Proposal for a Center for the determination of the

Causal Transcriptional Consequences of Human Genetic Variation (CTCHGV)

submitted
May 25, 2009
by

Professor George M. Church
Department of Genetics
Harvard Medical School

in response to Program Announcement Number
PAR-08-094

2. Specific Aims

The goal of our proposed Center for the Causal Transcriptional Consequences of Human Genetic Variation (CTCHGV) is to develop methods that will identify and characterize cause-effect relationships between human genome sequence variation and transcriptional networks, with specific focus on cis transcription. Recent genome wide association studies (GWAS) involving many human cohorts have improved our knowledge of human genetic variation and its relationship to human physiology and disease. Yet these developments are only early steps towards the detailed causal understanding of how genetic variation relates to phenotype needed to translate this knowledge to effective clinical practice. This is particularly so for variation in non protein coding regions which comprise 99% of the genome and which obey few known rules. As of this writing, 315 GWAS studies have uncovered 1439 SNPs that associate with ~200 human traits with p<1e-5 (59), 95% of which are in non coding regions. Most of these are tag SNPs in linkage disequilibrium with possibly causative SNPs as yet unidentified or even assayed in their subjects. The tag SNPs are typically common variations and the contribution to human health and disease of common vs. rarer variations is still debated. Meanwhile, ongoing sequencing of diverse populations and the growing number of sequenced individual human genomes yield an ever-increasing number of previously unseen and rare point, indel, and rearrangment variations. To move from association to cause in a manner that is not complicated by variation rarity, population sampling, and sequencing depth, CTCHGV will develop and demonstrate innovative techniques that establish cis variants' causal status by systematically and precisely varying cis sequences at single-nucleotide resolution using synthetic biology techniques, so that the effects of these variations on cis gene transcriptional level can be observed directly (Aim 1). Data generated by these methods will directly explain many GWAS findings of associations between cis variations and expression levels (40, 132, 165), and will enable refinement of hypotheses for disease causation where GWAS finds associations between cis regulatory loci and disease or phenotype. Moreover, to assist such refinement, CTCHGV will extend application of these new methods to human induced Pluripotent Stem cells (iPS) in order to make its methods able to explore the impact of cis variations in diverse human cell types representing different tissues (Aim 2). To achieve scalability in its methods for discerning sequence causality by systematically examining combinations of variations, CTCHGV will develop methods that operate with small samples of cells, including methods that assay many individual cells. Here, to determine causal cis variants requires only that the transcription of one gene—the cis gene—be assayed in small samples and in single cells. To extend beyond this and observe the systematic effects of variations, CTCHGV will therefore also develop new methods for obtaining transcriptome level information in single human cells, including both dispersed cells and in-situ structured tissues (Aim 3). Finally, CTCHGV will develop a number of innovative basic enabling technologies to achieve the scale and control over DNA synthesis and cell handling required to meet the goals above (Aim 4). As these technologies will have great impact and wide utility in biological research, CTCHGV will develop them with general usage in mind, in an open-source manner, and in collaboration with our many academic and business partners.

Aim 1: We will develop and demonstrate novel methods that identify and characterize natural cis variations that directly affect transcriptional activity in individual humans based on direct modification and testing of combinations of variants in gene regulatory regions in cell lines, and that can be applied to thousands of genes.

1.1: We will develop and demonstrate novel, high-efficiency methods to create human cell populations containing combinations of natural variations in gene regulatory regions, focusing on zinc-finger nuclease (ZFN)-mediated recombination of externally generated altered insert libraries, and direct modification of human cells using oligo-based methods.

1.2: We will demonstrate the identification of specific sets of variations that affect cis gene transcription by engineering many combinations of variations and directly observing their effects on transcription, and also by novel methods of assaying complex populations of combinatorially modified cells at a single-cell level.

1.3: We will assess the extent to which cis variants identified as causing altered transcript expression may operate through alternative mechanisms such as differential expression of RNA isoforms, differential transcript degradation, copy number variations, and epistatic marks.

1.4: We will analyze the relationship between our methods and results and those of Genome Wide Association Studies and characterize their complementary insights into the effects of variation.

Aim 2: We will adapt and extend Aim 1 methods to function in human induced Pluripotent Stem cells (iPS) and then use iPS to characterize the effect of cis regulatory region variations in a variety of derived cell types that represent different human tissues. We will engineer "marked allele" human iPS that are heterozygous in all exons of many genes that will enable analysis of allele-specific transcriptional and splicing effects in diverse cell types.
2.1: We will combine Aim 1 methods with automated techniques for iPS generation and maintenance to enable exploration of iPS with altered cis regulatory regions.
2.2: We will differentiate iPS generated in Aim 2.1 into diverse cell types that represent distinct human tissues and characterize the cell type-specific consequences of cis-regulatory variations.
2.3: We will engineer human iPS with "marked alleles" for 10-50 genes and demonstrate their use by characterizing allele-specific transcription and splicing in multiple tissues.

Aim 3: We will develop novel single-cell in-depth transcriptome assays scalable to millions of individual cells simultaneously in both structured tissues and dispersed cell samples, subject to sequencing capacity. These methods will be used to explore systematic transcriptional effects of genetic variations in different human cell types.
3.1: We will develop and optimize methods that pipeline in-situ single-cell cDNA synthesis to next generation sequencing in ways that preserve cell identity and that can be applied in parallel to 100s to 1000s of cells. We will investigate multiple techniques in support of these methods, including cell bar-coding, in-situ cell sequencing, and single-molecule in-cell sequencing, characterize their performance and limits, and select one for continued development and application.
3.2: We will use these single cell transcriptomics capabilities to characterize the transcriptional state differences in cells bearing artificial and natural variant combinations from Aim 1, and from cell types developed from iPS from different genetic backgrounds.

Aim 4: In support of Aims 1-3, we will develop innovative and widely applicable methods for high-throughput synthesis of long DNA constructs, highly efficient homologous recombination in human cells, and highly multiplexed single cell handling that enables sorting based on morphology.
4.1: We will develop a platform that integrates DNA synthesis and sequencing and uses sequence information to assure synthesis of DNA constructs with extremely low error rates.
4.2: We will improve ZFN–mediated homologous recombination in human cells by engineering a comprehensive zinc-finger archive, by developing novel methods of delivering ZFNs into cells, and by developing a "segmental genome replacement" strategy.
4.3: We will develop new high-throughput cell handling and sorting capabilities that can incorporate morphology information in addition to optical signals generated by markers, and which can operate on live cells.

Beyond the five years of our Center, we foresee the innovations we develop being applied at large scale by partner academic research centers such as the Broad Institute, as well as their adoption and further development by sequencing and synthesis companies with which we have close relationships (see Data and Materials Dissemination Plan). Our approaches will help biomedical research in general move beyond population-based associations to causal understanding. Their application to individual humans vs. populations will be critical for developing the knowledgebase required to promote and evaluate the effectiveness of personalized medicine.

Our world-class team has expertise in all the areas required for success in this project and has a track record of impact and innovation. Professor George Church (Harvard Medical School), CTCHGV's proposed director, has several times developed innovations that exhibited improvement factors of 10 or more in scale or power compared to contemporaneous commercial collaborators. Indeed, Professor Church led a prior Molecular Genomics and Imaging CEGS (MGIC) that consistently developed improved sequencing methods ~2 years ahead of commercial efforts which later adapted many of our innovations: Under him, MGIC demonstrated his initial polymerase colony (polony) methods in 2003 (116, 117), versions of which are now widely used commercially (Illumina, ABI), while in 2005 MGIC developed sequencing by ligation (155), which is now in use in ABI SOLiD. Another example is in DNA synthesis, where he has led the way in synthesis and use of complex oligo mixtures cleaved from arrays for large construct assembly and targeted sequencing, and where in the course of four years he has advanced from 4000 90-mer to 54000 150-mer oligo arrays (100,

174). Dr. J. Keith Joung (Harvard Medical School, Massachusetts General Hospital) is a leading expert on the development of zinc-finger nucleases for human cell engineering and gene targeting. He is the leader and co-founder of the Zinc Finger Consortium (http://www.zincfingers.org/), which was established to ensure and to promote continued research and development of engineered zinc finger technology. The Consortium is committed to developing a zinc finger engineering platform that is robust, user-friendly, and freely available to the academic scientific community. Dr. George Q. Daley's (Harvard Medical School, Children's Hospital, HHMI) work has transformed the field of stem cell development and differentiation. Recipient of numerous awards, including the first NIH Director's Pioneer Award, as well as major awards from the American Philosophical Society, Society for Pediatric Research, Burroughs Wellcome Fund, and the Leukemia and Lymphoma Society of America, Dr. Daley's work focuses on functional hematopoietic and germ cell elements from ES cells, and the genetic mechanisms that predispose to malignancy. Dr. Daley's lab was one of the first three world-wide to derive human iPS cells, and the first to produce a repository of patient-specific iPS cells (from 10 different disease conditions). Professor Kun Zhang (UCSD) developed innovative methods for long range haplotyping, single cell genome sequencing, targeted sequencing, and measurement of allele-specific expression, as a post-doc in the Church Lab, where he was a member of the MGIC team. He is currently working with Professor Church on methods for targeted exon sequencing in connection with an NHLBI grant (HLB08-004). In addition to these key personnel, CTCHGV will have additional support from experts in GWAS, genome wide screens in human cells, and companies offering sequencing support (see Letters of Support from David Altshuler, Steven McCarroll, Robert Plenge, Steven Elledge, and Complete Genomics, among others).

3. Background and Significance

3.1 Determining the causal consequences of natural human genetic variations: Rapid developments over the past 8 years led quickly from the completion of draft human genomes (97, 180) to the identification of common human genetic variations (146) and haplotypes (68), and then to the use of these variations to identify loci associated with human traits and disorders through Genome Wide Association Studies (GWAS) (188). GWAS, conceptualized early in the Human Genome Project (144), has become the leading method for linking genetic variation with human traits: As of this writing, 315 GWAS studies have uncovered 1439 SNPs that associate with ~200 human traits with p<1e-5 (59), 95% of which are in non coding regions. By their nature, GWAS do not identify the causal mechanisms linking variations with traits but only the associated variations themselves, but these may provide important clues about the causal networks underlying the traits. However, limitations of GWAS are frequently noted (4): The identified variations may not themselves be in the causal network for a trait but may only be tag SNPs in linkage disequilibrium with causative variations, and in practice GWAS can only be used to detect common SNPs with moderate effect sizes that can be cost-effectively genotyped in some few thousands of individuals. As a result, efforts are underway to make rarer mutations analyzable by GWAS (4, 67, 126), and to expand the GWAS paradigm to identify variations specifically associated with gene expression level (notably *via* expression Quantitative Trait Loci (eQTL) (23, 32, 40, 46, 165) and allele-specific expression (ASE) (152)) and to investigate the effects of Copy Number Variations (111, 164). These advances will refine our understanding of the causative networks underlying traits, but in the end they may still fall short of identifying the specific variations that are causal. This is because GWAS' power to dissect co-occurring human variations and genetic and environmental backgrounds is ultimately constrained by the spectra of natural human variation, population structure, and linkage disequilibrium that have been shaped by human evolutionary history. CTCHGV proposes to develop and demonstrate technologies that will enable such genetic covariates to be dissected in the specific domain of variations that affect cis gene expression levels. These technologies will be based on direct engineering of combinatorial modifications to cis regulatory region genotypes, followed by direct observation of the effects of these modifications on cis gene transcription levels. While these methods will not dissect direct links between variations and disease, they will help refine hypotheses concerning disease causation where GWAS finds associations between disease and regulatory variations. To assist this analysis, CTCHGV will extend its technologies to human induced Pluripotent Stem cells (iPS) so that the causal impacts of regulatory variation can be examined in diverse human cell types representing different tissues. In this way, CTCHGV will provide an essential complement to GWAS: While GWAS take us from phenotype to associated locus by deeply leveraging natural human variation, CTCHGV methods will dissect the roles of variations that are below its limen.

The methods that will be developed by CTCHGV in its five years of operation will specifically focus on the characterizing the effects of cis variations on transcriptional level (Aim 1) and some of their downstream consequences (Aim 3). Eventually, follow-on work will require methods that analyze the broad range of other molecular mechanisms, including effects of variations on RNA splicing, protein expression level and translational control, as well as the effects of epigenetic variation and imprinting. Here Aim 1.3 will assess a selection of these effects in specific contexts in an effort to quantify their importance and chart a path forward.

3.2 Single cell transcriptomics: Acquiring detailed information about the transcriptional state of individual cells is critical for understanding the development of complex organisms and the functions of structured, differentiated tissues. The leading methods available today involve the isolation of individual cells by disaggregation/dilution or microdissection, extraction of cell contents by lysing or micropipetting, followed by synthesis and subsequent amplification of cDNA to levels at which it can be assayed via RT-PCR or microarray (44, 83, 94). Though these methods have been applied successfully in numerous studies, they are subject to important limitations: (i) Cell isolation procedures have limited scalability and, where tissue cells must be disaggregated and diluted, may destroy structural information about a cell's location in a tissue. (ii) The need for very high pg-to-μg amplification of single cell mRNA increases the potential for introducing biases into sample transcript abundances. To enable transcriptomes to be obtained for large numbers of individual cells will require addressing both of these limitations. In common with most technology development aimed at increasing throughput, (i) is best addressed by miniaturization and parallelization of operations on single cells. Microarrays require large amounts of starting material and so exacerbate problem (ii). Researchers are increasingly turning to next-generation sequencing to supplant microarrays generally, resulting in such methods as RNA-Seq and PMAGE (84, 185), and this has recently been applied to single cells (169). While this will partly alleviate (ii), application of these methods in parallel to many single cells as required by (i) will still require further advances in sequencing technology.

The ideal solution would be one in which cell transcriptomes could be sequenced *in-situ*; this would confine and parallelize sequencing within single cell compartments, and also allow structured tissues to remain intact so that cell locations and relationships are preserved. The challenge to sequencing technology is that, at present, most next-generation sequencing methods must create and operate with spatially distinct, compactly localized amplicons of sample sequences dispersed on planar surfaces (154, 156); thus, new methods would be needed to create and interrogate these amplicons in existing cell volumes. (Note that the localized amplification needed by sequencers is distinct from the high gain in-solution amplification of cDNA required by microarrays; for multiplex single-cell sequencing, the latter should be avoided as it not only increases sequencing costs but may also re-introduce amplification biases.) However, possibilities exist (a) for creation of localized amplicons in stacks of very thin sections of cells that can be sequenced *in-situ*, and (b) for transcripts of individual cells to be labeled *in-situ* in a way that preserves cell identity so that these transcripts could be dispersed, locally amplified, and sequenced *ex-situ* as usual and re-assigned to their cell of origin. Also (c) single molecule sequencing methods (45, 55) can in principle avoid the need for local amplification, but interrogating individual transcripts in existing cell volumes is still beyond the reach of these methods as they are configured today, especially for (45). Aim 3 will explore these possibilities. Aim 3 work on (a) will greatly scale up prior methods developed in our own and other labs on *in-situ* localized amplification and detection of RNAs (163, 202) in the direction of detecting and quantifying many thousands of RNA species in cells. Assuming cells of ~10μm, a 1 mm^2 section of tissue that is one cell thick would contain 1e4 cells. If each cell is assumed to contain ~2e5 coding transcripts, sequencing 50bp tags of all coding transcripts in these cells would require 2e9 reads and 1e11 bp. Illumina has recently announced that its Genome Analyzer will be able to produce ~1e11 bp of sequence per run by the end of 2009 (66). As recent trends of increasing sequencing throughput per dollar are expected to continue, CTCHGV does not see sequencing capacity and cost as inherent limitations to the ability to sequence transcriptomes of all cells in a tissue sample of this size by the end of the five years of our proposed center. However, as the availability of sequencing capacity on this order is likely to be a practical consideration for many research projects for some time, CTCHGV will develop options for assaying of targeted transcriptome subsets vs. complete transcriptomes in single cells that will give researchers the ability to allocate available sequencing capacity as best fits their needs.

3.3 induced Pluripotent Stem Cells (iPS): Full understanding of the effects of natural variations in humans requires exploration of their impacts in different tissues. Expression Quantitative Trait Loci (eQTL) found by observing quantitative differences in gene expression in multiple individuals are numerous and highly heritable (40, 119, 150, 182), but these examine only limited numbers of tissue types, and it requires a large

number of tissue samples to reach adequate statistical power due to their generally weak individual effect in addition to measurement noise and other confounding factors (32, 78). Although eQTL mapped from different tissue types overlap (23, 24, 46, 121, 149), many regulatory pathways are known to be tissue- and cell type-specific. To address these limitations, the Genotype-Tissue Expression Project (125) has been launched to collect various tissue types from a large number of subjects. However, collection of diverse human tissues using surgical and tumor specimens is complex and affected by sampling and processing artifacts (32), and these issues must be overcome for a very large number samples to detect the weak effects of most eQTL.

By contrast, iPS technology (133, 167, 168, 195) allows biomedical researchers to derive cells representative of numerous tissues and cell-types *in vitro* from a single common source—a superficial skin biopsy—making iPS cells a powerful platform for studying individual differences in gene regulation without limitation of tissue or cell type. Although it is unclear whether conclusions based on iPS-derived tissue cell types can be generalized to primary tissues, their use offers undeniable practical advantages in terms of ease of cell type access, controllable purity of cell type vs. the complex composition of primary tissues, and ability to work directly with human cells vs. other species. Admitedly, this strategy may also face challenges, including epigenetic alterations in gene expression among iPS clones, heterogeneous cell differentiation, and changes in genome-wide DNA methylation (108, 113, 114, 161); moreover, there may be random mono-allelic gene expression in iPS clones, as reported for EBV-transformed lymphoblasts (50, 137). Finally, typical iPS reprogramming is accompanied by numerous random viral integration events (~20 per clone), possibly affecting the expression of nearby genes; however, we anticipate eliminating this variable by generating and studying transgene-free iPS cells (cf. (194)). We will explore use of allele-specific expression (ASE) to assess cis regulation in iPS-derived cells. This strategy helps control the effect of experimental variations on gene expression, which function predominantly in *trans* (102, 159), by using one of the expressed alleles as an internal control. Using our highly quantitative ASE measures (Preliminary Results, 4.2), we find that ASE is largely stable over the sources of variation described above. iPS differentiation can thus be expected to yield additional cell type-specific ASE that is not captured by use of adult somatic cell lines alone. On that basis we propose to use ASE analysis to observe and map *cis*-acting regulatory variation using human iPS cells (Aim 2).

3.4 Synthetic technologies in human cells: Decades of research have given us the tools to make multiple precise changes to the genomes of cells of model organisms, to the point where the synthesis and assembly of standardized parts for engineering complex pathways has become the defining goal of the new field of *synthetic biology* (43, 79, 174). While synthetic biology is spawning many useful applications in lower organisms, a key goal is to make its technologies operate efficiently in human cells where they can be used to analytically dissect the mechanisms underlying human traits and disease, and to implement pathway repairs that modulate disease and deliver them into targeted human cells via gene therapy. Efficient application of these synthetic technologies to human stem cells and iPS is a particularly important goal because such cells are self-renewing and therefore have potential for permanent therapeutic effect, and because they can be used to generate many cell types (cf. (54)). Motivated by these objectives, researchers have made considerable progress in engineering human cells to the point where gene therapy clinical trials have been performed or are under consideration for over 20 human disease areas (2, 3). The many key challenges that remain can be divided into a set that relate to improving the ability to apply synthetic technologies to human cells generally, and to a set that relate to achieving clinical success. While the latter involves critical problems such as efficient access and targeting of only specific cell types within the human organism, and identification of the genetic targets that must be altered to modulate disease, here we focus on the former—specifically, on the technical challenges of engineering human cells accurately and efficiently without making unwanted changes, assuming their accessibility. Three related areas can be identified that will receive focus and development by our proposed Center: (i) Because of the large size and complexity of the human genome, it is necessary either to create and introduce large fragments of DNA into human cells, or many smaller fragments of DNA into the cells, to achieve a desired result. This entails that large DNA fragments or many small DNA fragments must be generated with minimal error either by direct synthesis or by extracting and modifying the relevant regions of the native cells' genome. (ii) Efficient delivery systems are needed to introduce these fragments of DNA into the target cells. Currently, modified viral vectors are the most efficient delivery systems, but they can only typically package 5-25kbps (31) of DNA, and retroviral systems can randomly integrate DNA into the genome. While random integration is particularly concerning clinically for gene therapy, and is quite possibly the cause of development of cancers in the otherwise most successful gene therapy trial (19), it remains a concern in a general engineering context simply because it creates unwanted modifications in the genome. (iii) Where a

delivery simply introduces DNA into the cell and relies on native mechanisms to integrate it, such as electroporation, nucleofection (http://www.amaxa.com/) or lipofection, the efficiency of desirable low-error replacement of targeted DNA by homologous recombination (HR) and/or gene conversion (GC) is low compared to error-prone non-homologous end joining (NHEJ) and random integration. Estimates of native HR:NHEJ efficiencies vary from 1:30 to 1:40000 (191).

The state of the art in these three areas can be summarized. (i) While large DNA constructs can be synthesized *de novo*, in part due to developments by the Church Lab (174) whereby DNA oligonucleotides (oligos) synthesized on arrays can be assembled into units of 1000s of nucleotides, most human genome engineering will entail making single or multiple small changes precisely in native human DNA. Many labs have developed methods for making small changes in human DNA directly in human cells using variously-designed oligos and small DNA fragments (64, 166), while the Church Lab is nearing completion on a project that uses oligo-based methods to replace all 314 instances of the TAG stop codon in the *E. coli* genome with TAA stop codons (see Preliminary Results 4.4). Both oligo-based methods are relevant for CTCHGV and will be pursued in Aim 1 (section 5.1.1). Meanwhile, *de novo* synthesis of large DNA constructs will be relevant to generation of zinc finger nucleases (see (iii) below and section 5.4.1). (ii) Transfer of BAC-size fragments of DNA to human cells by modified bacteria has been reported by several labs and presents advantages of supporting transfer of very large DNA constructs with high integrity compared to viral and chemical/electrical methods (52, 98, 124, 135). (iii) The Joung Lab within CTCHGV and others have developed methods for design and use of zinc-finger nucleases (ZFNs) for targeting specific genomic locations for highly efficient HR. Here, fusion proteins are constructed between sets of three or four zinc-finger DNA binding domains that are designed to recognize particular nucleotide sequences, plus a type-IIS endonuclease domain (usually *FokI*) so that, when introduced into a cell, the fusion proteins dimerize and introduce a double stranded DNA break (DSB) at a selected unique genome location. The DSB is then repaired with high efficiency by HR vs. NHEJ using homologous DNA introduced into the cell. Targeted gene replacement efficiencies as high as 29% have been reported using these methods. A consortium of academic laboratories (The Zinc Finger Consortium; http://www.zincfingers.org) led by co-investigator Keith Joung has developed “open-source” reagents, protocols, and software that enable researchers to engineer their own ZFNs (48, 107, 134, 175). A company (Sangamo Biosciences, http://www.sangamo.com/index.php) has also developed their own platform for engineering custom zinc-finger proteins and access to this technology is available through Sigma-Aldrich at a price per zinc finger nuclease pair of $25,000 (115, 134).

Within Aims 1 and 4, CTCHGV plans to both use these methods and make extensive improvements in them in support of proposed Center goals. The prospect of success is high not only because CTCHGV labs have already made key contributions to these developments (see above), but because CTCHGV provides a focus for specifically improving the engineering aspects of human cell synthetic technologies apart from the clinical aspects of gene therapy: In particular, CTCHGV’s concentration will be on modifying potentially large (~100kb) human DNA constructs and introducing them into human cells and using ZFNs to efficiently drive HR, and also on using oligos to make multiple targeted small modifications to human DNA, in a research context that does not require simultaneously addressing tissue accessibility, cell targeting, or clinical impact. However, CTCHGV success will be immediately translatable to clinical gene therapy, both via CTCHGV’s improved engineering methods, generally, and because in Aim 2 CTCHGV also commits to making these methods work in human iPS (see 3.4 above; also cf. (54)).

3.5 Personalized medicine: The premise of personalized medicine is that genomic information can identify individuals with different profiles of disease risk, response to treatments, or susceptibility to side effects, and thus be used to stratify individuals to optimal treatment and surveillance regimes (13). Genetic risk screening, pharmacogenetics, and expression profile assays are already in use, with 1422 clinical genetic tests for disease (177) and 23 drugs with pharmacogenetic testing information on their labeling available as of this writing ((158), Table 5). While translation from research to clinic of such genomic tests depends on many factors such as education of practitioners and patients and the development of low cost and reliable assays, the success of personalized medicine will ultimately depend on their efficacy in predicting disease and improving treatment. The CTCHGV Aims have potential to improve such efficacy in two ways: First, by refining understanding of the causal consequences of human variation, CTCHGV methods will help identify with greater precision than GWAS which variations carried by any particular individual have consequences for disease and treatment, enabling improved accuracy and sensitivity of genomic tests. CTCHGV methods also have potential to themselves be the basis of new kinds of personalized tests that could directly inform

decisions about disease monitoring or treatment. For instance, we can envision creating iPS from hair follicles or skin fibroblasts from an individual with a family history of cancer, developing disease-relevant tissue cells from the iPS, and then creating populations of cells bearing single gene deletions covering hundreds to thousands of genes mimicking loss of heterozygosity, to see if phenotypes or transcription profiles related to the cancer appear; and we could subsequently help prioritize available treatments by identifying those that best relieve these phenotypes or profiles.

4. Preliminary Results

As CTCHGV research will apply and significantly extend methods and expertise developed by our former Molecular and Genomic Imaging CEGS (MGIC), we begin our discussion of preliminary results by mentioning MGIC areas of achievement that will be relevant to CTCHGV goals. MGIC made important contributions to development of targeted DNA (10, 100, 139) and RNA (101) sequencing, single cell genomics (200), long range haplotyping (176, 201), image analysis (7, 8, 83), instrumentation design for automation (172), stem cells (including induced Pluripotent Stem cells (iPS); see 4.2 below), and RNA splicing (203), RNA editing (101), microRNA (181), and methylation analysis (10). CTCHGV will also leverage MGIC's considerable achievements in next generation sequencing, which not only included the development and release of the open-source, commercially available Polonator instrument (172), but in numerous methods and reagents that have been incorporated into other commercial instruments (see Data and Materials Dissemination Plan for companies with which the Church Lab within MGIC established close collaborative relationships). While selected aspects of these MGIC-related developments will be described below, we will focus mostly on preliminary results regarding other developments relevant to CTCHGV; however, see References for a list of the 45 published or pending articles produced by MGIC. We also note that MGIC's track record in training and education is relevant to CTCHGV (see Training and Minority Action Plans).

4.1 Targeted sequencing of DNA and RNA: The Church and Zhang Labs have been leaders in development of targeted sequencing based on Molecular Inversion Probes (MIPs) or padlock probes. In these methods, padlock probe oligos that can be synthesized on an array are designed with ends that hybridize specifically to regions flanking thousands of target sequences of interest. In a single reaction including the template DNA, the oligos, and both polymerase and ligase, the polymerase extends the oligo 3' ends across the target sequences until the 5' end of the oligo is reached, at which point the polymerase-extended oligo is circularized. The circles are purified and common sequences built into the oligos are used to amplify the targets and as sequencing primers. The method can be applied equally well to genomic DNA or cDNA; an illustration of the application of the method to cDNA is shown below in Figure 4.2-1.

The goal of targeted sequencing is to reduce sequencing requirements by restricting sequence feature creation and coverage to only the targeted subset of the initial sequences provided, which could be as large as an entire genome or transcriptome. Since their initial development during the MGIC Center (139), the Church and Zhang labs have carried on development of padlock probe methods as part of NHLBI grant HL08-004 with Jon Seidman of Harvard Med School, the goal of which is to develop targeted sequencing of human exomes of Personal Genome Project (136) (see 4.9 below) and Framingham Heart Study subjects. In addition to padlock probe-based methods, the HL08-004 project is also developing hybridization-based enrichment of exonic targets from fragmented genomic DNA. Ongoing optimization of padlock probes has improved their efficiency by 10,000-fold (100). Other than the NHLBI exomes, the Church and Zhang labs have developed and demonstrated targeted capture and sequencing via padlock probes of many thousands of targets in parallel for assays of several biologically important phenomena, including measurement of allele-specific expression of transcripts (see 4.2 below), measurement of variation rates in CpG dinucleotides (100), genome-wide assessment of CpG methylation levels (10), and detection of RNAs subject to RNA editing (101).

While padlock probe-based sequencing is a strong interest of the Church and Zhang labs, CTCHGV will not hesitate to employ or combine their use with other methods where this is advantageous. For instance, one advantage of padlock probes is that they can help equalize coverage of targets of different abundance. In this way, measurements of allele-specific expression (ASE, see 4.2 below) of low abundance RNA targets can be made more precise than those based directly on non-targeted sequencing of transcripts. However, this same feature of padlock probes may be a source of bias when the object is to compare the relative expression levels of different transcripts. Therefore, CTCHGV may use RNA-Seq or P-MAGE (84), or gene expression microarrays in conjunction with padlock probes, where both relative transcript abundance and ASE must be

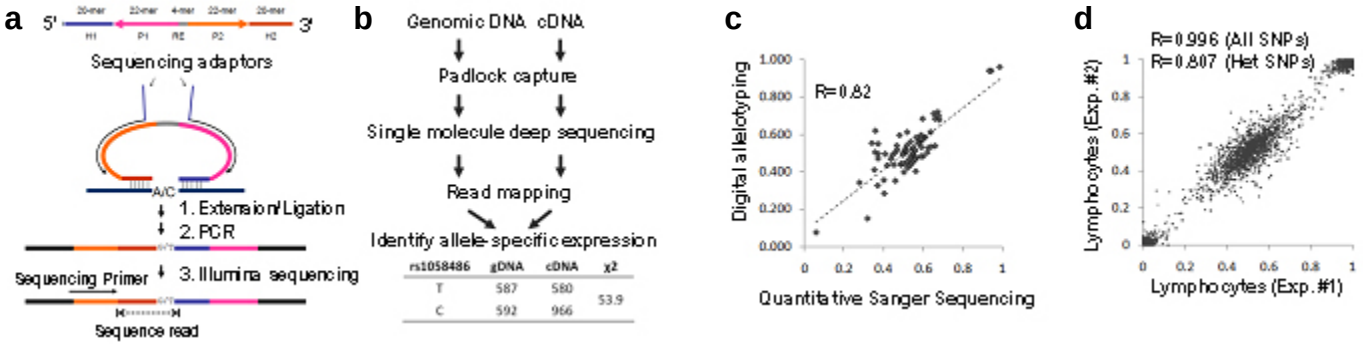


Figure 4.2-1. RNA allelotyping. (a) A schematic diagram for MIP capture and single-molecule sequencing. **(b)** Detection of allele-specific gene expression. **(c)** Comparison of allelic ratios measured by RNA allelotyping and quantitative Sanger sequencing. **(d)** Comparison of allelic ratios between technical replicates.

assessed (see, e.g., Specific Aim 1.3).

4.2 Allele specific expression (ASE) and long-range haplotyping: ASE: We recently adapted molecular inversion probes (MIPs) for digital quantification of RNA allelic ratios. To do this, we designed a library of 27,000 probes, each targeting a common SNP in a transcribed region (Figure 4.2-1a). All 27,000 SNPs were captured from both genomic DNA and cDNA in single-tube reactions, and the allelic ratios were determined by ultra-deep sequencing (Figure 4.2-1b). The capturing and sequencing protocols have been extensively optimized, such that the allelic ratios could be measured accurately (Figure 4.2-1c) and consistently (Figure 4.2-1d)

Haplotyping: *(i) Amplification of single human chromosome molecules.* We have extended the polymerase cloning method to amplification and sequencing of single human chromosomes. To do this, we trapped lymphocytes at the metaphase and extracted intact human chromosome molecules. The chromosome solution was then diluted to ~0.5 chromosome/reaction and amplified. The amplicons were then labeled and hybridized with a regular chromosome spread. The specific FISH signals indicated that large single chromosome fragments or intact chromosomes could be specifically amplified (Figure 4.2-2a,b). *(ii) Improvement of Multiple Displacement Amplification (MDA) for single molecule amplification.* Our original polymerase cloning method relies on MDA, which has an inherent limitation in the representation bias. Recently, it was reported that MDA using longer randomized primers (N9) in the presence of trehalose provides more even genome coverage (131). We have confirmed the reduced bias of this method by performing MDA on single human lymphocytes followed by genotyping with Illumina Infinium chips. The N9 primer exhibited slower amplification kinetics than the conventional N6 primer, probably due to lower priming efficiency. We found that a new LN9 primer containing partial locked nucleic acids has higher amplification efficiency and a lower level of background amplification (Figure 4.2-2c,d). We are currently assessing LN9-based genome coverage using Illumina genotyping and next-gen shotgun sequencing. *(iii) Post-normalization protocol.* Amplification bias on single template DNA molecules is unavoidable and leads to requirements for increased sequencing. We recently found that biased sequencing libraries can be normalized

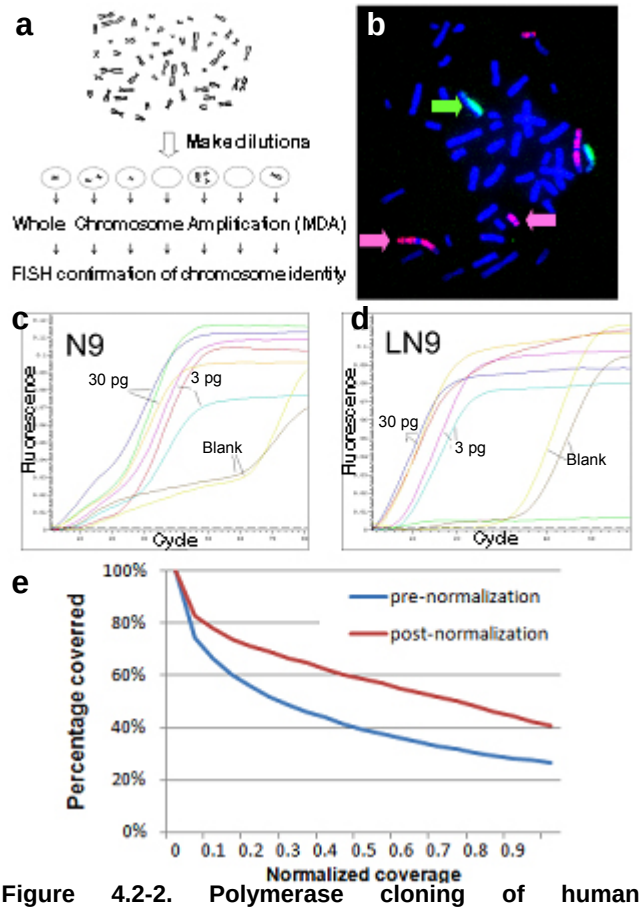


Figure 4.2-2. Polymerase cloning of human chromosome molecules. (a) MDA is performed on limited dilutions of human metaphase chromosome molecules. **(b)** FISH hybridization confirmed that one amplicon (purple) was from chromosome 6 and another amplicon (green) was from chromosome 19. **(c, d)** Real-time amplification curves with the N9 and LN9 primers. **(e)** Reduction of representation bias with post-amplification normalization.

prior to sequencing. The normalization procedure involves denaturing and slowly annealing the libraries, digestion of annealed sequences with a double-strand specific DNA nuclease, and enrichment of the single-stranded species with PCR. The percentage of sequences that have at least half of the average sequencing coverage was increased from 41% to 61% (Figure 4.2-2e).

4.3 Stem cells: The Daley Lab within CTCHGV is a world leader in stem cell research in all aspects of stem cell and iPS generation, differentiation, and analysis described in this proposal. Here we focus on Church Lab experience that is relevant to analysis of allele specific expression (ASE) in iPS and derived cells and the automation of iPS generation and maintenance procedures. In order to explore whether iPS cells and their derivatives can be used for cis-regulatory mapping, we derived iPS cells and performed allele-specific expression (ASE) analysis on pluripotent and differentiated cells. We captured 27,000 expressed exonic SNPs on 10,345 human genes using padlock probes (10, 100, 101, 139). Differential allele expression is then measured as a ratio between the numbers of the reads mapped to the two alterative alleles (reference vs. alternative allele). We found ASE patterns in iPS biological replicates to be highly reproducible. When each of these replicates was treated with trans-retinoic acid for 12 hours to induce differentiation, we observed significant changes in ASE, which we likewise observed in differentiating embryoid bodies (EBs). Despite large changes in epigenetics during iPS reprogramming (10), we find that ASE differences from primary fibroblasts are relatively small. Despite these variations, the overall ASE signature (up to 50%) was invariant among different cell types, culture conditions and cell batches (Figure 4.3-1). We estimate that 5-15% of genes may show differentiation-specific changes in ASE. Our results show conclusively that random allelic-bias and epigenetic influences are relatively small for iPS and iPS-derived cell lines, which can thus be used for reliable mapping of individual-specific cis-regulatory variants.

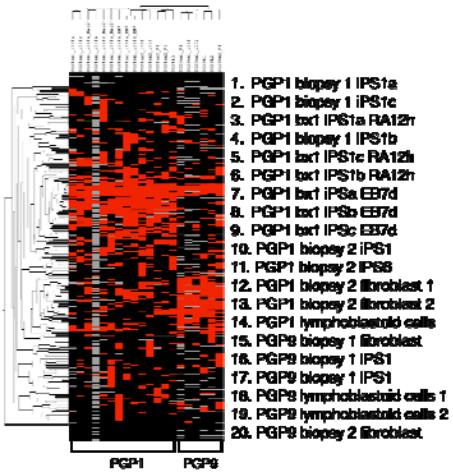


Figure 4.3-1. Hierarchical clustering of statistically significant allele-specific expression (ASE) in reprogrammed cells, showing that ~50% of overall ASE signature was invariant among different cell types, culture conditions and cell batches.

Because iPS-derived cell differentiation most closely mimicks embryonic development, the iPS transcriptome may not reflect the relevant expression signatures in adult tissues due to aging, tissue damage, and other factors. To trigger a diverse set of trans-acting regulators capable of teasing out cis-acting variants in adult processes, we partially reprogrammed primary fibroblasts using adenoviral pluripotency factors (pAdeno-OCT4, pAdeno-KLF4, qAdeno-MYC, qAdeno-SOX2). We found that many developmental trans-acting regulators were upregulated, particularly those for mesodermal (e.g., lymphocyte and muscle/skeletal) development. We also found that innate inflammation induced by adenoviral infection caused transcriptional changes consistent with immune system activation. Using this system, we were able to detect cis-variants that are relevant to adult medical disorders such as HIV-1 Rev binding protein, INFGR2 and SWAP-70, as well as developmental ASE information also obtained from iPS cells. Our results revealed that using adenoviral reprogramming may be informative for common adult medical disorders associated with tissue inflammation.

In our first steps to optimize and automate the reprogramming process, we have begun using retroviral mono-vectors to deliver the reprogramming factors and have begun adapting iPS cell culture on microcarriers in collaboration with Global Cell Solutions, Inc. Preliminary data suggests that iPS cells and hES can be maintained independently of feeder layers while retaining pluripotency for at least 2-3 passages. We are currently attempting nucleofection (Amaxa) and retroviral reprogramming on magnetic microcarrier beads in mini-bioreactors. Preliminary results using primary fibroblasts indicate that this method may be superior to normal electroporation using trypsinized and/or suspended cells. We have been adapting the bioreactor overflow for iPS reprogramming in a high-throughput manner, which will facilitate genome-wide engineering of pluripotent cell lines used in Aim 2.

4.4 Genome engineering, synthesis of large DNA fragments and combinatorial libraries: The genome engineering and synthesis needs of CTCHGV can be achieved through three main strategies: 1. *de novo* DNA synthesis, 2. Multiplex Automated Genome Engineering (MAGE) in *E. coli* or 3. Direct Multiplex Automated Genome Engineering in Human cell lines. Importantly, each strategy is rooted in technology that was recently developed or in development in the Church laboratory, uniquely positioning us to generate

combinatorial libraries of upstream cis variants.

4.4.1. de novo DNA synthesis: In prior work, we developed an inexpensive and high-throughput technology for large-scale DNA synthesis (174). In these experiments, we synthesized all 21 genes that encode the proteins of the *E .coli* 30S ribosomal subunit and mutated their DNA sequence to optimize their translation efficiency (Figure 4.4.1-1). In related work, we employed a circular assembly amplification method that significantly reduces DNA error to construct genes encoding a thermostable DNA polymerase (11). We maintain ongoing efforts to improve the fidelity and scale of DNA synthesis which will be further developed in CTCHGV Aim 4.1 and used to generate zinc finger nucleases (ZFNs) for Aim 1. Using a single 244,000 feature programmable DNA microchip, CTCHGV could generate all the ZFNs required to analyze 1000 genes in one subject (1000 genes x 5 loci/gene x 2 alleles/locus x 20 oligos/ZFN (see Overviews, sections 5 and 5.4)).

Figure 4.4.1-1 Multiplex DNA Synthesis or large DNA fragments (174)

4.4.2. Multiplex Automated Genome Engineering (MAGE). The Church Lab has pioneered the development of MAGE for large-scale programming of cells. MAGE simultaneously targets many locations on the chromosome for modification in a single cell or across a population of *E. coli* cells, thus producing combinatorial genomic diversity (Figure 4.4.2-1). In ongoing work in the Church Lab under the auspices of the Church Lab's Department of Energy Genomes-to-Life Center, we have been replacing all 314 instances of the TAG stop codon in the *E. coli* genome by TAA stop codons, thereby freeing up TAG for other possible uses. In this project, we divided the *E. coli* genome up into 32 ~145kbps segments and used MAGE to replace all coding TAGs with TAAs in each segment. At this time all segments have been completed and we are now in the process of joining them to generate a complete functioning *E. coli* genome that lacks the TAG codon. Genome construction is proceeding *via* a hierarchical series of conjugations of segment-bearing strains supplemented by suitable markers and selections that ensure that entire and not just partial segments are recombined. These techniques have direct relevance to methods we propose to apply in Aims 1.1 and 4.2 (see sections 5.1.1 and 5.4.2).

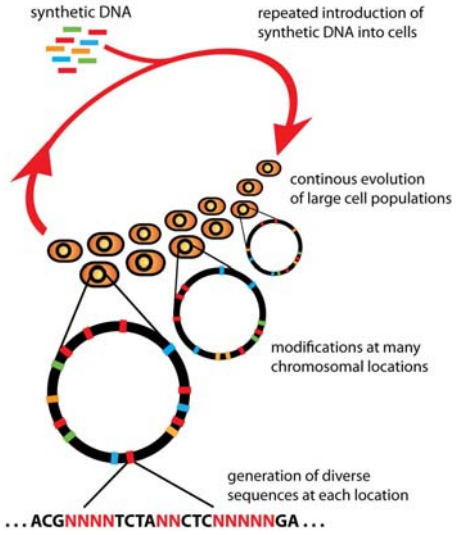


Figure 4.4.2-1 Multiplex Automated Genome Engineering (MAGE)

In another recent application, we used MAGE for large-scale programming and evolution of cells *in vivo* (184) to optimize the 1-deoxy-D-xylulose-5-phosphate (DXP) biosynthesis pathway in *E. coli* to overproduce the industrially important isoprenoid lycopene. As many as 24 genetic components in the DXP pathway were modified simultaneously using a complex pool of synthetic DNA, creating over 4.3 billion combinatorial genomic variants per day. We isolated variants with more than five-fold increase in lycopene production in less than 3 days, a significant improvement over existing metabolic engineering techniques. Since the process is cyclical and scalable, we constructed prototype devices that automate the MAGE technology to facilitate rapid and continuous generation of a diverse set of genetic changes (mismatches, insertions, deletions).

4.5 ZInc-finger nucleases (ZFNs) for improved homologous recombination (HR): OPEN (Oligomerized Pool ENgineering) is a rapid, publicly available zinc finger engineering method that was developed by the Joung lab (107) which has led academic efforts to advance engineered zinc finger technology ((134), also http://www.zincfingers.org). Like other combinatorial selection-based methods, OPEN identifies combinations of fingers that effectively deal with context-dependent DNA-binding effects. However, OPEN is simpler than other methods because it uses an archive of pre-selected zinc finger pools constructed to bind a variety of different 3 bp target "subsites". With the current set of zinc finger pools targeted to 66 subsites, OPEN can be used to target a sequence once every ~200 bp of random sequence. In addition, a large number of OPEN selections can be performed very rapidly – at present, two technicians in the Joung lab

can perform 48 selections in less than two months (M. Maeder, J. Foley, and J.K. Joung, unpublished data). OPEN is the only publicly available method which has been successfully used to create ZFNs that modify endogenous genes in human cells: Specifically, using OPEN ZFNs, the Joung lab and collaborators have used OPEN ZFN pairs to modify target sites in four endogenous human genes (*VEGF-A*, *HoxB13*, *CFTR, and PIG-A*) ((107) and unpublished research). Gene targeting/HR induced by OPEN ZFNs was so efficient that as many as four copies of *VEGF-A* could be modified in a single cell. In addition, the Joung lab (working with the Peterson lab at Massachusetts General Hospital) also successfully used OPEN in recent unpublished work to generate ZFN pairs for additional target sites in various endogenous zebrafish (*TfR2*, *dopamine transporter*, *telomerase*, *HIF*, and *gridlock*) and plant (*SuRB*) genes (48).

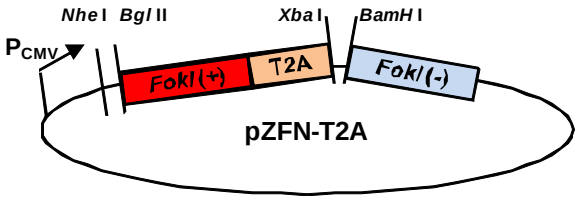


Figure 4.5-1. pZFN-T2A – dual ZFN expression vector

Direct comparisons in human cell-based assays show that OPEN is more effective and yields higher quality ZFNs than previously described "modular assembly" approaches (see Figure 2 in (107)). The higher success rate of OPEN is likely attributable to its greater sensitivity to context-dependent effects on DNA-binding among neighboring zinc-fingers that are largely ignored by modular assembly. In addition, three-finger ZFNs made by OPEN exhibit minimal toxicities in human cells compared to fully optimized four-finger ZFNs made using the complete algorithm-driven Sangamo platform (115). These findings are consistent with another recent report which demonstrated that a pair of three-finger ZFNs (made using a strategy similar to OPEN) was also no more toxic than fully optimized four-finger Sangamo ZFNs (140). Dimers of our three-finger OPEN ZFNs and Sangamo's four-finger ZFNs should recognize 18 and 24 base pair target sequences, respectively, and, assuming full specificity, these ZFNs should be capable of recognizing genome-unique sites.

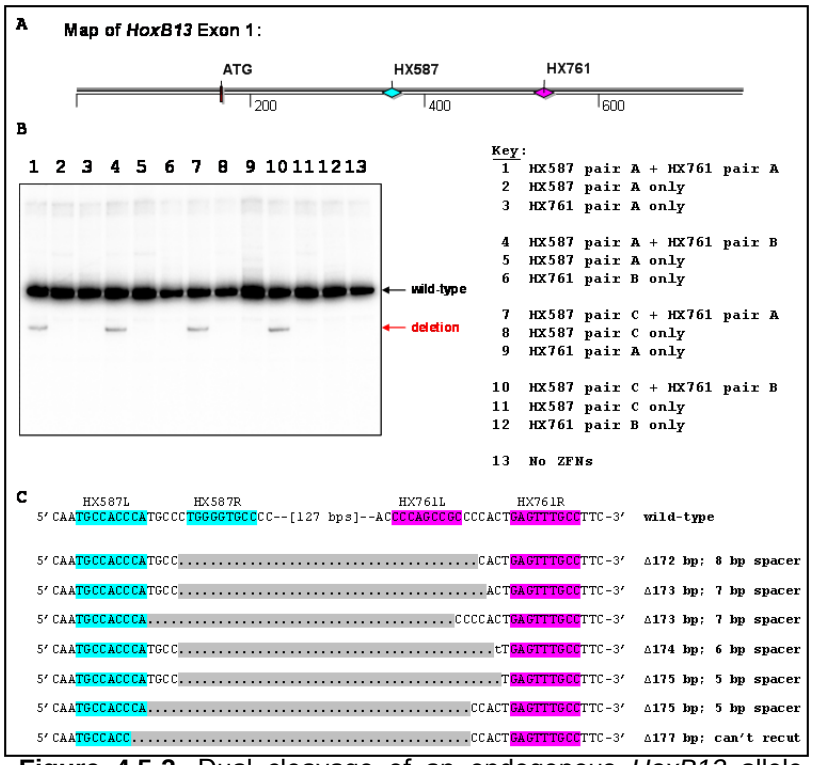


Figure 4.5-2. Dual cleavage of an endogenous *HoxB13* allele leads to deletion of intervening sequence. **A.** Map of human HoxB13 exon 1 with sites targeted by ZFNs. **B.** Limited-cycle PCR assay of genomic DNA from cells treated with ZFN combinations of ZFNs that cut at the HX587 and HX761 sites (107). **C.** DNA sequences of deletion alleles cloned from genomic DNA of cells treated with two ZFN pairs that cleave at the HX587 (blue) and HX761 (pink) sites. ZFN half-sites are highlighted for each full ZFN site: left (L), right (R). Deletions indicated in grey.

In CTCHGV Aim 4 (see section 5.4.2) we propose to improve the replacement of large segments of DNA required in Aim 1 (section 5.1.1) by using pairs of ZFNs that cut double stranded breaks (DSBs) at the flanks of the targeted segment, a strategy that should improve genome engineering capabilities generally. The Joung lab has designed and constructed a mammalian expression vector that efficiently co-expresses two ZFN monomers from a single coding transcript. In this vector, a strong CMV promoter drives expression of a single open reading frame encoding two ZFNs joined by a self-cleaving picornavirus T2A peptide (Figure 4.5-1). Previous work has shown that expression of two ZFNs joined in this way leads to efficient stoichiometric expression of the two ZFN monomers that are cleaved apart during translation at and by the intervening T2A peptide (42). In our version, the two *Fok*I cleavage domain coding sequences also harbor obligate heterodimer mutations which have been shown to reduce the toxicity of ZFNs due to their significant reduction of unwanted homodimer formation (115). To reduce recombination between the two *Fok*I sequences, we re-coded one *Fok*I monomer to make it as dissimilar to the other as

possible at the nucleotide level. Zinc finger arrays selected using OPEN can be excised and cloned in-frame into pZFN-T2A using the sites indicated in Figure 4.5-1. We tested the pZFN-T2A vector in human cells by using it to express a pair of ZFNs targeted to a site in the human *VEGF-A* locus (VF2468). Using a limited-cycle PCR/CEL-I enzyme mismatch detection assay that we and others have previously used to assess mutations introduced by ZFNs, we found that our vector could be used to efficiently express two ZFNs from a single vector (C. Ramirez & J.K. Joung, unpublished data).

Based on our success in simultaneously introducing two ZFN monomers into a cell, we reasoned that our pZFN-T2A vector should also enable introduction of two *pairs* of ZFNs (i.e., *four* ZFN monomers) into human cells. To demonstrate this, we used ZFN pairs targeted to two different sites in the endogenous human *HoxB13* gene—HX587 and HX761, for which the Joung lab had previously engineered pairs of ZFNs using OPEN selection (107) that each induced highly efficient non-homologous end joining (NHEJ)-mediated mutations at their respective target sites in human HEK293 cells. HX587 and HX761 are both present in human *HoxB13* exon 1 and are separated by ~180 bps (Figure 4.5-2A). We tested the hypothesis that the two pairs of ZFNs could both cleave a single *HoxB13* allele with the result that the intervening sequence might be deleted *via* rejoining of the two ends by NHEJ. After simultaneously transfecting cells with two pZFN-T2A vectors encoding pairs of HX587 and HX761 ZFNs and harvesting genomic DNA three days post-transfection, we performed limited-cycle PCR with primers flanking the two sites and found a PCR product from the doubly-transfected cells that was ~180 bp smaller than that from the wild-type allele (Figure 4.6-2B). Quantification suggests that the deletion occurs at a frequency of ~1 to 2%, although this may be an overestimate because smaller size deletion product might amplify more efficiently than the larger wild-type product. The deletion product was not visible in control experiments performed with cells transfected with only one of the two ZFN pairs or with no ZFNs (Figure 4.6-2B). Cloning and sequencing eight instances of the smaller size product revealed that they all harbored deletions of *HoxB13* sequence between the centers of the HX587 and HX761 ZFN sites, and many also exhibited additional variable length deletions on either side of each cleavage site, consistent with the hypothesis that rejoining of the two DSBs might occur via NHEJ (Figure 4.5-2C). Notably, we observed that all but one of the eight sequenced alleles should be capable of being re-cleaved with a ZFN pair comprising a LEFT HX587 and a RIGHT HX761 ZFN. Based on these results, we conclude that two OPEN ZFN pairs can efficiently cleave the same allele in human cells. In Aim 4 we will pursue the strategy of providing template DNA along with pairs of OPEN ZFNs to drive HR vs. NHEJ.

The Joung Lab (working with Bradley Bernstein at the MGH and the Broad Institute) has begun work on a method for unbiased genome-wide determination of off-target alterations to genomic DNA caused by of use of ZFNs. We have proposed a very similar method in Aim 1.2 (see section 5.1.2(iv)). In this method, Chip-Seq is used to identify all binding sites in the genome of a catalytically *inactive* version of a ZFN. Subsequently, the corresponding catalytically *active* ZFN is used and targeted sequencing (Prelminary Results 4.1, above) is used to look for actual DNA alterations at these sites (for details, see 5.1.2(iv)). At this time, the Joung and Bernstein Labs have taken their procedures to the point of verifying highly specific binding of catalytically *inactive* FLAG-tagged OPEN ZFNs targeted to site VF2468 in the human VEGF-A promoter. The inactive ZFNs included a previously described mutated version of *Fok*I (14). The inactive ZFNs were expressed as obligate heterodimers using pZFN-T2A (see Figure 4.5-1). Human K562 cells ($3x10^7$) were nucleofected (Amaxa) with this vector and genomic DNA harvested 24-hours post-transfection for ChIP with FLAG antibody (M2, Sigma). Initial qPCR results indicated 36-fold enrichment of the VF2468 site in the DNA from the ChIP compared with whole cell extract. ChIP DNA was then prepared for Illumina sequencing and qPCR repeated on the library DNA demonstrated 143-fold enrichment of the VEGF-A ZFN target site in ChIP DNA versus whole cell extract (Figure 4.5-3). The Joung Lab is awaiting results from actual Illumina sequencing.

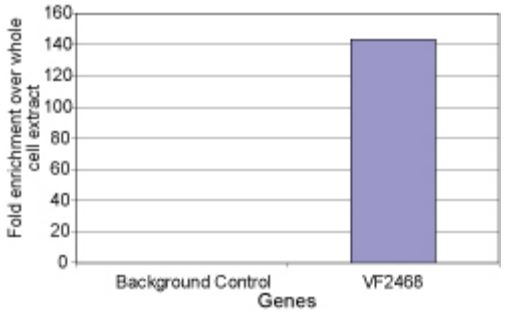


Figure 4.5-3. Enrichment of an OPEN ZFN VEGF-A target site in genomic DNA extracted via ChIP with anti-FLAG antibody (see text for details)

4.6 Polonator instrument: The Polonator instrument (Figure 4.6-1) was developed as a low-cost, open source sequencing platform in our MGIC CEGS, but will not be developed as such in CTCHGV. For routine high-throughput sequencing, CTCHGV will employ available commercial platforms such as the Illumina Genome Analyzer or the Roche 454 sequencer, or look to our collaborators (see Letter of Support from

Complete Genomics, Inc). However, CTCHGV Aims 1, 3, and 4 involve design and optimization of instrumentation for parallel single cell assays and for integrated DNA sequencing and synthesis. These can be usefully developed on the Polonator, which serves as a very general foundation for integrating flow cell-based cell handling, automated reagent handling, and integrated microscopy and image analysis. The Polonator may be used for sequencing when this needs to be integrated with new instrumentation. Here, we report that we have been developing four-color Sequencing by Synthesis (SBS) reversible terminator strategies to increase Polonator read lengths, as well as the capability to attach an ordered pattern of Rolling Circle Amplified sequences on the Polonator to increase sequencing feature density and throughput. A cyclic ligation strategy for increasing read length to 48 bases (24 from each of 5' and 3' ends) is also under development. We expect most of these improvements to be in place by the time CTCHGV funding becomes available. Our work on the Polonator, as well as on MAGE (see 4.4.2), has given us considerable expertise in instrumentation development, generally.

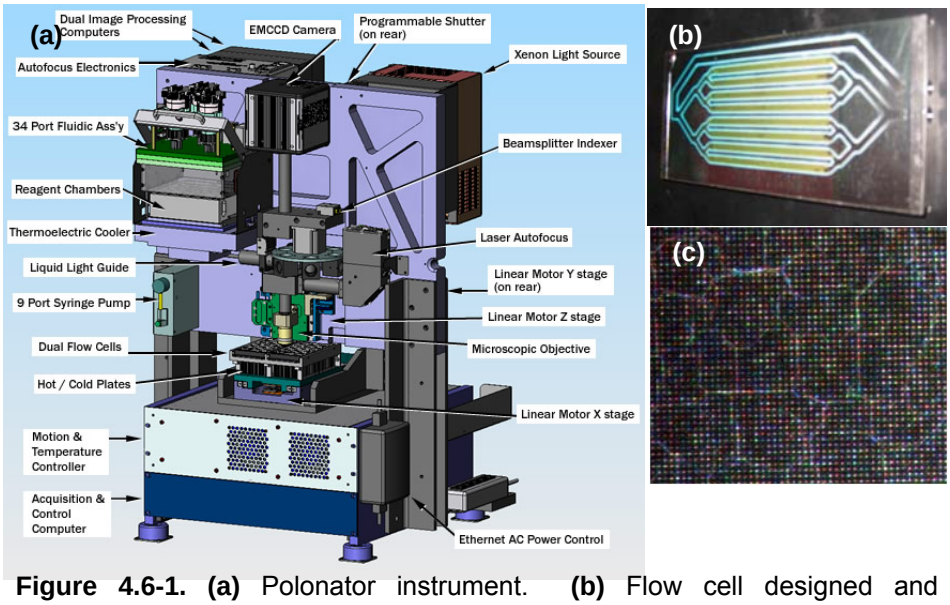


Figure 4.6-1. (a) Polonator instrument. **(b)** Flow cell designed and incorporated into Polonator to increase throughput and reduce runs costs by reducing reagent volumes and expenses. Overall dimensions: 150 X 70 X 8 (mm); lanes' "active area": 70 X 3.3 X 0.1 (mm) (.025 mm in testing). When loaded, the flow cell contains 0.5-1e9 1μm beads. **(c)** Polonies created by Rolling Circle Amplification (RCA) instead of on 1μm beads, deposited on a grid with 600 nm spot diameter and center to center spacing of 1700 nm. The grid was etched on a silicon wafer using standard photolithography. The image was obtained after a single Sequencing by Synthesis cycle on the Polonator using fluorescent reversible terminators (112, 190). Different colors represent the different bases incorporated during the cycle. Note that CTCHGV proposes to develop single cell transcriptomics using RCA polonies in Aim 3.

4.7 Single cell sequencing: The sequencing of a genome of an individual prokaryotic cell was accomplished by Kun Zhang while a member of the Church Lab (200), and comparable methods are used for long range haployting by the Zhang Lab (see section 4.2). Professor Zhang also has an R01 from NHGRI (R01HG004876) to develop a lab-on-chip device for single cell sequencing. These successes demonstrate CTCHGV capabilities applicable to single cell transcriptomics and other single cell assays (Aims 1.2, 3, 4.3).

4.8 Splice variant and methylation analysis: CTCHGV Aim 1.3 will validate the causality of cis variants identified as causing differential allelic transcription in part by exploring alternative explanations, including such phenomena as differential splicing and methylation. Church Lab experience in performing such analyses includes the following: In (203), comprehensive identification and quantification of alternative splicing for selected transcripts was performed in a single molecule gel polony framework, while (10) analyzes genome-wide methylation levels via two methods, sequencing of genomes cut by methylation-sensitive restriction enzymes, and targeted bisulfite sequencing of methylation sites (see section 4.1). Methylation is accurately measured by both measures. Versions of these procedures will be used to assess selected sets of potentially causative cis variants identified in Aim 1. RNA-Seq may be used to assess RNA splicing more globally via exon junction fragments; our development of the PMAGE (84) RNA sequencing procedure puts this within easy reach.

4.9 Personal Genome Project (PGP): The purpose of the PGP is to promote and organize the development of a set of human genome sequences and cell lines supplemented by phenotype information that can be used as research resources by the scientific community, as well as to increase public awareness of and participation in the shaping of personal genomics (28, 136). As comprehensive genomic and phenotypic

information is inherently identifying, a central aspect of the PGP has been the development of informed consent protocols and resources by which volunteers can educate themselves about the risks of making their data available and, at their option, consent to this (106). Towards this end, the PGP has worked closely with the Harvard Institutional Review Board (IRB) since 2004. Starting from initial approval for one participant to make available his data, the IRB has recently (February 2009) approved informed consent protocols that will enable up to 100,000 participants to volunteer their data. In the meantime, with the help of 10 initial participants (56, 60), preliminary exome and SNP data were released with phenotype information and cell lines made available through Coriell. Research community interest in the PGP has been high, with the consequence that many additional resources have been donated to the PGP, including computer equipment and software for managing the large data sets and automated processing that will be generated by the Project, as well as genomic services. Of particular interest, Complete Genomics, Inc. (http://www.completegenomics.com/) has committed to sequencing and making available up to 10 PGP diploid genomes (see Letters of Support).

4.10 Image and computational analysis: CTCHGV research will require sophisticated computational analysis in several key areas, including (i) image analysis, (ii) next-generation resequencing and sequence variant identification, (iii) RNA expression and ASE measurement, (iv) systems management and support for maintenance and computational analysis of large sequence and expression data sets, (v) instrumentation support, and (vi) algorithm development. The CTCHGV has substantial experience in all of these areas. (i) The Church Lab has developed sophisticated algorithms for feature calculation, morphological analysis, and classification of individual cells and cell samples (7, 8, 83). Image analysis tools have also been developed in support of gel (201, 203) and bead polonies (29, 155). (ii) The Church and Zhang labs are not only experienced with standard tools such as MAQ (99) for mapping sequence reads to target sequences and calling variants (10, 136), but have developed their own algorithms for mapping and variant calling (100), as well as for calling RNA editing sites (101). (iii) See section 4.2 for examples of CTCHGV work on ASE and RNA expression analysis. (iv) Sophisticated infrastructure and data management tools have been developed in support of the PGP (197); these tools will be available for CTCHGV. (v) Considerable software development for controlling instrumentation was built into the Polonator and will be directly usable where the Polonator is used as a framework for new CTCHGV instrumentation (see section 4.6). (vi) Initial frameworks for the algorithms that will identify causative cis variants and models of differential allelic expression are described in the Research Design Overview (section 5) and in Aim 1.2 (section 5.1.2).

5. Research Design and Methods

The goal of the CTCHGV is to develop new methods to identify cause-effect relationships between natural human genetic variations and the transcriptional states of cells, with focus on cis gene transcription. CTCHGV will do so by using synthetic biology methods to directly modify natural variations systematically in human cells to identify those combinations that result in changes in transcriptional state. We aim to develop scalable techniques that will allow combinations of variants to be analyzed for thousands of genes, and to use human induced Pluripotent Stem Cells (iPS) to generate a diverse set of cell-types. We will also develop techniques for assaying transcriptomes in many individual cells. Achieving these aims will require significant improvements to synthetic methods for generating DNA constructs for modifying human cell populations, to genetic engineering techniques for introducing and integrating these constructs into human cells efficiently, to analytic methods for simultaneously determining genetic and transcriptional state in individual human cells, and to the cell handling techniques that will integrate and automate these assays across thousands to millions of individual cells. Because CTCHGV aims only to develop and demonstrate our new technologies within our area of focus, but not to comprehensively apply them to large human populations, we plan to work with samples from a limited set of human individuals. We will use pre-existing, publicly available tissues and cell lines with potential to be transformed into iPS from HapMap, the PGP, the Framingham Heart Study, or other sources, with preference for samples for which comprehensive genome sequence is available. It will be important to start with *clonal* populations of cells from whatever source we use for reasons noted in Aim 1.2 (section 5.1.2(i)), and also to obtain diploid genome sequences of a large number of genes and their regulatory regions. Here, we will have support from our collaborator Complete Genomics, Inc (see Letter of Support).

Overview of CTCHGV Research Strategy, Numerical Targets, and Scope: The problem of identifying which natural cis variations causally affect cis gene transcription levels is important in two ways. First, knowing which specific variants causally affect transcription is important in itself (see Background and

Significance 3.1), as this information will be relevant to understanding specific gene functions in relation to specific phenotypes. Second, the methods needed to address the problem must effectively meet several general challenges in human biological research, so that these methods will immediately have important application elsewhere. Among these issues are: (i) To identify cis causal relationships in human cells requires accurate and efficient ways of engineering them. (ii) The human organism has hundreds of identifiably different tissues, and within them, many thousands of cell types. To properly characterize cis variation requires being able to assess the spectrum of human tissue types and the complex populations of individual cells within them. (iii) The immediate transcriptional effects of causally relevant cis variations only require consideration of a single gene – the cis gene – but these may have complex and cascading downstream transcriptional effects. The biological implications of cis variation thus require being able to track these effects in these complex populations of individual cells. (iv) To achieve these goals requires advances in technology that will enable accurate, *high-throughput* handling, manipulation, and observation of small cell populations and single cells. Broadly speaking, CTCHGV's four Specific Aims follow out these four imperatives, with Aim 1 simultaneously carrying out (i) and yielding the direct payoff of identifying specific causative cis variations for thousands of genes, while Aims 2-4 address (ii)-(iv), respectively. This overview of CTCHGV's research strategy thus starts with a careful look at Aim 1 and how it connects with Aims 2-4.

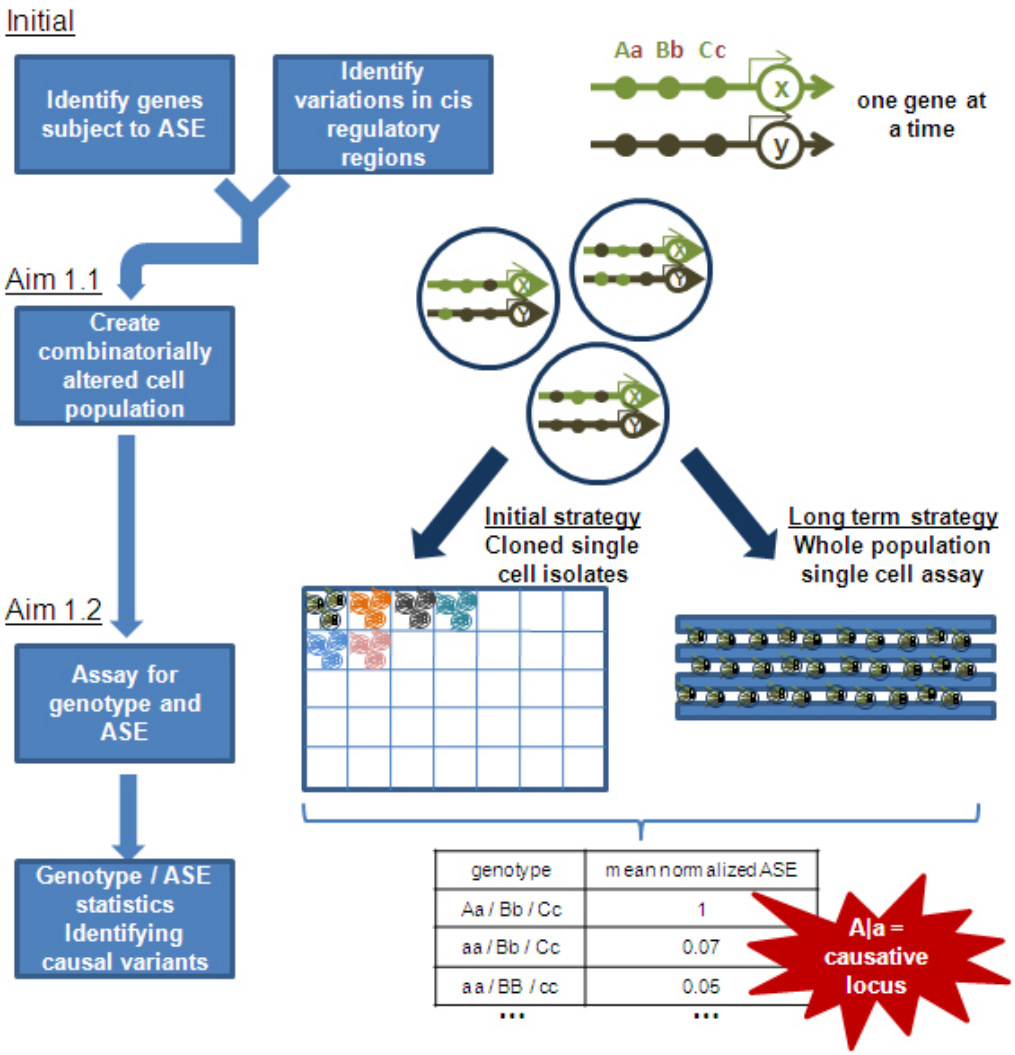


Figure 5-1: Overview of Aim 1 strategy for identifying causative cis variations. Initial Aim 1 work identifies genes subject to allele-specific expression (ASE), and, via next-gen sequence data, also identifies variations in regulatory regions, e.g., here the 100kpbs upstream region. Via Aim 1.1, cell populations are created for each gene bearing combinations of the variations identified for that gene. Via Aim 1.2, cells from this population are genotyped and assessed for ASE so that the specific loci and loci interactions that control ASE can be identified. The initial Aim 1.2 strategy will examine clonal outgrowths of individual altered cells from the population, while a longer term strategy will assay the entire mixed altered population at a single cell level. This strategy is executed one gene at a time for 100s to 1000s of genes.

Aim 1's strategy for identifying variations that causally affect transcription levels is depicted in Figure 5-1. We start by identifying genes in subject cell lines that exhibit allele-specific expression (ASE) as measured by differential expression of indicator alleles **x** and **y** in gene coding regions (sub-Aim 1.2, section 5.1.2(i)). We then identify cis variations in putative regulatory regions of the genes, considering not only SNPs but also small indels and other sequence variations. Assuming ~1 variation / 1000 bp, a regulatory region such as a 100kb upstream region may contain ~100 variants. Our starting assumption is that some of these variations may actually be causes of ASE, while many of the rest are associated with it by dint of being in the same haplotype block. Our task is to identify variants that are actually

causal. As a first cut, we will use information on conservation and transcription factor binding sites to reduce the number of variants we will analyze to ~5 or less (sub-Aim 1.2, section 5.1.2(i)). We will then use engineering techniques developed in sub-Aim 1.1 (section 5.1.1) to, in effect, break down the haplotype block so that the transcriptional effects of these variations can be observed independently, revealing which are causal. Specifically, we will generate subject cell lines in which these five variant loci are modified to at least individually assume all haplotypic states. For instance, if one of the loci is **A**|**a**, with **A** cis to the **x** coding indicator allele, and **a** cis to the **y** indicator allele, we will generate cells which are **AA**, **aa**, and **Aa** in both haplotypes, and we will similarly alter the other four variant loci. To engineer these changes, we will (1) extract gene regulatory region genomic sequences from the subject cell lines, (2) alter them in *E. coli* using MAGE techniques (see Preliminary Results, 4.4), and (3) re-introduce the altered regions back into the subject lines and induce homologous recombination to replace the native regulatory region alleles using Zinc Finger Nucleases (ZFNs) that are targeted to the regions (18, 107, 115). However, we will also develop oligo-based methods which will greatly simplify and expedite altered cell line generation (sub-Aim 1.1, section 5.1.1). To generate and optimize the many ZFNs that will be needed to analyze many genes, which may include allele-specific ZFNs, we will develop high-throughput synthesis and ZFN targeting optimization methods in Aim 4 (section 5.4).

Generating cell lines with the four haplotypes per locus for each of five loci individually will entail generating 20 altered cell lines. These altered lines can be made with ZFNs one locus at a time. However, with the MAGE- and oligo-based methods we will develop (sub-Aim 1.1, section 5.1.1), we can easily generate combinatorially altered cell populations that contain all 4^5=1024 haplotypes at once. Both strategies have advantages. If we work one locus at a time, we can identify individual loci that affect differential allele expression by themselves, and then investigate interactions systematically by manipulating specific pairs or triplets (etc.) of these loci. On the other hand, fewer ZFNs are needed to create combinatorially altered populations, and by presenting all possible combinations of the alleles at once, they also put us in position to acquire maximal information on possible interactions between the cis loci immediately—if they can be assayed efficiently (see below). A third strategy will be to create combinatorial populations and isolate sets of individually altered cells for clonal outgrowth to get a random sample of isogenic altered cell populations. We will develop each of these strategies and apply them as they best fit circumstances.

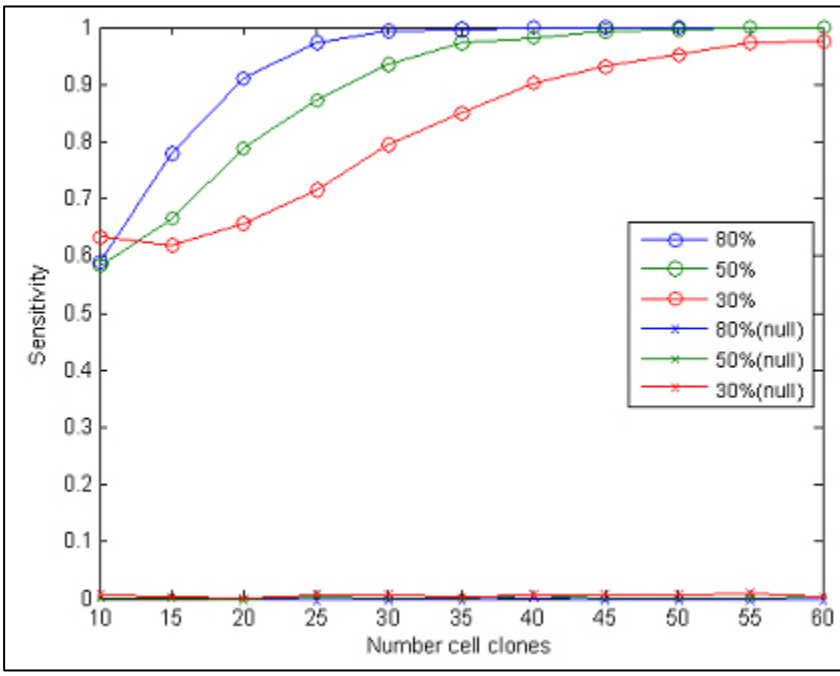


Figure 5-2: Sensitivity of detection of causality using assays of clonal altered cell lines, based on simulations involving 5 loci in a cis gene regulatory region. For locus **A**|**a**, allele **A** was assumed to cause its cis gene allele to be expressed at 2x the level of the **a** allele; other loci had no effect. Combinatorial cell populations were simulated in which all haplotypes occur with equal frequency whenever homologous recombination (HR) took place: HR efficiencies of 80%, 50%, and 30% were considered: Individual cells are isolated (number, X-axis), grown clonally, and assayed for genotype and gene expression level. Expression levels of the cis gene alleles phased with **A** and **a** were simulated subject to 20% Gaussian error. The A|a locus was identified as causal if gene expression level correlates with number of **A** alleles with p<.001 (corrected for 5 loci). "Sensitivity" = fraction of 1000 random simulations in which **A**|**a** was detected as causal. "null" lines: fraction of simulations in which non-causal locus **B**|**b** was identified as causal.

Having created the cis-locus altered cell lines, the next step is to assay them to see how specific alterations affect cis gene or allele expression. The most basic tests involve growing clonal populations from individual cells from the variously altered cell lines and identifying situations in which increased expression follows a particular allele of a particular locus (e.g., where gene expression is highest in **AA** lines, intermediate in **Aa** lines, and lowest in **aa** lines), or where ASE is abolished in cell lines homozygous for the locus (**AA** and

aa) but not in those in which the locus is heterozygous (**Aa**). Simulations (see Figure 5-2) show that with as few as 60 clonal populations grown from randomly selected individual cells from a combinatorially altered population, we should be able to identify which one of five cis loci cause two-fold differential allelic expression with confidence p<.001, with a sensitivity ≥ 97.5% and false positive rate < 1%, even if homologous recombination efficiency is assumed to be 30%, a value that has been achieved with current ZFN technology (18, 107, 115). We will pursue this strategy in Aim 1.2 (section 5.1.2), but we will also seek to go beyond it by looking for interactions between the cis loci. Here many models can be considered. e.g., cis loci can be additive, or upregulation might depend on specific allele phasing. These models can be distinguished by looking at the average profiles of differential expression over the various combinations of genotypes (see Table 5-1). We can acquire information on these profiles by using clonal isolates from combinatorially altered populations. However, these operations may become cumbersome when dealing not only with larger numbers of cells, but also with the large number of genes we wish to examine. As a longer term strategy, we will therefore develop an assay that enables the combinatorially altered population for a gene to be examined *as* a population, wherein genotypes and ASE levels will be assayed together in millions of altered individual cells, obviating the need for isolating and growing up many single cells (details in section 5.1.2 (ii)).

#A / #B	haplotypes	M1	M2	M3	M4
2/1	AB Ab	1	0	.5	0
1/2	AB aB	1	1	.5	1
1/1	AB ab Ab aB	.5	1	1	.83
1/0	Ab ab	0	1	.5	.67
0/1	aB ab	0	0	.5	0

Table 5-1: Profiles of ASE by genotypes for four models of cis variant interaction involving two cis variant loci (**A**|**a** vs. **B**|**b**), assuming cells from a completely random combinatorial population of all haplotypes. Models M1 **AB** necessary and sufficient to upregulate the cis allele. M2 **A** alone necessary and sufficient to upregulate cis allele. M3 **A** and **B** additively and equally upregulate their cis alleles. M4 **A** upregulates cis allele relative to **a**, while **B** modifies **A** such that having **A** and **B** in cis causes an additional 50% increase in upregulation of **A** vs. **a**. ASE measurement assumes that **AA** and **aa** yield zero ASE values and **Aa** has a positive value regardless of haplotype. Mean ASE levels are given normalized relative to maximum possible ASE level for any genotype. #A / #B: Genotypes given by the numbers of **A** and **B** alleles respectively. Genotypes homozygous for both loci are uninformative for ASE and not shown. Haplotypes consistent with each shown genotype are given.

The assays we perform will identify many cis variants that determine gene expression levels. These identifications will automatically have gone deeper than associations because their participation in the cause-effect chain has been directly interrogated in cell lines with constant genetic background save for precise engineering of a few variations. However, further analysis will be needed to determine the nature of the causation. A cis variant might cause a change of expression level by altering a transcription factor binding site (170), or it might alter the methylation or histone modification profile of the regulatory region. Or, the cis variant might not actually regulate expression level but, rather, alter the splice isoform profile of the cis gene (95), so that the changes in abundance of the coding region indicator alleles **x** and **y** by which ASE is measured might be due to differential splicing of the exon containing them. In Aim 1.3 (section 5.1.3) we will assess the prevalence of a variety of such phenomena in a representative set of original and altered CTCHGV cell lines.

As noted in Specific Aims and Background and Significance (section 3.1), our proposed strategy has a close relationship with GWAS. Our methods will go beyond GWAS by identifying variants that actually cause vs. associate with phenotypes, and will also avoid GWAS constraints on effect size and allele frequency. However, our methods will themselves be limited to finding variants that specifically cause changes in expression level, compared to GWAS which finds associations with disease and phenotype. However, because GWAS frequenty identify associations in non-coding regions, variations found by our methods to be causative of differential expression will generate and refine hypotheses stemming from GWAS associations. In this proposal, we make these relationships with GWAS explicit in two ways. First we use GWAS to help prioritize genes and variants that will be analyzed by our methods. Second, in Aim 1.4 we close the loop by assessing what it would take for GWAS to discover the effects we find without our methods. This analysis must, by its nature, focus on GWAS that examine expression levels vs. phenotypes, but these comparisons will illuminate how GWAS and the techniques we develop will complement each other.

The analysis of cis variants provided by Aim 1 will be constrained in several dimensions. While some scope limitations are described below, others will be addressed in other Aims. Because of the large amount of engineering that will be performed on cell lines in Aim 1, we will employ tractable cell lines that tolerate the conditions of engineering, and this constraint will limit the universe of expression profiles (including ASE profiles) under study. However, in Aim 2 (section 5.2), we will generate induced Pluripotent Stem Cells (iPS)

with alterations developed in Aim 1, which will enable us to explore their impact in iPS-derived cell types representing diverse human tissues. The alterations generated in Aim 1 will also be developed and analyzed one gene at a time (but are scalable to many genes). In Aim 2, we will apply Aim 1 techniques to develop complexly altered iPS in which *many genes* are altered at once. In its focus on cis variant causation, Aim 1 will mainly look at expression levels of one gene in relation to the variants it manipulates—the cis gene. In Aim 3 we will develop tools to examine transcriptome-level information in the individual cells of complex tissues and cell populations. These methods will put CTCHGV in position to examine downstream effects of of the cis variations we study. The need for Aim 4 developments in support of Aims 1-3 has already been noted, and Aim 4 projects will have wide applicability to biomedical research. These observations illustrate the high degree of integration and innovation in the CTCHGV proposal.

Numerical Targets and Scope Clarifications: The number of genes (Aim 1), cell types (Aim 2), and transcripts (Aim 3), we will actually analyze will depend on the success of our methods and cannot be predicted with certainty. However, based on our track record of innovating high-throughput methods (see Specfic Aims) and our experience with the relevant technologies, we prefer setting ambitious goals that may seem risky vs. more secure and unambitious goals, with the understanding that we will review goals and renegotiate them with NHGRI at the end of year 2 of the Center in the light of our own progress, that of our collaborators, and advances in the field generally. We also feel that only ambitious goals will allow us to adequately evaluate and demonstrate the *scalability* of our methods, and we believe scalability will be essential to follow-on application of our methods outside of the Center. With this general statement in mind, our intitial targets for main Aim initiatives are: 1000 genes analyzed for cis causality (Aim 1), 50 genes in three iPS-derived cell types (Aim 2), and 1000 transcripts per single cell (Aim 3, both directed and untargeted sequencing; see Aim 3, section 5.3); for Aim 4, see the end of section 5.4. Final and intermediate goals are discussed at the end of each Aim's Research Design section. The targets described here apply to "main" Aim directions but not necessarily to all sub-Aims, some of which deal with analysis of statistical or representative subsets of genes or variants, or demonstrations of related methods or phenomena. For instance, sub-Aim 1.3's analysis of mechanisms by which cis variants control expression is not intended to be performed for all 1000 targeted genes and variants, but only to a small representative subset of genes and variants.

Finally, having described our goals, here we clarify the scope of our Center by identifying several items that we specifically consider to be *out* of our scope: (a) We emphasize again that the 'causation' that we study is *causation by genetic variants of differential cis gene expression*, not causation of disease or organismal phenotype. (b) Nor do we attempt to systematically identify variants that cause differential expression in *trans*, although Aim 3 will allow us to track some downstream consequences of cis causal variants. (c) While ideally we should like to study actual primary human tissues, we will focus on iPS-derived cell types representing human tissues for the reasons indicated in Background and Significance, section 3.1. However, Aim 3 will feature a comparison between one iPS-derived cell type and primary tissue. Nor will we study effects in tissues in non-human animal models. (d) Although in Aim 1.4 we explore how GWAS can relate to our findings, we ourselves will not perform GWAS nor any large scale population analysis.

5.1: Aim 1: We will develop and demonstrate novel methods that identify and characterize natural cis variations that directly affect transcriptional activity in individual humans based on direct modification and testing of combinations of variants in gene regulatory regions in cell lines, and that can be applied to thousands of genes.

5.1.1: Aim 1.1: We will develop and demonstrate novel, high-efficiency methods to create human cell populations containing combinations of natural variations in gene regulatory regions, focusing on zinc-finger nuclease (ZFN)-mediated recombination of externally generated altered insert libraries, and direct modification of human cells using oligo-based methods.

We will develop two main methods for generation of combinatorially altered cells, both of which will use our oligo-based MAGE technology (see Preliminary Results, 4.4). Our principal strategy will be to use MAGE to alter regulatory regions in human BACs in *E. coli*, after which these altered regions will be re-introduced into human cells. The second will develop and apply a version of MAGE that operates directly in human cells. Both of these strategies will involve development of significant technology that will have wide applicability outside of CTCHGV. While the first will leverage a MAGE technology that works efficiently in *E. coli* (see Preliminary Results, 4.4), it will need to be integrated with substantially improved methods for transferring DNA fragments between human and *E. coli* cells (including entire altered regulatory regions of ~100kb), and for

efficiently inducing homologous recombination in human cells, in order to replace native with altered regulatory regions. Here we will focus on improving direct transfer by modified bacteria of BAC fragments to human cells, and on use of zinc-finger nucleases (ZFNs) to induce efficient recombination. For the second, we will adapt our MAGE method so that it works directly and efficiently in human cells (including, eventually, induced Pluripotent Stem cells (iPS); see Aim 5.2), taking into account other oligo-based methods for engineering human cells (see Background and Significance, 3.4, and 5.1.1 (ii) below). We will refer to the first strategy as MAGE-BAC/ZFN and the second as MAGE-human. The MAGE-BAC/ZFN strategy is illustrated in Figure 5.1.1-1. In terms of structure, we divide our research plan into four sections, putting the two MAGE sections first: (i) MAGE-BAC (the MAGE and human/*E.coli* transfer aspects of MAGE-BAC/ZFN; with plans for ZFN improvement in (iii)), (ii) MAGE-human, (iii) improvement of ZFN-mediated recombination, (iv) performance evaluations.

As noted in the Research Design Overview above, depending on circumstances, we will sometimes generate altered cells with specific sets of genotypes or haplotypes, and sometimes generate cell populations with combinatorially randomized genotypes or haplotypes. In general, MAGE-BAC/ZFN will enable all of these (see (i) below), while MAGE-oligo will be better suited to generation of specific or combinatorially randomized genotypes than haplotypes as it will not always be straightforward to target oligos to specific alleles.

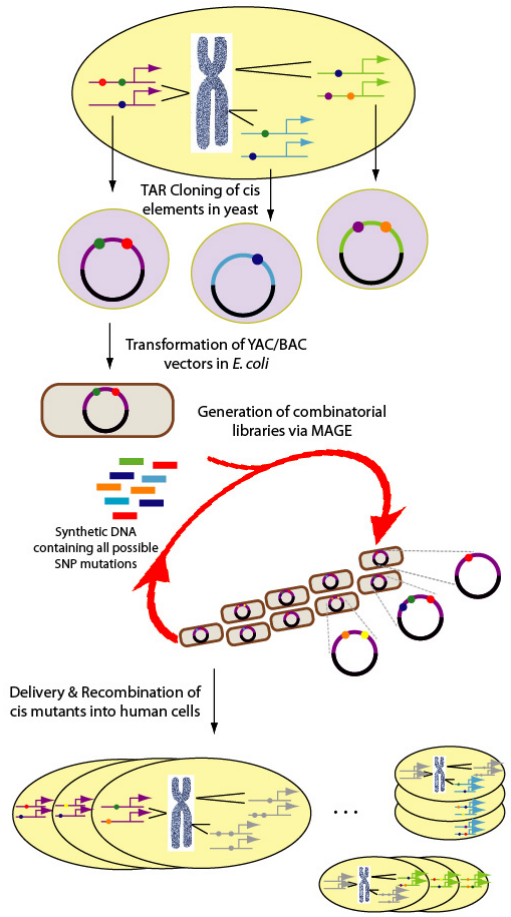


Figure 5.1.1-1: Illustration of MAGE-BAC/ZFN strategy for generating combinatorially modified human cells

5.1.1(i) MAGE-BAC: To create libraries of altered cis loci in regulatory regions of a gene, the regulatory fragments on which MAGE will be applied will first be isolated from our CTCHGV subject cell lines and moved into *E. coli* on BACs. After alteration, parts or wholes of regions altered for specific loci, or entire combinatorial libraries, will be transferred back to the original human cells to form cis-altered populations of cells. We describe the MAGE and transfer phases of this work here, while in 5.1.1(iii) below we describe how we will induce recombination of the re-introduced varied regions in the human cells.

5.1.1(i.a) Isolation of Specific Upstream Cis elements via TAR Cloning. We will clone the targeted regulatory region genomic DNA fragments from the subject cell lines onto shuttle YAC-BAC vectors using current methods of Transformation-Associated Recombination (TAR) cloning (88-90). Recent studies have shown that TAR cloning has been used successfully to isolate specific human DNA onto yeast artificial chromosomes (YACs) from human and mouse cell lines (87, 88, 91). Building on these methods (89), we plan to selectively clone all target cis elements by using shuttle YAC-BAC vectors with a 5' targeting-sequence (hook) and a common repeat (e.g., Alu) as a second targeting sequence. Thus, a library of all target cis elements will be constructed at the end of the TAR cloning process. Importantly, our YAC vectors generated by *in vivo* recombination in yeast will contain the F-factor origin of replication, permitting their propagation as BACs in *E. coli*.

5.1.1(i.b) MAGE-generated altered BAC Libraries. Having isolated cis elements on BACs *via* TAR cloning in *E. coli*, we will use our recently developed automated genome engineering methods ((184) and Preliminary Results, 4.4) to create alterations. Using MAGE, specific single-stranded DNA oligonucleotides (oligos), or pools of oligos, are introduced into an *E. coli* strain which initially contain an isogenic cis fragment derived from the human cell line, and this step may be performed repeatedly on *E. coli* strains derived from previous steps to eventually obtain an *E. coli* strain that contains cis elements with all the desired modifications. We will design oligos that specifically mutate targeted loci within the cis elements contained on the BACs. If we wish to change only a single locus, we introduce only the oligos corresponding to that locus; a small number of altered *E. coli* clones may need to be generated and assayed to identify one that contains the

desired alteration on the desired cis element allele. This strategy can then be serially iterated to make specific multi-locus alterations in the cis element. To simultaneously create combinatorial libraries of each cis element on each cis element allele, all we need do is introduce oligos that modify all loci at once. Upon completion of the MAGE process, the mutated *E. coli* population will have been altered to sample all possible combinations of the targeted cis loci. After transfer and recombination of the altered regions back into human cells, we will assess the performance of our methods as described below in 5.1.2 (iv). This information will allow us to tune the MAGE process so as to improve the efficiency of targeted alteration and the uniformity of combinatorial alterations.

5.1.1(i.c) Delivery of mutated DNA from altered BAC libraries of upstream cis elements into human cells. We plan to optimize the re-introduction of the mutated cis regions generated in (b) into human cells using a number of recently developed delivery and recombination methods: (c.1) Bacteria have been long been known to have the capacity to conjugally transfer genomic DNA to other bacteria (171), but more recently bacterial transfer to eukaryotes (58), including mammalian cells (186) has been reported. We will optimize this process for delivery of our altered BAC DNA from *E. coli* to human cells by strategic placement of the *oriT* sequence and selectable markers that flank the region of transfer, factors we have found to be important in our recent work on re-engineering the *E. coli* genome (Preliminary Results, 4.4). (c.2) In a second approach, we will utilize the GET recombination inducible homologous recombination system for the delivery of human genomic BAC clones into mammalian cells (124). The *E. coli* strain DH10B will be used as the host. We plan to introduce four genetic modifications: The λ-prophage from strain DY330 (193) and the *mutS*$^{-}$ gene deletion (33) will enable efficient oligo-mediated λ Red recombination to generate the desired mutations; while *asd*$^{-}$ gene deletion (52, 124) and expression of the *Yersinia pseudotuberculosis* invasin gene will enable DNA transfer. Expression of invasin renders *E. coli* competent to invade HeLa, COS-a and CHO cells by allowing the bacteria to bind to mammalian integrin receptors and trigger their internalization into primary vesicles. Once inside the cells, the *asd*$^{-}$ mutation causes diaminopimelic acid auxotrophy, leading to defective cell wall synthesis and death of the bacteria, making their DNA available to the human host cell (52, 124). (c.3) As an alternative approach to the bacterially-mediated transfer methods above, we will consider use of the HSV family of viruses for the infectious delivery of large BACs. Prior work has shown the successful viral packaging of 150 kb BACs with efficiency of delivery ranging from 25 to 100% into human MRC-5V2 and fibroblast cell lines (105, 183). Importantly, the high-capacity HSV-1 amplicon system permits the rapid transfer of mutated BAC libraries into an appropriate human cell line.

Finally, selectable markers will be used to select human cells to which modified DNA fragments have been delivered. However, these markers cannot be used to ensure integration of the modified fragments because that would require the markers to be integrated as well, and this would confound our objective of changing nothing but the ≤ 5 targeted cis variant loci per gene. Instead we will flank the modified within the modified sequence with the markers so that they will not be integrated during homologous recombination. This strategy has proved successful in our assembly of the *E. coli* genome out of ~145kbps modified fragments described in Preliminary Results, section 4.4.

Potential problems and alternatives: We do not anticipate problems for Tar cloning of regulatory regions (5.1.1(i.a)) as this is a well documented and widely used procedure, nor with performing MAGE on BAC fragments in *E. coli* (5.1.1(i.b)), as we have successfully applied MAGE techniques very effectively (see Preliminary Results 4.4). However, the introduction and homologous recombination of altered regulatory regions into human cells represent new ground and these processes may be inefficient. In that case: (1) It may prove difficult to re-introduce the very large altered BAC fragments intact into human cells. In that case we can break the fragments into smaller pieces and introduce them one at a time. This will likely require a ZFN to be designed for each piece of each fragment. (2) We can use a combination of selectable and counterselectable markers in such a way as to make seamless modifications, a strategy which works well in prokaryotes (104) and which was partially developed for *E. coli* by the Church Lab. This strategy has potential to improve both delivery and integration of large DNA fragments simultaneously. (3) Through our collaboration with the Elledge Lab (see Letter of Support), we can perform genome-wide RNAi and overexpression screens to identify factors that improve the efficiency of human cell delivery of DNA, and then use cell lines modified with the appropriate factors. (4) Finally, note that Aim 4.2 (section 5.4.2(iv)) lays out our plans to develop a segmental genome replacement strategy based on simultaneous use of *two* ZFNs per regulatory region.

5.1.1(ii) MAGE-human: The second strategy to generate human cells with modified gene regulatory

elements will be to develop a MAGE method that works directly in human cells (184). Site-specific gene modification can be achieved by targeting oligos or DNA fragments to the homologous genomic DNA sequence using chimeric RNA-DNA oligonucleotides (RDO), single-stranded oligodeoxynucleotide (ODN), small fragment homologous replacement (SFHR) and triple-helix forming oligonucleotides (TFO) (64). Efficiency of oligo recombination is the key metric in implementing a seamless and selection-free recombination system. This metric, combined with our expertise in achieving highly efficient oligo recombination in *E. coli*, directs our efforts towards similar oligo-directed recombination strategies directly in human cells. In *E. coli*, oligo recombination is mediated by the β protein of the λ Red recombination system, in which the oligo has been proposed to chromosomally integrate at lagging strand synthesis of DNA replication at 25% efficiency (33). Our MAGE technology improved this efficiency to greater than 30% with an ability to introduce multiple modifications simultaneously targeting many chromosomal loci (184). The MAGE technique iterates oligo-based changes through successive populations of cells, and can be used to both generate populations that are 100% modified at particular sites, or to generate combinatorial populations.

Oligonucleotide-mediated recombination has already been used to induce site-specific genetic modifications in select mammalian cell lines, including HEK-293, CHO and embryonic stem (ES) cells (36, 64, 128, 143), with efficiencies of ~0.03-5%. While the mechanisms of oligo recombination in mammalian cells are not known, these studies have revealed important design criteria that we propose to implement and enhance, including: (a) Similar to *E. coli*, the mismatch repair (MMR) pathway negatively affects oligo-directed recombination in human cells (36, 129). (b) Preferentially enhanced gene repair activity of antisense over sense oligos was observed in human cells, suggesting a link to transcription-coupled repair where gene repair by oligos occurs more efficiently when the target gene is actively transcribed (64). Moreover, (c) it has also been shown that chromosomal positioning effects have little or no influence on observed strand bias of oligos and that unmodified (e.g., no phosphorothioate bonds) antisense oligos exhibit 16-45-fold higher rates of modification than sense oligonucleotides (128, 129). (d) The Rad52 protein, mechanistically similar to the single-stranded DNA binding protein β from λ Red, can also be utilized to enhance the recombination of oligos during replication in mammalian cells (64). Given that heterologous ssDNA-binding protein homologs have been shown to function in *E. coli* (34), we also plan to test if these homologs (and others) can enhance oligo-directed recombination in human cells. To complement these efforts, we will also investigate enzymes and factors that are implicated in homologous recombination (HR) and MMR. For example, Rad51, a central enzyme of HR that polymerizes on ssDNA and assembles into helical nucleoprotein filaments, promotes both homology searches in dsDNA and exchange of DNA strands between ssDNA bound with the filament and the homologous dsDNA. Also, the Msh2 protein, a central factor in MMR, is involved in the inhibition of recombination between mismatched sequences, and only upon its deletion can oligos introduce mutation in mouse ES cells (36).

Finally, key to the success of MAGE in *E. coli* was optimization of the electroporation conditions used to deliver the oligos into the cells, and development of automation that iteratively cycled *E. coli* populations through periods of electroporation and recovery, so that modest initial per-cycle modification rates could be amplified significantly. We will develop similar optimization and automation for DNA delivery (see 5.1.1(i.c) above) and human cells.

Potential problems and alternatives: It is possible that oligo-based HR will remain inefficient even after exploring the options above. In that case, (1) we can perform genome-wide RNAi and overexpression screen to identify additional factors that will enhance efficiency with the help of our collaborators in the Elledge Lab (see Letters of Support). (2) We will also explore whether ZFNs can improve oligo-based HR. The latter has been reported to be enhanced by the presence of double stranded DNA breaks induced by I-*Sce*I meganuclease (143). Use of ZFNs with oligos would represent an HR strategy that would allow us to eliminate the Tar cloning and MAGE-BAC components above. However, if oligo-based HR remains intractable, we will pursue MAGE-BAC/ZFN as our exclusive strategy.

5.1.1 (iii) Zinc-finger nuclease (ZFN) mediated recombination: Engineered ZFNs present an attractive direction for recombining the sequences of MAGE-BAC-generated constructs and libraries into the regulatory regions of Aim 1 target genes. ZFNs function as dimers with each monomer composed of an engineered zinc finger array (typically consisting of three or four fingers) fused to a non-specific cleavage domain from the *Fok*I endonuclease (Figure 5.1.1-2a). The zinc finger arrays in ZFNs can be engineered to bind target DNA sequences of interest. Each individual zinc finger binds a 3 bp "subsite" and therefore a ZFN dimer can in principle recognize 18 or 24 bp target sites, depending on the number of fingers in each ZFN monomer. Each

ZFN monomer binds to a DNA "half-site" in the full target sequence and introduces a double-strand DNA break (DSB) in a "spacer" sequence between the half-sites. Repair of a ZFN-induced DSB by homologous recombination (HR) with an appropriately designed exogenous "donor template" can be used to introduce a

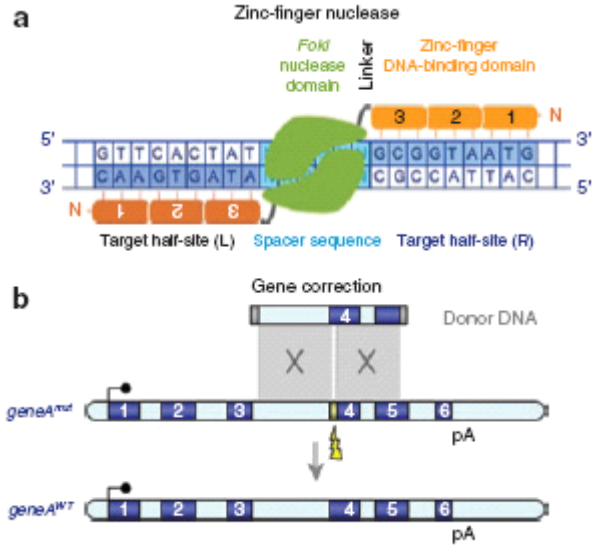


Figure 5.1.1-2. Engineered zinc-finger nucleases (ZFNs). (a) Architecture and application of ZFNs. A ZFN designed to create a DNA double-strand break (DSB) in the target locus comprises two monomer subunits. Each subunit contains three zinc-fingers (1-2-3), which recognize 9 base pairs within the full target site, and the *Fok*I endonuclease domain (green). Dimerization activates the nuclease, cutting the DNA in the spacer sequence separating the target halfsites (L) and (R). ZFN subunits comprising four zinc-fingers that recognize 12 base pairs have also been developed. (b) ZFN-mediated gene disruption and correction by homologous recombination (HR). A DSB (yellow flash) is introduced by the ZFN into mutant allele A^{mut} of a gene. The presence of donor wild-type DNA drives DSB repair through HR vs. error-prone non-homologous end joining, yielding a functional wild-type allele $geneA^{WT}$. Rather than repair genes, CTCHGV will use ZFNs to mutate cis gene regulatory regions in order to determine their causal role in allele-specific expression. Figure adapted from Cathomen, 2008 (see References).

specific mutation or insertion with very high efficiency near the break (Figure 5.1.1-2b). This method, known as ZFN-induced *gene targeting*, has been used successfully to alter endogenous genes in human cells with absolute efficiencies ranging from 1-50% (107, 118, 178). However, these performance levels are not consistent across all genes and cell lines tested to date, and many additional avenues for increasing efficiency remain to be explored. We will pursue several strategies for improvement that are close at hand in the context of this Aim (1.1). We note that longer-term improvements employing more advanced technologies will also be developed in Aim 4.2 (section 5.4.2 below).

One of the current challenges of targeted genome modifications in human cells is the low rate of native homologous recombination. Although this is greatly improved with ZFNs, error-prone non-homologus end joining (NHEJ) is still a relevant competing pathway of DNA repair which can introduce unwanted insertions and deletions at the break site. We will pursue multiple strategies to enhance homologous recombination and to minimize undesired NHEJ:

5.1.1(iii.a) siRNAs and overexpression cDNAs. The Elledge Lab at Harvard Medical School has already performed a screen which has identified siRNAs that can enhance double-strand DNA break-induced HR of a GFP reporter gene. With them (see Letters of Support), we will test whether siRNAs discovered in this screen can also improve ZFN-induced HR at endogenous human genes. The Joung Lab has validated ZFN pairs that can induce targeted, highly efficient HR events at the endogenous human *VEGF-A* (107) and *PIG-A* genes and also has plasmids encoding ZFNs targeted to the endogenous human *IL2Rγ* gene (178). siRNAs from the Elledge Lab screen will be tested at these three loci in a variety of different cell types including iPS cells. Cells will first be transfected with siRNAs and then will be transfected three days later with ZFN-encoding plasmids and donor template DNAs which introduce a restriction site. Four days post-transfection of the ZFN-encoding and donor template DNAs, the frequency of HR in the absence and presence of various siRNAs will be determined using quantitative restriction digest/limited-cycle PCR assays as previously described by the Joung lab (107). All PCR-based results will be confirmed by quantitative Southern blot assay as previously described (107). A screen similar to the one performed by the Elledge Lab could also be performed using cDNA overexpression libraries instead of collections of siRNAs. cDNAs identified by this approach could also be tested using the endogenous human gene HR assays described above.

In addition, we can also design a screen which identifies siRNAs or cDNA overexpression clones that inhibit mutagenic NHEJ. To set up this screen, we will construct a cell line which stably expresses a luciferase gene from a single integrated construct. In addition, we will use OPEN selections to engineer ZFN pairs targeted to sequences in the first quarter of the luciferase gene. To validate these ZFNs, we would demonstrate that they can be used to inactivate the luciferase gene in the cell line we create. These reagents could then be used to screen for either siRNA or cDNA clones that inhibit NHEJ. Specifically, we would first

transfect the luciferase-expressing cell line with siRNA or cDNA clone libraries and then introduce plasmids encoding the luciferase-specific ZFN pair. If an siRNA or cDNA clone inhibits NHEJ in the cell, this should result in less inactivation of the luciferase gene and therefore greater luciferase activity.

All factors identified from these screens would effectively transiently reprogram cells to become more "recombination competent", thereby increasing the efficiency of the targeted genome modifications and improving the specificity by limiting unwanted NHEJ-associated mutations.

5.1.1(iii.b) Small molecules. HR is most active in S-G2 and cell cycle arrest with vinblastine has enabled the highest rates of HR reported to date (107, 178). However, this drug also induces over 95% cell death when used in conjunction with ZFNs (M. Maeder et al., unpublished data). We will explore whether other agents that induce cell cycle arrest such as indirubin or hydroxyurea can also induce higher levels of ZFN-enhanced HR without the higher toxicity observed with vinblastine. In addition, we will also test whether small molecular inhibitors of NHEJ-specific components (e.g., the DNA-PK inhibitor NU7441) might reduce the mutation frequency due to NHEJ and enhance HR events. We will test the effects of these various small molecules using ZFNs and donor templates which introduce restriction sites at the endogenous human *VEGF-A* (107), *PIG-A*, and *IL2Rγ* (178) genes. Initially, we will use a restriction digest/limited-cycle PCR assay (previously validated by the Joung Lab) to quantify HR frequencies at these loci. For those compounds that show increased HR, we will also use limited-cycle PCR/sequencing assays to assess both HR and mutagenic NHEJ frequencies simultaneously. Our expectation is to identify compounds that increase HR and diminish mutagenic NHEJ. If these targeted approaches do not yield results, we can also perform unbiased screens of small molecule libraries to identify compounds that enhance ZFN-induced HR or diminish ZFN-induced NHEJ in cell-based screens similar to those described above for siRNAs or cDNA libraries. In the unlikely event that these screens are non-productive, we will explore inhibitor-free methods that use FACS to synchronize cell populations in S-G2 phase prior to introduction of ZFNs and donor templates.

5.1.1(iii.c) Longer donor DNA templates produced with MAGE. To date, ZFN-induced HR performed by the Joung Lab and others have used donor templates with relatively short homology arms (typically ~1.5 kb of total sequence). Studies of gene targeting in mouse and other organisms have used donor templates with significantly longer homology arms. Our expectation is that the use of donors with longer arms should lead to increased frequencies of ZFN-induced HR. We will explore whether 100kbps-sized donor DNA from YACs or BACs (introduced via conjugation) can improve the efficiency of HR. These longer donor DNAs will be made using MAGE technology (see section (i) above).

Potential problems and alternatives: Based on our extensive experience with ZFNs we expect considerable success using them to induce HR of altered gene regulatory regions. If the above steps do not lead to sufficient consistent improvements in HR efficiency, we will pursue the following additional methods (recalling that in Aim 4.2 we will also be exploring more advanced methods for achieving improvements): *(d) Tethering of donor DNA templates.* To stimulate faster kinetics of repair, we will create protein-based tethers comprising two zinc finger domains - one that binds the endogenous targeted allele and one that binds the donor DNA template. This approach may both enhance HR and permit use of lower levels of donor DNA, thereby potentially reducing the frequency of random integration of the donor template. *(e) Recombination "hot spots".* At present, it remains unknown whether chromatin state can affect the ability of ZFNs to enhance HR. Experiments at the human *HoxB13* gene performed by the Joung Lab suggest that an expressed, open-chromatin state may positively influence the efficiency of HR (107). To explore the hypothesis suggested by this observation, we will design ZFNs that target regions adjacent to DNaseI hypersensitive sites identified from genome-wide surveys. Our expectation is that such sites might serve as potential recombination hot spots for ZFN-induced HR. An understanding of the relationship between chromatin state and HR frequency will enable better choice of target selection to maximize gene targeting efficiencies.

5.1.1 (iv) Performance evaluations: For each method we develop above, we will measure recombination efficiencies and also recombination biases using subsets of samples and cell lines generated for a number of genes. It will be especially important to measure these for combinatorial cell populations, as knowledge of efficiencies and biases will be important for accurate simulations and statistical analysis that will guide our isolation of sets of clonally altered cells (e.g., see Figure 5-2 above), as well as the single cell genotyping / ASE assay we will develop in Aim 1.2 (section 5.1.2 below). When analyzing populations of cells in which multiple loci have been modified, we will use our published single molecule gel-polony method for long-range haplotyping (201), which will allow us to identify and phase modified loci for the cis regulatory region and coding regions up to the indicator SNPs. This method is ideally suited to giving us complete

information on the distribution of modifications across the regions.

For MAGE-BAC/ZFN, we will characterize the relative rate of random insertions and off-target modifications *vs.* targeted regulatory region replacements. Because our methods in section 5.1.2 look at many independently altered cells, random insertions and off-target events are not expected to have significant impact on our ability to analyze causality unless they occur with high incidence or operate with high bias on specific regions of the genome. Random insertions can be gauged easily by performing *in-situ* hybridizations with labeled probes from the regulatory regions: multiple integration events should be easily detectable as cells with more than two spots representing instances of the regions. We will characterize off-target events in a representative set of ZFNs by the following method: (a) For each ZFN, we will construct an inactivated version of ZFN (iZFN) that contains the same zinc finger arrays as the active one, but where these arrays are joined to a version of *Fok*I whose catalytic domain is inactive. Each iZFN should bind the same locations as its corresponding active ZFN (aZFN) but will not cut the DNA at these locations. (b) We will introduce each iZFN into one or more CTCHGV cell lines and then use Chip-Seq (74, 82, 179) to identify all locations in the genome to which the iZFN binds and to estimate the percent occupancy of the iZFN at these locations. (c) After identifying all target locations with significant iZFN binding, we will design padlock probes that target these locations using procedures in use in the Church Lab (see Preliminary Results, 4.1). (d) Finally, we will introduce the corresponding aZFNs into the CTCHGV cell lines used in (b) and perform targeted sequencing using the padlock probes designed in (c) to obtain actual genomic DNA sequences of locations likely occupied by the aZFN. We will examine these sequences for evidence of non-homologous end joining (NHEJ) events (see Preliminary Results, 4.5, and Figure 4.5-2) and to estimate the frequency of these events at each location. To estimate the frequency of off-target recombinations and NHEJ events relative to on-target recombinations, we will modify (d) by introducing template DNA into the cells along with the aZFN, where the template has a shorter region of homology with the aZFN's genomic target than the genomic sequences targeted by the padlock probe arms. This will ensure that, during the padlock probe capture reactions (Preliminary Results, 4.1), the probes will copy genomic DNA that contains junctions between the template and genomic site of template integration. These junctions will uniquely specify the locations of all template integrations. Note also that the Joung Lab is planning to submit an R01 grant for development of this method and has already obtained preliminary results for steps (a) and (b) (see Preliminary Results, section 4.5).

Potential problems and alternatives: Based on our experience with and Preliminary Results for the methods indicated, we do not anticipate significant problems gathering the performance data described above.

5.1.2: Aim 1.2: We will demonstrate the identification of specific sets of variations that affect cis gene transcription by engineering many combinations of variations and directly observing their effects on transcription, and also by novel methods of assaying complex populations of combinatorially modified cells at a single-cell level.

As described in the Research Design Overview, the basic idea for identifying which cis loci differentially affect transcription is simply to examine transcription levels of cis genes or alleles among cell lines with altered cis variants: those loci for which changes in the cis alleles have no effect can be eliminated from consideration, while those which do have an effect have causal relevance and can be analyzed for interactions with other loci. As noted in the Overview and in Figure 5-1, we will develop two methods for performing this analysis, a simple one based on clonal altered cell populations (the left branch in Figure 5-1), and a potentially more efficient and scalable method that analyzes entire combinatorial cell populations at once at a single cell level (the right branch). We describe these in turn in (ii) and (iii), after describing initial analyses we will perform to identify the sets of genes and cis variants we will consider for the course of the project.

5.1.2(i) Initial identification of genes and cis variations. The purpose of this component is to identify the genes and regulatory regions to be analyzed in all Aim 1.1-1.2 work and to develop associated resources. As already noted (see Research Design introduction), we will use pre-existing, publicly available tissues and cell lines with potential to be transformed into iPS from HapMap, the PGP, the Framingham Heart Study, or other sources. We will give preference to samples for which comprehensive genome sequence is available, particularly if the sequence is diploid. If diploid sequences are unavailable, we will obtain diploid sequences of a large number of genes and their regulatory regions, either by applying the long-range haplotyping methods described in Preliminary Results 4.2, or through our collaborator Complete Genomics, Inc (see Letter of Support). It will be important to use *clonal* populations of cells from these subjects as our methods depend extensively on analysis of differences in ASE, and different clones of the same cells have been observed to be

subject to high levels of random allele inactivation (50). The replicates we use should be separate cell lines cloned from the same subject so that they likely represent different clones. In Aim 1.3 we will compare these replicates to assess the impact of random allele inactivation may have on our results, generally. From the diploid genome sequences it will be straightforward to identify indicator alleles for all genes with heterozygous coding regions, and to identify all SNPs and other non-reference sequences in regulatory regions. Our selection of regulatory regions will be from among 100kbps regions upstream of the transcription start site, introns plus segments of adjacent exons, 100kbps downstream regions of documented transcription termination sites, or documented enhancer regions cis to but outside of these distance bounds. From the ~100 variants that might be expected by chance to be in any 100kb regulatory regions, we will use available information (e.g., from the USCS Genome Browser (81, 93)) on known transcription factor binding sites, conservation, GWAS associations, and other data, to prioritize variations and pick ~5 or less for follow-up analysis. In selecting genes, regulatory regions, and indicator alleles, we will attempt to leverage resources already developed in (199) where possible. We will likely pick one regulatory region for most genes, and will prioritize genes and regions that are known to be associated with human diseases or traits (taking into account cell types and GWAS that will be considered in Aim 2.2, section 5.2.2), which have little repetitive sequence, and for which unique priming sites with good hybridization properties can be designed for cis variants and indicator alleles.

Potential problems and alternatives: The only potential problem we can foresee is if delays arise in obtaining *diploid* sequences for our subject cell lines. In that case we would develop a provisional prioritized list of genes and variations based on imputed haplotypes, and refine this list as diploid sequences become available. As we have only targeted analysis of 50 genes by the end of year 2 (see Intermediate Goals below), there will be ample time to refine the list. Also, in 5.1.2(ii) below, it will be seen that considerable analysis can be done with knowledge only of genotypes vs. haplotypes of variant loci in regulatory regions.

5.1.2 (ii) Assay via clonal altered cell populations. The key elements of this strategy have already been described in the Research Design Overview; here we give only some additional details. The first step is to develop, for the gene at hand, a set of altered cell populations that are clonal for a sufficient set of combinations of cis locus modifications to be able to identify causation and interaction. If cell lines are altered for individual cis loci one at a time, one of the strategies described in Aim 1.1 (section 5.1.1), clonal populations for each combination will either be the direct outcomes of the modification procedure or will be easily created from them. For instance, if we use MAGE-BAC/ZFN to make a single variant **A**|**a** locus homozygous **AA**, the resulting population will be clonal, while if we use MAGE-human and supply **A** and **a** oligos together, we will get all four possible cis regulatory alleles: **A** and **a** on the allele with indicator SNP **x**, and similarly for indicator SNP **y**. Clonal populations for each set of haplotypes can be easily isolated. However, if instead of altering loci one at a time, we generate a population combinatorially modified for all loci, Figure 5-2 above and the surrounding discussion show that by isolating and growing out modest numbers of single cells (~60) from the population, we can get a sufficient set of populations to identify simple causality with near statistical certainty.

The great advantage of the clonal altered cell population strategy is the simplicity of the assays performed on each clonal population. All that is required is that the cis regulatory genotype of each clonal population be identified along with the expression levels of the cis gene. Given that at most five loci will be modified in any gene regulatory region, the genotypes can be identified by five allele-specific genomic DNA PCRs. Expression level can be analyzed either at the overall gene level, or at the allele-specific level; we will investigate both options. When considering overall gene expression relative to a locus **A**|**a**, significant correlation between overall gene expression for **AA** vs **Aa** vs **aa** genotypes identifies the locus as one in which the allele has causative significance for expression. When considering allele-specific expression levels (ASE), the test is to see whether ASE disappears in homozygous **AA** or **aa** genotypes but is present in **Aa** genotypes. In the simulation of Figure 5-2, the correlation strategy exhibited better sensitivity than a t-test comparing ASE levels between homozygous and heterozygous genotypes. However, this outcome may be dependent on assay variance and other factors, so both overall and ASE measures will be considered. Both may be measured by simple RT-PCRs, with ASE requiring PCRs that are specific to the indicator alleles in the coding regions. Notice that all of these statistical tests require knowledge of only cis region genotypes vs. haplotypes. This relieves us from having to do additional assays to learn the phasing of the various modifications.

Potential problems and alternatives: We anticipate no significant problems with the approach above, except that assessment of interactions between loci may require large numbers of isolated single cells. The approach developed in section 5.1.2(ii) below will be our principal effort for dealing with this possibility.

5.1.2 (ii) whole population assay: In Aim 4.3 (section 5.4.3) we describe plans to develop advanced cell-handling capabilities which will enable millions of cells to be arrayed on the surface of a flow cell, where they can be both assayed biochemically and morphologically via image analysis. Our plan is to use this capability to array a population that has been combinatorially modified for all five targeted cis loci and assay each cell individually for genotype and allele-specific expression (ASE). The rationale for this proposal is that, although the single cell assays may exhibit high error individually, this error can be overcome by aggregating measurements for millions of cells.

Specifically, after arraying the cells on the flow cell surface, fixing, and permeabilizing them, we will add reagents and oligos required to intracellularly amplify each of the $n \leq 5$ cis regulatory sites modified for the gene at hand and perform the multiplex amplification for all cells simultaneously. This step is then followed either by probing or small scale in-situ sequencing that identifies the alleles present at each regulatory site. Several methods will be examined and evaluated for sensitivity, sensitivity, complexity, and cost. The simplest method is to introduce 2*n* primers that amplify the *n* sites *in situ*, and then perform sequential one-base primer extensions using labeled bases, a method that was successfully applied in gel polonies (203). Another class of methods involves allele-specific amplifications within the cells using two allele-specific primers and a common second primer, where the 3' ends of the allele-specific primers correspond to the alleles and a distinct sequence barcode is affixed to the 5' ends of this primer. The sequence barcodes for each allele can then be interrogated by appropriate small scale sequencing as above, or by *in situ* hybridizations with labeled probes. Several forms of allele-specific amplification can be evaluated, including ordinary allele-specific PCR, padlock probes with nested common and allele-specific PCR primer sequences together, or padlock probes followed by rolling circle amplifications. In developing this protocol, consideration will be given to methods that make the DNA accessible, such as proteases, detergents, and, possibly fragmentation of genomic DNA (27). The read outs obtained from this step are used to classify the cell for the presence of an allele. If only one allele is detected, that allele will be considered to be present with copy number 2 and the other allele to be present with copy number 0. If both alleles are detected, copy numbers of 1 will be assumed for each. Duplicated or repetitive sequences will have been filtered out in initial selection of the genes of interest (see (i) above) and copy number variation is considered in section 5.1.3.

At this point, interrogation of the cis regulatory genotypes is complete, and the next step is to assay ASE for the gene transcript. The cells are now treated with DNAse to destroy the genomic DNA corresponding to the indicator alleles. Methods akin to those used above are now applied to amplify and interrogate the indicator SNPs present in the transcript coding regions that identify the transcript alleles, except that the amplification must begin with reverse transcription. Because the cis region loci have been randomly re-assigned, a genotype heterozygous in a locus **A|a** may be present in both haplotypes, such that the **A** allele may be cis to indicator SNP allele **x** in one cell but cis to the other indicator SNP allele **y** in another. Because we are not resolving the haplotypes in this assay, ASE must be measured in a symmetrical fashion such as abs(log(I_x/I_y)), where I_u represents signal for the transcript with indicator allele *u*.

parameters	P-value	loci under consideration				
		1	2	3	4	5
α=0.3	0.01	6710	24906	85122	282274	921708
δ=0.2	0.001	10972	38378	127448	414995	1337470
σ=5	0.0001	15353	52143	170587	550051	1760090
α=0.8	0.01	445	1148	2729	6293	14291
δ=0.1	0.001	727	1767	4080	9240	20715
σ=2.5	0.0001	1016	2398	5457	12240	27244

Table 5.1.2-1: Numbers of single cells (N) that must be analyzed for ASE and genotype from a combinatorially altered cell population using the single cell assay proposed in Aim 1.2 to identify, with various P-values, that one of up to five modified loci is responsible for ASE, assuming completely random haplotypes among cells that were successfully modified, that only one of five modified loci controls ASE, that all loci are independent, and two sets of performance parameters (see text for discussion and comparison).

Three factors will control the performance of this population assay: the efficiency of altered cell generation (α), the standard deviation of the error of single cell ASE measurements (σ), and the probability of single cell genotyping misreading a genotype (δ). We assume that the predominant genotype misreading error will be failure to detect one allele. This type of error is important because it potentially has a large impact on the statistics of comparing ASE differences between genotypes, for a locus should cause ASE only when it is heterozygous, not when it is homozygous, and this error makes heterozygous loci appear homozygous. Table

5.2.1-1 provides an estimate of the numbers of cells that must be evaluated to detect which of up to 5 loci may be causative of ASE with P-values ranging from 0.01 to 0.0001, given two sets of performance parameters: Table 5.2.1-1 (top rows) considers a conservative set of performance targets—an α of 30% that has already been achieved in some cases (see Research Design Overview and section 5.1.1), with high error single cell assessment (δ = 20% and σ = 5x maximum normalized ASE), while Table 5.2.1-1 (bottom rows) represents concerted improvement whereby α is improved to 80% while the error rates are cut in half. While the numbers of cells required to detect cis loci interacting according to models considered in Table 5-1 has not been specifically modeled, the low numbers indicated in Table 5.2.1-1 suggest that good statistics for such interactions are indeed achievable.

The computational analysis required by this assay is as follows: Images of single cells arrayed in the flow cell will be acquired and intensities, obtained in each cell for the various cis loci and allele-specific transcript labeled mini-sequencings or in-situ hybridizations, will be calculated as image analysis features using standard methods already in use in the Church Lab (e.g., (7)). Various additional features based on additional stains such as DAPI may be obtained to determine cell integrity, cell cycle state (which could alter ploidy), or flow cell position occupancy by other than a single cell. If images are acquired for cells arrayed in random positions (e.g., for cells arrayed randomly on slides), these other stains will be used to segment the image into cells and to exclude any segment that is not an isolated single cell. The multiple images acquired will be registered, and genotypes and ASE measures will be computed for each cell as described above and in section 5.1 Overview. Evaluation of genotypes and ASE will employ intensity (for genotypes) and intensity ratio (for ASE) thresholds developed from original CTCHGV subject cell lines that are clonal with respect to genotype. Meanwhile, estimations of α and the distributions of haplotypes generated in combinatorially modified populations will come from Aim 1.1 (section 5.1.1 (iv)), which will also provide information on random integrations vs. replacements of native cis regulatory regions. If random integrations are too common, additional steps will be taken to filter out cell segments that have excessive intensities from genotype images, or more than two genotype intensity maxima. Parameters δ and σ will be estimated from images acquired of non-modified original CTCHGV subject cell lines. Simulations using these parameters will be used to estimate the numbers of cells needed to distinguish between different cis interaction models (exemplified in Table 5-1).

Potential problems and alternatives: The intracellular assays present the key challenge. However, given that only a maximum of 6 loci (5 regulatory DNA and one transcript) need to be interrogated per cell, and that methods similar to what we require have already been developed (163, 202), we believe prospects for success are high, especially given that we will be developing more powerful *in situ* intracellular Rolling Circle Amplification methods in Aim 3 (section 5.3.1.2). The simultaneous querying of genotype and coding transcript will be a new element. Here we can explore modifications to the protocols described above that can eliminate possible complications. For instance, suitably designed ZFNs may be used to cut the genomic loci corresponding to indicator SNP alleles vs destroying all DNA via DNAse: Then, RT-PCR can be performed on the gene transcripts without incurring contaminating signal from the corresponding genomic sites, and without destruction of the amplicons created from the cis regulatory loci. Finally, image analysis can be exploited at many levels additional to those described above to filter out any cells for which signals representing genotype content and transcript level cannot be interpreted. For instance, if the readout of cis regulatory locus genotype described above based on detection of the presence of each allele proves error prone, we can add additional conditions such as requiring that, where only one allele has been detected in a cell, the intensity must be ~2x the intensity found for that allele in cells where both alleles have been detected.

5.1.3: Aim 1.3: We will assess the extent to which cis variants identified as causing altered transcript expression may operate through alternative mechanisms such as differential expression of RNA isoforms, differential transcript degradation, copy number variations, and epistatic marks.

Cis variants detected as causing allelic expression bias could actually operate through other mechanisms. It is also possible that observed ratio differences in allelic expression only arise in certain biological contexts. To address these issues we propose using a variety of established techniques to check for the contribution of each of these aspects to our data. These will include splice variants, copy number variants, epigenetic context and allele specific epigenetic differences, and possibly other factors. These phenomena could potentially affect both our measures and our conclusions. For instance, regarding measurement, apparent allele-specific expression (ASE) based on padlock capture of indicator SNPs in the two alleles of a gene could actually be caused by differential splicing vs. differential expression whereby one allele contains

more isoforms carrying its indicator SNP exon than the other (170) (see also (203)). Regarding conclusions, it could be that our identification of a regulatory region cis variant as a cause of differential allelic expression might actually be artifactual and instead, that the construct with which we introduced altered cis sequence disrupted normal methylation patterns. Our primary objective is to assay these phenomena in a subset of our samples to quantify the extent to which they may affect our measurements and our identification of causative cis variants, not to comprehensively assess their impact. This subset will include original, unaltered CTCHGV subject cell lines and a selection of altered samples cloned from our combinatorially modified populations for a small set of genes. We will consult with the Center for the Epigenetics of Common Human Disease CEGS regarding our investigations of epigenetic impacts.

To assess the extent to which measured ASE may be due to differential splicing, we will use RNA-seq (185) or the Affymetrix Human Exon Array (http://www.affymetrix.com/products_services/arrays/specific exon.affx#1_1) to detect splicing variants. More targeted and cost effective approaches, such as custom exon arrays, PCR or the padlock probe-based capture (100, 139), could be used for a selected number of genes. If alternative splicing contributes to the apparent ASE, we will expect to observe the exon (or part of the exon) carrying the SNP marker more frequently in the apparently more highly expressed allele than the other, and to observe more alternative splicing junctions that skip the exon in the less highly expressed allele. These experiments performed on altered sample clones will also reveal the extent to which cis variants we have identified as causative may actually alter splicing vs. expression. Measurements of ASE may also be perturbed by random allelic inactivation that differs between clones (50). We will assess this by comparing unaltered replicate CTCHGV sample cell lines, which should not represent the same clones.

Copy number variations (CNVs) in which gene and regulatory region alleles may be amplified or deleted may complicate inferences of causality based on statistical models such as illustrated in Tables 5-1 and 5.1.2-1. By contrast, our measures of ASE should be normalized for CNV. We will use comparative genomic hybridization (CGH) arrays (such as http://www.nimblegen.com/products/cgh/) or massively parallel sequencing (25) to detect copy number variation in a selected number of samples before and after recombination.

To assess for allelic bias due to differences in allele methylation or histone modification, we propose to detect epigenetic state in an allele specific manner. This can be done with targeted sequencing capturing locations that contain a heterozygous site of variation (e.g., a SNP). To detect DNA methylation, similar to our recently published methods (10, 37, 199), we will target ~200-bp regions in bisulfite-treated DNA containing an altered site, an unaltered variation site (see section 5.1.2(i)), and a CpG site. The target size of ~200-bp is chosen because it is within both the capability of padlock probes (139, 199), and the read length of the current Illumina paired-end sequencing platform. Allele specific differences in histone modification will be detected in a similar manner by applying padlock probes encompassing variation sites to chromatin immunoprecipitated (ChIP) DNA produced with an antibody to the histone modification of interest (e.g., H3K4 methylation or H3 acetylation). Allele-specific detection of ChIP DNA has already been successfully performed in a microarray context (110) and should be readily translated to sequencing-based methods. If only a very few such regions can be targeted in this manner, we will need to apply our haplotyping methods based on sequencing of dilute DNA preparations (see Preliminary Results and section 5.1.2) to bisulfite treated DNA to obtain the allele-specific methylation profiles. The regions of interest can be targeted by tiled padlock probes to reduce overall haplotype sequencing requirements. In our original, unaltered cell lines, these experiments will be informative as to whether initially measured ASE may have been due to different allelic methylation patterns. We can also trap high molecular weight genomic DNA in polyacrylamide gels, perform *in situ* bisulfite conversion and polony amplification, and then identify alleles and quantitate methylation levels using single base extensions, using the method of (203). It is also possible that causal cis variants only contribute to differences in expression in certain epigenetic contexts; for example, a relevant transcription factor may only bind in the context of particular histone modifications (53). In our altered sample clones, allele specific epigenetic measurements may identify cases where cis variants affect expression level by means of altering local methylation, or cases where recombination or MAGE oligos (see section 5.1.1) used to create the altered cell have locally disrupted epigenetic state in an incidental manner that is unrelated to introduced sequence variants. To distinguish between these alternatives, we must assess whether altered methylation travels with the introduced DNA or the cis variant.

Potential problems and alternatives: We anticipate no significant problems as the techniques described are well-established or methods with which CTCHGV investigators have considerable experience.

5.1.4: Aim 1.4 We will analyze the relationship between our methods and results and those of Genome Wide Association Studies and characterize their complementary insights into the effects of variation.

GWAS and other studies have identified associations between SNPs and gene expression levels (40, 121, 149, 150, 165, 182), and some of these findings will be used to prioritize genes we will analyze for causal regulatory variants (section 5.1.2 (i)). Here we will examine our findings from the GWAS side and ask what it would take for GWAS to be able to identify cis causal regulatory variants that we have identified. This analysis will clarify the sensitivities, specificities, and amounts of effort required for GWAS vs. the engineering methods developed here to discover and characterize cis regulatory variants controlling gene expression. Once CTCHGV has discovered a set of causative cis variants, we will attempt to estimate the population frequency of the variant, its haplotype block, and its effect size. For any variant that happens to be assayed on platforms designed for GWAS, its frequency should be easily assessed from available GWAS data, but most variants will likely not have been assayed. For these, we will examine HapMap samples. If sufficient sequence data are available (67) we will estimate the allele frequency from the sequences; otherwise we will measure the frequency from HapMap cell lines. To estimate effect size we must consider the tissue from which the variant was identified in the CTCHGV subject. If matching tissue data is available from genotyped samples from the CTCHGV subject's population (125), we will use it to estimate effect sizes and variances. Otherwise we will approximate the effect size based on the degree of differential expression by which the variant was identified by CTCHGV and apply available information to estimate variances (26). Finally, we will consider two models for GWAS. In the common variant model, we will assume that GWAS is performed with tag SNPS from commonly used array platforms. We will characterize the haplotype block containing the variant, idenfity the tag SNP in greatest linkage disequilibrium (LD) with it, and estimate the population sizes that would be needed to find an association between expression level and the tag SNP given the LD and effect size, using standard statistics and tools (9, 21, 35, 80, 85, 123, 142, 204, 205) for partial and for whole genome searches for cis effects (165). For the rare variant model, we will use the frequency of the variant itself, the effect size, and make comparable computations, this time considering corrections for limited candidate gene sets (126) in addition to partial and whole genome searches for cis effects.

As a second related analysis, we will consider that the expression level change identified for the variant is itself associated with a disorder with one of a fixed range of penetrances, and estimate the population sizes that would be needed to associate the variant with the disorder by the GWAS models above.

Finally, as noted above (Research Design Overview), CTCHGV will not itself conduct GWAS or population studies. However, CTCHGV will communicate with partner Centers and collaborators who do conduct GWAS (such as the Broad Institute) concerning cis variants found to causally impact gene expression level. Our partners will then be able to explore refined hypotheses for the phenotypes studied by the GWAS, and in turn may be able to assess population frequencies for the variant, furthering the analyses above.

Potential problems and alternatives: We anticipate no significant problems with this sub-Aim. It involves applying standard GWAS tools and methods to parameters determined by CTCHGV experiments.

Aim 1 Goals: Final goals: As noted in Research Design Overview, our final goal is the identification of single and/or combinations of natural variations in regulatory region sequence that control differential cis gene expression using the methods described above for 1000 genes. The survey of causal mechanisms in Aim 1.3, and the analysis of relationship between new methods and GWAS in Aim 1.4, will consider representative subsets of genes and variants. Intermediate goals: Again as noted in our Research Design Overview, we will evaluate progress at the end of year 2 of the Center and renegotiate goals as appropriate. We expect we will have processed ~50 genes by that time. Impacts: In addition to increasing biological knowledge about the regulation of any specific genes we have analyzed by identification of causal cis variants, CTCHGV will have developed methods for precise engineering of human cells through ZFN-mediated homologous recombination and oligo-mediated recombination that will have general and impactful application to the understanding of human disease, gene therapy, and personalized medicine.

5.2: Aim 2: We will adapt and extend Aim 1 methods to function in human induced Pluripotent Stem cells (iPS) and then use iPS to characterize the effect of cis regulatory region variations in a variety of derived cell types that represent different human tissues. We will engineer "marked allele" human iPS that are heterozygous in all exons of many genes that will enable analysis of allele-specific

transcriptional and splicing effects in diverse cell types.

Overview: The work of Aim 1 depends on the ability to engineer alterations into human cell lines efficiently. Thus, to achieve the goals of Aim 1, CTCHGV will use robust human cell types that tolerate the protocols that implement these engineering steps. To extend these methods to other human cell types via iPS, either these protocols will need to be modified for iPS, or modified Aim 1 cell lines will need to be transformed into iPS. Once CTCHGV develops methods for generating these iPS, we will use them to explore the effects of cis regulatory variants in different derived cell types. The justification for CTCHGV use of iPS-derived cell types vs. primary tissues from humans or other animal model organisms has been provided in Background and Significance 3.3 and the Research Design Overview above. These goals both imply a strong need for automating methods for generation, maintenance, and control of multiple populations of iPS, and efficient methods for differentiating them. This will be the focus of Aim 2.1, while Aim 2.2 will then fulfill the goal of exploring the effects of variations in different cell types. In Aim 2.3, we will use Aim 1 methods to make complex modifications of iPS and develop a potentially highly useful resource for the research community – the "marked allele" iPS. Through this we will both expand and discern the limits of how far one can engineer iPS.

5.2.1: Aim 2.1: We will combine Aim 1 methods with automated techniques for iPS generation and maintenance to enable exploration of iPS with altered cis regulatory regions.

While iPS reprogramming is now widely practiced and becoming more routine, high throughput and rapid generation of iPS cells for large functional studies will require improvements in efficiency and cost reductions. In the present case, we will utilize a retroviral "monovector" that expresses all four factors necessary for efficient iPS reprogramming from one polycistronic expression cassette (77, 153, 189). We are also incorporating small molecules to enhance iPS reprogramming, as reported, thereby enabling even higher reprogramming efficiency from hair, skin, blood (1, 30, 62, 157, 192). We will grow human keratinocytes or fibroblasts directly on 77 micron algae-based microcarriers containing magnetic beads (Global Cell Solutions). These microcarriers are controlled using a magnetic field during media changes, aeration, and stirring. We will develop instrumentation to couple the delivery of viral and small molecules to automated high-density cell culture for iPS reprogramming on the microcarriers. We will use Complex Object Parametric Analysis and Sorting (Union Biometrica) along with Tra 1-60 and Tra 1-81 cell surface markers to identify reprogrammed cells and sort them into microtiter plates. Note that, with reference to Aim 4.3 (section 5.4.3), iPS cells can be made flat for morphology-based selection by eliminating the alginate complex via transient chelation of Ca++ and Mg++. Importantly, we have observed that hES cells and iPS cells load and proliferate on the alginate complexes without differentiation, and that iPS colonies growing on alginate-based microcarriers can be frozen down without further manipulation. We will develop methods to automate primary cell isolation, iPS cell derivation, and iPS cell freezing and storage. This will enable rapid and affordable distribution of individualized iPS to researchers world-wide. As a proof-of-concept, we will take our original CTCHGV subject samples and a selection of samples modified by Aim 1 and reprogram, derive, and expand iPS cells simultaneously. This will also enable multiplexed exposure of iPS cells to a combinatorial library of differentiation factors (growth factors, genetic factors, small molecules) for directed *in vitro* differentiation and sorting, all directly on microcarriers. By the end of two years, we expect to have a highly efficient, automated platform for generating functional iPS cells and their derivatives, ready for distribution to the research community.

The foregoing will generate iPS with Aim 1 modifications from primary cell samples engineered in Aim 1. We will also attempt to apply Aim 1 techniques directly on iPS developed from original, non-modified, CTCHGV sample cell lines. This will involve testing and optimizing MAGE-BAC/ZFN techniques decribed in section 5.1.1(i.c) and 5.1.1(iii), and also MAGE-human techniques from 5.1.1(ii), on iPS cell lines.

Potential problems and alternatives: (a) While our collaborators have expanded and cultured hES cells on microcarriers for up to 2 weeks while maintaining pluripotency (see Preliminary Results 4.3), *in vitro* manipulations such as transfection, homologous recombination, and selection may affect their ability to maintain pluripotency. These operations may also result in chromosomal abnormalities. To control for this we will routinely sample for human pluripotency surface markers and karyotype clones, and only propagate those with intact pluripotency and normal karyotype. (b) Currently iPS and hES pluripotency is most effectively checked by 2D visualization under the microscope. A 3D culture system may make it difficult to image pure iPS colonies during culture. If so, we will maintain cells in the 3D system during reprogramming and move to ordinary 2D culture after reprogramming or when high definition imaging is necessary. We and our

collaborators will then explore traditional automation and robotics for analysis of the 2D cultures (e.g., using CompacT CellBase at http://www.automationpartnership.com/cb_ibcss/CBsystem_overview.htm).

5.2.2: Aim 2.2: We will differentiate iPS generated in Aim 2.1 into diverse cell types that represent distinct human tissues and characterize the cell type-specific consequences of cis-regulatory variations.

We will proceed on two tracks: (i) We will identify a limited number of genes and cis variant combinations from our Aim 1.1-1.4 set that are implicated in tissue-specific function and, using the methods of Aim 2.1 (section 5.2.1), create iPS populations from subject samples that were engineered in Aim 1 to contain these identified combinations of cis variants for these genes, and observe the effects of variations on cis gene transcription in iPS-derived cell types corresponding to these tissues. In selecting genes and variants for cell type analysis, we will consider the emerging data from large GWAS and ASE studies that identify regions and alleles associated with specific disease phenotypes, with the thought that CTCHGV findings indicating that particular variants cause changes in cis gene allele expression levels may have relevance to research into the corresponding diseases. Cardiovascular diseases, insulin resistance, and obesity are of particular interest not only because numerous associations have been reported (38, 49, 122, 173, 187), but also because methods for *in vitro* differentiation of human embryonic stem cells and iPS into corresponding cell types (cardiomyocytes, endothelial cells, adipocytes, and beta-islet cells) are well characterized. To ensure that CTCHGV can pursue these studies, genes and variants implicated by these studies will be prioritized in the initial selection of genes for which CTCHGV will engineer variations in Aim 1.2 (see section 5.1.1(i)). Once iPS cell populations engineered for combinations of variants for these genes are in hand, we will create clonal isolates of these cells and measure ASE of the corresponding cis genes as in Aim 1.2, and we will do this again after differentiation of the clones into our target cell types. We will compare these ASE profiles with each other and with the profiles already developed from the somatic cell-based populations assayed in Aim 1.2, and we will identify each clone and gene that shows reproducible changes in ASE after differentiation compared to its original unengineered or corresponding engineered original subject cell lines. These changes in ASE for the cis genes will be indicative of upstream regulation that depends on both regulatory region allele and cell type, against the genetic background of our original CTCHGV samples. For these genes we will attempt to identify transcription factors that bind to the engineered sites (170), look for evidence in the literature that they are differentially expressed in corresponding primary tissues, and assay for corresponding changes in expression in our differentiated and undifferentiated cell lines. If a biological effect is documented for the cis genes or the upstream factors, we will test for the biological consequences, including alterations in signaling pathways known to play a role in disease pathophysiology. (ii) Using the single cell ASE / genotyping analysis developed in Aim 1.2, *supplemented with transcriptional assays for tissue-specific expression markers*, we will attempt to simultaneously identify causative cis alleles in multiple cell types developed from the combinatorial iPS populations developed in (i), thus multiplexing Aim 1.2 over both cis variant combinations and cell types.

Potential problems and alternatives: We do not anticipate significant difficulties with (i) as these methods have been generally extensively tested using iPS and human and mammalian embryonic stem cell lines. (ii) If reliable cell type-specific expression markers are available, this will generate few difficulties additional to Aim 1.2(ii) (section 5.1.2(ii)).

5.2.3: Aim 2.3: We will engineer human iPS with "marked alleles" for 10-50 genes and demonstrate their use by characterizing allele-specific transcription and splicing in multiple tissues.

While GWAS/eQTL are currently used to identify genomic loci controlling RNA expression level, the methods of Aims 1.1-1.2 will enable dissection of cis regulatory control down to the nucleotide level, and Aims 2.1-2.2 will make these methods applicable to iPS cells and multiple cell types. Here we will further engineer human iPS cell lines using Aim 1.1-1.2 and develop additional methods that will enable measurement of the effects of sequence variations on RNA transcript structure, function, and cellular phenotype. Finding subtle ASE and expression profile variants with specific changes in allele-specific isoforms and/or function in multicellular environments will be important in itself and will greatly aid in identifying the causal ASE and ultimately the causal nucleotides. As a proof of concept, we propose creating a systematically marked allele-specific genome for a set of 10-50 genes in a subset of our CTCHGV iPS lines. For each gene, an indicator SNP distinguishing each allele will be engineered into a degenerate codon position in each exon where natural SNPs are not already present. Thus, every exon in both transcripts from these trait-associated loci will now be amenable to interrogation in an allele-specific manner. These changes will enable the exon distribution of

each transcript allele to be characterized, revealing the presence of allele-specific splicing regulation. These capabilities will enable us to investigate whether allele-specific RNA is cell type-dependent, which could provide many insights into the functional consequences of human variation. In conjunction with Aim 2.2 (section 5.2.2 above), we will incorporate cis variants relevant to cell type into these marked iPS cell lines to assess both ASE and *isoform profile* in cell lines differentiated into particular cell types. We will compare these results with isoform profile information obtained in Aim 1.3 from original CTCHGV sample tissues, to identify cell type specific isoform profile changes in our marked genes. Additionally, we will produce marked allele iPS lines for multiple CTCHGV samples and analyze these for differences in cell type-specific isoform profiles and ASE among individuals.

Potential problems and alternatives: Since "marked allele" cell lines can generated by successively engineering one gene at a time, there is no problem in principle to achieving this goal using the methods of sections 5.1.1 and 5.2.2. Marked exons may potentially affect mRNA secondary structure and, through this, mRNA processing (92). To limit this possibility we will computationally screen possible "marks" in each exon and implement only those that are predicted to have minimal impact on secondary structure (39, 109, 206).

Aim 2 goals: Final goals As noted in our Research Design Overview, we intend to analyze allele-specific expression using engineered iPS cell lines in 50 genes in three iPS-derived cell types. For sub-Aim 2.3, we will generate iPS lines with marked alleles for a collection of 50 genes in 3 subject cell lines. Intermediate goals Again, as noted in our Research Design Overview, we will evaluate progress at the end of year 2 of the Center and renegotiate goals as appropriate. We expect we will have engineered marked alleles in 5 genes in one subject iPS cell line at that time. Impacts Aside from the direct biological knowledge gained from our analyses of specific genes and subjects, CTCHGV-developed methods for automation of maintenance and differentiation of iPS lines, and for engineering iPS with precise genetic changes, will establish broadly enabling technology for research into disease processes in diverse tissues, and into gene therapy and personalized medicine. We foresee that the creation of useful "marked allele" iPS cell lines for *all* genes (vs. 50 demonstrated in CTCHGV) has high potential to become a research community goal, akin to the creation of yeast deletion strains for every yeast gene.

5.3: Aim 3: We will develop novel single-cell in-depth transcriptome assays that are scalable to millions of individual cells in both structured tissues and dispersed cell samples, subject to sequencing capacity. These methods will be used to explore systematic transcriptional effects of genetic variations in different human cell types.

Overview: While single cell technologies are available for gene expression and transcriptome analysis, they are limited by the need to isolate single cells and greatly amplify their minute DNA and RNA content. Laser capture microdissection has greatly improved single cell isolation, and microfluidics improves management of the extremely low concentrations of biological material, but it is still impractical and expensive to isolate and analyze more than a few cells at a time. Here, recent tremendous progress in "next generation" DNA sequencing technology presents significant opportunities, as these technologies have overcome similar limitations by miniaturizing, localizing, and parallelizing operations of DNA capture, DNA amplification, and signal detection. An attractive path forward is to truly integrate DNA sequencing with single cell methods by performing as many of these operations as possible within individual cells rather than in sequencing and preparatory instrumentation. As many different "next generation" methods are available, we will explore multiple approaches for integration. There are also two distinct possible objectives for single cell transcriptomics: (i) *undirected sequencing*, by which one hopes to sequence as many transcripts as possible in every cell in a completely unbiased way, and (ii) *targeted sequencing*, by which one seeks to detect and quantitate large, specific sets of transcripts in every cell, e.g., mRNAs associated with specific biological functions or transcriptional networks. These objectives are complementary, and the choice will depend on the biological problem at hand. We will attempt to develop methods that enable both objectives. The approaches that we will explore are illustrated in Figure 5.3-1, while our strategy is detailed in section 5.3.1. One set of approaches will integrate cDNA synthesis with the introduction of bar codes into cells with single cell resolution, so that the cDNAs bear a sequence tag identifying the cell of origin (see section 5.3.1.1). The cDNAs can be sequenced *via* normal next-gen sequencing. These bar coding approaches will be useful for undirected sequencing of small numbers of cells or targeted sequencing of up to millions of cells. Where structured tissues are under study, we will investigate methods for associating bar codes with cell location prior

to destruction of the tissue. A second approach will develop *in situ* cell sequencing using rolling circle amplification (RCA) of cDNA in either dispersed cells bound to a surface, or in thin tissue sections (see section 5.3.1.2). Finally, to alleviate potential limitations in *in situ* sequencing arising from amplification bias, molecular crowding, and RNA sub-localization, we will explore *in situ single molecule* sequencing (see section 5.3.1.3) by configuring high resolution optical capability into our current Polonator platform (see Preliminary Results, 4.6).

As noted in Background and Significance 3.2, while sequencing capacity is not an inherent limitation to single cell transcriptomes, it may be a practical consideration. Our strategy will be to develop methods by which researchers can control the sizes of the transcriptome subsets that they wish to interrogate, enabling them to use available sequencing capacity to assay large subsets in smaller numbers of cells, or small subsets in larger numbers of cells, as best suits their needs. The technical means for selecting transcriptome subsets will be the choice of the oligos that are used to capture and prime intra-cellular first strand cDNA synthesis. For *undirected sequencing*, the capture oligos will be based on polyT sequences. The 3' degenerate oligo polyT-V (V=A,C,G) will capture all mRNAs and and thus yield complete single cell transcriptomes, but smaller transcriptome subsets can be specified simply using more extended and specific 3' ends. For instance, use of polyT-AA, or of polyT-ACG, will yield ~1/12 and ~1/48 transcriptomes, respectively. By this means, smaller but still unbiased transcriptome subsets can be obtained from each cell, with the choice of transcriptome size and number of cells left for the researcher to decide based on available sequencing capacity. For *targeted sequencing*, capture oligos will be equimolar mixtures of specific sequences designed to target specific sets of transcripts. For such transcript sets, capture sequences will be chosen based on standard criteria such as uniqueness across transcripts, uniformity of Tm, and secondary structure, with attention to exon boundaries and alternative splicing profiles. The methods whereby capture oligos are created and the oligo mixtures that may be used will differ for the three sequencing approaches we will develop. In developing our sequencing

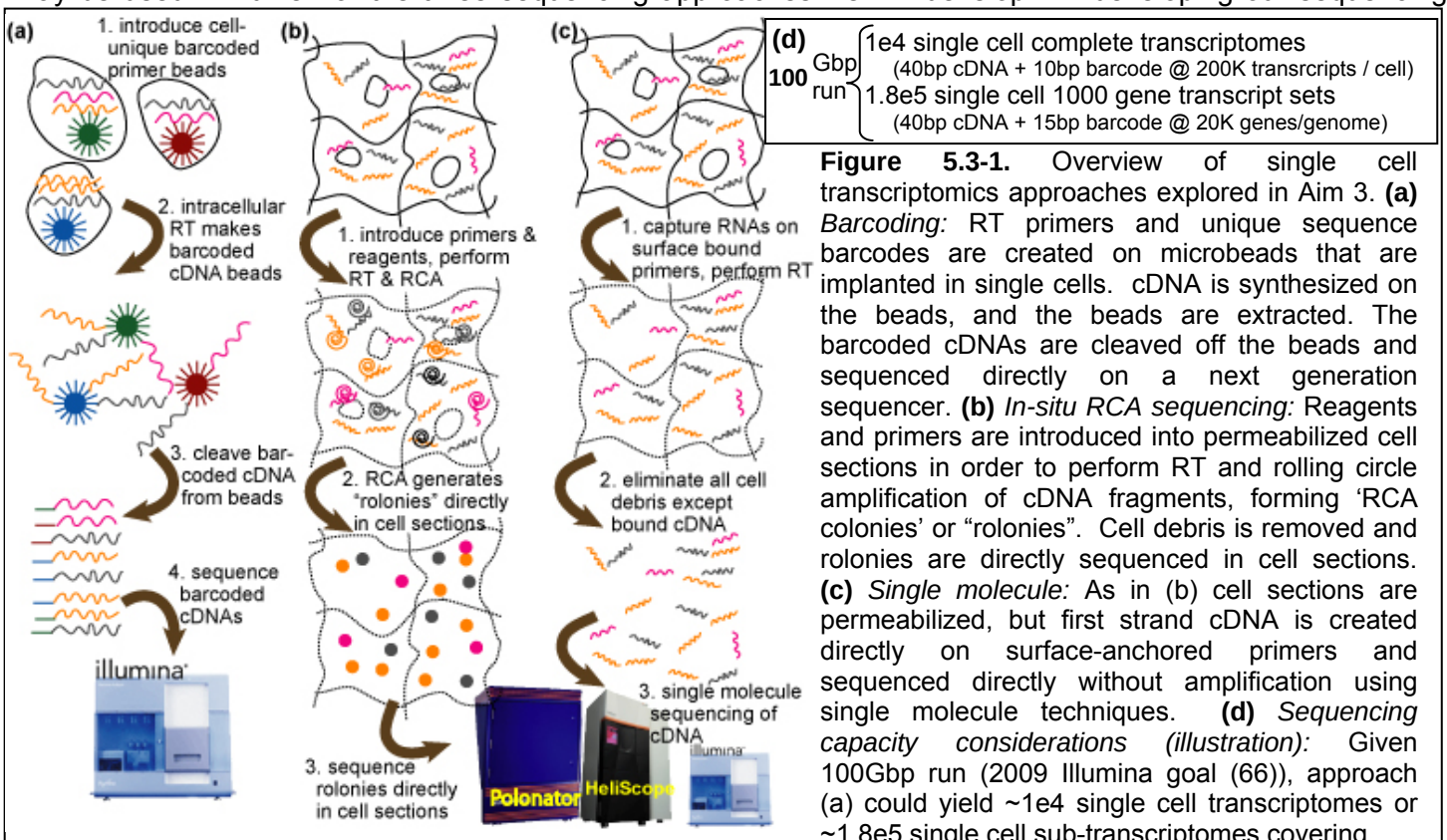


Figure 5.3-1. Overview of single cell transcriptomics approaches explored in Aim 3. **(a)** *Barcoding:* RT primers and unique sequence barcodes are created on microbeads that are implanted in single cells. cDNA is synthesized on the beads, and the beads are extracted. The barcoded cDNAs are cleaved off the beads and sequenced directly on a next generation sequencer. **(b)** *In-situ RCA sequencing:* Reagents and primers are introduced into permeabilized cell sections in order to perform RT and rolling circle amplification of cDNA fragments, forming 'RCA colonies' or "rolonies". Cell debris is removed and rolonies are directly sequenced in cell sections. **(c)** *Single molecule:* As in (b) cell sections are permeabilized, but first strand cDNA is created directly on surface-anchored primers and sequenced directly without amplification using single molecule techniques. **(d)** *Sequencing capacity considerations (illustration):* Given 100Gbp run (2009 Illumina goal (66)), approach (a) could yield ~1e4 single cell transcriptomes or ~1.8e5 single cell sub-transcriptomes covering 1000 specific genes. *Sequencing:* (a) can be performed on any current next gen sequencer, but (b) and (c) will require modified technology. We will develop on the Polonator and migrate to other instrumentation.

approaches, we will focus on undirected sequencing first, and then proceed to targeted sequencing. This will allow us to address the common problem of compartmentalizing mRNA capture in single cells first with simple capture oligos before proceding to more complex mixtures of targeted capture sequences. An illustration of how sequencing capacity might be allocated in different ways is given in Figure 5.3-1d.

Finally, we will develop single cell sequencing approaches for both dispersed single cells and for structured tissues. Here the different methods will differ with respect to how these sample types can be accommodated. We will generally begin testing and development with dispersed cells and move on to structured tissues, where, for convenience, we will do initial work on dispersed cells with human blood cell lines or disaggregated fibroblast cell lines, and initial work with structured tissues using convenient cultured cell lines that have been grown to confluence. As development proceeds, we will switch to using cell types and samples used in Aim 1 (section 5.1) and by Aim 3.2 (section 5.3.2 below).

We expect to develop the three approaches described in Figure 5.3-1 during the first 2 ½ years of our CEGS, and then to determine which to develop further (see *Evaluation of the three approaches* below). The selected approaches will be applied to biological problems under study in CTCHGV Aim 3 (section 5.3.2): Specifically, we will track transcriptome development of cells undergoing de-differentiation to iPS, or differentiation to distinct cell types from iPS. Single cell resolution is important here because only a small and unpredictable subset of cells achieve iPS de-differentiation, and iPS differentiation, likewise, exhibits a strong stochastic component. Examining individual cells may thus reveal molecular transitions that precede and predict these outcomes that may be hard to observe in any other way. Such observations may lead to new methods for efficient control of de-differentiation and differentiation pathways. Additionally, we will test these methods on structured tissues, comparing in-situ transcriptomes from primary human skin (a complex tissue with many cell types) with iPS cells differentiated into fibroblasts.

Evaluation of the three approaches: We expect all single cell transcriptomics approaches to exhibit trade offs between detection vs. accurate and precise quantitation of transcripts within individual cells. During our development of each approach, we will use common samples and measures that will allow us to compare performance according to these parameters. If a single method has superior performance for both detection and accuracy/precision, it alone will be picked for further development. If no one method is best, or if one works best for undirected sequencing while another works better for targeted sequencing, we may continue development with two methods. To gather the required detection and accuracy data, we will use a common dispersed cell line, and we will assay the transcriptome of this sample as an aggregate population using RNA-seq (84, 185). Using these data, we will define a set of transcripts ("measurement set") consistent with our capture primers that exhibit a range of expression levels expressible as copies/cell, including many with copy numbers of 1 or less. We will then conduct single cell transcriptome assays for both undirected and targeted sequencing on 50-100 cells of this population using each of our three methods. For each transcript in the measurement set, the number of cells in which it appears should approximate a Poisson distribution. We will measure the sensitivity of detection of our methods by assessing the extent to which low copy number transcripts appear as often as they should according to this distribution. We will measure the accuracy and precision of our methods by examining regressions between mean transcript levels across the 50-100 cells and their expression levels measured from the aggregate population, using transcripts in the measurement set with medium to high expression levels. The unexplained variance from these regressions will include contributions from actual stochastic differences between the individual cells and the imprecision of our single cell transcriptome methods. Assuming that actual stochastic differences will be similar for all samples, the method that exhibits the lowest unexplained variance will be the most precise. We will also assay transcriptomes of a number of single cells using microarrays using standard methods (44, 83, 94) and compare the average level of the transcripts observed in these microarrays against the average expression levels seen by our methods. Because both the amplification procedures used to obtain these microarray assays and our own single cell methods may each be subject to systematic biases, we will not use these to judge the fidelity of these approaches but, rather, to assess the presence and degree of any differential biases. We will also perform RNA-seq on individual cells using the technique of (169) and compare with the results of our methods.

5.3.1: Aim 3.1: We will develop and optimize methods that pipeline in-situ single-cell cDNA synthesis to next generation sequencing in ways that preserve cell identity and that can be applied in parallel to 100s to 1000s of cells. We will investigate multiple techniques in support of these methods, including cell bar-coding, in-situ cell sequencing, and single-molecule in-cell sequencing, characterize their performance and limits, and select one for continued development and application.

5.3.1.1 Single cell sequencing via bar coding: Bar coding is chiefly attractive because it can be used with any current next-generation sequencing capability. The main technical issues for bar coding are the generation and placement of unique bar codes in the individual cells to be sequenced. Our initial bar coding

approach will use emulsions to encapsulate millions of individual cells with single 1 μm beads displaying approximately one million bar-coded oligonucleotides bearing mRNA capture sequences. This method will be applicable to blood and to disaggregated tissues; a variant of the method applicable to structured tissues will be considered below in (v). Concentrations of beads and cells will be chosen such that there is an average of 1 bead and ≤ 1 cell per compartment. Following emulsion preparation, the cells will be lysed with heat and the mRNAs will hybridize to the bead-bound oligonucleotides. After mRNAs have been captured, reverse transcriptase (RT) will be introduced to generate first-strand cDNAs, coating the beads with cDNAs (see Figure 5.3.1.1-1) To introduce RT, the beads will either be extracted and re-emulsified in a solution containing RT, or bead-containing droplets will be fused with droplets containing RT. The emulsion will be broken, the beads collected, and the cDNA will be processed and sequenced by RNA-seq or PMAGE (84, 185).

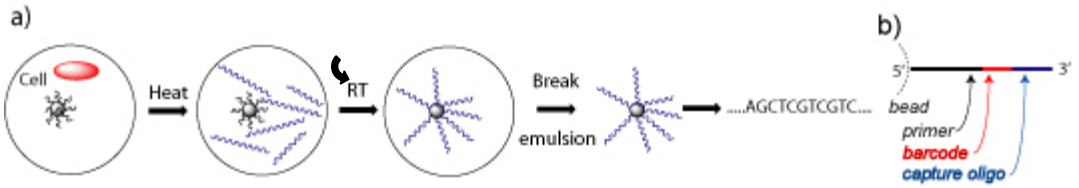


Figure 5.3.1.1-1. single cell mRNA capture and barcoding in emulsion. a) A water droplet containing a single cell and bead. The cell is lysed by heat, bound capture oligos bind to mRNA target sequences, reverse transcriptase is introduced and places the cDNA onto the bead, the emulsion is broken, and the cDNAs collected and sequenced. b) depiction of a bead-bound oligo. 5' primer region (black) allows subsequent amplification of captured cDNA, a bar-code (red) identifies transcripts belonging to the same cell, and the capture oligo (blue) captures the mRNA.

5.3.1.1 (i) Split pooled DNA synthesis on beads: To generate populations of 1 micron beads where each bead has a unique bar-code sequence that differs from others in the population, we will employ split-pooled oligonucleotide synthesis (Figure 5.3.1.1-2) (16). Synthesis will proceed on the bead surface rather than in controlled pore glass, thus ensuring that the oligonucleotides are displayed on the bead. As bead wetting and swelling properties in organic and aqueous media are important for oligo synthesis and mRNA capture, respectively, we will explore multiple bead surfaces; the swelling properties of polystyrene beads make them ideal for this application, but we will also explore mono-dispersed glass beads, and gold nanoparticles (see below). Oligonucleotides will be synthesized in the 5' to 3' direction on the bead surface, as opposed to the canonical 3' to 5' direction, thus allowing the bead-bound oligo to serve as a primer for reverse transcriptase. As bar-codes must be synthesized error free, we will develop a method for purification of the oligos while on the beads that will result in purities rivaling that of trityl-on RP-HPLC purification. This method exploits the exonuclease resistance of achiral phosphorodithioate linkages. By incorporating this linkage at only the 3' base of the oligonucleotide, lambda exonuclease treatment of the deprotected DNA will degrade incomplete oligos.

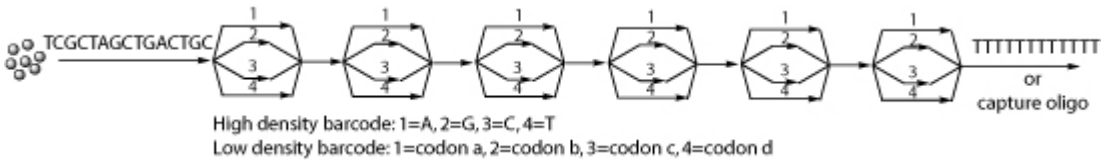


Figure 5.3.1.1-2. Split pooled synthesis of bar-coded oligonucleotides. 5' fixed sequence is grown on the beads. To apply the bar-code, beads are split into four pots and a reagent (single nucleotide or nucleotide triplet) is added to each. The beads are mixed and the process repeated. After addition of the bar-code, beads are pooled and poly-T or capture oligos are added.

We will explore high density and low density barcoding strategies, for readout *via* sequencing and hybridization, respectively. High density bar-codes comprise ordinary sequences to be decoded by standard sequencing and provide 2 bits of information for every base pair in the bar code. Low density bar codes will encode 2 bits of information per 3 base pair "codon" and are decoded by hybridization. Thus, each three nucleotide codon will have 4 variants. To enable 1000 cells to be addressed uniquely with $p<0.001$ requires high density barcodes of at least 5 bp and low density barcodes of at least 15 bp (five 3bp "codons"). Although low density barcodes are longer, they are not read during sequencing and so their decoding does not contribute to sequencing overhead. By contrast, high density barcodes *must* be sequenced and so yield 5 bp of overhead for each of potentially many millions of sequence reads. Use of low density barcodes is illustrated in Figure 5.3.1.1-3. The hybridization probes used against low density barcodes can be made by combinatorial

synthesis of labeled oligonucleotides from triplet phosphoramidites.

5.3.1.1 (ii) directed and undirected sequencing capture oligos. For undirected sequencing, capture oligos will be polyT with appropriate 3' suffixes as described above. For targted sequencing, specific capture oligos must be synthesized and then affixed to the beads after barcode generation. To do this, an equimolar mixture of the capture oligos will be prepared and this mixture will be ligated to the bead sequences using appropriate sets of splint oligos. To allow bead sequences that do not become extended by capture oligos to be degraded, achiral phosphorodithioates at the 3' ends of the capture oligos may be used in a manner similar to (i) above. In developing this approach, we will pay close attention to ligation efficiency and bias, as these factors will limit the size of the target transcript set and the accuracy of quantification. We will optimize these factors initially with commercially available mRNAs that we will mix in different proportions, and then move on to sets of transcripts that are parts of well studied transcriptional networks that are expressed in our samples, with particular attention to networks we expect will be relevant to our intended application in Aim 3.2 (section 5.3.2).

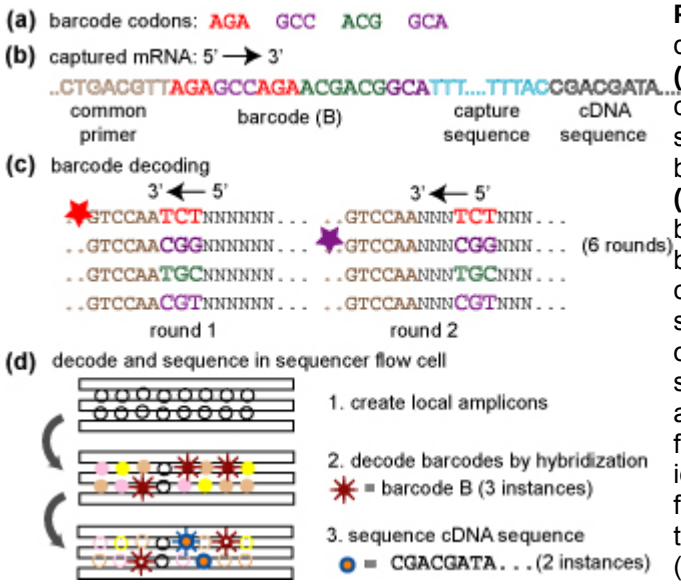


Figure 5.3.1.1-3. Illustration of low density barcodes. **(a)** Barcode codons. **(b)** Barcoded cDNA captured from single cell using undirected sequencing capture sequence polyT-AC. All cDNAs with this barcode ("B") come from the same cell. **(c)** Hybridization rounds used to decode barcodes. Starred primers correspond to barcode B. **(d)** Assignment of cDNA to cell using barcodes followed by sequencing of cDNAs. Local amplicons of the cDNAs are generated for sequencing. Barcode probing as in (c) assigns amplicons to cells. Sequencing from capture sequence or adaptor identifies cDNA. In this illustration, 3 features are assigned to barcode B, and two cDNAs have sequence of cDNA in (b), one of which is in cell B.

5.3.1.1 (iii) mRNA capture and barcoding. Using our emulsion PCR procedure (138), we will place cells and beads in a buffer containing reverse transcriptase, RNAse inhibitors, and dNTPs. Emulsions will be formed using the appropriate sets of oils in a single tube by controlled vortexing. Conditions will be optimized such that the average compartment size will be large enough to contain a single cell and bead. The cells will be lysed by heating the emulsion to 95 °C for 10 minutes, after which RT will be added as described above. To capture the mRNA's onto beads we will explore both isothermal reverse and thermocycling transcription. The former is preferred, as it ensures that each mRNA is captured once. However, thermocycling may be necessary to capture all mRNA's in a sample. Following mRNA capture, the emulsion will be broken and the beads collected and pooled for either whole or targeted transcriptome digital analysis.

5.3.1.1 (iv)Transcriptome digital analysis: To analyze the captured mRNAs in each cell, the cDNA will be cleaved using a frequent cutter such as *Nla*III similarly to previous work (84), keeping only the fragments attached to the beads. We will ligate a double stranded adapter to the other extremity of the cDNA, and use it as a common priming site for very limited PCR amplification in pairs with the corresponding common primer synthesized directly upstream of the bar-code. The library will be size selected and sequenced on Illumina GAII next generation platform. Sequencing of one end will convey transcript identity and expression levels. Low density barcodes will be decoded by hybridization (see (i) and Figure 5.3.1.1-3 above), while high density barcodes will be revealed by sequencing from the other end. Transcript abundances can then be obtained simply by counting numbers of sequence features per cell that map to the same gene. For undirected sequencing, transcripts will be mapped to the closest gene whose stop codon is 5' of the sequence read from the transcript in the orientation determined by cDNA capture and preparation protocols.

5.3.1.1 (v) Structured tissue bar-coding: To apply these barcoding methods to structured tissues requires a method of delivering the bar-coded capture oligos to cells that have not been disaggregated. Two possible methods are to use ligand-printed surfaces such as those that can be prepared with the use of a device such as the BioForce Nano eNabler (http://www.bioforcenano.com/index.php?id=295), or by shooting beads prepared as above into the tissue using a biolistics device. A biolistics approach would only label a

small fraction of cells, but the bar codes of these cells could be identified by using the hybridization strategy of (i) above. We will explore these options up to the point of developing an initial proof-of-concept experiment whose execution will depend on whether an appropriate device is available (as these devices have not been budgeted). A simple initial experiment would be to synthesize an array of barcodes on a microarray and attempt to create cDNAs from a permeabilized tissue section that is laid down on the array. As microarray features are typically larger than human cells, this would not support single cell transcriptomics, but could enable transcriptomics of small tissue regions.

Potential problems and alternatives: The greatest potential for difficulty is that coupling efficiency for 5'→3' oligonucleotide synthesis tends to be lower than for canonical 3'→5' synthesis. If these problems prove significant, we will explore polarity switching on solid surfaces, wherein the oligonucleotides are synthesized on the beads in the 3'→5' direction and then inverted *in situ* (96). Another strategy is to synthesize polymer phosphoramidites for all non-variable regions. This is widely practiced with triplet phosphoramidites and decreases the number of couplings required, resulting in higher synthesis fidelity.

5.3.1.2 *In situ* cell sequencing of rolling circle amplified cDNAs: By dint of its covalent linkage of amplified product, rolling circle amplification in various forms has been used to generate extremely compact amplicons primed off of specific genomic and mRNA sequences within individual cells, that can be used for both *in situ* genotyping and digital quantification of transcript abundances (163, 202). However, these applications have only considered very small numbers of loci at a time. Here we propose to greatly increase the scale and generality of these techniques to the point where transcriptome-level information can be obtained from large numbers of individual cells, and we will develop this approach both for dispersed cells and structured tissues. By combining microscopy and image analysis with suitable non-destructive tissue staining prior to *in situ* sequencing, this approach will enable integrated collection of data on cell morphology, protein content and localization, and transcriptome, as well as cell location in structured tissues. We will do our initial development with dispersed cells and move on to structured tissues (see (iii) below).

5.3.1.2 (i) In situ library preparation in dispersed cells: Using technology developed for binding cells to a surface described in section 5.4.3, we will capture cells in a dispersed fashion into our flow cell. We will then permeabilize the cells sufficiently to introduce capture primers and reverse transcription reagent, where capture primers are as described above. When specific capture oligos are used, it will be necessary to design them with a common sequence appended 5' to the specific capture sequence to serve as a sequencing primer; for undirected sequencing this is unnecessary as sequencing can be primed off the polyT sequence that is incorporated into each cDNA. Following annealing of the capture primers, we will conduct reverse transcription *in situ*, and then flow in RNase H to degrade the RNA component of the RNA/cDNA hybrids. The cDNAs will then be circularized inside the cells using T4 DNA ligase and short splint oligonucleotides to anneal to both ends of the cDNA, where the splint oligos are designed to hybridize to the common sequence part of the capture oligos at one end while they are degenerate at the other end. Non circularized material will then be digested using exonuclease I and III. The remaining circles will be ready for *in situ* rolling circle amplification (RCA) using phi29 DNA polymerase primed using polyA oligonucleotide. Molecular crowding of rolling circle-amplifiifed mRNA is expected to be a consideration of this approach and will be addressed in the following ways: (a) With dispersed, separated cells, the cells may be lysed to enable localized diffusion of the cDNAs prior to RCA. (b) Capture primers will be redesigned to capture fewer transcripts. (c) Where crowding is not extreme, different sets of sequencing primers may be introduced in the RCA step so that the effects of crowding can be overcome in the sequencing step by initiating sequencing based on one primer at a time (i.e,, instead of resolving the crowding on a spatial level, it is resolved by serializing it over time). An issue related to molecular crowding is that localized concentrations of mRNAs may exist in the cell, e.g., in RNA stress granules and P-bodies (5, 6, 47). Very concentrated RNA bodies may be difficult to resolve by these methods; however, detecting and counting them in many individual cells may be an important biological application of *in situ* transcriptome sequencing in its own right.

5.3.1.2 (ii) In situ sequencing and digital transcriptomics: Cells will be sequenced using our current single base extension or ligation chemistry on the Polonator platform (Preliminary Results, 4.6) using the common sequence incorporated into specific capture oligos, or polyT for undirected sequencing. Serial sequencing runs on the same cells using different sequencing primers may be required as just described in 5.2.1.2 (i). Transcript abundances can then be obtained by the mapping and counting methods described above in 5.3.1.1 (iv), and tested by application to mixtures of cell lines also described there.

5.3.1.2 (iii) In situ analysis of structured tissues: Since *in situ* cDNA library preparation in a tissue

section and in dispersed cells (see 5.3.1.2 (i) above) are similar from a technical point of view, we will test our procedures above on a complex tissue to study the various interactions between various cell types and their difference in gene expression. The methodology remains the same with tissue sections, except that we cannot lyse cells as a way or relieving molecular crowding as described in 5.3.1.2 (i). However, with tissue sections there is the option of creating stacks of thin sections of the tissue and sequencing them individually, reconstructing the full cells' transcriptomes by adding together the transcriptomes of the individual layers.

Potential problems and alternatives: It is possible that the splinting strategy for circularizing cDNAs described in 5.3.1.2(i) will not be efficient *in situ*. If so we will focus on padlock probe-based or similar strategies that have been used widely by the Church Lab in other contexts and demonstrated *in situ* at small scales by other groups (163, 202). This approach may complicate capture primer design.

5.3.1.3 Single-molecule *in situ* sequencing: The *in situ* sequencing method of 5.3.1.2 could be improved if the cDNA amplification step of 5.3.1.2 (i) could be avoided, as this would reduce the issue of molecular crowding (5.3.1.2 (i)) as well as reduce the potential for amplification bias. To explore this option, we will test single molecule sequencing of transcripts in cells. A single molecule sequencing instrument, the Heliscope, has been commercialized by Helicos (http://www.helicosbio.com/) (55) and another is in development by Pacific Biosciences (http://www.pacificbiosciences.com/index.php) (45). Both systems gather sequence information by tracking extensions by single labeled nucleotides of individual transcript molecules that have been captured on a surface, but neither has been developed for single molecule sequencing within cells. However, of the two platforms, the Heliscope is more amenable to this development because it does not require transcripts to be captured within Zero Mode Waveguides arrayed with special geometry on the surface. A Heliscope is available at Harvard. However, because changes in sequencer operation and software will likely be required, we will plan to outfit our Polonator system (see Preliminary Results, 4.6) with optics capable of detecting single molecule signals rather than work on a Heliscope itself. The open-source design of the Polonator will make it much easier to adjust components and software on the Polonator vs. the Heliscope.

5.3.1.3 (i) Biological preparation and sequencing overview: Briefly, our approach will be to layer a thin tissue section or dispersed cells in the flow cell from section 5.4.3 (also used above in section 5.3.1.2 (i)). Initial development will entail sequencing purified single molecules of RNA attached to the flow cell on the Polonator, followed by RNA populations that are anchored by hybridization to polyT-V oligonucleotides. Sequencing reactions will be primed directly off of the polyT-V oligonucleotides and proceed with reverse transcription using single base extension with fluorescent reversible terminator nucleotide technology (see Preliminary Results, 4.6). Targeted sequencing will be tested by capturing RNA molecules on polyT primers that are dideoxy-terminated and subsequent annealing with transcript-specific primers. Each nucleotide will be incorporated directly onto the growing cDNA primed off of the mRNA, incorporated nucleotides will be imaged at every cycle, and the fluorophore and terminator cleaved off to ready the molecule for the next nucleotide incorporation. Once capability is achieved, we will move into attaching dispersed cells on our flow cells, to be followed by 2 micron thin tissue sections as described in section 5.3.1.2. Sequencing of permeabilized cells will be performed similarly to what was previously described in 5.3.1.2. A key issue in sequencing captured transcripts in situ is to reduce background that may be caused by cell debris autofluorescence and labeled nucleotide adsorption. To reduce the impact of these factors, we will treat the cells with proteases and detergents after transcript capture to wash away debris, and avoid use of fluorescent labels on nucleotides whose emission wavelengths coincide with cell autofluorescence. Restriction of fluorescent labels on nucleotides will require increasing the number of sequencing cycles, as nucleotides can only be included in the same cycle if they have distinct labels. Heliscope sequencing employs a protocol by which a molecule can be sequenced twice (once in a forward and again in a reverse direction) to reduce sequencing error (55), and we will modify this technique as required to operate in our *in situ* conditions.

5.3.1.3 (ii) Technological requirement of single molecule sequencing: The fundamental requirement for single molecule sequencing is to use optics that enable single molecule resolution. To adapt the Polonator to the technological level necessary, we will redesign our current optical configuration from EPI-Fluorescence to exploit TIRF (Total Internal Reflection Fluorescence). The required hardware modifications needed to retrofit a Polonator for through objective TIRF are: (a) Replacement of the standard 20x Leica objective with 100x high numerical aperture objective, PL APO 100x 1.4NA or similar. The high NA is needed so that the angle of incidence is greater than or equal to the critical angle. (b) Insertion of opaque disk in the illumination path post-collimation pre-camera path so as to only permit fringe light rays from reaching the back aperture of the objective. c) Flowcell skirt addition via adhesive gasket to hold immersion fluid. Outside of the optics, other

components such as the optical train, stage motion, fluidics delivery, scanning capture, and algorithms will only need minor tuning to adapt to chemistry in cells. Software changes needed for implementation of single molecule sequencing should be minimal, although this depends on the ability to use reversible terminator nucleotides (vs (55)). If reversible terminator nucleotides are not successful, we will proceed with unterminated nucleotides and make software modifications required to analyze homopolymer additions, as in (55).

Potential problems and alternatives: Detection of base incorporation at a single molecule level in the presense of cell debris will be very challenging technically. If we cannot detect incorporation reliably after efforts to clean up cell debris, we will not pursue this strategy and focus exclusively on the approaches of section 5.3.1(i) and (ii).

5.3.2: Aim 3.2: We will use these single cell transcriptomics capabilities to characterize the transcriptional state differences in cells bearing artificial and natural variant combinations from Aim 1, and from cell types developed from iPS from different genetic backgrounds.

As noted in the Aim 3 overview (section 5.3), our strategy will be to develop and evaluate three single cell transcriptomics approaches in the first half of our CEGS, and proceed to further development in the context of demonstrations of the best approach(es) in the second half. This sub-Aim describes our plans for these demonstrations and their integration with CTCHGV Aims 1 and 2. We will proceed through four series of experiments that start with single cell transcriptome sequencing of dispersed cells and mixtures whose results can be confirmed easily by other means, and proceed to in situ transcriptomics on structured tissues.

5.3.2 (i) Preliminary experiments on cell mixtures We will create mixtures of CTCHGV cell lines that are expected to exhibit transcriptional differences in different fixed proportions, and gauge the extent to which we can observe single cell transcriptomes that bear these differences in approximately the same proportions. These experiments will both test the performance of our approaches, and will also help define optimal identities and sizes of the transcriptome subsets interrogated by our targeted and undirected sequencing methods. Among the mixtures we will consider are: (a) Original, unaltered CTCHGV cell lines corresponding to different tissues (if available). (b) Mixtures of original, unaltered CTCHGV cell lines from different subjects, *if* in the course of Aim 1 (especially Aim 1.3, section 5.1.3) we have observed transcriptional signatures that differ between samples. (c) An original, unaltered CTCHGV cell line, and the same cell line which has been modified (by techniques of Aim 1.1, section 5.1.1) to contain an integrated GFP gene. (d) The two cell lines from (c) where the GFP-containing cell line has additionally been modified by deletion of both copies of a major transcription factor. Here it will be of interest to see how well presence or absence of GFP correlates with expected changes in the transcriptional network controlled by the transcription factor. (e) An original, unaltered CTCHGV cell line grown in two conditions, e.g., CTCHGV fibroblast cell lines grown in the presence of vs. the absence of serum (70).

5.3.2(ii) Downstream consequences of cis regulatory variations We expect Aim 1 to identify cis variations that control key transcription factors. We will take original CTCHGV cell lines and/or versions of these cell lines altered in Aim 1 to maximize differences in expression of these factors, and first assess clonal outgrowths of these cells for mean expression levels over the entire transcriptome by normal array or RNA sequencing methods. From these data we will then identify a small number of transcripts that exhibit significantly different average expression levels, and assess a large number of cells of each population by in situ hybridizations targeted to these transcripts to characterize the *distribution* of expression levels in individual cells vs. the mean expression levels captured initially. Finally, we will perform single cell transcriptomic sequencing of these cell lines by our Aim 3.1 methods and assess the extent to which transcripts found to differ at a mean level over the population are also found to differ in mean level across the individual cells, and whether the distributions of individual cell transcript levels found within these cells correlates with the distribution found by in situ hybridization. Notice that these experiments will be performed on the individual cell lines separately, not on mixtures as in (i) above.

5.3.2(iii) Differentiation and de-differentiation of iPS Here we integrate our single cell transcriptomics methods with Aim 2. We will take a subset of the CTCHGV iPS cell lines that are being differentiated into cell types representing different tissues in Aim 2.2 (section 5.2.2), extract and preserve aliquots at different time points, and perform single cell transcriptome sequencing on these time point samples. We will also take original (or Aim 1-modified) CTCHGV cell lines that are being de-differentiated into iPS, similarly preserve aliquots at different time points, and perform transcriptome sequencing on these samples. The first set of these experiments should exhibit a progression of subpopulations of cells that ultimately assume

transcriptional characteristics of target cell types, but in view of the stochastic nature of differentiation, it may also exhibit subpopulations that do not. A key interest will be to look for any transcriptional characteristics that appear to anticipate the assumption of target cell type identity, as these may give insights into better ways of controlling differentiation. Similar considerations will apply to the de-differentiation experiments.

5.3.2(iv) in situ structured tissue In our main demonstration of in situ single cell transcriptomics of structured tissue, we will perform in situ transcriptome sequencing of primary human skin cells from a CTCHGV subject, and compare the results with single cell transcriptome sequencing of iPS cells from the same subject that have been differentiated so as to yield fibroblasts. As human skin is a very complex structured tissue, we will expect to see a range of distinct transcriptomes in the primary cells, only some of which correspond to the iPS-derived fibroblast transcriptomes. This experiment will be of interest because it will reveal the extent to which single cell transcriptomes vary across a primary sample, how much transcriptomes within a cell type within the sample may vary according to the locations of the cells in the sample, and how much iPS-derived cells of a type within the sample resemble their primary cell counterparts. Our ability to proceed with this experiment will depend on the availability of an appropriate tissue sample (see Research Design Overview and Aim 1.2(i), section 5.1.2(i)).

Potential problems and alternatives: Success on Aim 3.2 above will depend on our success in Aim 3.1, and we have therefore designed the applications above as a series of tests of Aim 3.1 methods that are graded in difficulty. We will take this series as far as we can and use any points of failure to inform further Aim 3.1 work on methods development, and thus move between Aims 3.1 and 3.2 iteratively.

Aim 3 goals: Final goals As noted in our Research Design Overview, we intend to demonstrate single cell transcriptomes (both targeted and undirected sequencing) for 1000 transcripts per cell. This demonstration will be on whichever of the three approaches we deem to be most promising half way through the CTCHGV five year period (see *Evaluation of the three approaches* above). Intermediate goals Again, as noted in our Research Design Overview, we will evaluate progress at the end of year 2 of the Center and renegotiate goals as appropriate. We expect we will have succeeded in interrogating 100 transcripts per single cell at that time by at least one of our approaches. Impacts CTCHGV-developed methods for obtaining single cell transcriptomic data will greatly broaden the ability to understand the distinct roles of the different cell types that participate in complex organisms in their actual, structured tissue contexts. Although single cell transcriptome-level information is obtainable today, current methods are not scalable to large numbers of cells and do not take advantage of the greatly increased throughput of next-generation sequencing. Additionally, CTCHGV's development of both targeted and undirected transcriptome sequencing methods will enable considerable flexibility in application and optimal utilization of sequencing capacity.

5.4: Aim 4: In support of Aims 1-3, we will develop innovative and widely applicable methods for high-throughput synthesis of long DNA constructs, highly efficient homologous recombination in human cells, and highly multiplexed single cell handling that enables sorting based on morphology.

Overview: Central to CTCHGV strategy for Aim 1 is the use of Zinc-Finger Nucleases (ZFNs) to support the engineering of regulatory sequences in human cells. As noted in Background and Significance section 3.4 and Research Design section 5.1.1(iii), the ability to engineer ZFNs targeted to specific genomic sites has matured to the point where both academic research consortia and companies are now generating customized ZFNs (134). CTCHGV will make use of these capabilities with efficiency improvements described in Aim 1.1 (section 5.1.1(iii)), but in support of its broader Aim 1 goal of making these techniques scalable to thousands of genes, here in Aim 4 we will apply our expertise in synthetic biology and zinc finger engineering to two projects (Aims 4.1 and 4.2) that will improve both scalability of synthesis and the targeting range of ZFNs generally. Meanwhile, Aim 4.3 will use elements of the system proposed in Aim 4.1 to improve cell handling that is needed in Aims 1.2 (section 5.1.2(ii)) and Aim 3, in a way that will enable a new form of cell sorting that expands the capabilities of FACS. All three of these projects involve development of highly innovative technology that will have very wide application in biomedical research generally in addition to their supporting roles in CTCHGV.

Aim 4.1 (section 5.4.1) will address scalability and accuracy of synthesis of ZFNs. ZFNs comprise two subunits, each of which is a fusion of three to four tandem Zinc-Finger domains (ZF domains) that enable specific recognition of a DNA sequence with an endonuclease *FokI* (section 5.1.1(iii)). Individual Cys_2-His_2 ZF domains are each about 30 residues long, with specificity mainly conferred by six residues in or adjacent to the

domain's α-helix (72, 130). DNA that codes for these 30 residue domains can be synthesized as single DNA oligonucleotides (oligos), so that, in the simplest scenario, to synthesize genes coding for a ZFN subunit targeted to a specific site would require synthesis of three to four site-specific oligos and a small set of splice oligos, followed by enzymatic assembly with common DNA scaffold components that code for the rest of the subunit. Since two subunits are required for a ZFN, synthesis of 1000 specific ZFNs entails ~2000 such operations, each requiring ~10 site-specific oligos (including splice oligos). We (see (174)) and others have recently turned to release and enzymatic assembly of oligos from oligo chips as a low cost method of implementing such synthesis tasks. Currently these methods experience yield and accuracy limitations due to the considerable crosstalk during annealing (and ligase or polymerase) assembly reactions that arises from the release of massive mumbers of oligos into small numbers of pools. In the context of ZFN synthesis, we will develop technology for massively parallel hierarchical synthesis that, in stages, sequesters oligos and subsequently assembled DNA fragments that correspond to thousands of individual target DNA constructs into separate compartments, so that assembly of these constructs is not hindered by crosstalk. This technology will innovatively integrate on-chip oligo synthesis and assembly, DNA sequencing, and light-directed release of arrayed fragments.

Aim 4.2 (section 5.4.2) will address the comprehensiveness, specificity, and efficiency of ZFNs. Each ZF domain of a ZFN subunit recognizes 3 base pairs, and specificities for recognition are subtlely different for each of the ZF domains linked in tandem in a ZFN subunit. To be able to specifically target any half site in the genome, we need libraries that cover all 4^9=262,144 possible 9 bp sites. To achieve this, we will use ribosome display methods (196) to sample all possibilities from very large combinatorial libraries engineered to cover this target sequence space.

One of the technological elements in Aim 4.1 is that we will array micron sized features (polony beads) in a flow cell with a light-labile chemical anchor, analyze them (specifically, sequence the DNA on them), and then, based on the results of analysis, release specific sets of beads together by directing light on them. With suitable modifications, these methods can be applied to cells as well as microbeads, and Aim 4.3 will develop these modifications. The cell arraying component will have immediate application in Aim 1.2 (section 5.1.2(ii)), where it will improve image analysis of cells, and also Aim 3.1 (5.3.1.2(i)). Meanwhile, the extra modification of using light-labile chemical anchors for the cells will enable a novel FACS alternative that will allow cells to be sorted not only by markers and general optical properties, but also by morphological features revealed by image analysis. We will demonstrate this capability on an application that integrates other CTCHGV Aims.

Aspects of Aims 4.1-3 will be developed on the Polonator (see Preliminary Results, 4.6), which provides the microscopy and image analysis, sequencing, flow cell control, programming, and open source access and compatibility that allow us to modify and integrate the new elements quickly and easily. However, we will make any modifications open source and use our close collaborations with companies (see Data and Materials Dissemination) to encourage incorporation of our innovations into commercial products.

5.4.1: Aim 4.1: We will develop a platform that integrates DNA synthesis and sequencing and uses sequence information to assure synthesis of DNA constructs with extremely low error rates.

A high level view of one version of the platform we propose is given in Figure 5.4.1-1. As is done today (e.g., (174)), the sequences of the large DNA constructs to be generated are first analyzed to determine a set of construction oligos with appropriate size, overlaps, common primer sequences, and Tms for correct amplification and self-assembly (Figure 5.4.1-1(a)), and these oligos are chemically synthesized on a single stranded DNA oligonucleotide array (as in (174)). The assembly process is illustrated in Figure 5.4.1-1(b). The construction oligos are cleaved from the array and amplified clonally on microbeads using emulsion PCR (155). The microbeads are then arrayed on a flow cell for sequencing, but here the beads are attached to the surface using light-labile linkers so that particular microbeads can be subsequently released by direction of narrow beams of light on them. The beads are then sequenced to simultaneously locate oligos that are part of the same large DNA construct, and to verify that the oligo sequences are error free. For each large construct, light is then directed to its sequence-verified oligo beads so that these can be flowed out of the flow cell and captured into an independent compartment for subsequent multiplex assembly. In place of the microbeads used for illustration in Figure 5.4.1-1, oligos or assembly products could be covalently immobized (e.g. using EDC/NHS chemistry) and amplified from single molecules using polymerase (or ligase) chain reactions – thermal cycling (PCR) or isothermally (e.g. RCA, hRCA, SDA, HDA, PWGA (http://www.biohelix.com/technology.asp)) using zero, one or two immobilized specific or general primers or no

primers at all (as in PWGA). The resulting polymerase colonies (polonies) can be sequenced by any of the "next-generation" DNA sequencing chemistries -- e.g. polymerase with FL-dNTPs (117) or ligase with 5-mers to 9-mers (155), MPSS (15), SBH, etc. Instead of release of polonies with correct sequence for subsequent assembly, polonies which have the incorrect sequence could also be selectively destroyed or released -- e.g. via photo-caged nitrobenzyl linkages. For very large constructs, the process is amenable to iteration in that oligos can first be assembled into construct fragments, and the fragments then combined for subsequent assembly. We will do initial development on the Polonator to simplify integration of sequencing and light direction (see 5.4.1(iii) below), but will consult with commercial providers of compatible instrumentation to abet technology transfer.

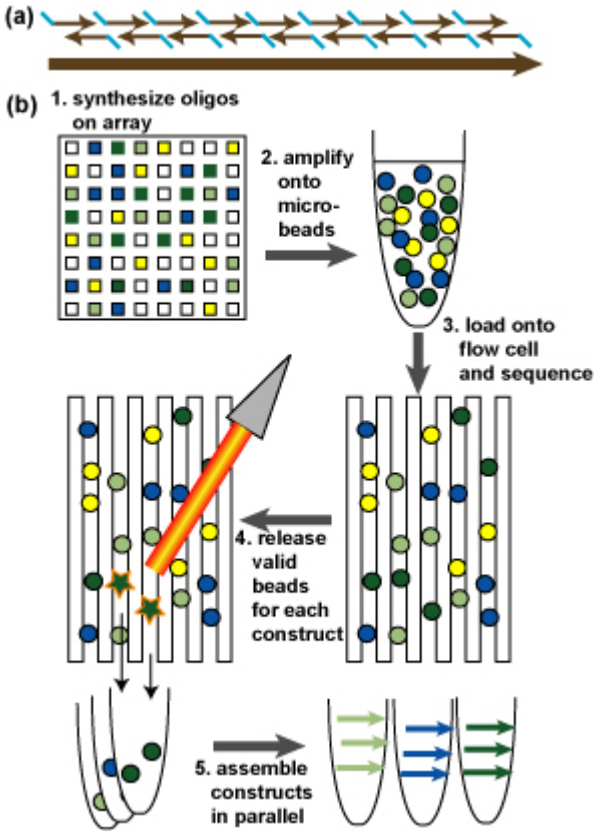
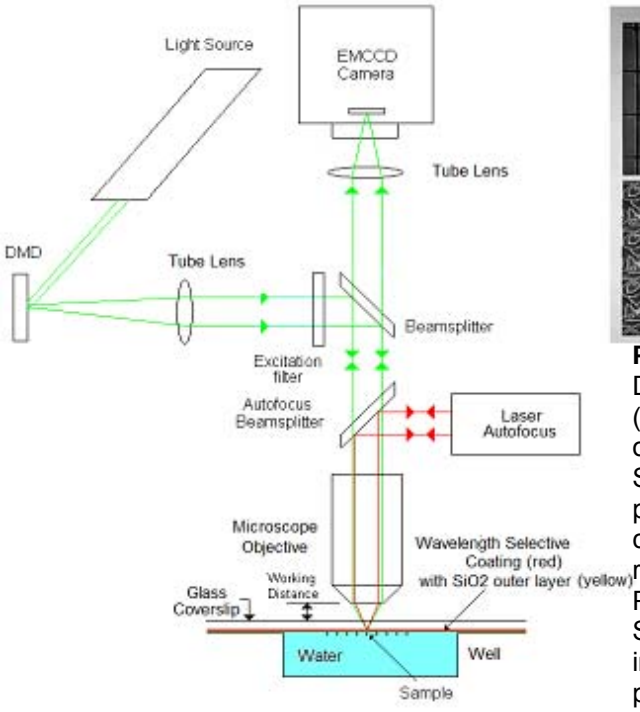


Figure 5.4.1-1. Schematic for one way of integrating DNA sequencing and synthesis for high-throughput reduced-error synthesis of large constructs. **(a)** Large DNA construct is analyzed into oligos with appropriate overlaps, uniqueness, Tms, as needed. **(b)** Processing pathway from synthesis of oligos on array for multiple constructs (represented by different colors) to multiplex synthesis. Amplfication in (2) is illustrated as emulsion PCR as in (155). Microbeads are loaded onto flow cell using light-labile chemical attachments (see text) and sequenced on the flow cell (3). For each construct, light is directed to microbeads with sequence-validated oligos for the construct for release and capture (4). Assembly of all constructs then proceeds in parallel (5).

These steps overcome crosstalk between masses of oligos by separately collecting and assembling oligos and fragments that are part of the same construct. The sequencing step also overcomes the high rate of error incurred during chemical synthesis of oligos on the array, which, at ~0.5% error per addition (12) can be ~33% for synthesized 80-mers. While current methods, such as mismatch-sensitive hybridization (174), mutS binding (17), MutHSL cleavage near mismatches (160), and direct cleavage at mismatches (11) allow synthesis and assembly errors to be incorporated and then allow them to be filtered out, the method outlined in Figure 5.4.1-1 avoids their incorporation in the first place.

Development of this method will involve small modifications to the Polonator bead and flow-cell preparation prototocols to support loading the beads to the flow cell with light-labile chemistry. Towards this end, phosphoramidites containing the photo-labile nitrobenzyl group will be incorporated into the oligonucleotides that are used to "cap" the bead-teathered DNA with the appropriate attachment chemistry. The Polonator may also need minor modification to prevent entry of stray light that could inadvertently release beads. More substantial modifications must be made to support the release of sequences by light directon. Suitable control of light direction can be achieved either by using a Digital Micro-mirror Device (DMD) or a Liquid Crystal Display (LCD). For simplicity, cost, and attainability, we will focus on the DMD approach with initial testing on the Polonator. The array will be used in conjunction with the standard Polonator illuminator and an appropriate photo cleavable (360nm) linker (see Figure 5.4.1-2). For this approach to work, the illumination path must be modified to allow the image of the DMD to be projected on the substrate. The optical path is modified as follows: a) Assemble a Polonator filter block consisting of a standard 50/50 beamspliter and 360nm excitation filter. The 50/50 beamsplitter is used instead of a 100% mirror facilitating focus and alignment of the DMD to the camera CCD array. b) Place the filter cube in the Polonator filter wheel to allow patterned illumination on the substrate, and c) Insert a tube lens before the 360nm excitation filter allowing the image of the DMD array to be collimated. d) Place the DMD at the focal plane of the tube lens and illuminate the array at an angle with the current 300 watt xenon source. The light reflects off the DMD and is collimated by the tube lens: collimated light then reflects off the 50/50 beam splitter onto the specimen and back up to the camera. The shutter and motion axis allow this selective release to be accomplished over the full area of the

substrate.



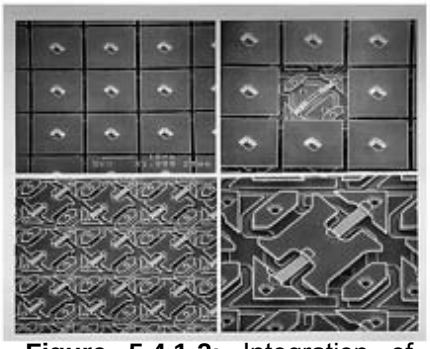


Figure 5.4.1-2: Integration of Digital Micro-mirror Device (DMD) array with the Polonator optical components. *Left:* Schematization of the optical path of light for DMD array control allowing selective release of cells from the Polonator flow cell. *Right:* Scanning Electron Microscope image of DMD mirrors and pivoting structure (Texas Instruments).

Use of this technology to generate large numbers of ZFNs required for other CTCHGV Aims After the DNA sequencing and bead release components have been successfully integrated, we will implement additional automation to manage the synthesis of large numbers of ZFN proteins, particularly for Aim 1. We will attach an autosampler which will take beads that are released into the flow-cell volume and feed them into 384 well plates. The same autosampler can then be used as the platform for hierarchical synthesis. The main issue is that the liquid volume after flushing the flow cell will be about 300μL, more than will fit in the well. To accommodate this we can use filtration and apply vacuum while dispensing into the well to remove excess liquid. An alternative would be to use magnetic separation, which would require design of a concentrating chamber. This could be developed using microfluidics. Once automation has been developed, we will design and order the oligonucleotide arrays required for building the ZFNs and proceed with actual synthesis.

Potential problems and alternatives: As the DNA sequencing, DNA synthesis, and DMD technologies are already individually well developed, the main issues that will arise are with integration, and we have laid out our key approaches above. As ZFNs are very simply structured, the problem of synthesis is particularly simple, as construction of each ZFN can be accomplished by the addition of a small set of oligos to standard fragments encoding the rest of the protein. The only other novel component to be integrated is new bead attachment chemistry. Here the main issue is likely to be amount of non-specific absorption to the surface, and we expect we can reduce this easily with different coatings.

5.4.2: Aim 4.2: We will improve zinc-finger nuclease (ZFN)–mediated homologous recombination in human cells by engineering a comprehensive zinc-finger archive, by developing novel methods of delivering ZFNs into cells, and by developing a "segmental genome replacement" strategy.

5.4.2.i: Engineering a comprehensive zinc finger archive: As noted in Preliminary Studies, section 4.5, ZFNs are dimers, the monomers of which each contain tandem arrays of three zinc fingers, which, at full specificity, would be enough to uniquely specify sites in the human genome; however, to date, OPEN zinc finger pools (see Preliminary Studies, 4.5) have been constructed for all three bp subsites of the form 5'GNN at all positions in a three-finger domain (48 pools) and for a smaller number of the 5'TNN subsites (18 pools) ((107) and M. Maeder, J. Foley, & J.K. Joung, unpublished). This limits the targeting range of OPEN to finding potential ZFN sites on average only once every 200 bp, the same range that is available commercially *via* the CompoZr™ zinc finger engineering service from Sigma (http://www.editforthebetter.com/FAQs.aspx). As gene targeting efficiency drops off with increased distance between ZFN-induced double stranded breaks (DSBs) and the desired alteration in mammalian cells (41, 162), improved targeting capability will be needed if ZFNs are to be used widely for homologous recombination (HR)-mediated targeting of gene cis regulatory regions in Aim 1 (section 5.1.1) and many other needs of the research community. We propose to complete the archive of OPEN zinc finger pools and to perform a comprehensive series of selection experiments to identify three-

finger arrays for every possible 9 bp target site. Our overall strategy is described in Figure 5.4.2-1. Completion of this library will result in a very substantial advance in the ability of the academic community to engineer and use ZFNs by providing a publicly available source of pre-engineered zinc finger arrays. Currently, even in the Joung lab, where OPEN was developed, it requires 1.5 FTEs about eight weeks to perform selections for 48 ZFN half-site targets, while the cost of CompoZr™ service is high at $25,000 per pair.

5.4.2 i.A Construction of a comprehensive set of OPEN pools: In addition to the 48 GNN and 18 TNN subsites already supported (see above), the Joung lab has already begun to construct pools for all 48 ANN subsites (16 x 3 finger positions) in collaboration with the lab of Daniel Voytas at the University of Minnesota. For Aim 4.2, we will therefore undertake construction of all 48 CNN subsites and the remaining 20 TNN

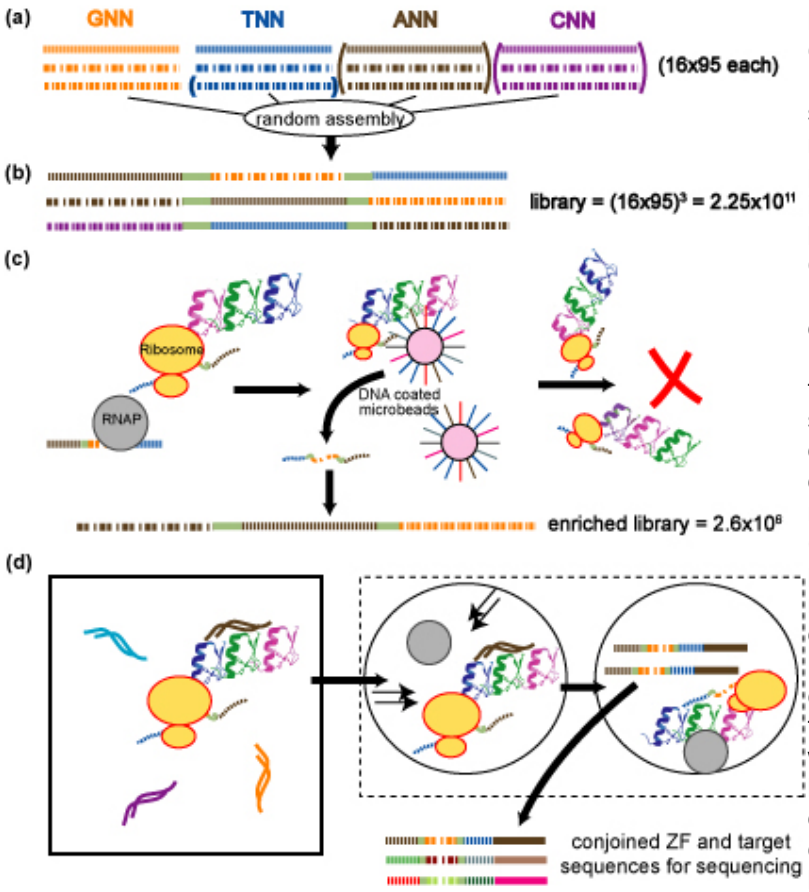


Figure 5.4.2-1. Strategy for ZFN improvement. **(a)** Completion of OPEN pools for all 3bp DNA sites. Each colored, dashed line represents a coding sequence specifying a zinc finger domain that recognizes a DNA triplet. Stacks of domains in parenthesis remain to be completed (section 5.4.2 i.A). **(b)** The pools of (a) will be combined into a library that codes all possible three zinc-finger OPEN pool combinations (section 5.4.2 i.B). **(c)** Ribosome display is used on the library of (b) to express all three zinc finger domains (left), and the RNA/Ribosome/ zinc finger triplet complexes are then exposed to microbeads coated with all double stranded DNA (dsDNA) 10mers (center). RNA extracted from beads will be used to generate the enriched library (bottom), while complexes that do not attach to beads will be washed away (right). (section 5.4.2 i.B) **(d)** Ribosome display is performed on the enriched library from (c) (box on right), and the RNA/Ribosome/zinc finger triplets exposed again to all dsDNA 10mers. RNA/Ribosome/zinc finger triplet/dsDNA complexes are purified, and an emulsion is created that contains 1 complex per compartment along with primers and enzymes (left circle of dashed box) for RT and PCR. RT, and then short overlap extension PCR (see section 5.4.2 i.C) are conducted in the emulsion (right circle of dashed box) to conjoin the bound dsDNA with cDNA from the complex. The conjoined DNA is extracted and sequenced (bottom). Each sequence specifies a zinc finger triplet and a dsDNA bound by its corresponding protein.

subsites. Together these additional pools will enable engineering of three-finger proteins for all possible 9 bp target sites, a substantial improvement over the current OPEN targeting range of one site every ~200 bp. Selections for the additional OPEN finger pools will proceed in a staged fashion, six at a time. Based on previous experience, we anticipate that one technician can obtain six new pools in approximately 4 weeks and therefore perform 68 selections in approximately one year. We will identify new OPEN zinc finger pools using the same randomized libraries and protocols we used to isolate the original pools (107). The existing master randomized zinc finger libraries are in a standard framework consisting of three tandem repeats of the middle finger of the murine transcription factor Zif268 in which the recognition helix residues have been altered. For each library, recognition helix residues –1, 1, 2, 3, 5, and 6 in one of the three fingers were randomized using 24 codons (degenerate sequence 5'VNS3'; V=G, A, or C; S=G or C) encoding 16 amino acids (excluding cysteine and the aromatics). The theoretical complexity of each library is therefore 24^6= ~2 x 10^8 members. Each library has already been converted into infectious M13 phage particles as previously described (63). For each selection that yields surviving colonies, we will pick 10 clones, isolate plasmid DNA, and determine the amino acid sequences of their recognition helices. A finger pool selection will be deemed successful if the recognition helix sequences of the 10 clones resemble each other but reveal few if any identical sequences.

We will archive successful pools of 95 clones each as previously described (107). Following successful completion of all GNN, ANN, TNN, and CNN pools, we will determine the sequences of all zinc fingers in these collections using high throughput sequencing.

Potential problems and alternative approaches: We do not anticipate significant difficulties performing selections for the 68 additional 5'TNN and 3'CNN pools because the Joung lab is the inventor of the protocol and has already successfully isolated 76 pools for 5'GNN and 5'TNN subsites. We have previously described how we handle failed selections which yield ony a small number of highly similar sequences (107). If some selections for 5'TNN or 5'CNN target subsites fail to yield surviving colonies at all, we can consider: (a) Using different randomized zinc finger libraries with stronger binding "anchor" fingers (M. Maeder & J.K. Joung, unpublished observations). (b) Constructing more diverse libraries. Our current 8 year old library used 24 codons to code 16 amino acids and does not contain all possible zinc finger variants. We can now use triplet phosphoramidites (Glen Research) encoding specific codons for amino acids to make libraries with all 20 possible amino acids using only 20 codons. Not only would this library be more diverse but it would be smaller in size (20^6=6.4 x 10^7 vs. 24^6=~1.9 x 10^8). These libraries would be constructed using the Joung Lab's ABI DNA synthesizer. (c) Using different zinc finger frameworks. Instead of using the middle finger from the murine Zif268 protein as a framework for the randomized finger, we will consider alternative framework fingers from other naturally occurring proteins.

5.4.1 i.B Ribosome display to create a rarified/enriched zinc finger library (Figure 5.4.2-1 b and c): We will use ribosome display to interrogate a very large library of zinc finger variants constructed from the comprehensive set of OPEN finger pools isolated in section 5.4.2 i.A above. This large library will be constructed by randomly recombining the 64 pools for each finger position into all possible three-finger array combinations using PCR-based methods previously described by the Joung lab (63, 107). The maximum theoretical complexity of this library will be (64 pools x 95 members/pool)^3 = 2.25 x 10^{11}, a size that could be reasonably constructed using standard ribosome display techniques (196). This large combinatorial library will be interrogated to identify members that possess specific DNA-binding activity. To do this, we will incubate the ribosome display library of zinc finger arrays with a randomized library of all possible 10 bp DNA sequences fused to magnetic beads. This randomized library will include ~1.05 x 10^5 DNA sequences. Following equilibration for 1 hour under conditions similar to those used in phage display (69), the beads with bound zinc finger arrays will be harvested and washed with buffer to remove residual unbound proteins. RNA will be eluted from the beads, reverse transcribed into DNA, amplified by PCR, and a portion sent for high-throughput sequencing to verify that enrichment for sequences has occurred such that there are on average only 10 zinc finger arrays (vs. the original 95^3 = 8.6 x 10^5) derived from each of the possible 64^3=262,144 possible combinations of finger pools. Assignment of a given zinc finger array to a particular combination of zinc finger pools will be based on the sequence information of pool clones determined in section 5.4.2 i.A above. If necessary, this selected pool of finger arrays will be converted again into a ribosome display library and interrogated with the randomized DNA site library for additional enrichment.

Potential problems and alternative approaches: We do not anticipate any difficulties constructing the large multiple pool library because the Joung lab will possess all of the required zinc finger pools and has extensive experience in building large combinatorial libraries (63, 75, 107). If the ribosome display approach described does not work, we will use the bacterial two-hybrid (B2H) selection system developed by co-I Joung as an alternative approach. This system can interrogate libraries with an upper limit of ~10^9 in size, requiring that we build and perform selections on 225 libraries composed of combinations of finger pools targeted to ~1200 different 9 bp sites. For each target site, we will isolate 10 zinc finger arrays from the selection. Pooling these 10 x 262,144 candidates will construct the equivalent of the rarified/enriched zinc finger library proposed above.

5.4.1 i.C Determining the DNA binding specificities of arrays from the rarified/enriched library (Figure 5.4.2-1d): To determine the DNA-binding specificities of the ~2.6 x 10^6 zinc finger arrays in the rarified/enriched library, we will again perform ribosome display using the enriched library from B above. A purified solution of the RNA/ribosome/zinc finger array complexes will be created, and a mixture of all possible double stranded DNA 10mers will be added, allowing the zinc finger array complexes to bind to their DNA targets. The complexes of RNA/ribosome/zinc finger array/bound dsDNA will then be extracted. Next, we will use a variant of our emulsion PCR procedures (138) in combination with a published Short Overlap Extension (SOE) PCR protocol (57) to conjoin the RNA sequence in each complex with the dsDNA bound by the zinc finger array. In brief, an oil-water emulsion containing the purified complexes will be created such that most compartments

contain at most one complex. The complexes will be denatured and reverse transcriptase added to synthesize cDNA from the RNA (using procedures similar to those in section 5.3.1.1), and RNase H will be added to degrade the RNA. Primers and polymerase will then be added to enable a limited PCR reaction that, with overlaps built into the dsDNA and into the enriched library from section 5.4.1 i.B (now represented in the cDNA), will create DNA fragments in which the target dsDNA and the cDNA sequence coding for the zinc finger array are joined. The fragments will be extracted and sequenced on a next generation sequencer. Each fragment will describe a zinc finger array and a dsDNA sequence to which the array bound.

As an alternative to ribosome display, we can use an *in vitro* compartmentalization approach and next generation sequencing. *In vitro* compartmentalization will be used to couple ~2.6 x 10^7 zinc finger arrays (to ensure 10-fold oversampling of the total sequence space) from the rarified/enriched library in section 5.4.1 i.B to 1 μm beads, using an adaptation of the published method of (51): Specifically, DNA fragments that encode fusions of finger arrays to an HA epitope tag, will be coupled to beads such that either only one or no DNA molecule is attached to each bead. These beads will also be coupled, *via* a protein A linkage, to a monoclonal antibody against the HA epitope tag. An emulsion will then be created where each droplet contains no more than a single bead, and *in vitro* transcription and translation will be performed, resulting in beads coated with both the DNA and the protein corresponding to a zinc finger array, The beads will be loaded into a Polonator machine (Preliminary Results, 4.6) flow cell and the DNA will be sequenced to identify the zinc finger arrays on each bead. The beads will then be serially interrogated with labeled clonal DNA fragments, each bearing a single 10 bp binding site, revealing the zinc finger arrays that bind the sites. While means exist to partially parallelize fragments, this system will not have the high throughput of our ribosome display technique. Another alternative is to use B2H as developed in the Joung laboratory. This is again low throughput compared to ribosome display, but we estimate that in a single selection we can fully characterize the DNA-binding specificities of 1000 zinc finger arrays, and by performing ~7500 such selections we can comprehensively probe the DNA-binding specificities of a very large percentage of zinc finger arrays in the rarified/enriched library.

Potential problems and alternative approaches: Variants of both ribosome display and *in vitro* compartmentalization have been tried in the context of zinc finger selections (65, 151). Among issues raised by these studies are: *(a) Non-specific binding of zinc finger arrays:* We are not overly concerned about non-specific binding because the fingers in the pools have already come through B2H selection and therefore have reasonable specificity. The enrichment step in 5.4.1 *i*.B can also be iterated to ensure better specificity. *(b) Non-specific binding of mRNAs to ribosomes, microbeads, dsDNAs:* We can use reverse transcriptase to double strand mRNA that is not bound to ribosomes to reduce the occurrence of RNA secondary structures that encourage non-specific mRNA binding. Studies (65, 151) each involve selection of zinc finger arrays for a *single* target site; thus a potential issue for our approach is *(c) Cross talk between 4^9 target sequences and $2.6x10^6$ zinc finger arrays:* If cross talk proves to be an issue, we can divide the 4^9 target sequences into *N* pools P_1, P_2, … P_N, and the zinc finger arrays into pools A_1, A_2, …, A_N whose predicted targets are in P_1, P_2, … P_N, respectively. We can then reduce cross talk between pools P_i by using microbeads coated with double stranded DNA from the other *N*-1 pools to filter out ribosome display complexes generated from A_i. The methods of 5.4.1 *i*.C could then be applied within each of the *N* pools individually. Pools can be combined during the sequencing phase. For instance, we could create 1024 pools P_i of the form $F_1F_2F_3F_4F_5$NNNN where each of the F_j is fixed as A, C, G, or T in a pool. To completely identify a zinc finger array and its bound target requires sequencing at most 63 bp (18bp for each of three zinc fingers + 9bp for the target) across 4-7 distinct sequence stretches.

5.4.2.ii Development of novel ZFN delivery methods: The use of ZFNs to improve engineering of human cells requires careful control of the activity of the ZFNs in the cells. Enough ZFN must be present to effect replacement of the targeted genomic element by the provided DNA template, but too much ZFN activity is cytotoxic (18). Achieving adequate control over ZFN activity can be difficult when ZFNs are generated by transfected or integrated expression constructs, as these at best enable control over timing and quantity of protein induction vs actual level or activity. An attractive alternative is to deliver calibrated quantities of externally provided ZFN proteins directly into the cells instead of trying to control their intracellular production. This can be achieved by the use of protein transduction domains (PTDs) that can penetrate directly into cells. PTDs harbor a high density of basic amino acid residues (arginine and lysine), which are critical for their transduction function (22, 76). Proteins as large as 110 kDa coupled to a PTD have been transduced into a variety of different cell types and systemic injection of such fusion proteins has demonstrated the effectiveness

PTD-mediated protein delivery *in vivo*. Numerous active PTDs have been described including Penetratin, polylysine, polyarginine, Tat, VP22, Syn B1, FGF-4, anthrax toxin derivative 254-amino acids (aa) peptide segment, diphtheria toxin'R' binding domain, MPG (HIV gp41/SV40 Tag NLS), pep-1, WR peptide, and exotoxin A (see references in (22, 76)).

The use of PTDs to successfully improve delivery of functional Cre recombinase into mammalian cells, both *in vitro* and *in vivo*, has already been demonstrated (73, 103), including into human embryonic stem cells (127), in which recombination efficiencies of 90-100% have been reported (103, 127). We propose to adapt the procedures of (103) to determine a set of PTDs that efficiently transduce ZFNs we generate in Aims 1 and 4.1-4.2 into CTCHGV cell lines. In (103), eleven combinations of PTD domains fused to Cre were tested, and efficiency was measured by reconstitution of an inactive integrated GFP construct, after taking into account factors such as protein yield from recombinant *E. coli*, solubility, fusion protein size and charge, and the conditions and concentrations in which the fusion proteins were provided. Different PTD combinations reportedly varied by factors of as much as 8 in performance. We will apply similar procedures to a set of 5-10 ZFNs, considering as additional variables the quantity and method of delivery of template DNA provided (a component not required by Cre recombinase), using disrupted GFP reporter elements that can be repaired by the template as in section 5.1.1(iii.a).

We will also consider targeted proteolysis as a strategy for controlling intracellular ZFN levels. While PTDs control entry of ZFNs into cells, these methods control their elimination. It has recently been reported that a ZFN fused with an N-terminal degradation tag (based on ubiquitin or the FKBP12 protein) can exhibit equivalent gene targeting efficiency but less toxicity than the corresponding untagged ZFN (141). In this strategy inhibitors are used initially to suppress tag-induced degradation and open a window for ZFN operation.

Potential problems and alternatives: We do not anticipate problems testing these techniques as the methods of (103) and (141) are well documented and accessible. For targeted degradation, PROTACs (PROteolysis TArgeting Chimeras) (147, 198) offer an alternative strategy. Here small molecules or peptides are used to cause a bait domain on a target protein to localize to an E3 enzyme for protein ubiquitination and degradation, an approach similar to one developed in the Church Lab whereby small molecules are used to target proteins directly to proteosome subunits (71). If consistent improvements cannot be achieved, we will rely on existing methods for ZFN expression and accept the lower efficiency that results from ZFN toxicity.

5.4.2.iii. Development of a ZFN-induced "segmental genome replacement" strategy: We propose to develop a novel sequence replacement strategy which will use two ZFN pairs to introduce a pair of DSBs flanking a region of genomic DNA to be altered. Our hypothesis is that this doubly broken stretch of genomic sequence can be repaired by a "donor template" which harbors the desired altered sequence flanked by homology arms composed of sequence adjacent to the two DSBs (Figure 5.4.2-1) We envision that if both ZFN-induced DSBs are introduced into the same allele, the cell might repair the two ends with the donor template as if they came from a single DSB, thereby replacing the original sequence between the two DSBs with the altered sequence from the donor template (Figure 5.4.2-1). In principle, this strategy will allow researchers to completely alter the sequence between the two ZFN-induced DSBs, thereby enabling more complex gene targeting alterations such as exon replacement, and to perform gene targeting even when it is not possible to design ZFNs for a target site close enough to the desired alteration site to achieve high efficiency HR.

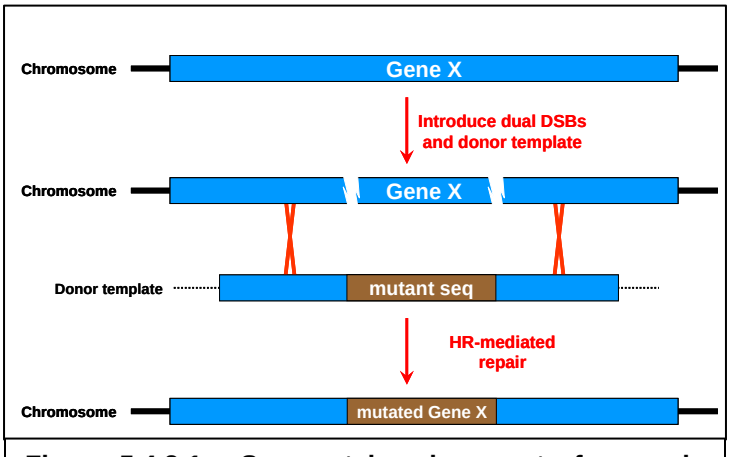


Figure 5.4.2-1 Segmental replacement of genomic sequence by ZFN "dual cut" gene targeting.

A key requirement of this strategy is that it requires efficient introduction of plasmids encoding two ZFN pairs (four ZFN monomers) into a single cell. This can be accomplished by using vectors which express pairs of obligate heterodimeric ZFNs as a single peptide joined by a self-cleaving picornavirus T2A peptide. As noted in Preliminary Studies, section 4.5, we have built a version of such a T2A plasmid that permits rapid and easy shuttling of zinc finger arrays into this vector by simple restriction digest, and have confirmed that our vector can successfully express a functional ZFN dimer. The success of our proposed approach will depend on the efficiency with which the

two ZFN-induced DSBs can be created on the same allele. In this regard, we note that the results cited in Preliminary Results, 4.5 show that co-expression of two pairs of ZFNs targeted to two different sites in the *HoxB13* gene led to deletion of the intervening ~180 bp sequence in ~1% to 2% of the alleles, with evidence that the deletion event is caused by NHEJ-mediated repair of the two DSB ends. This result strongly suggests that a significant percentage of alleles can be "double-cleaved" when two pairs of ZFNs are expressed in the same cell. It is likely that the efficiency of "dual cut" gene targeting will be lower than that observed with existing standard "single-cut" ZFN strategies. Here we will explore whether the long homology arms that will be generated in Aim 1 (section 5.1.1) can boost the rate of HR-mediated segment replacement.

Although the studies described in Preliminary Results, 4.5 suggest that two pairs of heterodimeric ZFNs can be expressed in a single cell, a high degree of cell death was observed in these experiments. This toxicity may be due to the formation of unwanted heterodimer species because the pair of ZFNs are not orthologous in their dimerization specificities. An important requirement for successfully developing the segmental genome replacement approach will therefore be the identification of *Fok*I nuclease domains with orthologous, obligate heterodimeric interaction specificities. To create such domains, we will use a combination of iterative structure-guided design and functional testing as recently described by Miller and colleagues at Sangamo Biosciences (115).

Potential problems and alternatives: Our testing strategy is clear and unproblematic. If this segmental strategy cannot be made to work efficiently, we will rely on the improvements in ZFN HR developed in Aim 1.1. We note that the Church Lab has been developing a successful segmental genome replacement strategy in *E. coli* using selectable markers positioned near the flanks of the template DNA regions to be integrated as part of the work described in Preliminary Results, 4.4.2.

5.4.3: Aim 4.3: We will develop new high-throughput cell handling and sorting capabilities that can incorporate morphology information in addition to optical signals generated by markers, and which can operate on live cells.

In Aims 1.2 and in Aim 3 we propose to develop assays for analyzing up to millions of individual cells for genotype and allele-specific expression (ASE) information, and for *in situ* transcriptome analysis (see sections 5.1.2(ii) and 5.3.1.2). These assays require means of arraying and probing or sequencing within individual cells that are similar in nature to sequencing that is currently performed on arrayed microbeads such as in (155). The main difference is in the need to attach cells vs. microbeads to a flow cell surface and to incorporate treatments (e.g. permeabilization) to the cells that allow these targets to be accessed. Here we develop a system that supports these methods, but also extend it to incorporate aspects of the capability developed in Aim 4.1 (section 5.4.1) above, by which cells may be anchored and selectively released by light-labile chemistry and analysis-based light direction. The result will be a general purpose system for analyzing and sorting cells that enables analysis of morphology in addition to both surface and intracellular molecular content, and for selective release of cells based on morphology and content. The ability of sort cells based on morphology as well as molecular content will be significant expansion of the capabilities of FACS. Here we describe plans for: (i) arraying of cells in a flow cell for image analysis for Aims 1.2 and 3 above, which require neither selective release of cells nor that the cells be maintained alive, (ii) changes needed to support selective release, (iii) considerations needed to support live cells. Then we will describe (iv) demonstrations we will perform of these capabilities.

5.4.3 (i) arraying of cells for sections 5.1.2(ii) and 5.3.1.2: Key requirements are set by the need to use image analysis to analyze cells for sequence-related signals. This requires that the cells be present in a monolayer on a planar surface to ensure uniform focus, and that they be sufficiently well separated that image features can be assigned without error to the proper cells. While these requirements can be met minimally by immobilizing a dilute, disaggregated suspension of cells on a glass slide, our preferred approach is to array cells in a pattern on a flow cell in order to reduce incidental contacts and overlaps between cells. Techniques must be chosen carefully such that cell density and spacing are easily controlled, chemical structures (e.g. DNA and RNA) are not damaged, and attachment chemistries can withstand the forces and time (possibly 3-7 days) required for analysis. To this end, we propose to explore multiple capture and fixation techniques.

Development will begin with the construction of a flow-cell suitable for cell capture. As in Aim 4.1 (section 5.4.1), we will use our Polonator system (see Preliminary Results, 4.6) as a test instrument for development, with expansion to commercial providers of compatible instrumentation to abet technology transfer. Our current flow cell design has 8, 3.3 mm by 70 mm lanes, giving a total surface area of 1848 mm^2

(1.848e9 μm^2). The glass surface of our flow cell will be patterned in the appropriate attachment chemistry using standard photoresist-based lithographic methods such that the attachment chemistry will appear as 5 – 10 μm^2 areas with a center-to-center spacing distance of 15 μm. Assuming a 10 μm cell diameter, this design will achieve a density of 1,026,666 cells per lane (> 8 million/flowcell) with 5 μm spacing between features.

We will explore multiple chemistries to anchor the cells to the flow cell, including both covalent and noncovalent attachment regimes. Each attachment chemistry has inherent advantages and disadvantages. We will test a subset of options by arraying and fixing cells, followed by a single round of probing/sequencing and simulation of a full 4 – 7 day run, and ending with a subsequent round of probing/sequencing. Those methods that prove viable will display accurate readouts and sufficient signal intensity for both the first and last probe/sequence run with minimal loss of cells throughout the simulated run. Our preferred embodiment uses covalent attachment chemistry as they are often stronger than noncovalent interactions. The Bertozzi group recently developed a cell arraying technique that uses an azide functionalized sugar derivative displayed in polysaccharides on the cellular surface and resulted in minimal alterations to cell morphology and proliferation compared to antibody and ligand bound cells (61). The azide functionality can be used with multiple "click" chemistries to functionalize the cell surface with various attachment chemistries including ssDNA. biotin, amines, aldehydes, or direct linkage to alkynal or phosphine derivatized surfaces (Figure 5.4.3-1, Table 5.4.3-1). (86)

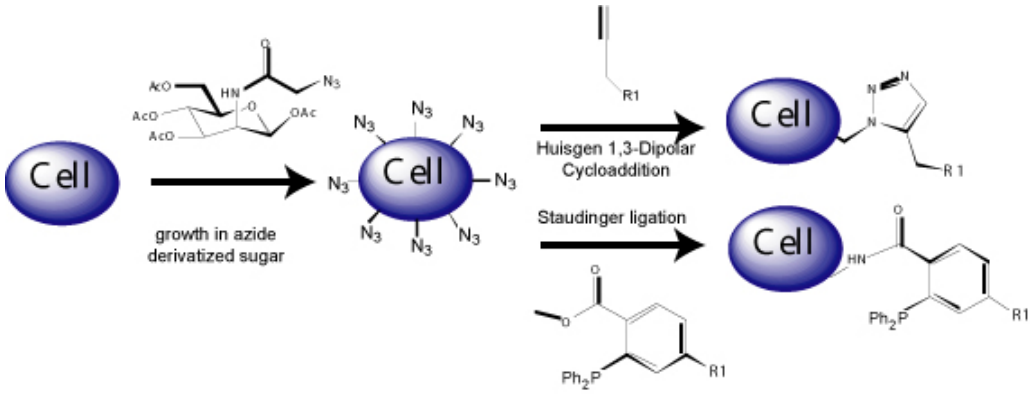


Figure 5.4.3-1. "click" chemistry capture of cells. Cells are grown in the presence of the azido sugar, which is displayed on the surface of cells. Azide groups undergo Huisgen cycloadditions or Staudinger ligations (148) to hetero-bifunctional linkers or solid surfaces. Table 5.4.3-1 describes combinations of hetero-bifunctional linkers with various functional groups (R1), surface coatings, and crosslinking agents.

Glass	R1	Reagent
streptavidine	biotin	None
Amine	Amine	BS3
Amine	aldehyde	$NaBH_4$
Aldehyde	Amine	$NaBH_4$
Alkynal	none	Huisgen
Phosphine	none	Staudinger
ssDNA	ssDNA	None

Table 5.4.3-1. Attachment chemistries

Many of the proposed crosslinking chemistries have been shown to be adequate attachment chemistries for multiplexed DNA sequencing, including streptavidine/biotin, amine/amine with a homo-bifunctional NHS-ester crosslinking reagent, amine/aldehyde reductive aminations, and dsDNA formation (61, 155). While noncovalent, the hybridization of ssDNA displayed on the cell surface with that arrayed in the flowcell adds the advantage that cells can be targeted to specific portions of the array, thus allowing for multiple samples to be probed or sequenced on the same flowcell (20).

In addition to the azide-based cell capture strategies, various noncovalent attachments between arrayed antibodies or ligands have been used (e.g. concanavalin A, laminin, fibronectin) (120). These proteins can be arrayed to an aldehyde derivatized surface via reductive amination (145). However, it may be found that these molecular structures do not hold their native conformation, and thus their binding affinity, throughout the multiple heating and cooling cycles associated with high-throughput DNA sequencing (61). In addition, the high cost associated with antibodies favors click chemistry based approaches.

5.4.3 (ii) selective release of cells: To enable selective release, we functionalize the cell surface for ssDNA attachment, and use a flow cell to which complementary DNA has been attached to the surface *via* nitrobenzyl or other photo-labile chemistry. Release of desired cells can then be accomplished by directing 360nm light to cells, as in section 5.4.1 above.

5.4.3 (iii) considerations for use with live cells: The key requirements are: (a) Cells must be anchored to the flow cell and assayed so that the morphology or phenotype that is to be observed is not altered prior to the point of observation. (b) The conditions and duration of the assay must be sufficiently mild that the cells

will survive through the point of release. These requirements exclude use of stains or hybridizations that fix or significantly perturb the cell, including many that are often used for morphological analysis of such as labeled phalloidin or anti-tubulin antibodies, which allow visualization of the actin microskeleton and microtubules; however, internal structures can be visualized by use of cells containing constructs for fluorescent-protein fusions with these or other structural proteins. The ability to sort cells based on such features goes beyond FACS, which is restricted to analyzing optical properties of cells and surface-bound labels (which, of course, can also be employed by the proposed system). To constrain the times needed for analysis of potentially millions of cell images, morphological features should also be easily computable. Finally, downstream processing of live cells released and captured off the array must analyze aspects of the cell that influenced morphology at the time of observation and which will survive the subsequent process of sorting; these could be genotypes, or they could be transcriptional states that survive or regenerate after the assay. These considerations are taken into account in our proposed demonstrations:

5.4.3 (iv) proposed demonstrations: For (i) above, use of cell array and assay capabilities in support of Aims 1.2 and 3.1 (sections 5.1.2(ii) and 5.3.2.1) will comprise an actual application vs. a demonstration. For analysis and sorting of live cells (iii), we propose: (a) We will use a cell line with fluorescently labeled histone proteins and a labeled translocatable protein such as NF-κB, and sort cells based on degree of localization of the translocatable protein. Determining the degree of translocation to the nucleus is simple from an image analysis point of view, which will shorten the duration of the assay. We will test cell sorting in two ways. (a.i) We will use cells of two different genotypes, one of which constitutively changes the level of localization of the translocatable protein, apply a mixed population to the array, sort based on localization level, and verify that sorting has successfully segregated cells by the genotypes. (a.ii) We will use cells of a single genotype and instead localize a ligand to part of the array that stimulates translocation. Here we will verify that morphology-based sorting successfully segregates cells based on the locations in which the ligand was present in the array. (b) We will attempt to recapitulate (a.i) at the level of RNA vs. genotype by using a cell population that is clonal except that a subpopulation overexpresses a factor that changes localization levels. The test will be to see if sorting successfully segregates the subset of cells that overexpress the factor, and that the RNA levels of this factor are stable through of recover from the conditions of the assay.

Potential problems and alternatives: A key issue for live cells is that they must be allowed to attach via their native mechanisms to suitable ligands to avoid anoikis. This is accomplished above by use of native ligands attached to the surface via a photocleavable substrate (preferably DNA, but also alternatives given in 5.4.2 (ii)). However, once live cells are on the surface, they may begin to migrate and attach to each other. To avoid this, we will lay down the ligand in grids. If this is insufficient, we will explore ways of generating direct cross-links between cell surface proteins and the surface, in effect “leashing” the cells to their locations.

Aim 4 goals: Final goals Our targets are: For Aim 4.1, we will synthesize 1000 complete ZFN proteins using the platform we develop. For Aim 4.2, we will complete the OPEN zinc finger pools and characterize the binding specificity of 10,000 triplet zinc finger arrays from the rarified library using the described ribosome display system. For Aim 4.3, we will build the image analysis and cell arraying required for Aims 1.2 and 3 on the Polonator, and demonstrate cell sorting by morphology. Intermediate goals As noted in our Research Design Overview, we will evaluate progress at the end of year 2 of the Center and renegotiate goals as appropriate. By that time we expect that: *Aim 4.1:* We will have demonstrated the ability to selectively release and capture microbeads based on DNA sequences on them required of the system in Aim 4.1. *Aim 4.2:* We expect to have completed all of the OPEN pool selections, and to have successfully tested the compartmentalized SOE-PCR that conjoins RNA bound to ribosomes with target DNA of zinc finger triplets, for simple mixtures of triplets *vs* the full combinatorial library. *Aim 4.3:* Cell handling for Aim 1.3 and 3 will have been completed. Impacts*:* The integrated DNA sequencing and synthesis platform will greatly improve the ability to create complex libraries of large DNA constructs and will likely be adopted commercially. Completion of the OPEN zinc finger pools and development of improved ZFN targeting and delivery techniques will put effective human cell genetic engineering in the hands of the research community, where it will broadly support biomedical research generally, and gene therapy in particular. The ribosome display experiments of Aim 4.2 will generate an extremely large data base of zinc finger array specificities and will provide unparalleled opportunities to develop new computational methods for designing arrays with new binding specificities. Extending cell sorting technology to incorporate cell morphology along with cell staining characteristics will be the basis of a new form of high-throughput screening that will have broad application in research.

Bibliography and References

MGI CEGS Publication Bibliography

The following is a list of publications of the MGI Center broken down by category.

- MGIC 2004-2009 publications (excludes MGI first year (2003) publications): 44
- MGIC submitted: 1
- MGIC conference proceedings: 1
- MGIC electronic : 1
- MGIC 2003 publications: 8

MGIC 2004-2009 publications (44)

Aach, J. and G. M. Church (2004). "Mathematical models of diffusion-constrained polymerase chain reactions: basis of high-throughput nucleic acid assays and simple self-organizing systems." J Theor Biol **228**(1): 31-46.

Bakal, C., J. Aach, G. Church and N. Perrimon (2007). "Quantitative morphological signatures define local signaling networks regulating cell morphology." Science **316**(5832): 1753-6.

Ball, M. P., J. B. Li, Y. Gao, J. Lee, E. LeProust, I.-H. Park, B. Xie, G. Q. Daley and G. M. Church (2009). "Targeted and whole-genome methylomics reveals gene-body signatures in human cell lines." Nat Biotechnol **27**(4): 361-368.

Church, G. M. (2006). "Genomes for all." Sci Am **294**(1): 46-54.

Church, G. M., G. J. Porreca, R. C. Terry and M. Lares (2008). "High-Speed Imaging for DNA Sequencing." Biophotonics (http://www.photonics.com/Content/ReadArticle.aspx?ArticleID=33989).

Conrad, C., J. Zhu, C. Conrad, D. Schoenfeld, Z. Fang, M. Ingelsson, S. Stamm, G. Church and B. T. Hyman (2007). "Single molecule profiling of tau gene expression in Alzheimer's disease." J Neurochem **103**(3): 1228-36.

Dantas, G., M. O. Sommer, R. D. Oluwasegun and G. M. Church (2008). "Bacteria subsisting on antibiotics." Science **320**(5872): 100-3.

Kim, D. S., S. E. Ross, J. M. Trimarchi, J. Aach, M. E. Greenberg and C. L. Cepko (2008). "Identification of molecular markers of bipolar cells in the murine retina." J Comp Neurol **507**(5): 1795-810.

Kim, J. B., G. J. Porreca, L. Song, S. C. Greenway, J. M. Gorham, G. M. Church, C. E. Seidman and J. G. Seidman (2007). "Polony multiplex analysis of gene expression (PMAGE) in mouse hypertrophic cardiomyopathy." Science **316**(5830): 1481-4.

Lee, C. S., L. Y. Tee, S. Dusenbery, T. Takata, J. P. Golden, B. A. Pierchala, D. I. Gottlieb, E. M. Johnson, Jr., D. W. Choi and B. J. Snider (2005). "Neurotrophin and GDNF family ligands promote survival and alter excitotoxic vulnerability of neurons derived from murine embryonic stem cells." Exp Neurol **191**(1): 65-76.

Lee, H. S., J. L. Sherley, J. J. Chen, C. C. Chiu, L. L. Chiou, J. D. Liang, P. C. Yang, G. T. Huang and J. C. Sheu (2005). "EMP-1 is a junctional protein in a liver stem cell line and in the liver." Biochem Biophys Res Commun **334**(4): 996-1003.

Lee, S. I., D. Pe'er, A. M. Dudley, G. M. Church and D. Koller (2006). "Identifying regulatory mechanisms using individual variation reveals key role for chromatin modification." Proc Natl Acad Sci U S A **103**(38): 14062-7. PMC ID: PMC1599912.

Leparc, G. G. and R. D. Mitra (2007). "Non-EST-based prediction of novel alternatively spliced cassette exons with cell signaling function in Caenorhabditis elegans and human." Nucleic Acids Res **35**(10): 3192-202. PMC ID: PMC1904267.

Leparc, G. G. and R. D. Mitra (2007). "A sensitive procedure to detect alternatively spliced mRNA in pooled-tissue samples." Nucleic Acids Res **35**(21): e146. PMC ID: PMC2175357.

Li, J. B., Y. Gao, J. Aach, K. Zhang, G. V. Kryukov, B. Xie, A. Ahlford, J.-K. Yoon, A. M. Rosenbaum, A. Wait-Zaranek, E. LeProust, S. Sunyaev and G. M. Church (2009). "Multiplex padlock capture and sequencing reveal human hypermutable CpG variations." Genome Res: in press.

Li, J. B., E. Y. Levanon, J.-K. Yoon, J. Aach, B. Xie, E. LeProust, K. Zhang, Y. Gao and C. G.M. (2009). "Genome-wide Identification of Human RNA Editing Sites by Parallel DNA Capturing and Sequencing." Science: in press.

Lunshof, J. E. (2006). "Personalized medicine: new perspectives – new ethics?" Personalized Med **3**(2): 187-194.

Mikkilineni, V., R. D. Mitra, J. Merritt, J. R. DiTonno, G. M. Church, B. Ogunnaike and J. S. Edwards (2004). "Digital quantitative measurements of gene expression." Biotechnol Bioeng **86**(2): 117-24.

Nardi, V., T. Raz, X. Cao, C. J. Wu, R. M. Stone, J. Cortes, M. W. Deininger, G. Church, J. Zhu and G. Q. Daley (2008). "Quantitative monitoring by polymerase colony assay of known mutations resistant to ABL kinase inhibitors." Oncogene **27**(6): 775-82.

Pare, J. F. and J. L. Sherley (2006). "Biological principles for ex vivo adult stem cell expansion." Curr Top Dev Biol **73**: 141-71.

Porreca, G. J., J. Shendure and G. M. Church (2006). "Polony DNA sequencing." Curr Protoc Mol Biol **Chapter 7**: Unit 7 8.

Porreca, G. J., K. Zhang, J. B. Li, B. Xie, D. Austin, S. L. Vassallo, E. M. LeProust, B. J. Peck, C. J. Emig, F. Dahl, Y. Gao, G. M. Church and J. Shendure (2007). "Multiplex amplification of large sets of human exons." Nat Methods **4**(11): 931-6.

Rambhatla, L., S. Ram-Mohan, J. J. Cheng and J. L. Sherley (2005). "Immortal DNA strand cosegregation requires p53/IMPDH-dependent asymmetric self-renewal associated with adult stem cells." Cancer Res **65**(8): 3155-61.

Rieger, C., R. Poppino, R. Sheridan, K. Moley, R. Mitra and D. Gottlieb (2007). "Polony analysis of gene expression in ES cells and blastocysts." Nucleic Acids Res **35**(22): e151. PMC ID: PMC2190707.

Schwartz, D., M. F. Chou and G. M. Church (2009). "Predicting protein post-translational modifications using meta-analysis of proteome scale data sets." Mol Cell Proteomics **8**(2): 365-79. PMC ID: PMC2634583.

Shendure, J., R. D. Mitra, C. Varma and G. M. Church (2004). "Advanced sequencing technologies: methods and goals." Nat Rev Genet **5**(5): 335-44.

Shendure, J., G. J. Porreca, N. B. Reppas, X. Lin, J. P. McCutcheon, A. M. Rosenbaum, M. D. Wang, K. Zhang, R. D. Mitra and G. M. Church (2005). "Accurate multiplex polony sequencing of an evolved bacterial genome." Science **309**(5741): 1728-32.

Shendure, J. A., G. J. Porreca and G. M. Church (2008). "Overview of DNA sequencing strategies." Curr Protoc Mol Biol **Chapter 7**: Unit 7 1.

Sherley, J. L. (2007). "Commentary: Facing up to the feasibility of ANT-OAR." Stem Cell Rev **3**(1): 66-7.

Sherley, J. L. (2008). "All good cells come from cells." Nat Cell Biol **10**(3): 248.

Tannenbaum, E., J. L. Sherley and E. I. Shakhnovich (2004). "Imperfect DNA lesion repair in the semiconservative quasispecies model: derivation of the Hamming class equations and solution of the single-fitness peak landscape." Phys Rev E Stat Nonlin Soft Matter Phys **70**(6 Pt 1): 061915.

Tannenbaum, E., J. L. Sherley and E. I. Shakhnovich (2005). "Evolutionary dynamics of adult stem cells: comparison of random and immortal-strand segregation mechanisms." Phys Rev E Stat Nonlin Soft Matter Phys **71**(4 Pt 1): 041914.

Tannenbaum, E., J. L. Sherley and E. I. Shakhnovich (2006). "Semiconservative quasispecies equations for polysomic genomes: the haploid case." J Theor Biol **241**(4): 791-805.

Turner, D. J., J. Shendure, G. Porreca, G. Church, P. Green, C. Tyler-Smith and M. E. Hurles (2006). "Assaying chromosomal inversions by single-molecule haplotyping." Nat Methods **3**(6): 439-45.

Vigneault, F., A. M. Sismour and G. M. Church (2008). "Efficient microRNA capture and bar-coding via enzymatic oligonucleotide adenylation." Nat Methods **5**(9): 777-779.

Wang, H., M. Johnston and R. D. Mitra (2007). "Calling cards for DNA-binding proteins." Genome Res **17**(8): 1202-9. PMC ID: PMC1933518.

Wei, L., L. Cui, B. J. Snider, M. Rivkin, S. S. Yu, C. S. Lee, L. D. Adams, D. I. Gottlieb, E. M. Johnson, Jr., S. P. Yu and D. W. Choi (2005). "Transplantation of embryonic stem cells overexpressing Bcl-2 promotes functional recovery after transient cerebral ischemia." Neurobiol Dis **19**(1-2): 183-93.

Willerth, S. M., K. J. Arendas, D. I. Gottlieb and S. E. Sakiyama-Elbert (2006). "Optimization of fibrin scaffolds for differentiation of murine embryonic stem cells into neural lineage cells." Biomaterials **27**(36): 5990-6003. PMC ID: PMC1794024.

Willerth, S. M., T. E. Faxel, D. I. Gottlieb and S. E. Sakiyama-Elbert (2007). "The effects of soluble growth factors on embryonic stem cell differentiation inside of fibrin scaffolds." Stem Cells **25**(9): 2235-44. PMC ID: PMC2637150.

Xian, H. and D. I. Gottlieb (2004). "Dividing Olig2-expressing progenitor cells derived from ES cells." Glia **47**(1): 88-101.

Xian, H. Q., K. Werth and D. I. Gottlieb (2005). "Promoter analysis in ES cell-derived neural cells." Biochem Biophys Res Commun **327**(1): 155-62.

Zaranek, A. W., W. Clegg, Vandewege and G. M. Church (2008). Free Factories: Unified Infrastructure for Data Intensive Web Services USENIX Annual Technical Confrerence, Boston, MA.

Zhang, K., A. C. Martiny, N. B. Reppas, K. W. Barry, J. Malek, S. W. Chisholm and G. M. Church (2006). "Sequencing genomes from single cells by polymerase cloning." Nat Biotechnol **24**(6): 680-6.

Zhang, K., J. Zhu, J. Shendure, G. J. Porreca, J. D. Aach, R. D. Mitra and G. M. Church (2006). "Long-range polony haplotyping of individual human chromosome molecules." Nat Genet **38**(3): 382-7.

MGIC submitted (1)

Bakal, C., M. Baym, J. Aach, B. Berger and N. Perrimon (2009). "Systems Modeling of the Nonlinear Rho Signaling Network (submitted to Cell)."

MGIC conference proceedings (1)

Forest, C. R., S. E. Ross and G. M. Church (2008). DNA Sequencing By Ligation On Surface-Bound Beads In A Microchannel Environment (Paper ID No. 0261). microTAS, San Diego.

MGIC electronic(1)

Terry, R., G. Porreca, K. McCarthy and G. M. Church (2008) "Polonator Instrument http://www.polonator.org "

MGIC 2003 publications (8)

Adams, L. D., L. Choi, H. Q. Xian, A. Yang, B. Sauer, L. Wei and D. I. Gottlieb (2003). "Double lox targeting for neural cell transgenesis." Brain Res Mol Brain Res **110**(2): 220-33.

Lee, H. S., G. G. Crane, J. R. Merok, J. R. Tunstead, N. L. Hatch, K. Panchalingam, M. J. Powers, L. G. Griffith and J. L. Sherley (2003). "Clonal expansion of adult rat hepatic stem cell lines by suppression of asymmetric cell kinetics (SACK)." Biotechnol Bioeng **83**(7): 760-71.

Merritt, J., J. R. DiTonno, R. D. Mitra, G. M. Church and J. S. Edwards (2003). "Parallel competition analysis of Saccharomyces cerevisiae strains differing by a single base using polymerase colonies." Nucleic Acids Res **31**(15): e84. PMC ID: PMC169973.

Mitra, R. D., V. L. Butty, J. Shendure, B. R. Williams, D. E. Housman and G. M. Church (2003). "Digital genotyping and haplotyping with polymerase colonies." Proc Natl Acad Sci U S A **100**(10): 5926-31. PMC ID: PMC156303.

Mitra, R. D., J. Shendure, J. Olejnik, O. Edyta Krzymanska and G. M. Church (2003). "Fluorescent in situ sequencing on polymerase colonies." Anal Biochem **320**(1): 55-65.

Qu, Y., S. Vadivelu, L. Choi, S. Liu, A. Lu, B. Lewis, R. Girgis, C. S. Lee, B. J. Snider, D. I. Gottlieb and J. W. McDonald (2003). "Neurons derived from embryonic stem (ES) cells resemble normal neurons in their vulnerability to excitotoxic death." Exp Neurol **184**(1): 326-36.

Xian, H. Q., E. McNichols, A. St Clair and D. I. Gottlieb (2003). "A subset of ES-cell-derived neural cells marked by gene targeting." Stem Cells **21**(1): 41-9.

Zhu, J., J. Shendure, R. D. Mitra and G. M. Church (2003). "Single molecule profiling of alternative pre-mRNA splicing." Science **301**(5634): 836-8.

References cited in this proposal

1. Aasen T, Raya A, Barrero MJ, Garreta E, Consiglio A, Gonzalez F, Vassena R, Bilic J, Pekarik V, Tiscornia G, Edel M, Boue S, Belmonte JC. 2008. Efficient and rapid generation of induced pluripotent stem cells from human keratinocytes. *Nat Biotechnol* 26: 1276-84.
2. Aiuti A, Bachoud-Levi AC, Blesch A, Brenner MK, Cattaneo F, Chiocca EA, Gao G, High KA, Leen AM, Lemoine NR, McNeish IA, Meneguzzi G, Peschanski M, Roncarolo MG, Strayer DS, Tuszynski MH, Waxman DJ, Wilson JM. 2007. Progress and prospects: gene therapy clinical trials (part 2). *Gene Ther* 14: 1555-63.
3. Alexander BL, Ali RR, Alton EW, Bainbridge JW, Braun S, Cheng SH, Flotte TR, Gaspar HB, Grez M, Griesenbach U, Kaplitt MG, Ott MG, Seger R, Simons M, Thrasher AJ, Thrasher AZ, Yla-Herttuala S. 2007. Progress and prospects: gene therapy clinical trials (part 1). *Gene Ther* 14: 1439-47.
4. Altshuler D, Daly MJ, Lander ES. 2008. Genetic mapping in human disease. *Science* 322: 881-8.
5. Anderson P, Kedersha N. 2006. RNA granules. *J Cell Biol* 172: 803-8. PMC ID: PMC2063724.
6. Anderson P, Kedersha N. 2008. Stress granules: the Tao of RNA triage. *Trends Biochem Sci* 33: 141-50.
7. Bakal C, Aach J, Church G, Perrimon N. 2007. Quantitative morphological signatures define local signaling networks regulating cell morphology. *Science* 316: 1753-6.
8. Bakal C, Baym M, Aach J, Berger B, Perrimon N. 2009. Systems Modeling of the Nonlinear Rho Signaling Network (submitted to Cell).
9. Balding DJ. 2006. A tutorial on statistical methods for population association studies. *Nat Rev Genet* 7: 781-91.
10. Ball MP, Li JB, Gao Y, Lee J, LeProust E, Park I-H, Xie B, Daley GQ, Church GM. 2009. Targeted and whole-genome methylomics reveals gene-body signatures in human cell lines. *Nat Biotechnol* 27: 361-8.
11. Bang D, Church GM. 2008. Gene synthesis by circular assembly amplification. *Nat Methods* 5: 37-9.
12. Behlke MA, Devor EJ. 2005. *Chemical Synthesis of Oligonucleotides (http://www.idtdna.com/Support/Technical/TechnicalBulletinPDF/Chemical_Synthesis_of_Oligonucleotides.pdf)*, Integrated DNA Technologies

13. Bell J. 2004. Predicting disease using genomics. *Nature* 429: 453-6.
14. Bitinaite J, Wah DA, Aggarwal AK, Schildkraut I. 1998. FokI dimerization is required for DNA cleavage. *Proc Natl Acad Sci U S A* 95: 10570-5. PMC ID: PMC27935.
15. Brenner S, Johnson M, Bridgham J, Golda G, Lloyd DH, Johnson D, Luo S, McCurdy S, Foy M, Ewan M, Roth R, George D, Eletr S, Albrecht G, Vermaas E, Williams SR, Moon K, Burcham T, Pallas M, DuBridge RB, Kirchner J, Fearon K, Mao J, Corcoran K. 2000. Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays. *Nat Biotechnol* 18: 630-4.
16. Brenner S, Williams SR, Vermaas EH, Storck T, Moon K, McCollum C, Mao JI, Luo S, Kirchner JJ, Eletr S, DuBridge RB, Burcham T, Albrecht G. 2000. In vitro cloning of complex mixtures of DNA on microbeads: physical separation of differentially expressed cDNAs. *Proc Natl Acad Sci U S A* 97: 1665-70. PMC ID: PMC26493.
17. Carr PA, Park JS, Lee YJ, Yu T, Zhang S, Jacobson JM. 2004. Protein-mediated error correction for de novo DNA synthesis. *Nucleic Acids Res* 32: e162. PMC ID: PMC534640.
18. Cathomen T, Joung JK. 2008. Zinc-finger nucleases: the next generation emerges. *Mol Ther* 16: 1200-7.
19. Cavazzana-Calvo M, Fischer A. 2007. Gene therapy for severe combined immunodeficiency: are we there yet? *J Clin Invest* 117: 1456-65. PMC ID: PMC1878528.
20. Chandra RA, Douglas ES, Mathies RA, Bertozzi CR, Francis MB. 2006. Programmable Cell Adhesion Encoded by DNA Hybridization *Angew. Chem. Int.* 45: 896-901.
21. Chapman JM, Cooper JD, Todd JA, Clayton DG. 2003. Detecting disease associations due to linkage disequilibrium using haplotype tags: a class of tests and the determinants of statistical power. *Hum Hered* 56: 18-31.
22. Chauhan A, Tikoo A, Kapur AK, Singh M. 2007. The taming of the cell penetrating domain of the HIV Tat: myths and realities. *J Control Release* 117: 148-62. PMC ID: PMC1859861.
23. Chen Y, Zhu J, Lum PY, Yang X, Pinto S, MacNeil DJ, Zhang C, Lamb J, Edwards S, Sieberts SK, Leonardson A, Castellini LW, Wang S, Champy MF, Zhang B, Emilsson V, Doss S, Ghazalpour A, Horvath S, Drake TA, Lusis AJ, Schadt EE. 2008. Variations in DNA elucidate molecular networks that cause disease. *Nature* 452: 429-35.
24. Cheung VG, Spielman RS, Ewens KG, Weber TM, Morley M, Burdick JT. 2005. Mapping determinants of human gene expression by regional and genome-wide association. *Nature* 437: 1365-9.
25. Chiang DY, Getz G, Jaffe DB, O'Kelly MJ, Zhao X, Carter SL, Russ C, Nusbaum C, Meyerson M, Lander ES. 2009. High-resolution mapping of copy-number alterations with massively parallel sequencing. *Nat Methods* 6: 99-103. PMC ID: PMC2630795.
26. Choy E, Yelensky R, Bonakdar S, Plenge RM, Saxena R, De Jager PL, Shaw SY, Wolfish CS, Slavik JM, Cotsapas C, Rivas M, Dermitzakis ET, Cahir-McFarland E, Kieff E, Hafler D, Daly MJ, Altshuler D. 2008. Genetic analysis of human traits in vitro: drug response and gene expression in lymphoblastoid cell lines. *PLoS Genet* 4: e1000287. PMC ID: PMC2583954.
27. Christian AT, Pattee MS, Attix CM, Reed BE, Sorensen KJ, Tucker JD. 2001. Detection of DNA point mutations and mRNA expression levels by rolling circle amplification in individual cells. *Proc Natl Acad Sci U S A* 98: 14238-43. PMC ID: PMC64666.
28. Church GM. 2006. Genomes for all. *Sci Am* 294: 46-54.
29. Church GM, Porreca GJ, Terry RC, Lares M. 2008. High-Speed Imaging for DNA Sequencing. *Biophotonics (http://www.photonics.com/Content/ReadArticle.aspx?ArticleID=33989)*.
30. Claassen DA, Desler MM, Rizzino A. 2009. ROCK inhibition enhances the recovery and growth of cryopreserved human embryonic stem cells and human induced pluripotent stem cells. *Mol Reprod Dev*.
31. Conrad C, Gupta R, Mohan H, Niess H, Bruns CJ, Kopp R, von Luettichau I, Guba M, Heeschen C, Jauch KW, Huss R, Nelson PJ. 2007. Genetically engineered stem cells for therapeutic gene delivery. *Curr Gene Ther* 7: 249-60.
32. Cookson W, Liang L, Abecasis G, Moffatt M, Lathrop M. 2009. Mapping complex disease traits with global gene expression. *Nat Rev Genet* 10: 184-94.
33. Costantino N, Court DL. 2003. Enhanced levels of lambda Red-mediated recombinants in mismatch repair mutants. *Proc Natl Acad Sci U S A* 100: 15748-53. PMC ID: PMC307639.
34. Datta S, Costantino N, Zhou X, Court DL. 2008. Identification and analysis of recombineering functions from Gram-negative and Gram-positive bacteria and their phages. *Proc Natl Acad Sci U S A* 105: 1626-31. PMC ID: PMC2234195.
35. de Bakker PI, Yelensky R, Pe'er I, Gabriel SB, Daly MJ, Altshuler D. 2005. Efficiency and power in genetic association studies. *Nat Genet* 37: 1217-23.
36. Dekker M, Brouwers C, te Riele H. 2003. Targeted gene modification in mismatch-repair-deficient embryonic stem cells by single-stranded DNA oligonucleotides. *Nucleic Acids Res* 31: e27. PMC ID: PMC152881.
37. Deng J, Shoemaker R, Xie B, Gore A, LeProust EM, Antosiewicz-Bourget J, Egli D, Maherali N, Park IH, Yu J, Daley GQ, Eggan K, Hochedlinger K, Thomson J, Wang W, Gao Y, Zhang K. 2009. Targeted bisulfite sequencing reveals changes in DNA methylation associated with nuclear reprogramming. *Nat Biotechnol* 27: 353-60.

38. Diabetes Genetics Initiative of Broad Institute of Harvard and Mit LUaNIoBR, Saxena R, Voight BF, Lyssenko V, Burtt NP, de Bakker PI, Chen H, Roix JJ, Kathiresan S, Hirschhorn JN, Daly MJ, Hughes TE, Groop L, Altshuler D, Almgren P, Florez JC, Meyer J, Ardlie K, Bengtsson Bostrom K, Isomaa B, Lettre G, Lindblad U, Lyon HN, Melander O, Newton-Cheh C, Nilsson P, Orho-Melander M, Rastam L, Speliotes EK, Taskinen MR, Tuomi T, Guiducci C, Berglund A, Carlson J, Gianniny L, Hackett R, Hall L, Holmkvist J, Laurila E, Sjogren M, Sterner M, Surti A, Svensson M, Svensson M, Tewhey R, Blumenstiel B, Parkin M, Defelice M, Barry R, Brodeur W, Camarata J, Chia N, Fava M, Gibbons J, Handsaker B, Healy C, Nguyen K, Gates C, Sougnez C, Gage D, Nizzari M, Gabriel SB, Chirn GW, Ma Q, Parikh H, Richardson D, Ricke D, Purcell S. 2007. Genome-wide association analysis identifies loci for type 2 diabetes and triglyceride levels. *Science* 316: 1331-6.
39. Ding Y. 2006. Statistical and Bayesian approaches to RNA secondary structure prediction. *RNA* 12: 323-31. PMC ID: PMC1383571.
40. Dixon AL, Liang L, Moffatt MF, Chen W, Heath S, Wong KC, Taylor J, Burnett E, Gut I, Farrall M, Lathrop GM, Abecasis GR, Cookson WO. 2007. A genome-wide association study of global gene expression. *Nat Genet* 39: 1202-7.
41. Donoho G, Jasin M, Berg P. 1998. Analysis of gene targeting and intrachromosomal homologous recombination stimulated by genomic double-strand breaks in mouse embryonic stem cells. *Mol Cell Biol* 18: 4070-8. PMC ID: PMC108991.
42. Doyon Y, McCammon JM, Miller JC, Faraji F, Ngo C, Katibah GE, Amora R, Hocking TD, Zhang L, Rebar EJ, Gregory PD, Urnov FD, Amacher SL. 2008. Heritable targeted gene disruption in zebrafish using designed zinc-finger nucleases. *Nat Biotechnol* 26: 702-8. PMC ID: PMC2674762.
43. Drubin DA, Way JC, Silver PA. 2007. Designing biological systems. *Genes Dev* 21: 242-54.
44. Eberwine J, Kacharmina JE, Andrews C, Miyashiro K, McIntosh T, Becker K, Barrett T, Hinkle D, Dent G, Marciano P. 2001. mRna expression analysis of tissue sections and single cells. *J Neurosci* 21: 8310-4.
45. Eid J, Fehr A, Gray J, Luong K, Lyle J, Otto G, Peluso P, Rank D, Baybayan P, Bettman B, Bibillo A, Bjornson K, Chaudhuri B, Christians F, Cicero R, Clark S, Dalal R, Dewinter A, Dixon J, Foquet M, Gaertner A, Hardenbol P, Heiner C, Hester K, Holden D, Kearns G, Kong X, Kuse R, Lacroix Y, Lin S, Lundquist P, Ma C, Marks P, Maxham M, Murphy D, Park I, Pham T, Phillips M, Roy J, Sebra R, Shen G, Sorenson J, Tomaney A, Travers K, Trulson M, Vieceli J, Wegener J, Wu D, Yang A, Zaccarin D, Zhao P, Zhong F, Korlach J, Turner S. 2009. Real-time DNA sequencing from single polymerase molecules. *Science* 323: 133-8.
46. Emilsson V, Thorleifsson G, Zhang B, Leonardson AS, Zink F, Zhu J, Carlson S, Helgason A, Walters GB, Gunnarsdottir S, Mouy M, Steinthorsdottir V, Eiriksdottir GH, Bjornsdottir G, Reynisdottir I, Gudbjartsson D, Helgadottir A, Jonasdottir A, Jonasdottir A, Styrkarsdottir U, Gretarsdottir S, Magnusson KP, Stefansson H, Fossdal R, Kristjansson K, Gislason HG, Stefansson T, Leifsson BG, Thorsteinsdottir U, Lamb JR, Gulcher JR, Reitman ML, Kong A, Schadt EE, Stefansson K. 2008. Genetics of gene expression and its effect on disease. *Nature* 452: 423-8.
47. Eulalio A, Behm-Ansmant I, Izaurralde E. 2007. P bodies: at the crossroads of post-transcriptional pathways. *Nat Rev Mol Cell Biol* 8: 9-22.
48. Foley JE, Yeh JR, Maeder ML, Reyon D, Sander JD, Peterson RT, Joung JK. 2009. Rapid mutation of endogenous zebrafish genes using zinc finger nucleases made by Oligomerized Pool ENgineering (OPEN). *PLoS ONE* 4: e4348. PMC ID: PMC2634973.
49. Frayling TM, Timpson NJ, Weedon MN, Zeggini E, Freathy RM, Lindgren CM, Perry JR, Elliott KS, Lango H, Rayner NW, Shields B, Harries LW, Barrett JC, Ellard S, Groves CJ, Knight B, Patch AM, Ness AR, Ebrahim S, Lawlor DA, Ring SM, Ben-Shlomo Y, Jarvelin MR, Sovio U, Bennett AJ, Melzer D, Ferrucci L, Loos RJ, Barroso I, Wareham NJ, Karpe F, Owen KR, Cardon LR, Walker M, Hitman GA, Palmer CN, Doney AS, Morris AD, Smith GD, Hattersley AT, McCarthy MI. 2007. A common variant in the FTO gene is associated with body mass index and predisposes to childhood and adult obesity. *Science* 316: 889-94. PMC ID: PMC2646098.
50. Gimelbrant A, Hutchinson JN, Thompson BR, Chess A. 2007. Widespread monoallelic expression on human autosomes. *Science* 318: 1136-40.
51. Griffiths AD, Tawfik DS. 2006. Miniaturising the laboratory in emulsion droplets. *Trends Biotechnol* 24: 395-402.
52. Grillot-Courvalin C, Goussard S, Huetz F, Ojcius DM, Courvalin P. 1998. Functional gene transfer from intracellular bacteria to mammalian cells. *Nat Biotechnol* 16: 862-6.
53. Guccione E, Martinato F, Finocchiaro G, Luzi L, Tizzoni L, Dall' Olio V, Zardo G, Nervi C, Bernard L, Amati B. 2006. Myc-binding-site recognition in the human genome is determined by chromatin context. *Nat Cell Biol* 8: 764-70.
54. Hanna J, Wernig M, Markoulaki S, Sun CW, Meissner A, Cassady JP, Beard C, Brambrink T, Wu LC, Townes TM, Jaenisch R. 2007. Treatment of sickle cell anemia mouse model with iPS cells generated from autologous skin. *Science* 318: 1920-3.

55. Harris TD, Buzby PR, Babcock H, Beer E, Bowers J, Braslavsky I, Causey M, Colonell J, Dimeo J, Efcavitch JW, Giladi E, Gill J, Healy J, Jarosz M, Lapen D, Moulton K, Quake SR, Steinmann K, Thayer E, Tyurina A, Ward R, Weiss H, Xie Z. 2008. Single-molecule DNA sequencing of a viral genome. *Science* 320: 106-9.
56. Hayden EC. 2008. Give me my genome (October 21, 2008). In *Nature* (*http://www.nature.com.ezp-prod1.hul.harvard.edu/news/2008/081021/full/news.2008.1182.html*)
57. Heckman KL, Pease LR. 2007. Gene splicing and mutagenesis by PCR-driven overlap extension. *Nat Protoc* 2: 924-32.
58. Heinemann JA, Sprague GF, Jr. 1989. Bacterial conjugative plasmids mobilize DNA transfer between bacteria and yeast. *Nature* 340: 205-9.
59. Hindorff LA, Junkins HA, Mehta JP, Manolio TA. 2009. A Catalog of Published Genome-Wide Association Studies. Available at: www.genome.gov/26525384. Accessed May 12, 2009.
60. Holden C. 2008. Genomes for the world. *Science* 322: 509.
61. Hsiao SC, Crow AK, Lam WA, Bertozzi CR, Fletcher DA, Francis MB. 2008. DNA-Coated AFM Cantilevers for the Investigation of Cell Adhesion and the Patterning of Live Cells. *Angew. Chem. Int.* 47: 8473-7.
62. Huangfu D, Osafune K, Maehr R, Guo W, Eijkelenboom A, Chen S, Muhlestein W, Melton DA. 2008. Induction of pluripotent stem cells from primary human fibroblasts with only Oct4 and Sox2. *Nat Biotechnol* 26: 1269-75.
63. Hurt JA, Thibodeau SA, Hirsh AS, Pabo CO, Joung JK. 2003. Highly specific zinc finger proteins obtained by directed domain shuffling and cell-based selection. *Proc Natl Acad Sci U S A* 100: 12271-6. PMC ID: PMC218748.
64. Igoucheva O, Alexeev V, Yoon K. 2004. Oligonucleotide-directed mutagenesis and targeted gene correction: a mechanistic point of view. *Curr Mol Med* 4: 445-63.
65. Ihara H, Mie M, Funabashi H, Takahashi F, Sawasaki T, Endo Y, Kobatake E. 2006. In vitro selection of zinc finger DNA-binding proteins through ribosome display. *Biochem Biophys Res Commun* 345: 1149-54.
66. Illumina Corp. 2009. Illumina Presents Development Roadmap for Scaling its Genome Analyzer Innovations to substantially increase output, decrease cost, and expand applications (http://investor.illumina.com/phoenix.zhtml?c=121127&p=irol-newsArticle&ID=1252407&highlight=).
67. International 1000 Genomes Consortium. 1000 Genomes Project (http://www.1000genomes.org/).
68. International HapMap C. 2005. A haplotype map of the human genome. *Nature* 437: 1299-320. PMC ID: PMC1880871.
69. Isalan M, Choo Y. 2001. Engineering nucleic acid-binding proteins by phage display. *Methods Mol Biol* 148: 417-29.
70. Iyer VR, Eisen MB, Ross DT, Schuler G, Moore T, Lee JC, Trent JM, Staudt LM, Hudson J, Jr., Boguski MS, Lashkari D, Shalon D, Botstein D, Brown PO. 1999. The transcriptional program in the response of human fibroblasts to serum. *Science* 283: 83-7.
71. Janse DM, Crosas B, Finley D, Church GM. 2004. Localization to the proteasome is sufficient for degradation. *J Biol Chem* 279: 21415-20.
72. Jantz D, Amann BT, Gatto GJ, Jr., Berg JM. 2004. The design of functional DNA-binding proteins based on zinc finger domains. *Chem Rev* 104: 789-99.
73. Jo D, Nashabi A, Doxsee C, Lin Q, Unutmaz D, Chen J, Ruley HE. 2001. Epigenetic regulation of gene structure and function with a cell-permeable Cre recombinase. *Nat Biotechnol* 19: 929-33.
74. Johnson DS, Mortazavi A, Myers RM, Wold B. 2007. Genome-wide mapping of in vivo protein-DNA interactions. *Science* 316: 1497-502.
75. Joung JK, Ramm EI, Pabo CO. 2000. A bacterial two-hybrid selection system for studying protein-DNA and protein-protein interactions. *Proc Natl Acad Sci U S A* 97: 7382-7. PMC ID: PMC16554.
76. Kabouridis PS. 2003. Biological applications of protein transduction technology. *Trends Biotechnol* 21: 498-503. PMC ID: PMC2597147.
77. Kaji K, Norrby K, Paca A, Mileikovsky M, Mohseni P, Woltjen K. 2009. Virus-free induction of pluripotency and subsequent excision of reprogramming factors. *Nature*.
78. Kang HM, Ye C, Eskin E. 2008. Accurate discovery of expression quantitative trait loci under confounding from spurious and genuine regulatory hotspots. *Genetics* 180: 1909-25. PMC ID: PMC2600931.
79. Kelly JR, Rubin AJ, Davis JH, Ajo-Franklin CM, Cumbers J, Czar MJ, de Mora K, Glieberman AL, Monie DD, Endy D. 2009. Measuring the Activity of BioBrick Promoters Using an In Vivo Reference Standard. *J Biol Eng* 3: 4.
80. Kendziorski CM, Chen M, Yuan M, Lan H, Attie AD. 2006. Statistical methods for expression quantitative trait loci (eQTL) mapping. *Biometrics* 62: 19-27.
81. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D. 2002. The human genome browser at UCSC. *Genome Res* 12: 996-1006. PMC ID: PMC186604.
82. Kharchenko PV, Tolstorukov MY, Park PJ. 2008. Design and analysis of ChIP-seq experiments for DNA-binding proteins. *Nat Biotechnol* 26: 1351-9. PMC ID: PMC2597701.

83. Kim DS, Ross SE, Trimarchi JM, Aach J, Greenberg ME, Cepko CL. 2008. Identification of molecular markers of bipolar cells in the murine retina. *J Comp Neurol* 507: 1795-810.
84. Kim JB, Porreca GJ, Song L, Greenway SC, Gorham JM, Church GM, Seidman CE, Seidman JG. 2007. Polony multiplex analysis of gene expression (PMAGE) in mouse hypertrophic cardiomyopathy. *Science* 316: 1481-4.
85. Klein RJ. 2007. Power analysis for genome-wide association studies. *BMC Genet* 8: 58. PMC ID: PMC2042984.
86. Kolb HC, Finn MG, B. SK. 2001. Click Chemistry: Diverse Chemical Function from a Few Good Reactions. *Angew. Chem. Int.* 40: 2004–21.
87. Kouprina N, Campbell M, Graves J, Campbell E, Meincke L, Tesmer J, Grady DL, Doggett NA, Moyzis RK, Deaven LL, Larionov V. 1998. Construction of human chromosome 16- and 5-specific circular YAC/BAC libraries by in vivo recombination in yeast (TAR cloning). *Genomics* 53: 21-8.
88. Kouprina N, Kawamoto K, Barrett JC, Larionov V, Koi M. 1998. Rescue of targeted regions of mammalian chromosomes by in vivo recombination in yeast. *Genome Res* 8: 666-72. PMC ID: PMC310736.
89. Kouprina N, Larionov V. 2006. Selective isolation of mammalian genes by TAR cloning. *Curr Protoc Hum Genet* Chapter 5: Unit 5 17.
90. Kouprina N, Larionov V. 2008. Selective isolation of genomic loci from complex genomes by transformation-associated recombination cloning in the yeast Saccharomyces cerevisiae. *Nat Protoc* 3: 371-7.
91. Kouprina N, Leem SH, Solomon G, Ly A, Koriabine M, Otstot J, Pak E, Dutra A, Zhao S, Barrett JC, Larionov V. 2003. Segments missing from the draft human genome sequence can be isolated by transformation-associated recombination cloning in yeast. *EMBO Rep* 4: 257-62. PMC ID: PMC1315894.
92. Kudla G, Murray AW, Tollervey D, Plotkin JB. 2009. Coding-sequence determinants of gene expression in Escherichia coli. *Science* 324: 255-8.
93. Kuhn RM, Karolchik D, Zweig AS, Wang T, Smith KE, Rosenbloom KR, Rhead B, Raney BJ, Pohl A, Pheasant M, Meyer L, Hsu F, Hinrichs AS, Harte RA, Giardine B, Fujita P, Diekhans M, Dreszer T, Clawson H, Barber GP, Haussler D, Kent WJ. 2009. The UCSC Genome Browser Database: update 2009. *Nucleic Acids Res* 37: D755-61.
94. Kurimoto K, Yabuta Y, Ohinata Y, Saitou M. 2007. Global single-cell cDNA amplification to provide a template for representative high-density oligonucleotide microarray analysis. *Nat Protoc* 2: 739-52.
95. Kwan T, Benovoy D, Dias C, Gurd S, Provencher C, Beaulieu P, Hudson TJ, Sladek R, Majewski J. 2008. Genome-wide analysis of transcript isoform variation in humans. *Nat Genet* 40: 225-31.
96. Kwiatkowski M, Fredriksson S, Isaksson A, Nilsson M, Landegren U. 1999. Inversion of in situ synthesized oligonucleotides: improved reagents for hybridization and primer extension in DNA microarrays. *Nucleic Acids Res* 27: 4710-4. PMC ID: PMC148770.
97. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, Funke R, Gage D, Harris K, Heaford A, Howland J, Kann L, Lehoczky J, LeVine R, McEwan P, McKernan K, Meldrim J, Mesirov JP, Miranda C, Morris W, Naylor J, Raymond C, Rosetti M, Santos R, Sheridan A, Sougnez C, Stange-Thomann N, Stojanovic N, Subramanian A, Wyman D, Rogers J, Sulston J, Ainscough R, Beck S, Bentley D, Burton J, Clee C, Carter N, Coulson A, Deadman R, Deloukas P, Dunham A, Dunham I, Durbin R, French L, Grafham D, Gregory S, Hubbard T, Humphray S, Hunt A, Jones M, Lloyd C, McMurray A, Matthews L, Mercer S, Milne S, Mullikin JC, Mungall A, Plumb R, Ross M, Shownkeen R, Sims S, Waterston RH, Wilson RK, Hillier LW, McPherson JD, Marra MA, Mardis ER, Fulton LA, Chinwalla AT, Pepin KH, Gish WR, Chissoe SL, Wendl MC, Delehaunty KD, Miner TL, Delehaunty A, Kramer JB, Cook LL, Fulton RS, Johnson DL, Minx PJ, Clifton SW, Hawkins T, Branscomb E, Predki P, Richardson P, Wenning S, Slezak T, Doggett N, Cheng JF, Olsen A, Lucas S, Elkin C, Uberbacher E, Frazier M, Gibbs RA, Muzny DM, Scherer SE, Bouck JB, Sodergren EJ, Worley KC, Rives CM, Gorrell JH, Metzker ML, Naylor SL, Kucherlapati RS, Nelson DL, Weinstock GM, Sakaki Y, Fujiyama A, Hattori M, Yada T, Toyoda A, Itoh T, Kawagoe C, Watanabe H, Totoki Y, Taylor T, Weissenbach J, Heilig R, Saurin W, Artiguenave F, Brottier P, Bruls T, Pelletier E, Robert C, Wincker P, Smith DR, Doucette-Stamm L, Rubenfield M, Weinstock K, Lee HM, Dubois J, Rosenthal A, Platzer M, Nyakatura G, Taudien S, Rump A, Yang H, Yu J, Wang J, Huang G, Gu J, Hood L, Rowen L, Madan A, Qin S, Davis RW, Federspiel NA, Abola AP, Proctor MJ, Myers RM, Schmutz J, Dickson M, Grimwood J, Cox DR, Olson MV, Kaul R, Raymond C, Shimizu N, Kawasaki K, Minoshima S, Evans GA, Athanasiou M, Schultz R, Roe BA, Chen F, Pan H, Ramser J, Lehrach H, Reinhardt R, McCombie WR, de la Bastide M, Dedhia N, Blocker H, Hornischer K, Nordsiek G, Agarwala R, Aravind L, Bailey JA, Bateman A, Batzoglou S, Birney E, Bork P, Brown DG, Burge CB, Cerutti L, Chen HC, Church D, Clamp M, Copley RR, Doerks T, Eddy SR, Eichler EE, Furey TS, Galagan J, Gilbert JG, Harmon C, Hayashizaki Y, Haussler D, Hermjakob H, Hokamp K, Jang W, Johnson LS, Jones TA, Kasif S, Kaspryzk A, Kennedy S, Kent WJ, Kitts P, Koonin EV, Korf I, Kulp D, Lancet D, Lowe TM, McLysaght A, Mikkelsen T, Moran JV, Mulder N, Pollara VJ, Ponting CP, Schuler G, Schultz J, Slater G, Smit AF, Stupka E, Szustakowski J, Thierry-Mieg D, Thierry-Mieg J, Wagner L, Wallis J, Wheeler R, Williams A, Wolf YI, Wolfe KH, Yang SP, Yeh RF, Collins F, Guyer MS, Peterson J, Felsenfeld A, Wetterstrand KA, Patrinos A, Morgan MJ, de Jong P, Catanese

JJ, Osoegawa K, Shizuya H, Choi S, Chen YJ, International Human Genome Sequencing C. 2001. Initial sequencing and analysis of the human genome. *Nature* 409: 860-921.

98. Laner A, Goussard S, Ramalho AS, Schwarz T, Amaral MD, Courvalin P, Schindelhauer D, Grillot-Courvalin C. 2005. Bacterial transfer of large functional genomic DNA into human cells. *Gene Ther* 12: 1559-72.
99. Li H, Ruan J, Durbin R. 2008. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res* 18: 1851-8. PMC ID: PMC2577856.
100. Li JB, Gao Y, Aach J, Zhang K, Kryukov GV, Xie B, Ahlford A, Yoon J-K, Rosenbaum AM, Wait-Zaranek A, LeProust E, Sunyaev S, Church GM. 2009. Multiplex padlock capture and sequencing reveal human hypermutable CpG variations. *Genome Res* in press.
101. Li JB, Levanon EY, Yoon J-K, Aach J, Xie B, LeProust E, Zhang K, Gao Y, G.M. C. 2009. Genome-wide Identification of Human RNA Editing Sites by Parallel DNA Capturing and Sequencing. *Science* in press.
102. Li Y, Alvarez OA, Gutteling EW, Tijsterman M, Fu J, Riksen JA, Hazendonk E, Prins P, Plasterk RH, Jansen RC, Breitling R, Kammenga JE. 2006. Mapping determinants of gene expression plasticity by genetical genomics in C. elegans. *PLoS Genet* 2: e222. PMC ID: PMC1756913.
103. Lin Q, Jo D, Gebre-Amlak KD, Ruley HE. 2004. Enhanced cell-permeant Cre protein for site-specific recombination in cultured cells. *BMC Biotechnol* 4: 25. PMC ID: PMC529453.
104. Link AJ, Phillips D, Church GM. 1997. Methods for generating precise deletions and insertions in the genome of wild-type Escherichia coli: application to open reading frame characterization. *J Bacteriol* 179: 6228-37. PMC ID: PMC179534.
105. Lufino MM, Edser PA, Wade-Martins R. 2008. Advances in high-capacity extrachromosomal vector technology: episomal maintenance, vector delivery, and transgene expression. *Mol Ther* 16: 1525-38.
106. Lunshof JE, Chadwick R, Vorhaus DB, Church GM. 2008. From genetic privacy to open consent. *Nat Rev Genet* 9: 406-11.
107. Maeder ML, Thibodeau-Beganny S, Osiak A, Wright DA, Anthony RM, Eichtinger M, Jiang T, Foley JE, Winfrey RJ, Townsend JA, Unger-Wallace E, Sander JD, Muller-Lerch F, Fu F, Pearlberg J, Gobel C, Dassie JP, Pruett-Miller SM, Porteus MH, Sgroi DC, Iafrate AJ, Dobbs D, McCray PB, Jr., Cathomen T, Voytas DF, Joung JK. 2008. Rapid "open-source" engineering of customized zinc-finger nucleases for highly efficient gene modification. *Mol Cell* 31: 294-301. PMC ID: PMC2535758.
108. Maherali N, Sridharan R, Xie W, Utikal J, Eminli S, Arnold K, Stadtfeld M, Yachechko R, Tchieu J, Jaenisch R, Plath K, Hochedlinger K. 2007. Directly reprogrammed fibroblasts show global epigenetic remodeling and widespread tissue contribution. *Cell Stem Cell* 1: 55-70.
109. Mathews DH, Sabina J, Zuker M, Turner DH. 1999. Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. *J Mol Biol* 288: 911-40.
110. Maynard ND, Chen J, Stuart RK, Fan JB, Ren B. 2008. Genome-wide mapping of allele-specific protein-DNA interactions in human cells. *Nat Methods* 5: 307-9.
111. McCarroll SA. 2008. Extending genome-wide association studies to copy-number variation. *Hum Mol Genet* 17: R135-42.
112. Meng Q, Kim DH, Bai X, Bi L, Turro NJ, Ju J. 2006. Design and synthesis of a photocleavable fluorescent nucleotide 3'-O-allyl-dGTP-PC-Bodipy-FL-510 as a reversible terminator for DNA sequencing by synthesis. *J Org Chem* 71: 3248-52.
113. Mikkelsen TS, Ku M, Jaffe DB, Issac B, Lieberman E, Giannoukos G, Alvarez P, Brockman W, Kim TK, Koche RP, Lee W, Mendenhall E, O'Donovan A, Presser A, Russ C, Xie X, Meissner A, Wernig M, Jaenisch R, Nusbaum C, Lander ES, Bernstein BE. 2007. Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature* 448: 553-60.
114. Milani L, Lundmark A, Nordlund J, Kiialainen A, Flaegstad T, Jonmundsson G, Kanerva J, Schmiegelow K, Gunderson KL, Lonnerholm G, Syvanen AC. 2009. Allele-specific gene expression patterns in primary leukemic cells reveal regulation of gene expression by CpG site methylation. *Genome Res* 19: 1-11. PMC ID: PMC2612957.
115. Miller JC, Holmes MC, Wang J, Guschin DY, Lee YL, Rupniewski I, Beausejour CM, Waite AJ, Wang NS, Kim KA, Gregory PD, Pabo CO, Rebar EJ. 2007. An improved zinc-finger nuclease architecture for highly specific genome editing. *Nat Biotechnol* 25: 778-85.
116. Mitra RD, Butty VL, Shendure J, Williams BR, Housman DE, Church GM. 2003. Digital genotyping and haplotyping with polymerase colonies. *Proc Natl Acad Sci U S A* 100: 5926-31. PMC ID: PMC156303.
117. Mitra RD, Shendure J, Olejnik J, Edyta Krzymanska O, Church GM. 2003. Fluorescent in situ sequencing on polymerase colonies. *Anal Biochem* 320: 55-65.
118. Moehle EA, Rock JM, Lee YL, Jouvenot Y, DeKelver RC, Gregory PD, Urnov FD, Holmes MC. 2007. Targeted gene addition into a specified location in the human genome using designed zinc finger nucleases. *Proc Natl Acad Sci U S A* 104: 3055-60. PMC ID: PMC1802009.

119. Morley M, Molony CM, Weber TM, Devlin JL, Ewens KG, Spielman RS, Cheung VG. 2004. Genetic analysis of genome-wide variation in human gene expression. *Nature* 430: 743-7.
120. Mrksich M. 2002. What can surface chemistry do for cell biology? *Curr Opin Chem Biol* 6: 794-7.
121. Myers AJ, Gibbs JR, Webster JA, Rohrer K, Zhao A, Marlowe L, Kaleem M, Leung D, Bryden L, Nath P, Zismann VL, Joshipura K, Huentelman MJ, Hu-Lince D, Coon KD, Craig DW, Pearson JV, Holmans P, Heward CB, Reiman EM, Stephan D, Hardy J. 2007. A survey of genetic human cortical gene expression. *Nat Genet* 39: 1494-9.
122. Myocardial Infarction Genetics C, Kathiresan S, Voight BF, Purcell S, Musunuru K, Ardissino D, Mannucci PM, Anand S, Engert JC, Samani NJ, Schunkert H, Erdmann J, Reilly MP, Rader DJ, Morgan T, Spertus JA, Stoll M, Girelli D, McKeown PP, Patterson CC, Siscovick DS, O'Donnell CJ, Elosua R, Peltonen L, Salomaa V, Schwartz SM, Melander O, Altshuler D, Ardissino D, Merlini PA, Berzuini C, Bernardinelli L, Peyvandi F, Tubaro M, Celli P, Ferrario M, Fetiveau R, Marziliano N, Casari G, Galli M, Ribichini F, Rossi M, Bernardi F, Zonzin P, Piazza A, Mannucci PM, Schwartz SM, Siscovick DS, Yee J, Friedlander Y, Elosua R, Marrugat J, Lucas G, Subirana I, Sala J, Ramos R, Kathiresan S, Meigs JB, Williams G, Nathan DM, MacRae CA, O'Donnell CJ, Salomaa V, Havulinna AS, Peltonen L, Melander O, Berglund G, Voight BF, Kathiresan S, Hirschhorn JN, Asselta R, Duga S, Spreafico M, Musunuru K, Daly MJ, Purcell S, Voight BF, Purcell S, Nemesh J, Korn JM, McCarroll SA, Schwartz SM, Yee J, Kathiresan S, Lucas G, Subirana I, Elosua R, Surti A, Guiducci C, Gianniny L, Mirel D, Parkin M, Burtt N, Gabriel SB, Samani NJ, Thompson JR, Braund PS, Wright BJ, Balmforth AJ, Ball SG, Hall AS, Wellcome Trust Case Control C, Schunkert H, Erdmann J, Linsel-Nitschke P, Lieb W, Ziegler A, Konig I, Hengstenberg C, Fischer M, Stark K, Grosshennig A, Preuss M, Wichmann HE, Schreiber S, Schunkert H, Samani NJ, Erdmann J, Ouwehand W, Hengstenberg C, Deloukas P, Scholz M, Cambien F, Reilly MP, Li M, Chen Z, Wilensky R, Matthai W, Qasim A, Hakonarson HH, Devaney J, Burnett MS, Pichard AD, Kent KM, Satler L, Lindsay JM, Waksman R, Epstein SE, Rader DJ, Scheffold T, Berger K, Stoll M, Huge A, Girelli D, Martinelli N, Olivieri O, Corrocher R, Morgan T, Spertus JA, McKeown P, Patterson CC, Schunkert H, Erdmann E, Linsel-Nitschke P, Lieb W, Ziegler A, Konig IR, Hengstenberg C, Fischer M, Stark K, Grosshennig A, Preuss M, Wichmann HE, Schreiber S, Holm H, Thorleifsson G, Thorsteinsdottir U, Stefansson K, Engert JC, Do R, Xie C, Anand S, Kathiresan S, Ardissino D, Mannucci PM, Siscovick D, O'Donnell CJ, Samani NJ, Melander O, Elosua R, Peltonen L, Salomaa V, Schwartz SM, Altshuler D. 2009. Genome-wide association of early-onset myocardial infarction with single nucleotide polymorphisms and copy number variants. *Nat Genet* 41: 334-41.
123. Nannya Y, Taura K, Kurokawa M, Chiba S, Ogawa S. 2007. Evaluation of genome-wide power of genetic association studies based on empirical data from the HapMap project. *Hum Mol Genet* 16: 2494-505.
124. Narayanan K, Warburton PE. 2003. DNA modification and functional delivery into human cells using Escherichia coli DH10B. *Nucleic Acids Res* 31: e51. PMC ID: PMC154239.
125. National Institutes of Health. 2009. Genotype-Tissue Expression Project (http://nihroadmap.nih.gov/GTEx/).
126. Nejentsev S, Walker N, Riches D, Egholm M, Todd JA. 2009. Rare variants of IFIH1, a gene implicated in antiviral responses, protect against type 1 diabetes. *Science* 324: 387-9.
127. Nolden L, Edenhofer F, Peitz M, Brustle O. 2007. Stem cell engineering using transducible Cre recombinase. *Methods Mol Med* 140: 17-32.
128. Olsen PA, Randol M, Luna L, Brown T, Krauss S. 2005. Genomic sequence correction by single-stranded DNA oligonucleotides: role of DNA synthesis and chemical modifications of the oligonucleotide ends. *J Gene Med* 7: 1534-44.
129. Olsen PA, Solhaug A, Booth JA, Gelazauskaite M, Krauss S. 2009. Cellular responses to targeted genomic sequence modification using single-stranded oligonucleotides and zinc-finger nucleases. *DNA Repair (Amst)* 8: 298-308.
130. Pabo CO, Peisach E, Grant RA. 2001. Design and selection of novel Cys2His2 zinc finger proteins. *Annu Rev Biochem* 70: 313-40.
131. Pan X, Urban AE, Palejev D, Schulz V, Grubert F, Hu Y, Snyder M, Weissman SM. 2008. A procedure for highly specific, sensitive, and unbiased whole-genome amplification. *Proc Natl Acad Sci U S A* 105: 15499-504. PMC ID: PMC2563063.
132. Park CC, Ahn S, Bloom JS, Lin A, Wang RT, Wu T, Sekar A, Khan AH, Farr CJ, Lusis AJ, Leahy RM, Lange K, Smith DJ. 2008. Fine mapping of regulatory loci for mammalian gene expression using radiation hybrids. *Nat Genet* 40: 421-9.
133. Park IH, Zhao R, West JA, Yabuuchi A, Huo H, Ince TA, Lerou PH, Lensch MW, Daley GQ. 2008. Reprogramming of human somatic cells to pluripotency with defined factors. *Nature* 451: 141-6.
134. Pearson H. 2008. Protein engineering: The fate of fingers. *Nature* 455: 160-4.
135. Perez-Luz S, Abdulrazzak H, Grillot-Courvalin C, Huxley C. 2007. Factor VIII mRNA expression from a BAC carrying the intact locus made by homologous recombination. *Genomics* 90: 610-9.
136. Personal Genome Project. 2009. http://www.personalgenomes.org/.

137. Plagnol V, Uz E, Wallace C, Stevens H, Clayton D, Ozcelik T, Todd JA. 2008. Extreme clonality in lymphoblastoid cell lines with implications for allele specific expression analyses. *PLoS ONE* 3: e2966. PMC ID: PMC2494943.
138. Porreca GJ, Shendure J, Church GM. 2006. Polony DNA sequencing. *Curr Protoc Mol Biol* Chapter 7: Unit 7 8.
139. Porreca GJ, Zhang K, Li JB, Xie B, Austin D, Vassallo SL, LeProust EM, Peck BJ, Emig CJ, Dahl F, Gao Y, Church GM, Shendure J. 2007. Multiplex amplification of large sets of human exons. *Nat Methods* 4: 931-6.
140. Pruett-Miller SM, Connelly JP, Maeder ML, Joung JK, Porteus MH. 2008. Comparison of zinc finger nucleases for use in gene targeting in mammalian cells. *Mol Ther* 16: 707-17.
141. Pruett-Miller SM, Reading DW, Porter SN, Porteus MH. 2009. Attenuation of zinc finger nuclease toxicity by small-molecule regulation of protein levels. *PLoS Genet* 5: e1000376. PMC ID: PMC2633050.
142. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PI, Daly MJ, Sham PC. 2007. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* 81: 559-75. PMC ID: PMC1950838.
143. Radecke S, Radecke F, Peter I, Schwarz K. 2006. Physical incorporation of a single-stranded oligodeoxynucleotide during targeted repair of a human chromosomal locus. *J Gene Med* 8: 217-28.
144. Risch N, Merikangas K. 1996. The future of genetic studies of complex human diseases. *Science* 273: 1516-7.
145. Roberts C, Chen CS, Mrksich M, Martichonok V, Ingber DE, Whitesides GM. 1998. Using Mixed Self-Assembled Monolayers Presenting RGD and (EG)3OH Groups To Characterize Long-Term Attachment of Bovine Capillary Endothelial Cells to Surfaces. *J Amer Chem Soc* 120: 6548–55.
146. Sachidanandam R, Weissman D, Schmidt SC, Kakol JM, Stein LD, Marth G, Sherry S, Mullikin JC, Mortimore BJ, Willey DL, Hunt SE, Cole CG, Coggill PC, Rice CM, Ning Z, Rogers J, Bentley DR, Kwok PY, Mardis ER, Yeh RT, Schultz B, Cook L, Davenport R, Dante M, Fulton L, Hillier L, Waterston RH, McPherson JD, Gilman B, Schaffner S, Van Etten WJ, Reich D, Higgins J, Daly MJ, Blumenstiel B, Baldwin J, Stange-Thomann N, Zody MC, Linton L, Lander ES, Altshuler D, International SNPMWG. 2001. A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* 409: 928-33.
147. Sakamoto KM. 2005. Chimeric molecules to target proteins for ubiquitination and degradation. *Methods Enzymol* 399: 833-47.
148. Saxon E, Bertozzi CR. 2000. Cell surface engineering by a modified Staudinger reaction. *Science* 287: 2007-10.
149. Schadt EE, Molony C, Chudin E, Hao K, Yang X, Lum PY, Kasarskis A, Zhang B, Wang S, Suver C, Zhu J, Millstein J, Sieberts S, Lamb J, GuhaThakurta D, Derry J, Storey JD, Avila-Campillo I, Kruger MJ, Johnson JM, Rohl CA, van Nas A, Mehrabian M, Drake TA, Lusis AJ, Smith RC, Guengerich FP, Strom SC, Schuetz E, Rushmore TH, Ulrich R. 2008. Mapping the genetic architecture of gene expression in human liver. *PLoS Biol* 6: e107. PMC ID: PMC2365981.
150. Schadt EE, Monks SA, Drake TA, Lusis AJ, Che N, Colinayo V, Ruff TG, Milligan SB, Lamb JR, Cavet G, Linsley PS, Mao M, Stoughton RB, Friend SH. 2003. Genetics of gene expression surveyed in maize, mouse and man. *Nature* 422: 297-302.
151. Sepp A, Choo Y. 2005. Cell-free selection of zinc finger DNA-binding proteins using in vitro compartmentalization. *J Mol Biol* 354: 212-9.
152. Serre D, Gurd S, Ge B, Sladek R, Sinnett D, Harmsen E, Bibikova M, Chudin E, Barker DL, Dickinson T, Fan JB, Hudson TJ. 2008. Differential allelic expression in the human genome: a robust approach to identify genetic and epigenetic cis-acting mechanisms regulating gene expression. *PLoS Genet* 4: e1000006. PMC ID: PMC2265535.
153. Shao L, Feng W, Sun Y, Bai H, Liu J, Currie C, Kim J, Gama R, Wang Z, Qian Z, Liaw L, Wu WS. 2009. Generation of iPS cells using defined factors linked via the self-cleaving 2A sequences in a single open reading frame. *Cell Res* 19: 296-306.
154. Shendure J, Mitra RD, Varma C, Church GM. 2004. Advanced sequencing technologies: methods and goals. *Nat Rev Genet* 5: 335-44.
155. Shendure J, Porreca GJ, Reppas NB, Lin X, McCutcheon JP, Rosenbaum AM, Wang MD, Zhang K, Mitra RD, Church GM. 2005. Accurate multiplex polony sequencing of an evolved bacterial genome. *Science* 309: 1728-32.
156. Shendure JA, Porreca GJ, Church GM. 2008. Overview of DNA sequencing strategies. *Curr Protoc Mol Biol* Chapter 7: Unit 7 1.
157. Shi Y, Desponts C, Do JT, Hahm HS, Scholer HR, Ding S. 2008. Induction of pluripotent stem cells from mouse embryonic fibroblasts by Oct4 and Klf4 with small-molecule compounds. *Cell Stem Cell* 3: 568-74.
158. Shin J, Kayser SR, Langaee TY. 2009. Pharmacogenetics: from discovery to patient care. *Am J Health Syst Pharm* 66: 625-37.
159. Smith EN, Kruglyak L. 2008. Gene-environment interaction in yeast gene expression. *PLoS Biol* 6: e83. PMC ID: PMC2292755.
160. Smith J, Modrich P. 1997. Removal of polymerase-produced mutant sequences from PCR products. *Proc Natl Acad Sci U S A* 94: 6847-50. PMC ID: PMC21247.

161. Srivastava M, Hsieh S, Grinberg A, Williams-Simons L, Huang SP, Pfeifer K. 2000. H19 and Igf2 monoallelic expression is regulated in two distinct ways by a shared cis acting regulatory region upstream of H19. *Genes Dev* 14: 1186-95. PMC ID: PMC316622.
162. Stark JM, Pierce AJ, Oh J, Pastink A, Jasin M. 2004. Genetic steps of mammalian homologous repair with distinct mutagenic consequences. *Mol Cell Biol* 24: 9305-16. PMC ID: PMC522275.
163. Stougaard M, Lohmann JS, Zajac M, Hamilton-Dutoit S, Koch J. 2007. In situ detection of non-polyadenylated RNA molecules using Turtle Probes and target primed rolling circle PRINS. *BMC Biotechnol* 7: 69. PMC ID: PMC2203993.
164. Stranger BE, Forrest MS, Dunning M, Ingle CE, Beazley C, Thorne N, Redon R, Bird CP, de Grassi A, Lee C, Tyler-Smith C, Carter N, Scherer SW, Tavare S, Deloukas P, Hurles ME, Dermitzakis ET. 2007. Relative impact of nucleotide and copy number variation on gene expression phenotypes. *Science* 315: 848-53.
165. Stranger BE, Nica AC, Forrest MS, Dimas A, Bird CP, Beazley C, Ingle CE, Dunning M, Flicek P, Koller D, Montgomery S, Tavare S, Deloukas P, Dermitzakis ET. 2007. Population genomics of human gene expression. *Nat Genet* 39: 1217-24.
166. Suzuki T. 2008. Targeted gene modification by oligonucleotides and small DNA fragments in eukaryotes. *Front Biosci* 13: 737-44.
167. Takahashi K, Tanabe K, Ohnuki M, Narita M, Ichisaka T, Tomoda K, Yamanaka S. 2007. Induction of pluripotent stem cells from adult human fibroblasts by defined factors. *Cell* 131: 861-72.
168. Takahashi K, Yamanaka S. 2006. Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. *Cell* 126: 663-76.
169. Tang F, Barbacioru C, Wang Y, Nordman E, Lee C, Xu N, Wang X, Bodeau J, Tuch BB, Siddiqui A, Lao K, Surani MA. 2009. mRNA-Seq whole-transcriptome analysis of a single cell. *Nat Methods* 6: 377-82.
170. Tao H, Cox DR, Frazer KA. 2006. Allele-specific KRT1 expression is a complex trait. *PLoS Genet* 2: e93. PMC ID: PMC1475705.
171. Tatum EL, Lederberg J. 1947. Gene Recombination in the Bacterium Escherichia coli. *J Bacteriol* 53: 673-84. PMC ID: PMC518375.
172. Terry R, Porreca G, McCarthy K, Church GM. 2008. Polonator Instrument http://www.polonator.org
173. Thorgeirsson TE, Geller F, Sulem P, Rafnar T, Wiste A, Magnusson KP, Manolescu A, Thorleifsson G, Stefansson H, Ingason A, Stacey SN, Bergthorsson JT, Thorlacius S, Gudmundsson J, Jonsson T, Jakobsdottir M, Saemundsdottir J, Olafsdottir O, Gudmundsson LJ, Bjornsdottir G, Kristjansson K, Skuladottir H, Isaksson HJ, Gudbjartsson T, Jones GT, Mueller T, Gottsater A, Flex A, Aben KK, de Vegt F, Mulders PF, Isla D, Vidal MJ, Asin L, Saez B, Murillo L, Blondal T, Kolbeinsson H, Stefansson JG, Hansdottir I, Runarsdottir V, Pola R, Lindblad B, van Rij AM, Dieplinger B, Haltmayer M, Mayordomo JI, Kiemeney LA, Matthiasson SE, Oskarsson H, Tyrfingsson T, Gudbjartsson DF, Gulcher JR, Jonsson S, Thorsteinsdottir U, Kong A, Stefansson K. 2008. A variant associated with nicotine dependence, lung cancer and peripheral arterial disease. *Nature* 452: 638-42.
174. Tian J, Gong H, Sheng N, Zhou X, Gulari E, Gao X, Church G. 2004. Accurate multiplex gene synthesis from programmable DNA microchips. *Nature* 432: 1050-4.
175. Townsend JA, Wright DA, Winfrey RJ, Fu F, Maeder ML, Joung JK, Voytas DF. 2009. High Frequency Modification of Plant Genes Using Engineered Zinc Finger Nucleases. *Nature*: advanced on-line print: April 29, 2009.
176. Turner DJ, Shendure J, Porreca G, Church G, Green P, Tyler-Smith C, Hurles ME. 2006. Assaying chromosomal inversions by single-molecule haplotyping. *Nat Methods* 3: 439-45.
177. University of Washington (Seattle). 2009. genetests.org (http://www.genetests.org/). pp. Quote from web page (03/24/09): 475 GeneReviews, 1,158 Clinics, 607 Laboratories testing for 1,707 Diseases: 1,422 Clinical, 285 Research
178. Urnov FD, Miller JC, Lee YL, Beausejour CM, Rock JM, Augustus S, Jamieson AC, Porteus MH, Gregory PD, Holmes MC. 2005. Highly efficient endogenous human gene correction using designed zinc-finger nucleases. *Nature* 435: 646-51.
179. Valouev A, Johnson DS, Sundquist A, Medina C, Anton E, Batzoglou S, Myers RM, Sidow A. 2008. Genome-wide analysis of transcription factor binding sites based on ChIP-Seq data. *Nat Methods* 5: 829-34.
180. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA, Gocayne JD, Amanatides P, Ballew RM, Huson DH, Wortman JR, Zhang Q, Kodira CD, Zheng XH, Chen L, Skupski M, Subramanian G, Thomas PD, Zhang J, Gabor Miklos GL, Nelson C, Broder S, Clark AG, Nadeau J, McKusick VA, Zinder N, Levine AJ, Roberts RJ, Simon M, Slayman C, Hunkapiller M, Bolanos R, Delcher A, Dew I, Fasulo D, Flanigan M, Florea L, Halpern A, Hannenhalli S, Kravitz S, Levy S, Mobarry C, Reinert K, Remington K, Abu-Threideh J, Beasley E, Biddick K, Bonazzi V, Brandon R, Cargill M, Chandramouliswaran I, Charlab R, Chaturvedi K, Deng Z, Di Francesco V, Dunn P, Eilbeck K, Evangelista C, Gabrielian AE, Gan W, Ge W, Gong F, Gu Z, Guan P, Heiman TJ, Higgins ME, Ji RR, Ke Z, Ketchum KA, Lai Z, Lei Y, Li Z, Li J, Liang Y, Lin X, Lu F, Merkulov GV, Milshina N, Moore HM, Naik AK, Narayan VA, Neelam B, Nusskern D, Rusch DB, Salzberg S,

Shao W, Shue B, Sun J, Wang Z, Wang A, Wang X, Wang J, Wei M, Wides R, Xiao C, Yan C, Yao A, Ye J, Zhan M, Zhang W, Zhang H, Zhao Q, Zheng L, Zhong F, Zhong W, Zhu S, Zhao S, Gilbert D, Baumhueter S, Spier G, Carter C, Cravchik A, Woodage T, Ali F, An H, Awe A, Baldwin D, Baden H, Barnstead M, Barrow I, Beeson K, Busam D, Carver A, Center A, Cheng ML, Curry L, Danaher S, Davenport L, Desilets R, Dietz S, Dodson K, Doup L, Ferriera S, Garg N, Gluecksmann A, Hart B, Haynes J, Haynes C, Heiner C, Hladun S, Hostin D, Houck J, Howland T, Ibegwam C, Johnson J, Kalush F, Kline L, Koduru S, Love A, Mann F, May D, McCawley S, McIntosh T, McMullen I, Moy M, Moy L, Murphy B, Nelson K, Pfannkoch C, Pratts E, Puri V, Qureshi H, Reardon M, Rodriguez R, Rogers YH, Romblad D, Ruhfel B, Scott R, Sitter C, Smallwood M, Stewart E, Strong R, Suh E, Thomas R, Tint NN, Tse S, Vech C, Wang G, Wetter J, Williams S, Williams M, Windsor S, Winn-Deen E, Wolfe K, Zaveri J, Zaveri K, Abril JF, Guigo R, Campbell MJ, Sjolander KV, Karlak B, Kejariwal A, Mi H, Lazareva B, Hatton T, Narechania A, Diemer K, Muruganujan A, Guo N, Sato S, Bafna V, Istrail S, Lippert R, Schwartz R, Walenz B, Yooseph S, Allen D, Basu A, Baxendale J, Blick L, Caminha M, Carnes-Stine J, Caulk P, Chiang YH, Coyne M, Dahlke C, Mays A, Dombroski M, Donnelly M, Ely D, Esparham S, Fosler C, Gire H, Glanowski S, Glasser K, Glodek A, Gorokhov M, Graham K, Gropman B, Harris M, Heil J, Henderson S, Hoover J, Jennings D, Jordan C, Jordan J, Kasha J, Kagan L, Kraft C, Levitsky A, Lewis M, Liu X, Lopez J, Ma D, Majoros W, McDaniel J, Murphy S, Newman M, Nguyen T, Nguyen N, Nodell M, Pan S, Peck J, Peterson M, Rowe W, Sanders R, Scott J, Simpson M, Smith T, Sprague A, Stockwell T, Turner R, Venter E, Wang M, Wen M, Wu D, Wu M, Xia A, Zandieh A, Zhu X. 2001. The sequence of the human genome. *Science* 291: 1304-51.

181. Vigneault F, Sismour AM, Church GM. 2008. Efficient microRNA capture and bar-coding via enzymatic oligonucleotide adenylation. *Nat Methods* 5: 777-9.
182. Visscher PM, Hill WG, Wray NR. 2008. Heritability in the genomics era--concepts and misconceptions. *Nat Rev Genet* 9: 255-66.
183. Wade-Martins R, Smith ER, Tyminski E, Chiocca EA, Saeki Y. 2001. An infectious transfer and expression system for genomic DNA loci in human and mouse cells. *Nat Biotechnol* 19: 1067-70.
184. Wang HH, Isaacs FJ, Forest CR, Sun ZZ, Xu G, Church GM. 2009. Programming cells by multiplex genome engineering and accelerated evolution *Nature* in press.
185. Wang Z, Gerstein M, Snyder M. 2009. RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet* 10: 57-63.
186. Waters VL. 2001. Conjugation between bacterial and mammalian cells. *Nat Genet* 29: 375-6.
187. Watkins H, Farrall M. 2006. Genetic susceptibility to coronary artery disease: from promise to progress. *Nat Rev Genet* 7: 163-73.
188. Wellcome Trust Case Control C. 2007. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* 447: 661-78.
189. Woltjen K, Michael IP, Mohseni P, Desai R, Mileikovsky M, Hamalainen R, Cowling R, Wang W, Liu P, Gertsenstein M, Kaji K, Sung HK, Nagy A. 2009. piggyBac transposition reprograms fibroblasts to induced pluripotent stem cells. *Nature*.
190. Wu J, Zhang S, Meng Q, Cao H, Li Z, Li X, Shi S, Kim DH, Bi L, Turro NJ, Ju J. 2007. 3'-O-modified nucleotides as reversible terminators for pyrosequencing. *Proc Natl Acad Sci U S A* 104: 16462-7. PMC ID: PMC2034218.
191. Yanez RJ, Porter AC. 1998. Therapeutic gene targeting. *Gene Ther* 5: 149-59.
192. Ying QL, Wray J, Nichols J, Batlle-Morera L, Doble B, Woodgett J, Cohen P, Smith A. 2008. The ground state of embryonic stem cell self-renewal. *Nature* 453: 519-23.
193. Yu D, Ellis HM, Lee EC, Jenkins NA, Copeland NG, Court DL. 2000. An efficient recombination system for chromosome engineering in Escherichia coli. *Proc Natl Acad Sci U S A* 97: 5978-83. PMC ID: PMC18544.
194. Yu J, Hu K, Smuga-Otto K, Tian S, Stewart R, Slukvin, II, Thomson JA. 2009. Human induced pluripotent stem cells free of vector and transgene sequences. *Science* 324: 797-801.
195. Yu J, Vodyanik MA, Smuga-Otto K, Antosiewicz-Bourget J, Frane JL, Tian S, Nie J, Jonsdottir GA, Ruotti V, Stewart R, Slukvin, II, Thomson JA. 2007. Induced pluripotent stem cell lines derived from human somatic cells. *Science* 318: 1917-20.
196. Zahnd C, Amstutz P, Pluckthun A. 2007. Ribosome display: selecting and evolving proteins in vitro that specifically bind to a target. *Nat Methods* 4: 269-79.
197. Zaranek AW, Clegg W, Vandewege, Church GM. 2008. *Free Factories: Unified Infrastructure for Data Intensive Web Services* Presented at USENIX Annual Technical Confrerence, Boston, MA
198. Zhang D, Baek SH, Ho A, Lee H, Jeong YS, Kim K. 2004. Targeted degradation of proteins by small molecules: a novel tool for functional proteomics. *Comb Chem High Throughput Screen* 7: 689-97.
199. Zhang K, Li JB, Gao Y, Egli D, Xie B, Deng J, Li Z, Lee J, Aach J, Leproust E, Eggan K, Church GM. 2009. Digital RNA Allelotyping Reveals Tissue-specific and Allele-specific Gene Expression in Human (submitted to Nature Methods).
200. Zhang K, Martiny AC, Reppas NB, Barry KW, Malek J, Chisholm SW, Church GM. 2006. Sequencing genomes from single cells by polymerase cloning. *Nat Biotechnol* 24: 680-6.

201. Zhang K, Zhu J, Shendure J, Porreca GJ, Aach JD, Mitra RD, Church GM. 2006. Long-range polony haplotyping of individual human chromosome molecules. *Nat Genet* 38: 382-7.
202. Zhou Y, Calciano M, Hamann S, Leamon JH, Strugnell T, Christian MW, Lizardi PM. 2001. In situ detection of messenger RNA using digoxigenin-labeled oligonucleotides and rolling circle amplification. *Exp Mol Pathol* 70: 281-8.
203. Zhu J, Shendure J, Mitra RD, Church GM. 2003. Single molecule profiling of alternative pre-mRNA splicing. *Science* 301: 836-8.
204. Zondervan KT, Cardon LR. 2004. The complex interplay among factors that influence allelic association. *Nat Rev Genet* 5: 89-100.
205. Zondervan KT, Cardon LR. 2007. Designing candidate gene and genome-wide case-control association studies. *Nat Protoc* 2: 2492-501.
206. Zuker M, Stiegler P. 1981. Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. *Nucleic Acids Res* 9: 133-48. PMC ID: PMC326673.