brain region. In each region, a large percentage of ASD-implicated genes had no detectable transcription (< 1 RPKM). E.g. in dorsolateral prefrontal cortex (40/219 or 18%) lack detectable expression.
  - Regional enrichment: the cerebellum and frontal cortex contained the greatest number of highly expressed "Autism genes" and the temporal cortex had the greatest number of "Epilepsy genes", whereas Schizophrenia gene expression distributed more evenly throughout the brain.
  - The developing hippocampus had the fewest ASD candidate genes expressed at high levels, and none were specific for the hippocampus.
  - The cerebellum contained a unique set of six Autism candidate genes that were not highly expressed in any other brain region.
  - Only one gene (Gabrb3) was specific to the frontal cortex, and it was only present at high levels in the ventrolateral prefrontal cortex. Interestingly, this gene lies in the 15q11?C13 imprinted region implicated in Prader-Willi and Anglemen Syndromes, and is one of the most reproducible loci identified in ASD GWAS.
- Highly expressed genes: we focused on genes in the top three expression tiers as genes that are significantly highly expressed as compared to all ASD-implicated genes. This yielded 32 genes for Autism, 42 for Epilepsy and 212 for Schizophrenia.
  - Nine Autism genes were highly expressed in all brain regions examined. Their temporal expression profiles were mostly constant across developmental stages, except for Fabp7, which exhibited drastic differential expression.
  - Gene ontology enrichment of the 32 highly expressed Autism genes: four new GO categories representing two significant processes: immune system regulation and apoptosis
  - GO enrichment of the highly expressed Schizophrenia genes yielded a much different set of processes, mostly implicating cellular morphogenesis, but none involving the immune response
- Network-based enrichment analysis: using the curated Ingenuity Pathway Analysis (IPA) database.
  - For Autism: corticotrophin releasing hormone signaling, g-protein and phospholipase C signaling, and neutrophil cytokine signaling.
  - Connectivity networks derived from the enriched gene set compared to those derived from all Autism-associated genes: cytokine signaling molecules.
  - Similar analysis comparing ASD-associated gene networks specific to brain regions did not result in a significant clustering by region, nor were there central network nodes

- Cell-type specific expression: Human Protein Atlas database for the 32 highly expressed Autism genes. Many highly expressed ASD genes are mainly detected in glia not neurons, and/or in specific layers of the cerebellum.
- Autism transcriptome data: in each study only $\tilde{5}\%$ of genes that are significantly different between ASD and control brains were previously implicated in ASD.
- Discussion: role of immune component in ASD:
  - Autoantibodies: Various autoimmune phenomena including autoantibodies to neural antigens and maternal-fetal cross-reactive neural antibodies
  - Inflmmatory response: Post-mortem brain tissue from ASD patients shows increased microglial density in grey matter, an activated morphology, and secretion of a cytokine profile consistent with a pro-inflammatory state, most prominent in the cerebellum.
  - Two prevailing theories of the role of immune function in ASD: one suggests exogenous factor(s) stimulate neuro-inflammation during development, while the other postulates autoimmune activation causes ASD pathology
  - Alternatively, as glia are increasingly implicated in normal formation of synaptic connectivity, it is possible that genomic aberrations ultimately funnel through core signaling pathways of glial cells to disrupt formation of neural networks independent of an inflammatory mechanism. Evidence: (1) a number of recent reports demonstrated that these same cytokine signaling pathways are central to proper brain development; (2) NF-κB pathway is important in synaptic plasticity independent of an inflammatory mechanism.

Epigenomic Landscapes Reflect Neuronal Diversity [Heinkoff, Neuron, 2015]

- Goal: high-resolution cell-type-specific epigenomic map for a mammalian brain.
- The INTACT method is based on affinity purification of nuclei using magnetic beads, where cell type specificity is achieved by expression of a nuclear envelope protein under control of a cell-type-specific promoter.
- Diversity between cell types: (1) 2,000 genes with 2-fold differences between each cell type pair. (2) ATAC-seq: only 13.4% of the 320,000 regions of accessible DNA were shared between neuronal cell types.

Chromosome conformation elucidates regulatory relationships in developing human brain [Won and Geschwind, Nature, 2016]

- Experiment: Hi-C in mid-gestation developing human cerebral cortex during the peak of neurogenesis and migration. CP (post-mitotic neurons) and GZ (NPCs).
- Validation of Hi-C interactions: limit to interactions within TADs. (1) DHS activity: Hi-C interacting enhancers and promoters show higher correlation. (2) eQTL: eQTLs and associated gene pairs show higher Hi-C interactions (OR = 3.2).
- Linking GWAS SNPs to target genes using Hi-C: Extended Figure 7. Obtain credible set SNPs: for SNPs in coding and promoters (2kb), target genes are known. For each remaining SNP, consider 10kb bin, and interactions within 1Mb. Interaction profiles in 10kb bins and do FDR correction. Q: only consider promoters in defining interactions?
- Results of linking GWAS SNPs to genes: most are far from SNPs. From distal SNPs, about 2000 genes, enriched in relevant GOs.
- Examples: DRD2, CHRNA2 (Figure 3d), SNPs in credible set, their interaction profiles with all genes in 1Mb range.

- Combined analysis with CMC data: hi-C interaction supported by eQTLs. FOXG1 (Figure 4): validated the SNP by reporter and CRISPR.
- **Lesson**: validation of Hi-C data, enrichment of interactions in eQTL-gene pairs; correlations of enhancer/promoter activity.

Genome-wide, integrative analysis implicates microRNA dysregulation in autism spectrum disorder [Wu and Geschwind, NN, 2016]

- DE miRNA in ASD vs. controls: 3 regions (FC, TC and cerebellum), 56 ASD cases and 42 controls. Analysis of 700 miRNAs (including 147 previously unknown ones), and test use LMM correcting for potential confounders. Results: 58 miRNAs, and widely dys-regulated in ASD cases.
- WGCNA analysis of miRNA: find modules and eigen-genes, and identify several modules associated with ASD status.
- Enrichment of putative ASD risk genes in targets of candidate miRNAs: candidate miRNAs are up- or down-regulated miRNAs and those in the three ASD modules. Predict targets using TargetScan, adding additional criteria, strongest targets or conservation. Enrichment of various ASD-related gene sets.
- Functional relationship of miRNA and target expression: for many candidate miRNAs, their target mRNAs change expression in the direction as expected.
- Experimental studies of specific miRNAs: DEX miRNAs, targets show expected changes. Demonstrate that the miRNAs regulate ASD-related genes.
- Lesson: WGCNA, building modules, inspecting modules by correlating with covariates (what may drive co-expression), eigen-genes, using modules to define interesting gene groups (e.g. correlated with trait of interest).
- Lesson: find candidate miRNAs from expression analysis, use significantly DEX miRNAs, and the ones in disease-associated co-expression modules. To identify disease-associated module, use WGCNA, and correlate eigen-gene with disease status.

Open Chromatin Profiling in hiPSC-Derived Neurons Prioritizes Functional Noncoding Psychiatric Risk Variants and Highlights Neurodevelopmental Loci [Forrest and Duan, Cell Stem Cell, 2017]

- Experiment: iPSC, iPSC neuron at 30 days and at 41 days. ATAC-seq profiling.
- Chromatin accessibility map: (1) GO of cell type-specific OCRs: consistent with cell types. (2) Spatial distribution: 80% OCRs are distal (more than 5kb). However, in shared OCRs, the majority are 0-5kb. Core promoters are constitutively active, while cell-specific expression are controlled by distal enhancers.
- Enrichment of SCZ GWAS variants in OCRs: 125 index SNPs, and obtain 3,507 SNPs in high LD. Assess the number of these SNPs in OCRs, and compare with permutation (random 125 SNPs, matching the number of LD proxies). 2-fold enrichment of ND-41, and 1.8 fold of ND30, and 1.7 fold of iPSC vs. 1.4 fold of LCL.
- Enrichment of SCZ GWAS variants in TF footprints using PIQ: comparing with LCL, iPSC footprints show 1.2 fold enrichment of GWAS SNPs, and Nd30, ND41 show 1.5 - 2 fold enrichment (note: here using LCL as background). Among 3,509 SNPs (LD $>$ 0.8), about 100 are within 100bp of TF footprints, PIQ $>$ 0.9.
- Lesson: core promoters (5kb) may not be very cell-type specific, and distal enhancers are more important for tissue-specificity.

An xQTL map integrates the genetic architecture of the human brain's transcriptome and epigenome [Ng and De Jager, NN, 2017]

- Data: eQTL, haQTL (H3K9ac, active promoters), meQTL in 400 brains. For mQTL, use 5kb for cis-SNPs. To call xQTL, use Spearman correlation, then Bonferroni correction.
- Method: π_1 estimation (used in several analyses), only use top SNP for each locus. Empirical null of π_1: from random subsets - only SNPs not overlapping xQTL are used in this.
- Cross-tissue comparison: (1) high sharing with blood eQTL (pi1 0.6), some fraction of brain eQTL have large p-values in blood eQTL (Figure 2b). Example of brain-specific eQTL (Figure 2d). Note: high sharing with blood is likely due to presence of immune cells in adult brain. (2) Lower sharing with adipose, liver, 0.2-0.5.
- xQTL sharing: for each xQTL SNP, consider only its nearest feature. Ex. mQTL vs.eQTL: ascertain mQTL, then use only the nearest cis-eQTL. Estimate pair-wise sharing (Figure 3d): 0.4-0.6.
- Mediation analysis (Figure 4): only on SNPs associated with all three phenotypes (10K SNPs). use Bayes networks to compare several models, whether epigenetic trait mediates expression, or the opposite or independent model (IM). In only 9% cases, find epigenetic-regulation model (EM), and 85% IM.
- Cell-type specific xQTL (Figure 5b): estimate cell proportions (one marker gene per cell type), then test whether the effects of xQTL depend on cell proportions. Found a small number (20) of cell-type specific xQTL.
- GWAS: (1) Enrichment of heritability of many traits in xQTL SNPs: from psychiatric to IBD. Note: just SNPs passing threshold. (2) xQTL-weighted GWAS: p-value of a SNP is scaled by a parameter, which depends on the annotation (xQTL or not). The parameters are chosen by a procedure similar to cross-validation. With this xQTL weighting, find 18 new loci for SCZ.
- Lesson: possible to estimate cell proportions with bulk tissue RNA-seq data.
- Discussion: when pi1 is large, the pi1 estimation method may not work well.

The Dynamic Landscape of Open Chromatin during Human Cortical Neurogenesis [Torre-Ubieta and Geschwind, Cell, 2018]

- Data: ATAC-seq and RNA-seq of germinal zone (GZ) and cortical plate (CP) from about 5-10 donors each.
- Change of chromatin accessibility in GZ vs. CP: about 15-20K differential accessibility (DA) peaks in GZ and CP. GO analysis of GZ > CP peaks suggests neurogenesis, and of CP > GZ suggests neuron functions.
- Correlation of change of chromatin accessibility and gene expression: (1) A few cases: quantitative change of chromatin accessibility in promoters. (2) Correlation of promoter peak changes (between GZ and CP) and expression changes.
- Linking open chromatin regions (OCRs) to target genes: enhancer-gene correlation is higher if supported by Hi-C, but modest change. Correlation of enhancer-gene across samples: 0.37 (best correlated enhancers), and 0.42 if limit to Hi-C interactions.
- Identifying enriched motifs in GZ > CP peaks and CP > GZ peaks: 97 and 26 motifs identified.
- Human-gain enhancers (HGEs): obtain the list of HGEs from [Reilly and Noonan, Science, 2015]. Enrichment of DA peaks in HGEs. Link HGEs with target genes using GZ > CP peaks and CP > GZ peaks. Expression of target genes suggest enrichment of radial glia (RG) cells, based on scRNA-seq fetal brain.

- GWAS enrichment analysis: stratified LDSC show enrichment in GZ > CP enhancers for SCZ, education attainment, intracranial volume and neuroticism (10 fold? Figure 7C), but not other traits such as autism and ADHD. No enrichment for CP > GZ peaks. Adult brain ATAC-seq peaks: enriched in SCZ and education attainment (similar degree), but not ADHD, ICV, neuroticism.

- Lesson: Hi-C interactions weakly correlate with functional interactions (inferred from correlated activities).

Evaluation of chromatin accessibility in prefrontalcortex of individuals with schizophrenia [Bryois and Crawford, NC, 2018]

- Data: 130 SCZ cases and controls from CommonMinds, ATAC-seq data in DLPFC.

- ATAC-seq data and evaluation: (1) comparison with other brain ATAC-seq datasets. 14% overlap with fetal brain Roadmap data, and 30-60% overlap with other adult data. (2) Comparison with ATAC-seq of other cell types: for each peak, obtain the number of cell types it occurs; then the histogram. Only 15% peaks are adult brain specific.

- OCRs are enriched with heritability: 7 fold enriched in SCZ and 3 in cognitive (comparison with other annotations), but not in other traits (education). ATAC + conserved: 36 fold enriched in SCZ, and 20 in education and coginitive trait.

- Differential accessibility between cases and controls: found only 3 DA peaks.

- Chromatin QTL (cQTL): 6200 SNPs,10% are in the peaks. No enrichment in GWAS loci. Mapping: use fastQTL, treating each peak as a gene. Test association of a peak with SNPs in 5kb or 50kb windows. Most associations are within 2kb of peaks (Figure 4a, 4b). Retain only the strongest caQTL per peak.

- Comparison of cQTL and eQTL: (1) About 23% eQTLs are also caQTLs: randomly choose an eQTL per gene, then obtain its caQTL p-values and estimate the proportion by FDR package. (2) discordant directions, 30% of times, and effect sizes weakly correlate, $r = 0.21$.

- Co-localization analysis: 8 regions with colocalization prob. > 0.9 (Table 2). Using Hi-C, find many chromatin interactions for a SNP that is chQTL and eQTL (multiple OCRs?)

- Remark: heritability enrichment in OCRs and lack of cQTL enrichment are puzzling. Possible explanations: (1) cQTL enrichment analysis is not done properly, e.g. handling LD; (2) lack of power: number of independent cQTL is too small.

Cell-specific histone modification maps in the human frontal lobe link schizophrenia risk to the neuronal epigenome [Girdhar and Akbarian, NN, 2018]

- Data: DLPFC and ACC of about 30 samples, H3K4me3 and H3K27ac (total of 157 libraries). Separate into neuronal and non-neuronal cells (using markers), and do ChIP-seq separately.

- Comparison of cell type specific epigenome: for two brain regions, > 90% overlap. About 30-40% of neuronal peaks are in non-neuronal cells.

- S-LDSC analysis of 18 phenotypes (Figure 3): Strongest in SCZ, then other brain-related phenotypes, but not AIDs or LOAD. Enrichment limited to neuronal peaks.

- Decomposing variance components of epigenome: cell types are major drivers, brain regions little.

- hQTL mapping: sample size 11-17 per data type (cell type and histone mark), 7000 or so hQTL in H3K27ac and 2000 hQTL in H3K4me3 with Rasqual. Limit to cis-hQTL analysis: 10kb regions of 90k peaks.

- Overlap of hQTL with GWAS: show a few SCZ GWAS loci (Figure 5). Plotting p-values from GWAS and p-values of hQTL in the same region: many shared sites. Remark: multiple SNPs in LD can drive shared signals in hQTL and GWAS.
- Epigenomic variation between cell types (Figure 6): differential activity analysis finds a large number of peaks different between neuronal and non-neuronal cells, and 500 H3K4me3 and 10K h3K27ac DA peaks between two regions of neurons. Multiple DA peaks are close to NPD genes.

Comprehensive functional genomic resource and integrative model for the human brain [Wang and Gerstein, Science, 2018]

- Resources: PFC from adult brain samples, about 1000 normal, 500 SCZ, 200 Bipolar, 44 ASD. RNA-seq in all samples, and H3K27ac in 400.
- Transcriptome variation (Figure 2A-B): (1) use scRNA-seq, about 15K cells to cluster and obtain cell-type specific (9 Ex. subtypes, 8 In subtypes, etc) transcriptome. (2) NMF for bulk expression (sample by gene): the factors match the cell-type specific transcriptome from single cell analysis.
- Enhancer/transcription comparison (Figure 3C,E): Reference Component Analysis (RCA) on transcriptome, clear separation of brain and other tissues. RCA on epigenome, much less separation.
- QTL mapping: 2000 cQTL using 200 samples. 2.5M eQTL using 1800 samples. Also a small number of cell fraction QTL. Comparison of multiple QTL: cQTL vs eQTL, $\pi_1 = 0.89$, fQTL vs. eQTL, only 0.11.
- GRN: (1) Adult brain Hi-C in reference: only 31% are in fetal brain Hi-C. Define physical linkage of enhancer-promoters in the same TAD. (2) TF-enhancer and TF-promoter linkage: elastic net of gene expression vs. TF expression, then presence of TFBS in the enhancer of the gene. The TFs are enriched with brain functions.
- GWAS candidate genes: 140 loci, several criteria of defining SNP-gene targets, eQTL, CRE-gene linkage (if CRE contains a binding site of TF regulator), Hi-C. 300 high confidence (at least two supports) genes.
- Deep Structured Phenotype Network (DSPN) (Figure 7): L0 input layer (SNP), L1 enhancers, genes and co-expression modules, L2 hidden and L3 phenotype. DSPN-full using all data greatly improves prediction of SCZ, BSD and ASD comparing with linear model. DSPN-impute (intermediate layers imputed) modestly improves prediction.
- Lesson: RCA is a compromise of PCA and t-SNE, capturing local structure while preserving global distance meaningfully.
- Remark: possible explanation of why epigenomes of different tissues are more similar than transcriptome: suppose expression of a gene is determined by multiple enhancers, acting synergistically, then each enhancer may have a small difference across tissues, but gene expression would vary more.
- Remark: GRN construction, rich information of TF-gene expression correlation.
- Remark: modeling the intermediate steps from genotype to phenotype can improve predictions.

Lost in Translation: Traversing the Complex Path from Genomics to Therapeutics in Autism Spectrum Disorder [Sestan and State, Neuron, 2018]

- Discussion of why ASD genes have specific phenotypes (rather than vision, for example): different sensitivity of different parts/layers of brain/neurons to perturbation.
- Remark: an interesting problem is: how "robustness" of a neuron network depends on its architecture, e.g. the number of layers.

What the placenta could reveal about autism [Spectrum News, 2019]

- How maternal-placenta-interface (MPI) affects autism: immune activation in mid-pregnancy can alter the mouse placenta, depriving the fetus of oxygen and stunting the growth of neuronal precursors in the fetal brain.
- Remark: a general path of how environment may affect the risk of ASD: env > immune system > placental > O2 or other important variables for development > embryo brain.

Editing of RNA may play sizeable role in autism [Spectrum News, 2019]

- RNA-seq from the postmortem brains of 35 autistic and 34 typical people: significantly less editing in patients. Differences are especially common in RNA from genes linked to autism and synapse genes.
- Role of FMRP: binds to ADAR, and to the editing sites. Significantly less editing in patients with FMRP mutations.

Big brains may hold clues to origins of autism [Spectrum News, 2019]

- Up to 20 percent of children with autism have early brain overgrowth. Usually notice in a few months after birth.Children with autism who show early brain overgrowth have 67 percent more neurons in the prefrontal cortex than controls do.
- Abnormal proliferation of neural progenitors. Between weeks 7 and 20 of gestation, these precursor cells divide rapidly. Abnormal proliferation and enlarged brain found in mutations of 16p11.2 (KCTD13), CHD8, WDFY3.
- Early brain overgrowth tends to affect the frontal and temporal lobes (front and sides) of the cerebral cortex.
- Cerebral overgrowth may cause changes in the proportions of different cell types in the affected brain areas, e.g.abundance of excitatory neurons relative to inhibitory neurons.

Transcriptomic and open chromatin atlas of high-resolution anatomical regions in the rhesus macaque brain [Yin and Yu (Dan Xie), review for NComm, 2019]

- Data: 52 brain regions from 8 monkeys. 400 RNA-seq and 26 ATAC-seq datasets.
- Clustering of RNA-seq samples: for cortex regions, clustering by individuals; for lower brain structures, clustering by regions. Also regions spatially close tend to cluster. Interpretation: (1) Cortex regions are more similar between each other, so do not cluster by regions; (2) Shared neurons and axons in close regions.
- WGCNA clusters: 55 co-expression modules. Some modules are specific to cortex, cerebellum, olfactory bulb, hypothalamus, etc.
- Individual difference in gene expression: focus on cortex regions. The top DE genes are enriched for MHC complex and oxidoreductase activity, mitochrondrial genes.
- Age difference: young vs. mid-aged. (1) Medulla oblongata (MO, brain stem): MHC II. In Brain stem: young brains enriched in synaptic / GPCR signaling; and old brains immune. (2) Pituitary gland: young, cell differentiation and neurogenesis; old, signaling. (3) Cortex: old, immune, MHC II. Young: rheumatoid arthritis (RA). (4) Striatum (mortor and reward system): old, meiosis genes, suggesting adult neurogenesis. Young: RA and asthma and antigen presentation.
- Gender difference: pituitary gland show large difference.

- Novel transcripts: only 30% of reads mapped to exons. Many new transcripts: about 30% (9000), 18% intergenic, 12% intronic. About 200 are homologous to human proteins. Some novel transcripts are highly region specific: e.g. in one lobe only.
- Novel lncRNAs: 2,800. Some modules enriched for TF motifs: function of lncRNAs as sponges of TFs.
- Comparison with human: in human, similar patterns, but AMY and HIP samples cluster with cortex samples from the same donor. This suggests evolutionary shifting of AMY and HIP towards cortex-like.
- Lesson: clustering patterns may differ in tissues/regions, for some, individual variation more important, for others, tissue/region variation may be important.

Mapping cis-regulatory chromatin contacts in neural cells links neuropsychiatric disorder risk variants to target genes [Song and Shen, NG, 2019]

- QC and data summary: four iPSC derived cells, excitary, hippocampal, astrocytes, motor neurons. 80% interactions are within 160kb. About 40% are promoter-promoter interactions.
- Validation: Weak enrichment of distal promoter interacting regions (PIRs) in enhancers.
- Cell type specific PIRs and motif analysis: (1) from cell type specific PIRs: map to promoters, then do GO enrichment: some neuron/astrocyte differentiation, some metabolic processes. (2) Motif analysis: using PIRs that are also OCRs. Find some cell-type specific enrichment, e.g. TBR1 only in hippocampal neurons. FOSL1/L2 strongest in astrocytes.
- Vista enhancers: show application of pcHiC data, assigning target genes.
- GWAS: (1) Enrichment analysis: putative fine-mapped SNPs and LD proxy, then test enrichment in PIRs. Enriched in ASD (all four cell types), and Excitatory enriched in AD. (2) Assign gene targets of SNPs/LD proxy: only $< 10\%$ cases assigned to nearest genes.
- GWAS case (Figure 5e): several SCZ associated SNPs in a region with ATAC-seq (about 10kb). The fragment interacts with DRD2 promoter (20kb away). ABE of the SNP changes DRD2 expression. Remark: there are other SNPs in the region (within 100kb) that do not interact with DRD2 promoter. It is unclear what the height of "SCZ SNPs track mean.
- Allele-specific chromatin interactions (PIRs): a significant number of SNPs showing allelic bias in pcHi-C (Figure 6a).
- Lesson: in Hi-C data, many are not enhancers (only 2-fold enrichment) or OCRs. Important to intersect PIRs with regulatory annotations.

Genetic Control of Expression and Splicing in Developing Human Brain Informs Disease Mechanisms [Walker and Geschwind, Cell, 2019]

- Data: fetal brain eQTL and sQTL (200 samples). Mid-gestation, NPC proliferation and migration.
- Characterization of eQTLs: enrichment of fetal brain ChromHMM annotations, about 2-3 fold. TF enrichment: using ENCODE ChIP-seq data (consensus of multiple tissue types) - filter with brain expression, and brain ChIP-seq of CHD8 and SMARCC2 (Figure 2H).
- Characterization of sQTLs: only 22% of eQTLs affect intron usage. Distance and LD of eSNP and sSNP of the same genes.
- Contribution of stage-specific eQTLs to NPDs: for SCZ, prenatal eQTLs about 3.7%, adding adult eQTL (PsychENCODE) 6.6% - showing prenatal and adult are mostly independent.
- Co-expression modules: from WGCNA, a few modules, eQTLs of the genes are enriched with GWAS.

- SCZ TWAS: 62 genes and 91 introns. 8 of them also significant by SMR. Most of the TWAS genes from prenatal are different from adult brain.
- Remark: linking expression modules to traits, using eQTLs of the genes.

5.4.1 Neurodevelopment and Neuron Circuitry

How to map the circuits that define us [Nature, 2017]

- **Paradigm** of mapping brain connectome and understanding behavior: (1) Map the connectiome: cross-sections, trace connections. (2) Map the circuits: activation patterns, and how relate to behavior, e.g. different neuronal activation at different behaviors (fly head movement). (3) Manipulation of circuits: optogenetics.
- Some model systems: fly larva, 15K neurons. Zebrafish: 100K neurons.
- Neurodevelopment: e.g. in zebrafish, neurons stayed together in bundles, and took mirror-image routes on each side of the animal. What guides them?
- Maintaining neuron circuits: things such as ion channels and receptors are replaced?
- **Key technical challenge**: how to measure activities of many neurons at the same time? Current techniques: measure proxies, e.g. calcium release from neurons.

A dont eat me immune signal protects neuronal connections [Nature, 2018]

- Background: during synaptic pruning, neurons express C1q and CR3 will be recognized by microglia and pruned (eat me).
- Neurons expressing CD47 send a dont eat me signal to microglia. In general, during tissue homeostasis and wound healing, healthy cells need to be protected from immune cells.

5.5 Development and Differentiation

Enhancer dynamics during development [personal notes]:

- Ref: Epigenetic priming of enhancers predicts developmental competence of hESC-derived endodermal lineage intermediates [Wang and Sander, CSC, 2015], Neurobiological functions of transcriptional enhancers [Nord and West, NN, 2020].
- Pluripotent enhancers: maintaining Pluripotency state in stem cells. Become decommissioned and silenced later in development.
- Lineage-specific enhancers: step-wise assembly model, first become poised, then become activated when cells commit to a specific lineage.
- Activity-dependent enhancers: become poised/permissive during cell lineage/type commitment, but not activate gene expression. Upon appropriate stimuli, become activated. Ex. neuron activity-dependent enhancers. Some immune cell enhancers that response to cytokine, may also belong to this category.
- Changes happen when cells commit to a fate: pluripotent enhancers become decommissioned. Some lineage-specific enhancers become activated and drive expression of genes important for this cell type, and other lineage-specific enhancers become poised (potentially become multiple lineages later). Activity-dependent enhancers become poised to respond to future signals.

- Question: during differentiation, a small number of lineage specific TFs can drive some enhancers to be poised, and others to be activated. How is this achieved?

Using single-cell RNA-seq to study transcriptome change during development [personal notes]

- Problem: suppose we are interested in a certain subpopulation of cells (known markers), and we would like to study its transcriptome dynamics in a time-course experiment.
- Basic strategy: clustering analysis on all cells, then identify which cluster(s) correspond to the cell type of interest. In the selected cluster, compare transcriptome at different time points.
- Correcting for batch effects: if cells are clustered by batches, rather than cell types, this suggests a batch effect.
- Discovery of novel subtypes: there may be multiple subtypes in the cell type of interest. This can be discovered by checking if there are multiple clusters matching the cell type of interest. We can then analyze transcriptome dynamics in each subtype separately.
- Limitations of this approach: the clustering approach is based on the assumption that cell types do not change much. However, if the cell type changes significantly during development, then its descendant cells may not cluster, and it would be hard to study its transcriptome dynamics.
- Possible idea: assigning a cell to a cell type using a pre-defined model (marker or classifier), assuming this assignment is invariant across time points.

5.5.1 Cellular Differentiation and Reprogramming

The fate of cell reprogramming [Karagiannis & Yamanaka, NM, 2014]

- Cell differenitation metaphor:
  - A cell in the pluripotent state can be thought to sit at the top of a hill from which signals pull it down into its differentiated state. Cell reprogramming to the pluripotent state involves both raising the cell back up the hill and, at the same time, deactivating the signals that pull it back down.
  - The reprogramming process acts as a perturbation that successfully reprograms the cell only if the cell is perturbed into an ideal state. This state, however, is dynamic, and cells can transition to other states. The resulting heterogeneity lead to undesirable outcomes.
- The lineage bias of particular clones means that differentiating an iPSC into a lineage different from its origin is not trivial. Possible strategy: direct conversion of somatic cells.
- Omics data for reprogramming and mathematical modeling. Beter understanding of reprogramming heterogeneity.
- Cellular reprogramming for disease modeling: best applicable to diseases due to specific genomic mutations because reprogramming resets the epigenome.

Epigenetic Priming of Enhancers Predicts Developmental Competence of hESC-derived Endodermal Lineage Intermediates [Wang and Sander, Cell Stem Cell, 2015]

- Experiment: iPSC > Definitive Endoderm (DE) > Gut tube (GT) > Posterior Foregut (FG) > Pancreatic endoderm (PE). Profile H3K4me1 and H3K27ac.
- Epigenetic changes during differentiation: most of enhancers have H3K4me1, only 20% or so have H3K27ac (and H3K4me1). Verify activity of enhancers: those with H3K27ac have higher transcription by GRO-seq.

- Clusters of poised enhancers that become induced in PE: Cluster I: stage-specific H3K4me1 and H3K27ac. Cluster II (majority of enhancers): poised, H3K4me1 in certain stages, but no H3K27ac. Cluster III: constitutive for both H3K4me1 and H3K27ac. In cluster I: some sub-clusters, most notably, active H3K4me1 in GT/FG/PE and active H3K27ac only in PE. These are poised enhancers.
- What about other poised enhancers? Hypothesis: these are enhancers later activated in other lineages, including liver, lung. Show this is true: associated with liver/lung genes, enriched with relevant motifs, and tissue-specific enhancers (from differentiated cells).
- Developmental competency: depend on enhancer states. Only GT cells, but not DE cells respond to inductive signals.
- Role of TFs/step-wise assembly model of enhancers: in DE > GT transition, pioneer factors (FOXA1/A2) opens up chromatin. Then in later stages, inductive signals, e.g. PDX1 activates chromatin.
- Pancreatic islet cell enhancers: existing in poised state in PE, and activated in late stage.
- Lesson: step-wise assembly model of enhancer activation during development, first, become poised enhancers, driven by pioneer factors, then active enhancers, driven by lineage-specific TFs. Why this model? Early on, the cell lineages are not committed yet (e.g. may depend on neighboring cells), so prepare for multiple possible lineages (poised states).

5.5.2 Early Development

Conceptual links between DNA methylation reprogramming in the early embryo and primordial germ cells [Seisenberger, COCB, 2013]

- Background: origin of germ cells. Primordial germ cells (PGC) migrate from primitive streak through gut to the developing gonad (about 4 weeks in human). Then PGC undergo meosis, followed by cellular differentiation, to produce sperms or eggs.
- Major problem: how does fertilized egg quickly change its epigenome and transcriptome to start the process of development?
- Main events in early embryo and rationale: in early embryo, need to remove methylation so that cells become pluripotent; then need remethylation to to turn off inappropriate genes as cells start to differentiate. In the process, certain regions need to be protected against demethylation: imprtinting (retain epigenetic memory), transposable elements, etc. Note: global demethylation may allow cells to quickly change its state, so this needs a different mechanism from typical cellular differentiation.
- Main events in PGC and rationale: second wave of demethylation, as we need to completely remove all history of differentiation as well as epigenetic memory (not useful for germline). Remethylation: to establish the cells fate as germ cells.
- Dynamics of DNA methylation reprogramming in early embryo: from fertilized egg to epiblast. At start: 90% methylation in sperm and 40% in oocyte. (1) Erasure of methylation: active in sperm and passive loss in oocyte (no maintenance). (2) Targeted maintenance: repeats, imprinting control regions (ICRs), germ-cell specific CGIs.
- Dynamics of DNA methylation reprogramming in PGC: first stage from E8.5, complete demethylation, but ICRs, repeat, germ-cell CGIs are demethylated more slolwly. Second stage: after migration to gonad, from E10.5, complete demethylation. A small number of CGIs retain their methylation, leading to transgenerational inheritance.

- Possible mechanisms for methylation changes: (1) Active demethylation: Tet oxidization (5mC to 5hmC and 5fC, 5caC); also the NER (repair) pathway. (2) Passive demethylation: exclude maintenance DNMTs, including DNMT1 and Np95. (3) Targeted maintenance: Zfp57 and Kap1 that recruit DNMTs. Global protein: Stella, maternal factor. (4) Remethylation: by de novo DNMTs, DNMT3a and DNMT3b.
- Mechanisms of global erasure of DNA methylation: (1) Early embryo: first active, then passive. (2) PGC: a possible model is that it is largely passive, but methylation in particular regions rely on active demethylation (hence depend on Tets) e.g. ICRs.
- Erasure of targeted maintenance of DNA methylation in PGC: likely active demethylation (oxidation) in the regions, then 5hmCs cannot be recognized by maintenance mechanisms.
- Q: Why targeted maintenance of germ-cell specific CGIs? When is targeted maintenance established?

The DNA methylation landscape of human early embryos [Guo and Qiao, Nature, 2014]

- Hypothesis/model: during early development, from fertilization to implantation, gene expression changes are accompanied/enabled by the change of methylation.
- Data: RRBS or WGBS, from sperm/egg to zygote to 2,4,8 cells to ICM to post-implantation.
- Main pattern of methylome changes (Figure 1B): dramatic demethylation from sperm/egg to 2-cell stage, then remethylation from ICM to post-implantation.
- Methylation pattern along genes (Figure 1A): lower methylation in TSS, and after TES.
- Compare methylome changes between male and female: paternal methylation decreases much faster than maternal methylation.
- DMR between sperms and eggs: sperm-specific DMRs are enriched with H3K4me1 enhancers that are active in other tissues - possible function in suppress activation of these enhancers (may interfere with sperm development).
- Imprinting: one example of maternal ICRs (0 methylation in sperm), methylation level 50
- Methylation changes in specific regions: histone marks data from human ESC (similar to ICM). (1) H3K4me3 (promoter) regions: low methylation; (2) H3K27me3 (repression): low methylation, actually strong negative correlation between H3K27me3 and methylation across stages/cells, suggesting that the two are complementary mechanisms of repression; (3) H3K9me3 (heterochromatin): no difference.
- Correlation of methylation and gene expression: negative correlation, from -0.2 to -0.4 (post-implantation). Remethylation genes: enriched in translation, RNA processing.
- Methylation at transposable elements: demethylation early, but remethylation later. Help repress transposable elements.
- Lesson: How methylation may be related to gene expression? Lower methylation in active promoters (and CGIs) and likely enhancers. Complementary with H3K27me3 (repressive mark).
- Lesson: How ESC transcriptome is achieved? Promoters: low methylation level established very early (in germ cells). H3K27me3: low methylation levels, established in germ cells.

The landscape of accessible chromatin in mammalian preimplantation embryos [Wu and Xie, Nature, 2016]

- Background: Zygotic genome activation (ZGA), development comes under the exclusive control of the zygotic genome (maternal to zygotic transition). Involve degradation of maternal products. Minor ZGA: 2-cell stage. Major ZGA: after minor ZGA.

- ATAC-seq experiment: use CRISPR to deplete mtDNA. 2-8 cell, ICM stages, and mESC.
- Validation of ATAC-seq data: biological replicates, comparison with DHS, H3K27ac, correlation of promoter ATAC-seq and expression.
- Allele-specific chromatin landscape: similar between maternal and paternal chromosomes. In contrast to DNA methylome.
- Unique properties of accessible chromatin in preimplanation: enrichment in TES and in repeats (transcribed), very different from somatic cells.
- Potential regulators of early development via ATAC-seq analysis: motif enrichment in peaks at each stage. GREAT analysis: chromatin regulators are enriched in early stages, and development gene in late stages.
- Master regulator analysis by integrating ATAC-seq and RNA-seq: MARINa, for each TF, find its targets by ATAC-seq peaks, and divide them into positive and negative groups (using scRNA-seq data, correlation with TF expression). For each transcriptome data, find DE genes, and test if DE genes are enriched with targets (positive or negative) of a TF.
- Chromatin state in minor ZGA: during minor ZGA, chromatin is in a permissive state, found large open chromatin domains associated with promiscuous transcription (of repeats). In Major ZGA, chromatin changes to repressive state (Figure 5d).
- Remark: MARINa is similar to [Jiang and Liu, PNAS, 2015]. Both test TF targets in transcriptome of a condition/sample, without using [TF] itself.
- Lesson: for most regulators, their targets can be either positive or negative.

5.5.3 Embryonic Stem Cells

TRN in pluoropotent stem cells [Boiani & Scholer, NRMCB, 2005]

- Aim: the TRN underlying the cell fate determination of ESCs (how to respond to extraceullular signals).
- Inner cell mass (ICM or primitive ectoderm) and ESCs: ESCs are derived from ICM. They are similar, but not equal. Ex. Lif is required for maintaining pluropotency in vitro, but Lif-/- has normal embryonic development. The phenotypes of gene targeting in vitro and invo are in Figure. 3.
- Siganling pathways of ESCs:
  - LIF: (i) essential in maintaining undifferentiated state of mouse ESC in vitro, but only able to sustain ESCs in the presence of serum, suggesting other factors. LIF does not prevent the differentiation of human ESCs. (ii) In vivo mouse embryos that lack LIF can develop to a stage subsequent to ESC derivation. (iii) LIF signaling pathway: differentiation of human ESCs26LIF $\rightarrow$ binding of LIFR-gp130 heterodimer $\rightarrow$ JAK activation $\rightarrow$ STAT3 activation.
  - BMP4: (i) essential in maintaining undifferentiated state of mouse ESC in vitro. BMP4 seems to be the serum-derived factor in ESC cultures that conatin serum + LIF. (ii) Mouse embryos without BMP4 develop past the stage at which ESCs can be derived. (iii) In the presence of LIF, BMP4 activates Smad4, which in turn activates Id (inhibitor of differentiation).
  - Wnt: activation of the canonical WNT pathway maintains the undifferentiated phenotype in both mouse and human ESCs, and sustains expression of OCT4, REX1 and Nanog in the absence of supplemented LIF.
- Oct4:

  - Expressed only in germline cells. Downregulated in the differentiated versus the undifferentiated state of stem cells.
  - Oct4-/-: development stop at ICM, cannot derive ESCs.
  - Relative amount of Oct4 determines cell fate in vitro, as overexpression of Oct4 leads to differentiation.
  - Increase or decrease target gene expression, depending on the flanking sequences (presence of co-factors).

- Sox2:
  - Expressed at ESC, also neural stem cells.
  - Sox2-/-: defective ICM, cannot derive ESCs.
  - Direct interaction between Oct4 and Sox2: the complex in the enhancer sequence of Fgf4.
  - Foxd3 has a phenotype similar to Sox2, and also directly interacts with Oct4.

- Nanog:
  - Expressed at early stages, reduced expression later, suggesting that the function of Nanog in germ cells is progressively extinguished as they mature.
  - LIF-independent ability of cell renewal and pluropotency (Nanog is identified in LIF-independent undifferentiated stem cells). Removal of Nanog restores LIF dependence.
  - Overexpression of Nanog $\rightarrow$ BMP4 independence.
  - Nanog-/-: lack of ICM.

- Combinatorial signals for pluropotency:
  - Key elements of ESC signaling and transcriptional regulation: LIF-gp130 singaling pathway; Oct4 and Nanog expression.
  - Oct4 and Nanog: independent pathways. (i) Overexpression of Oct4 $\rightarrow$ differentiation while overexpression of Nanog enhances self renewal. (ii) Oct4 suppresses trophectoderm differentiation, while Nanog suppresses differentiation into primitive endoderm. (iii) Oct4 is required for Nanog function, but Nanog expression is normal in Oct4-/-.
  - Nanog and STAT3: function in parallel, but may target common genes. (i) Nanog does not activate STAT3 expression, and STAT3 does not activate Nanog expression. (ii) STAT3 enhances self-renewal of Nanog-overexpressing cells.
  - Oct4 and STAT3: no direct interaction because Oct4 overexpression cannot rescue differentiation of STAT dominant negative cells.

Mouse preimplementation [Wang & Zernicka-Goetz, Dev Cell, 2004; Wang & Powers, Reproductive Biomedicine, 2005]

- Problem: identify genes that are important for preimplementation development via expression profiling.
- Background:
  - One-color array (Affymetrix): designed to measure the expression level of mRNA, relative to other genes in the same sample, not the same gene in different samples (two-color array).
  - Possible to detect the presence of genes in a sample with one-color array: by a threshold level (e.g. 500) or by the Present Call algorithm of Affymetrix.
- Results:

  - Quality control: the Pearson correlations of transcriptomes between replicates. Found to be ¿ 0.95 most of times.
  - Comparison of stages: assess the overall transcriptome change between adjacent stages, find the time point when the expression changes fastest.
  - Validation of expression profiling: the expression patterns of known stage-specific genes conform to expectation.
  - Expression of maternal transcripts: identify maternal transcripts by a pre-defined pattern (high expression in early stage relative to later stages, 2-fold change) and the similarity with known maternal genes; analysis of 57 maternal genes: some are candidate polarity determination genes (from fly orthologs); etc.
  - Oct4 target genes: identify two patterns: (i) some are targets of cofactor Sox2, sharing the expression pattern with Sox2 (high expression later); (ii) other target genes (similar to a known target) are consistently expressed
  - Important signaling pathways (Wnt, BMP, Notch) are active in preimplementation (genes are present); and genes of some important processes, cell cycle, apoptosis, etc. are also active.

- Lessons:
  - Important to validate the quality of expression data. One idea is to compare replicates.
  - Compare the transcriptomes to understand the relations of cell types/stages/strains/etc.
  - Analysis of gene-phenotype association by stage-specific genes identified through: (i) defined stage-specific pattern; (ii) similarity of expression pattern with known stage-specific genes.

Oct4, Sox2, Nanog targets in human ESCs [Boyer & Young, Cell, 2005]

- Problem: how the self-renewal and pluripotency of ESC are created/maintained in terms of gene regulation? Example, genes important for differentiation are suppressed.
- Background: ES cells are derived from the inner cell mass (ICM) of developing blastocycst. The evidences of Oct4, Sox2 and Nanog for ESC include:
  - Disruption of Oct4 and Nanog $\rightarrow$ differentiation of ICM and ES cells to trophectoderm and extra-embryonic endoerm, respectively.
  - Overexpression of Oct4 in ES cells leads to a phenotype siilar to loss of Nanog.
  - Oct4 can heterodimerize with Sox2 to affect the expression of several genes in mouse ES cells.

  The two distinct phenotypes of ESC: self-renewal (can be maintained in culture); and pluripotency (can differentiate into any cell type).
- Methods:
  - ChIP-chip of 3 proteins: 18K genes, -8k to +2k promoter sequence, 60-mer oligonucleotide.
  - Validation of ChIP-chip for determing the targets using yeast TFs: FP $< 1\%$, FN $\approx 20\%$.
- Results:
  - Oct4, Sox2 and Nanog coocupy many target genes: 353 genes bound by all three. Greater than 90% regions bound by Oct4 and Sox2 are also bound by Nanog. The distances between the bound regions of these TFs are close in most cases.
  - Among 353 target genes: up-regulated genes (from expression stuides) are enriched with ES-releted TFs, components of TGF-β and Wnt pathways; down-regulated genes are enriched with TFs important in cell differentiation.

  - Feedforward loop: Oct4 and Sox2 regulates Nanog, then they together control other genes; autoregulatory loop: Oct4, Sox2 and Nanog bind the promoters of their own genes.
  - Extended regulatory network: see Figure 5. The three genes ativate genes in chromatin remodeling and modification, ES cell TFs, TGF-β signaling; and repress TFs involved in differentiation of endoderm, ectoderm, mesoderm, extra-embryonic/placental.

- Remark: it's not clear how the three genes acitivate some targets while repressing the other ones.

Oct4 and Nanog TRN in mouse ESCs [Loh & Ng, NG, 2006]

- Aim: the TRN in ESCs centered on Oct4 and Nanog.
- Methods: ChIP-PET
- Results:
  - Binging site distribution: for Oct4, 5' proximal (10kb) - 19%; 5' distal (10 - 100kb) - 13%; intron - 41%; 3' proximal (10kb) - 7.5%; 3' distal (10 - 100kb) - 9.8%; exon - 2.3%.
  - Oct4 and Nanog binding configuration: about 45% of Oct4 targets are bound by Nanog. Also independent binding of Nanog and Oct4.
  - Nanog motif: CATT containing, confirmed by EMSA
  - Roles of Oct4 and Nanog in controlling gene expression: (i) Induced differentiation of ESC; (ii) RNAi of Oct4 and Nanog; (iii) Overexpression of Nanog. The conclusions:
    * Main targets: diverse clasess, in particular, TFs, growth factors, signaling molecules, DNA damage response sensors and suppressors of lineage-specific genes.
    * Oct4 and Nanog can activate or repress transcription. Dominant role in activating ES cell-specific genes.
    * Genes such as Trp53bp1 and Mycn that are bound by Nanog but are not regulated by it, as observed through RNAi experiments.
  - Esrrb (nuclear hormone receptor) and Rif1: knockdown cells become fibroblast-like.
  - Comparison with human targets [Boyer05]: only 9.1% of Oct4 and 13% of Nanog targets overlap between the two studies.

RNAi of mouse ESC [Ivanova & Lemischka, Nature, 2006]

- Methods: test the effect of removing one gene by self-renewal assay: shRNA + GFP (reduced GFP if reduced self-renewal).
- Genes required for self-renewal: Nanog, Oct4, Sox2, Esrrb, Tbx3 and Tcl1 (transcriptional co-factor).
- Overexpression of Nanog complements Esrrb, Tbx3, Tcl1, Dppa4, but not Oct4 and Sox2.
- Functional roles:
  - Oct4 blocks trophectoderm differentiation
  - Nanog blocks endodermal differentiation. In addition, Nanog downregulation induces the expression of markers for trophectoderm and epiblast-derived lineages, namely mesoderm, ectoderm and neural crest cells. Therefore, Nanog seems to be a global regulator that represses multiple differentiation programmes.
  - Sox2 functions to repress the development of trophectoderm and epiblast-derived lineages.
  - Esrrb and Tbx3 are necessary to block the differentiation into mesoderm, ectoderm and neural crest cells

  - Tcl1 seems to be even more restricted as it appears to repress only a subset of neural crest genes.

- Remark: the function of a TF in ESC can be understood as the functions of its targets: differentiation of different lineages.

Interactome in mouse ESC [Wang & Orkin, Nature, 2006]

- Methods: first find targets of Nanog and Rex1, then find the secondary targets of these genes (only those verified).

- Results:

  - Verified targets of Nanog and Rex1: Dax1, Nac1, Zfp281, Oct4. Dax1 and Sal4 knockdown cells lose pluropotency.
  - Most genes of the interactome are co-regulated, and specifically down-regulated on ES cell differentiation
  - The network is linked to several cofactor pathways involved in transcriptional repression, including: the histone deacetylase NuRD (P66b and HDAC2), polycomb group (YY1, Rnf2 and Rybp) and SWI/SNF chromatin remodelling (BAF155) complexes.

GRN in embryonic stem cells [Zhou & Wong, PNAS, 2007]

- Problem:

  - How the pluripotency of ESC is maintained? Or speak differently, how the expression pattern controlling ESC (e.g. genes X = on,Y = on,Z = off) is maintained?
  - What are the target genes of the known regulators of ESC? Synergistic interactions among TFs?

- Methods:

  - Data: (i) expression data: oct4-sorted cells, i.e. the expression of all genes in 2 cell types, high-oct4 and low-oct4; (ii) ChIP-chip data of oct4, etc.
  - Identification of putative targets of oct4: intersection between the genes whose expression correlates with oct4 (i.e. signicantly affected by oct4 level in terms of fold change), and the genes which are near to some oct4-bound region in ChIP-chip
  - Identification of motif (TF) interactions: a TF interacts with oct4 if its BSs are enriched with oct4-bound region vs expectation under Poisson background distribution (hypergeometric test) + optional conservation filtering (i.e. consider only enrichment within conserved region, defined as the top 20% of the region using PhastCons score)

- Results:

  - Of 24 compiled motifs: in Oct4 dataset, the enriched ones: Oct4, Sox2, STAT3, Esrrb, LRH-1, and Otx2; in Nanog dataset, the enriched ones: Oct4, Sox2, Esrrb, Nanog, STAT3, Sall4, LRH-1.
  - Validation of predicted motif interactions: through expression pattern (coexpression) and the other known functions. Example, Esrrb is highly expressed in Oct4-sorted cells, and an essential gene for self-renewal of ESC via RNAi experiments.

Zfx in embroynic and hematopoietic stem cells [Galan-Caridad & Reizis, Cell, 2007]

- Background: Zfx is a zinc-finger TF, up-regulated in several types of stem cells.

- Results:

  - Zfx is required for self-renewal of mouse ESCs, and dispensable for the growth and differentiation of their progeny.

- Zfx controls Tbx3 and Tcl1: overexpression of Zfx leads to higher expression of Tbx3 and Tcl1, but not Nanog, Oct4. In Zfx-/- cells, the components of the Nanog/Oct4/Sox2 transcriptional network were only minimally affected, but Tbx3 and Tcl1 are down-regulated. In addition, Tbx3 and Tcl1 promoters contain multiple consensus sites of Zfx.

Klf4 in mouse ESC [Jiang & Ng, NCB, 2008]

- Background:
  - Klf4 is dispensable in maintaining pluropotency.
  - Klf2, Klf4 and Klf5 are down-regulated in ES cell differentiation.
- Methods: Evaluating ES cell differentiation: alkaline phosphatase staining and observations of morphological change.
- Results:
  - Simultanesous depletion of Klf2, 4, 5 by RNAi leads to loss of pluropotency.
  - Redundency of binding by Klf2,4,5: depletion of one Klf can be compensated for by the binding of other Klfs to the common target sites.
  - Similar binding specificites of Klf2, 4 and 5. Knockdown of all 3 Klfs leads to abolished enhancer activity of Nanog and other genes.
  - Knockdown of all Klfs change expression of ESC-specific and differentiation genes.

Integration of external signaling pathways with the core transcriptional network in embryonic stem cells [Chen & Ng, Cell, 2008]

- Background: the genes known or expected to be important for self-renewal and/or pluripotency of ESC include:
  - Regulators (with genetic evidence): Oct4, Sox2, Nanog, Esrrb, Zfx
  - Signaling pathways: Smad1 (BMP pathway), STAT3 (LIF pathway). Both BMP and LIF pathways are necessary to maintain ES cells
  - Reprogramming factors: Klf4, c-Myc (these two and Oct4, Sox2 are sufficient to reprogram fibroblasts to induced pluripotent stem cells, iPS), n-Myc
  - E2F1: important for regulating cell cycle progression
  - CTCF: transcriptional insulation

  In addition, test the histone deacetylase p300 and another regulator Suz12.
- Methods:
  - ChIPSeq data analysis: each position in the genome will be assigned a number - the number of overlapping DNA fragments that include this position. For each +/-500 bp window, find the peak. Choose only peaks that satisfy: (i) cutoff determined by Monte Carlo simulation: randomly extract 200 bp fragments, and get the distribution of intensities of all fragement (use peak intensity for each fragment). FDR threshold 0.05; and (ii) fold change ≥ 5 comparing with the negative control (in the random control, peaks with high-intensity can be found in specific genomic regions like satellite repeats, so need additional filtering).
  - Moitf finding: the motifs are found by running de novo motif-discovery algorithm on the top 500 peaks (+/- 100bps).
  - ChIPSeq validation: ChIP-qPCR analysis.

  - Multiple transcription factor binding Loci (MTL): iteratively merge any two TFBSs if they are less than 100bp apart. The significance is tested by: for any motif, the number of sites under Poisson distribution. The p-values of different motifs are combined using Fisher's method.
  - Regulatory network: for any TF, find all its targets using both ChIPSeq data (genes that are close to a peak) and RNAi based expression data (genes whose expression change significantly in RNAi experiment of that TF). A gene is target in either dataset if it is ranked high in both datasets.

- Results:
  - MTL analysis: find 3583 total MTLs, about 20% at promoter regions. Two groups of cooccurrances: Nanog, Sox2, Oct4, Smad1 and STAT3 as a group and n-Myc, c-Myc, E2f1 and Zfx as a group.
  - Enhancer ability of the MTLs: 25 fragements from Nanog-Oct4-Sox2 cluster can drive reporter expression while 8 fragments from the Myc cluster cannot.
  - Interaction between TFs Nanog-Oct4-Sox2 and Smad1, STAT3 signaling: Smad1 and STAT3 share many targets with Nanog, Oct4 and Sox2; depletion of Oct4 leads to reduction in Smad1 and STAT3 binding, but perturbation of the two pathways do not affect Oct4 binding $\Rightarrow$ Oct4 helps Smad1 and STAT3 binding.
  - The function of p300: targets co-occur with Nanog-Oct4-Sox2 cluster $\Rightarrow$ Nanog-Oct4-Sox2 recruit p300 to their binding region.
  - Integrating binding and expression data (Fig. 6): cluster genes by their binding profiles to all the TFs. I: Oct4, Sox2, Nanog, Smad1, STAT3; II: c-Myc, n-Myc; III: n-Myc, Klf4, Esrrb, Tcfcp2l1, Zfx, E2F1; IV: Suz12; V: none. Divide all genes into 3 groups: upregulated in ESC, downregulated in ESC, and not differentially expressed. Classes I and II are associated with up-regulated genes; class IV are associcated with down-regulated genes; class III are split between the two.

Extended TRN in mouse ESCs [Kim & Orkin, Cell, 2008]

- Methods:
  - Technology: bioChIP-chip for measure protein binding.
  - Sequenes: -8k to +2kb of promoter sequences.
  - Factors: (i) reprogramming factors: Oct4, Sox2, Klf4, c-Myc; (ii) other genes: Nanog, Dax1, Rex1, Zpf281, Nac1 (from [Wang06] paper of ESC interactome).

- Results:
  - Targets of the TFs: most BSs are close to TSS. Myc and Rex1 colocalize, while the other seven colocalize (Fig. 3E).
  - Regulation of expression of target genes: single targets are more likely to be inactive or down-regulatd; targets of multiple factors are more likely to be up-regulated. And the two types of targets show different GO enrichments.
  - Histone modification: cMyc targets show a different pattern. Distince function of cMyc in ESC: positive regulation of proliferation, negative regulation of differentiation and regulatory of chromosal accessibility of other factors.
  - TRN: Klf4 $\rightarrow$ Oct4, Sox2, Myc; Oct4, Sox2 $\rightarrow$ Nanog.

Transcription factor binding dynamics during human ES cell differentiation [Tsankov & Meissner, Nature, 2015]

- Background: cross-species comparison of OCT4 and NANOG targets showed that only 5% of regions are conserved and occupied across species.

- Motivation: define the roles of TFs during ES differentiation to three germ layers. The approaches include: change of TF expression, change of TF targets, and enrichment of TF binding in superenhancers in each cell types.
- Data: diff. of ES to an intermediate state (dMS) and then three germ layers (dEN, dME, dEC). Obtain histone marks, RNA-seq, Bis-seq and ChIP-seq of 38 TFs.
- TF binding dynamics: comparison of successive time points or between two layers. Static (similar targets), dynamic (different targets), enhanced (more) or suppressed (fewer). Example: NANOG and CTCF are static, GATA4 and SMAD4 very dynamic.
  - GATA4 and OTX2 binding not only divergent (different targets), but also change pattern/histone marks. In dME 36% of all GATA4 binding sites occur in promoters, compared to only 13.6% in dEN.
- TF clusters to understand the relation between TFs: defined by overlap of targets.
- Use super-enhancers (Extended H3K27Ac domains) to define lineage regulators:
  - In ES cells, core regulators OCT4, SOX2, NANOG (called OSN) and OTX2 binding is highly enriched at super-enhancers.
  - Endoderm: many of the core regulators bound at ES cell super-enhancers also occupy dEN super-enhancers, including OSN, OTX2, SMAD1, TCF4, and SMAD2/3.
  - Mesoderm: GATA4 and SMAD1 were the most highly enriched factors at dME super-enhancers.
  - Ectoderm: OTX2 enrichment at super-enhancers (known to regulate neuronal subtype specification in the midbrain)
- Use poised enhancers to find lineage regulators: in dEN, the H3K27ac domains were mostly devoid of known endoderm TFs, so test H3K4me1 domain. Found GATA4 binding is enriched.
- Questions:
  - Superenhancers during ES cell differentiation?
  - Valid of testing TF binding enrichment in superenhancers to study the TF role? Couterexample: GATA4 not enriched in H3K27ac defined enhancers in endoderm.
- Lesson: the general question is to define lineage regulators, the challenge is that the importance of TF (causal influence to differentiation) may not be clear from the observed data. It would be helpful to know what genes are important for a lineage. The approach would include a combination of: change of TF expression, change of targets, and enrichment at specific enhancers (super-enhancers or enhancers of known importance)

5.5.4 Occlusion Model

The "occlusis" model of cell fate restriction. [Lahn, BioEssays, 2011]

- Cell development: Yang - cell fate determination, whereby differentiating cells progressively acquire phenotypic identities of specialized cell types. Yin - cell fate restriction, whereby cells progressively lose their potential for all but the lineage to which they are committed. Goal of this work is the mechanism of cell fate restriction.
- Occlusis model:

  - Each gene exists in two states, competent or occluded. In the competent state, a gene can be activated when appropriate milieu is present; in the occluded state, the gene is irreversibly silenced (during normal development). Note that a competent gene may be silenced in a cell by lack of activator or presence of repressor.
  - Cell fate restriction during differentiation is achieved by progressive shift from competent to occluded states. Not all lineage-inappropriate genes need to be occluded however, for example, the occlusion of upstream regulators may be sufficient to block the expression of all downstream genes.
  - During reproduction, global deocclusion takes place, such that occluded genes throughout the genome are reset to the competent state.

- Support of occlusis model: X-inactivation, imprinting and results from cell fusion experiments.

- Occlusion vs. epigenetic silencing:

  - Epigenetic silencing focus on silenced genes: however, some of these genes may be turned off by trans- mechanism (lack of activator or presence of repressor). The occlusis model, on the other hand, distinguish genes silenced in trans- and in cis- (occlusion).
  - Epigenetic modification at CREs/genes: DNA methylation, histone modification, etc. are associated with gene silencing, but the causality is not established. It is possible that first occlusion is established, then epigenetic modification is used to mark the genes (e.g. for maintainence of the occluded state).

- Elaborations and predictions of the omdel:

  - It is possible that a gene is occluded in one milieu but competent in another.
  - Imprinting: the occluded state is maintained even during deocclusion; so it can be viewed as "super-occlusion".
  - Establishing and maintaining occluded state may involve different mechanisms: e.g. X-inactivation requires XIC (X-inact. center), but maintainence does not.
  - Deocclusion and reprogramming: cells have natural deocclusion ability (used during reproduction), so if touching this ability using artifically expressed master regualtors, we can reprogram cells. In contrast, trans-differentiation is generally difficult.
  - Deocclusion capacity should cease to operate at the onset of lineage differentiation. Epiblast stem cells (EpiSCs) are pluripotent stem cells, but it probably have lost its deocclusion capacity (X-inactivation has started). To test this, differentiated cells could be fused with EpiSCs to see if occluded genes would fail to be deoccluded in the fusion.

- Experimental study of occlusion:

  - Cell fusion experiment: to identify occluded genes, use one cell as reprogrammer cell, and the other as responder cell (to be studied). We look at genes expressed in reprogrammer cells, but silent in responder cells before fusion. After fusion, three scenarios: (1) active in reprogrammer and silent in responder - occluded genes; (2) active in both cells - activatable genes (silent in trans in responder cells before fusion); (3) silent in both cells - exclusion, no conclusion drawn.
  - Using mouse and rat cells (rather symmetric). The reason is that we need to know the origin of chromosome (so they should be genetically distinct). Using different strains of the same species: often too few variations in key developmental genes. Also can fuse different types of cells, e.g. ESC with differentiated cells.
  - Positive control: to demonstrate that an occlused gene can be expressed if cis-silencing mechanism is removed (not due to species incompatibility, etc.), introduce the gene in BAC and test its expression.

  - Testing progressive shift of occluded states: follow differentiation of cell types, the occlused states in the progenitor cells should be maintained in the progeny cells. Ex. use neural stem cell (NSC) $\rightarrow$ neuron system.

- Lessons:
  - Gene silencing can be caused by cis- or trans- mechanisms, and they are conceptually different. For instance, a trans-silenced gene can be activated in another physilogical condition. The cells need both to implement stable cell lineage, and the flexibilty of responding to different conditions.
  - Deocclusion: a natural capacity of cells, but only in germline/ESC. Lost early in development.

Analysis of the occlusis model: [personal notes]

- General goal: identify key linegage-determining events, both Yin and Yang, and use these events to explain the patterns of transcrtipome changes in reprogramming and cell fusion experiments; and explain the properties of reprogramming and trans-differentiation.
- Silenced genes: three types of silenced genes in a given cell type:
  - Occluded genes: cis-silenced.
  - Targets of occlused genes: not cis-silenced themselves, but because of the lack of activators (occluded), they are effectively "dead" in this cell type. These genes can be rescued if the appropriate activators are introduced in trans.
  - Inducible genes: with the correct conditions, they can be activated again.
- Analysis of fusion: we consider fusion between two different cell types A (responder) and B. Before fusion, in A, the lineage-determining factors of B is occluded, and some targets of the B-related factors are silenced, but not occluded. After fusion, in A, the B-factors are still silenced (occluded), but some targets of B factors will be expressed (activatable). General patterns:
  - Whether we can detect occluded genes in A depends on whether these genes are expressed in the reprogramming cells. For example, when fused A and B, we can detect occluded factors of B, but not another cell type C.
  - Normally, the expression of lineage-determining factors and their targets (all are tissue-specific genes) are coupled. But in fusion experiment, they may become decoupled because some of the target genes may be silenced in trans, and thus become activatable.
- Hypothesis/predictions of the model:
  - Experimental proof of occluded genes in a cell type: in the more differentiated progeny, or in a different condition (e.g. treatment of cytokine, hormone), the occluded cannot be activated.

Systematic mapping of occluded genes by cell fusion reveals prevalence and stability of cis-mediated silencing in somatic cells [Looney & Lahn, GR, 2014]

- Motivation: to ascertain an occluded gene in a cell type (mouse fibroblast, 129TF), we fuse the cell to another cell type where the gene is expressed. However, only a small subset of silent genes in 129TF is expressed in another cell type, so need to fuse 129TF with many other cell types (rat).
- Defining expression (or not) and gene status: the number of transcripts per diploid genome (TPG). We deemed a gene expressed if it produces ≥ 2 TPG, and silent if < 0.2 TPG.
  - "Occluded" genes in 129TF are defined as silent in 129TF both before and after fusion, and active in fusion partner cells.

  - "Activatable" genes in 129TF are defined as silent in 129TF before fusion but expressed after fusion, with the additional condition that the expression level from 129TF in fused cells is $> 10\%$ of the level from the fusion partner.

- Comprehensive mapping of occluded genes: The 12 fusions between 129TF and rat cells each uncovered 85-246 occluded genes in 129TF (average 155), while the numbers of activatable and extinguished genes in 129TF are lower. In total, 730 are occluded, 346 activatable, and 850 extinguished in at least one of the rat cell types.
  - Estimation of the total number of occluded genes: 739 occluded genes in a total of 1800 informative silent genes in 129TF cells (6171 silent genes in total), so the total number of occluded genes is about 2000 - 3000.

- Regulatory and functional features of occluded genes:
  - Promoters: occluded genes are much more likely to have CpG island promoters (associated with housekeeping genes) than TATA box promoters (associated with tissue-specific genes), whereas activatable genes are almost equally likely to have these two types of promoters.
  - GO analysis: occluded genes enriched for regulatory functions, and include many key factors controlling important cellular and developmental processes. Activatable genes are enriched for immune system functions, transport, response to stimulus, and various metabolic processes.

- Robustness of fusion and cell identify: correlation of transcriptome of 129TF cells before and after fusion is higher than the correlation of both cells in the fused cell. The results are robust to fusion ratio, post-fusion culture time, DNA replication, and whether fused cells are analyzed at the population level or clone level.

- Chromatin patterns in 129TF cells (regions from 2 kb upstream of the TSS):
  - Histone marks: the two silenced categories (occluded and activatable) show distinct patterns from expressed genes. Occluded and activatable genes are somewhat different, but not much.
  - DNA methylation: generally higher in silenced genes, as expected. The mCG density around TSS for occluded genes is about twice that for activatable genes, which is partly due to occluded genes having greater CG density around TSS and and partly due to occluded genes having greater mCG density per CG in said region.

- Combined signatures using multiple chromatin marks and DNA methylation: (using fraction of genes positive for a certain mark)
  - Activatable genes are significantly more likely to be positive for active marks and hypomethylation as compared with occluded genes, whereas occluded genes are significantly more likely to be hypermethylated (about 50%, comparing with 20-25% in activatable genes).
  - A subset of occluded genes is highly enriched around TSS for H3K9me3 and H3K27me3 along with DNA hypermethylation. For activatable genes, a subset shows enrichment for multiple active marks and depletion for silent marks - likely transcriptionally poised genes.
  - Using machine learning algorithm to classify the two groups of genes, achieving high accuracy.

- Role of DNA methylation in occlusion: treat the cells with DNA methyltransferase inhibitor: the great majority of occluded genes were unaffected while a small minority were noticeably affected, with 11% of the genes being activated to levels over the ≥ 2 TPG threshold. Furthermore, occluded genes activated by AdC persisted in the active state even 1 wk after AdC removal.

- Role of histone deacetylation: histone deacetylation plays a role in implementing the silencing of at least a subset of occluded genes (similar to DNA methylation), but it exerts a highly transient effect (genes remain silenced once the drug is removed) and therefore appears unlikely to be involved in the memory of occlusion. Furthermore, DNMT inhibitor and HDAC inhibitor target different sets of genes.

5.6 Aging

Aging Research?Where Do We Stand and Where Are We Going? [Guarente, Cell, 2014]

- Aging at the cellular level: many cellular defects including
  - DNA damage in the nucleus and mitochondria.
  - Mitochondrial dysfunction leading to increased production of reactive oxidative species (ROS) and decreased production of ATP.
  - Oxidative damage to proteins and other macromolecules.
  - Protein misfolding and aggregation,
  - Protein glycation.
  - Induction of proinflammatory cytokines.
  - Telomere shortening.
  - Cell senescense.

  In addition to tissue-autonomous aging, also brain helps govern aging of many organs.
- Sirtuin pathway: NDA+-dependent protein deacetylase. Mediates some of the phenotypes of calorie restrictino (CR). The role of NAD^+:
  - NAD+/NADH ratio activate sirtuins and help drive effects of CR.
  - NDA+ level declines with aging. Possible reason: DNA damage leads to activation of one DNA repair enzyme, which consumes NAD+. May be also related to circadian clock.
- MAPK and TOR pathway: when energy is limiting, TOR is downregulated and MAPK is activated, which activates stress-response pathways, stimulates mitophagy to improve mitochondrial quality.
- Other approaches to studying aging: parabiosis, genetics of centenarians, genomics study of aging:
  - "we may have but scratchedthe surface of what bioinformatics canprovide in identifying new genes andpathways important in human aging, aswell as allowing for the knowledge wehave already gained to be applied in amore effective, personalized way".
  - "bioinformatics will play a substantial role in the progress of aging research".
  - Transcriptome, epigenomic, proteome anaysis of individuals spanning a wide age range.

  Model of aging (testing genes and pathways) using iPSC.

Identification and Application of Gene Expression Signatures Associated with Lifespan Extension [Cell Metabolism, 2019]

- Background: life-span extension interventions include mutations, growth hormone receptor KO, CR, methionine restriction (MR). ITP study: use genetically heterogenous mouse population.
- Background: effects often vary with sex.
- Experiment: liver expression data of young mice, 8 interventions. Note: use young mice data because expression changes in old mice are consequence of life-span extension.
- Comparison with DE gene lists: ribosome and drug metabolism common for all interventions. Some are specific, TCA cycle, ox. phosphorylation, fatty acid oxidation.
- Feminizing effects of interventions: observed for some, however, they do not explain the effect on life-span (e.g. those having large feminizing effects have no effect on life-span).

- Identifying common signatures: fix LMM, treating each intervention as random effects. However, few genes found, and relax the assumptions.
- Lesson: examples of transcriptome signatures of a phenotype (1) Different ways of leading to a phenotype (life-span): via different interventions. (2) Usually some pathways are shared, some unique. (3) The effects may depend on environment (gender).

A meta-analysis of genome-wide association studies identifies multiple longevity genes [NC, 2019]

- GWAS: 10K/3K cases, meta-analysis. Associations: APOE, two variants associated with longer or shorter lifespan. GPR78: lower lifespan. Previously reported associations: only FOXO and CDKN2A/B were replicated.
- TWAS: 14 genes, 8 are near APOE4.
- Genetic correlation: negative with CAD, T2D, insulin. Positive with year of education, paternal/mother age of death, age of first birth.
- Remark: genetic correlation may be hidden by several factors; also if sharing fraction is small, may not detect significant correlation.

5.7 Pregnancy and Birth

Organs, tissues and cells important for pregnancy:

- Major parts of female reproductive system: uterus, ovary, uterine tube, endometrium, myometrium, cervix, vagina.
- Endometrium: the inner epithelial layer, along with its mucous membrane, of the mammalian uterus. Consists of epithelia and stromal cells.
- Myometrium: middle layer of the uterine wall, consisting mainly of uterine smooth muscle cells. Its main function is to induce uterine contractions. The myometrium stretches during pregnancy to allow for the uterus to become several times its non-gravid size, and contracts in a coordinated fashion during the process of labor.
- Endometrium stromal cells: connective tissue cells of any organ, for example in the uterine mucosa (endometrium). Fibroblasts and pericytes are among the most common types of stromal cells. Undifferentiated fibroblast, upon progestone, becomes stromal cells. They have immune functions, communicating with immune cells.
- Cells in the placental membrane: (1) Maternal placenta (decidua): stromal cells and many immune cells (CD4 and CD8 T-cells). (2) Fetal placenta (chorion): some immune cells and trophoblast (first differentiated cells from fertilized eggs, providing nutrients to embryo). In experiments, sorting of maternal placental membrane to obtain four main cell types and then do genomic experiments.
- Trophoblast: multiple functions, immune, detoxification. Distinct from most cells. TCM: medium from cultured trophoblasts. Molecules secreted by these cells.

Physiological changes during pregnancy

- Growing placenta. Decidualization of endometrium: response of maternal cells to the hormone progesterone. the glands and blood vessels in the endometrium further increase in size and number. Vascular spaces fuse and become interconnected, forming the placenta.
- Expansion of uterus: block myometrium contraction

- Inhibition of maternal immune response: immune tolerance. Labor: can be understood as hyperinflammation.
- Metabolic adaption of mothers.

Model of immune system in pregnancy and labor [personal notes]

- Principle: immune system has distinct roles in pregnancy: (1) Adaptive immunity: create immune tolerance. (2) Innate immunity: implantation, and labor/parturition.
- When pregnancy starts: (1) Inflammation is triggered, which helps implantation. This process is likely mediated by SASP. Meanwhile, senescence cells are cleared up by NK cells, creating a balance. (2) When immune cells move to placental (maternal), they also change to create immune tolerance. Possible mechanism: PDL-1 subpopulation expanded.
- After labor, PDL1 subpopulation is also induced: likely explanation, need PDL1 to shut down hyperinflammation which triggers labor.

Hormones during pregnancy

- Progesterone is a potent agonist of the nuclear progesterone receptor (nPR). Progesterone binds to and behaves as a partial agonist of the glucocorticoid receptor (GR), albeit with very low potency (EC50 >100-fold less relative to cortisol). Mutual antognization between PGR and GR.
- Progesterone has a number of physiological effects that are amplified in the presence of estrogens. Estrogens through estrogen receptors (ERs) induce or upregulate the expression of the PR.
- Progesterone has many roles relating to the development of the fetus:
  - Progesterone converts the endometrium to its secretory stage to prepare the uterus for implantation
  - During implantation and gestation, progesterone appears to decrease the maternal immune response to allow for the acceptance of the pregnancy. More generally, provide support of the endometrium to provide an environment conducive to fetal survival.
  - Progesterone decreases contractility of the uterine smooth muscle. This is often called the "progesterone block" on the myometrium. Toward the end of gestation, this myometrial-quieting effect is antagonized by rising levels of estrogens, thereby facilitating parturition.
  - Progesterone inhibits lactation during pregnancy. The fall in progesterone levels following delivery is one of the triggers for milk production.
  - A drop in progesterone levels is possibly one step that facilitates the onset of labor
- With few exceptions, the concentration of estrogens in maternal blood rises to maximal toward the end of gestation.
- Two of the principle effects of placental estrogens are:
  - Stimulate growth of the myometrium and antagonize the myometrial-suppressing activity of progesterone. In many species, the high levels of estrogen in late gestation induces myometrial oxytocin receptors, thereby preparing the uterus for parturition.
  - Stimulate mammary gland development. Estrogens are one in a battery of hormones necessary for both ductal and alveolar growth in the mammary gland.

Evolution of pregnancy in mammals [MOD]:

- Evolution of pregnancy: (1) Maternal provisioning (Matroptropy) evolve first (2)l ive bith (viviparity) in marsupial, loss of eggshell and yolk. (3) Placental: in eutherian mammals, with maternal immutolrance, decidualization, etc. This enables prolonged pregnancy. (4) Primates evolve invasive placentas, spotaneous decidaulization, and an divergent parturition signal.
- For most mammals, laboring is triggered by progestone drop. Progestone recept (PR) has two isoforms: switch of one isoform to the other leads to laboring.
- Old world monkeys and human: very invasive placenta. Also progestone level does not change in laboring. However, sign of positive selection (large number of AA changes in activation domain) in one PR isoform.

Finding pregnancy-related genes through comparative genomics [Vinny Lynch, 2016]

- Transcriptome in pregnant endometrium in multiple mammals: pregnant endometrium cells. Classifying expression of a gene as 0 or 1 using a mixture model. Then find the genes that show changes in old-monkey and primates.
- Gained genes in eutherian mammals: enriched with absent NK cells, abnormal uterin artery morphology, allograft genes (mouse KO phenotypes). IL15, CSF1 promote traophoblast (mediate recuritment of NK and MP). PD-1 ligands, etc. that suppress production of inflammation of cytokines, and recruitment of regulatory T cells.
- Lost genes: neurological functions, behavior defects. Explanation: genes are enriched with iron transporters/channels: shell formation.
- Progesterone induces anti-infmamatory cytokines. Also expression of TSG in relevant cells. Connecton with cancer: invasiveness may correlate with tumor.
- Gained genes in primates: pro-inflammatory cytokines.
- Human-specific genes: GR process from Reactome. Many genes involved in corticortiod biosynthesis, progestone biosynthesis and CRH.
- Remark: the genes found in this way may be related to adaption, which is different from the PTB phenotype. Ex. some may be related to change of pelvis size and bipedalism.

PRPLE study (Project for Preterm Labor Epegenetics) [Carole Ober, 2016]

- Experiment: from the placenta membrane after birth, separate cells by flow cytometry, then do genomic profiling. Separation into maternal (decidua) and fetal (chorion).
- Characterizing cell types: more T-cells and macrophage in maternal and more CD8-T cells and mononcytes in fetal.

B cells in pregnancy and PTB [Kang Chen, 06/2017]

- Choriodecidua of women undergoing spontaneous PTL harbored functionally altered B cell population
- Inflammation can trigger PTB. Treat mice with LPS: in WT mice, normal. But in B-cell deficient mice, PTB; and neutrophil infiltration.
- The B-cell protection mechanism is independent of IL10, TGF-beta or IL35.
- Possible mechanism: B-cell deficient mice have normal levels of circulating progestone after LPS, but have reduced PIBF1 (multiple isoforms, some in nucleus, and the secretable form is a cytokine) in the uterus.

- PIBF1 protects against inflammation-induced PTB: theraputetic PIBF1 protects against LPS-induced PTB in B-deficient mice.
- PIBF1 expression is induced by IL-33: promoter analysis.
- PTB patients have diminished expression of IL33 receptor.
- B cells have function not only in Ab production, but also maintenance of tissue homeostatsis (response to stress signal such as IL33).
- Model: uterine stress (e.g. infection, inflammation) $\rightarrow$ IL33 $\rightarrow$ PIBF1 $\rightarrow$ IL4Ralpha, which affect PTB.

GWAS of PTB [Lou, 07/2017]

- Data information: discovery cohort from GWAS. Other cohorts: National Birth Cohort: 4000 samples (mother/kid). MoBa cohort. FIN: Finish data. 2000 pairs (available).
- Genetic evidence of PTB: the biggest risk factors of PTB include previous PTB, sister having PTB. Classical genetics: 30-40
- 6 loci associated with gestational length and PTB: EBF1, WNT4, EEFSEC, AGTR2. 23andme study (European). No robust association in fetal GWAS.
- EBF1: (early B-cell factor 1) adipoctye differentiation and development, B-cell lineage activation and maintanence. Stress would modify EBF1.
- EEFSEC: Selenium metabolism. Reduced Selenium is associated with PTB risk (e.g. African country with highest PTB rate, also lowest Selenium). Involved in making Selenium proteins, which are important for antioxidative response.
- WNT4: development of female reproductive system, and control decidualization in endometrium. Protective for PTB, but risk for endometriosis.
- AGTR2: angiotensin 2 receptor. Modulating uteroplacentral circulation.
- ADCY5: membrane-bound adenylyl cyclase. Cell energy, metabolism. Associated with birth weight.
- WNT4 locus: variant affect ESR1 binding (estrogen receptor). Minor allele (protective) binds ESR1 more strongly.
- Capture Hi-C analysis: EBF1 locus interaction with EBF1 in LCL. EEFSEC locus interaction with GATA2 (both heart and LCL).
- New GWAS datasets: Avon data, Finish and other dbGaP data. Total more than 20K samples. Adding 9000 more samples: four new loci.
- Haplotype score analysis: consider related traits, derive their polygenic scores and study their relation with PTB. Blood glucose scores effect on birth weight and gestational age. Compare maternal transmitted, non-transmitted and paternal transmitted. Some interactions here: allele from father and mother have different effects?
- Hazardous risk analysis: see when the protective/risk effects occur. Ex. some risk factor (SNP?) only increases the risk of very early PTB, but not later.
- WES in families with recurrent PTB. Northern Finish population: good for genetics (less environmental effects).

- Top genes: GR, ER signaling, AMPK signaling. AR, HSPA1L (Hsp90). HSPA1L: new phyosphorylation site. Same rare variant associated with IBD, reducing chaperon activity.

Placental transcriptome co-expression analysis reveals conserved regulatory programs across gestation [Buckberry, BMC Genomics, 2017]

- RNA-seq on 16 samples. First-time mothers, placental villlia tissues (mostly fetal tissue).
- WGCNA: most variable genes. 13 highly connected modules. A gene can be included in multiple modules.
- Eigengene: first PC of a module and represent the weighted average of gene expression. Eigengene explains 40-80% of variance. For each module, define hub genes (highly correlated with eigen-gene).
- M10 is enriched for Y chr. genes, and significantly different between males and females. M3 correlated with fetal birth weight.
- Modules are conserved in mouse, E11.5. Five modules are conserved. Strongest is M3.
- ZNF423 and EBF1 are important for module 3: motif analysis in 10kb regions. EBF1 and ZNF423 often together: motifs similar. For both TFs: expression highly correlated with eigengene of M3.
- Enrichment of PTB and PE (preeclampsia) genes. PTB gene list from a website (gene expression changes). M12 are significantly up-regulated in PE.
- **Lessons**: WGCNA essentially a clustering algorithm, but also create eigengenes, which can be correlated with other covariates (e.g. sample characteristics). To find important modules: enrichment with known phenotypic genes. Identifying important regulators of key modules.

Genomic characterization of pregnancy-related cells and their hormone response [MOD Theme 1, Oct, 2017]

- RNA-seq experiment: Placental membrane cell extraction: decidua (maternal) and chorion (fetal) separation. RBC lysis treatment. Obtain 5-6 primary decidual stromal cell lines: from term membrane. dTL1, etc. (decidual, term-labor)
- Comparison of cultured vs. ex vivo cells: composition of decidual stromal cells (DSC) vs. decidual mesochymal stem cells (DMSC). In cultured cells, loss of DSCs. RNA-seq in sorted cells (uncultured or ex vivo) similar to cultured cell lines: so cultured cells a good representation.
- Primary decidual stromal cells: retain the ability to decidualize upon progestone treatment? Decidualization by cAMP confirmed. DE genes: GO analysis not clear, but a number of genes involved in decidualization, e.g. FOXO1, STAT-X.
- Effect of TCM: TCM from 3 placental membranes (Cytotrophobast cells). After treatment: 300 DEGs in 6h, 1200 DEGs in 48 hours. TCM Promotes expression of immune genes: very different from cAMP/PG treatment, e.g. immune regulation, wound healing.
- Epigenomic experiments: assess reproducibility between sites (ATAC-seq, chromatin marks): PCA plot, show that samples are clustered by biology, rather than by sites.
- DNase-seq: MA plot (log2-fold change vs. vs intensity). Among 100K peaks, 6K open and 1.6K closed upon decidualization.
- ChIP-seq of H3K4me1/3 and H3K27ac. Only 1K show promoter changes upon decidualization. 12K show H3K27ac changes.

- Correlation of gene expression changes: in H3K4me3, 25 fold enrichment of DEGs in the same direction. H3K27ac: 5 fold enrichment of nearest genes.
- Motif analysis: PGR, Foxo1, Fosl2 motifs. Q: only in DE peaks?
- Hi-C experiment: interaction distance median is 170kb, of 200K interactions. Only 10% in nearest gene, 10% each skip 1,2,3,4 genes, and 50% more than 4 genes. Hi-C: GR induction, then measure Hi-C changes and transcriptome changes. Only 5% or so interactions change strength.
- Tissue-specific expression level correlates with number of interactions (in promoters) in Hi-C experiment.
- STARR-seq: for enhancers (defined by chromain marks), about 10% show activity in STARR-seq.
- GWAS: top SNP near both EBF1 and CLINT1 (about 500Kb each), found interactions with both. Another SNP: many interactions with GATA2 instead of the Selenium gene.
- Bioinformatics pipeline: (1) Use DESeq2, first do log-transformation. Control for batch effect. (2) Differential peak analysis: count reads in merged peaks in each sample. Then do RNA-seq pipeline. (3) Downstream analysis: HOMER to find motifs. (4) Hi-C: PE reads. Call interactions using CHICAGO.
- Functional QC of bioinformatic results: correlation of interactions with expression, motif search, correlation betwen number of enhancers and expression.
- **Lesson**: (1) Obtaining a model of the system of interest: cell extraction and sorting. Note: sorting and the related steps may damage cells. (2) Evaluating if the model faithfully represents the system: using scRNA-seq to characterize the cells, comparison of expression profiles with ex vivo results, whether retaining the ability to respond to stimuli.
- **Lesson**: reproducibility of experiments important. A related issue is whether biological variations (e.g. individual variations) overwhelm the true variation of interest (e.g. response to treatment).
- Lesson: chromatin interactions likely do not change significantly during hormone response.