Video Article

Screening for Functional Non-coding Genetic Variants Using Electrophoretic Mobility Shift Assay (EMSA) and DNA-affinity Precipitation Assay (DAPA)

Daniel E. Miller[1], Zubin H. Patel[1,2,3], Xiaoming Lu[1,3], Arthur T. Lynch[1], Matthew T. Weirauch*[1,4], Leah C. Kottyan*[1]

[1]Center for Autoimmune Genomics and Etiology, Cincinnati Children's Hospital

[2]Medical Scientist Training Program, University of Cincinnati

[3]Immunology Graduate Program, University of Cincinnati

[4]Divisions of Biomedical Informatics and Developmental Biology, Cincinnati Children's Hospital

*These authors contributed equally

Correspondence to: Leah C. Kottyan at Leah.Kottyan@cchmc.org

URL: https://www.jove.com/video/54093
DOI: doi:10.3791/54093

Keywords: Genetics, Issue 114, EMSA, DAPA, non-coding genetic variants, SNPs, SNP function, transcription factors, gene regulation, nuclear lysate, GWAS

Date Published: 8/21/2016

Citation: Miller, D.E., Patel, Z.H., Lu, X., Lynch, A.T., Weirauch, M.T., Kottyan, L.C. Screening for Functional Non-coding Genetic Variants Using Electrophoretic Mobility Shift Assay (EMSA) and DNA-affinity Precipitation Assay (DAPA). *J. Vis. Exp.* (114), e54093, doi:10.3791/54093 (2016).

Abstract

Population and family-based genetic studies typically result in the identification of genetic variants that are statistically associated with a clinical disease or phenotype. For many diseases and traits, most variants are non-coding, and are thus likely to act by impacting subtle, comparatively hard to predict mechanisms controlling gene expression. Here, we describe a general strategic approach to prioritize non-coding variants, and screen them for their function. This approach involves computational prioritization using functional genomic databases followed by experimental analysis of differential binding of transcription factors (TFs) to risk and non-risk alleles. For both electrophoretic mobility shift assay (EMSA) and DNA affinity precipitation assay (DAPA) analysis of genetic variants, a synthetic DNA oligonucleotide (oligo) is used to identify factors in the nuclear lysate of disease or phenotype-relevant cells. For EMSA, the oligonucleotides with or without bound nuclear factors (often TFs) are analyzed by non-denaturing electrophoresis on a tris-borate-EDTA (TBE) polyacrylamide gel. For DAPA, the oligonucleotides are bound to a magnetic column and the nuclear factors that specifically bind the DNA sequence are eluted and analyzed through mass spectrometry or with a reducing sodium dodecyl sulfate polyacrylamide gel electrophoresis (SDS-PAGE) followed by Western blot analysis. This general approach can be widely used to study the function of non-coding genetic variants associated with any disease, trait, or phenotype.

Video Link

The video component of this article can be found at https://www.jove.com/video/54093/

Introduction

Sequencing and genotyping based studies, including Genome-Wide Association Studies (GWAS), candidate locus studies, and deep-sequencing studies, have identified many genetic variants that are statistically associated with a disease, trait, or phenotype. Contrary to early predictions, most of these variants (85-93%) are located in non-coding regions and do not change the amino acid sequence of proteins[1,2]. Interpreting the function of these non-coding variants and determining the biological mechanisms connecting them to the associated disease, trait, or phenotype has proven challenging[3-6]. We have developed a general strategy to identify the molecular mechanisms that link variants to an important intermediate phenotype – gene expression. This pipeline is specifically designed to identify modulation of TF binding by genetic variants. This strategy combines computational approaches and molecular biology techniques aimed to predict biological effects of candidate variants *in silico*, and verify these predictions empirically (**Figure 1**).

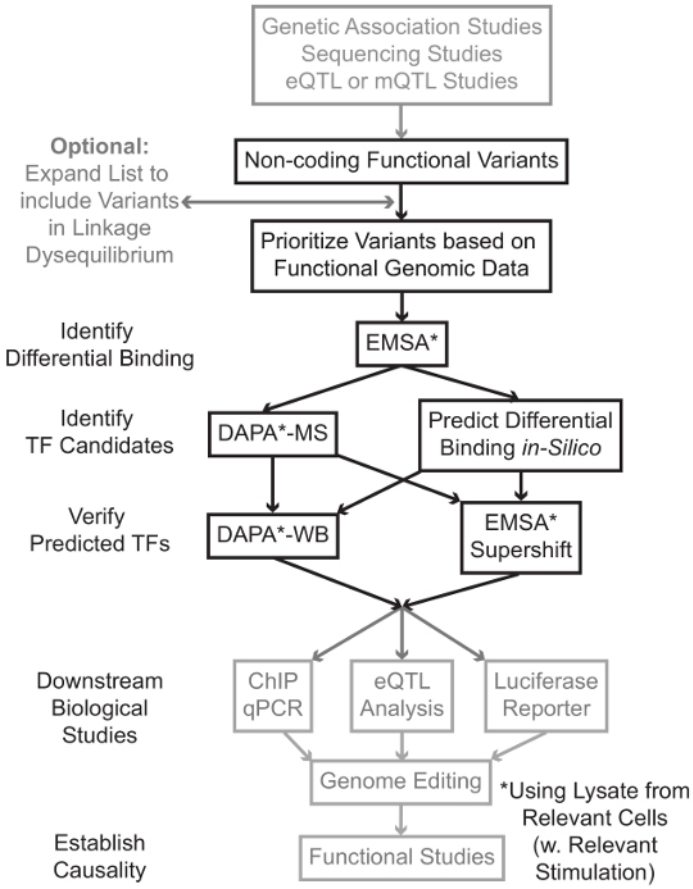

Figure 1: Strategic approach for the analysis of non-coding genetic variants. Steps that are not included in the detailed protocol associated with this manuscript are shaded in grey. Please click here to view a larger version of this figure.

In many cases, it is important to begin by expanding the list of variants to include all those in high linkage-disequilibrium (LD) with each statistically associated variant. LD is a measure of non-random association of alleles at two different chromosomal positions, which can be measured by the r^2 statistic[7]. r^2 is a measure of the linkage disequilibrium between two variants, with an $r^2 = 1$ denoting perfect linkage between two variants. Alleles in high LD are found to co-segregate on the chromosome across ancestral populations. Current genotyping arrays do not include all known variants in the human genome. Instead, they exploit the LD within the human genome and include a subset of the known variants that act as proxies for other variants within a particular region of LD[8]. Thus, a variant without any biological consequence may be associated with a particular disease because it is in LD with the causal variant-the variant with a meaningful biological effect. Procedurally, it is recommended to convert the latest release of the 1,000 genomes project[9] variant call files (vcf) into binary files compatible with PLINK[10,11], an open-source tool for whole genome association analysis. Subsequently, all other genetic variants with LD r^2 >0.8 with each input genetic variant can be identified as candidates. It is important to use the appropriate reference population for this step- *e.g.*, if a variant was identified in subjects of European ancestry, data from subjects of similar ancestry should be used for LD expansion.

LD expansion often results in dozens of candidate variants, and it is likely that only a small fraction of these contribute to disease mechanism. Often, it is infeasible to experimentally examine each of these variants individually. It is therefore useful to leverage the thousands of publically available functional genomic datasets as a filter to prioritize the variants. For example, the ENCODE consortium[12] has performed thousands of ChIP-seq experiments describing the binding of TFs and co-factors, and histone marks in a wide range of contexts, along with chromatin accessibility data from technologies such as DNase-seq[13], ATAC-seq[14], and FAIRE-seq[15]. Databases and web servers such as the UCSC Genome Browser[16], Roadmap Epigenomics[17], Blueprint Epigenome[18], Cistrome[19], and ReMap[20] provide free access to data produced by these and other experimental techniques across a wide range of cell types and conditions. When there are too many variants to examine experimentally, these data can be used to prioritize those located within likely regulatory regions in relevant cell and tissue types. Further, in cases where a variant is within a ChIP-seq peak for a specific protein, these data can provide potential leads as to the specific TF(s) or co-factors whose binding might be affecting.

Next, the resulting prioritized variants are screened experimentally to validate predicted genotype-dependent protein binding using EMSA[21,22]. EMSA measures the change in the migration of the oligo on a non-reducing TBE gel. Fluorescently labeled oligo is incubated with the nuclear lysate, and binding of nuclear factors will retard the movement of the oligo on the gel. In this manner, oligo that has bound more nuclear factors will present as a stronger fluorescent signal upon scanning. Notably, EMSA does not require predictions about the specific proteins whose binding will be affected.

Once variants are identified that are located within predicted regulatory regions and are capable of differentially binding nuclear factors, computational methods are employed to predict the specific TF(s) whose binding they might affect. We prefer to use CIS-BP[23,24], RegulomeDB[25], UniProbe[26], and JASPAR[27]. Once candidate TFs are identified, these predictions can be specifically tested using antibodies against these TFs

(EMSA-supershifts and DAPA-Westerns). An EMSA-supershift involves the addition of a TF-specific antibody to the nuclear lysate and oligo. A positive result in an EMSA-supershift is represented as a further shift in the EMSA band, or a loss of the band (reviewed in reference[28]). In the complementary DAPA, a 5'-biotinylated oligo duplex containing the variant and the 20 base-pair flanking nucleotides are incubated with nuclear lysate from relevant cell type(s) to capture any nuclear factors specifically binding the oligos. The oligo duplex-nuclear factor complex is immobilized by streptavidin microbeads in a magnetic column. The bound nuclear factors are collected directly through elution[29,48]. Binding predictions can then be assessed by a Western blot using antibodies specific for the protein. In cases where there are no obvious predictions, or too many predictions, the elutions from variant pull-downs of the DAPA experiments can be sent to a proteomics core to identify candidate TFs using mass-spectrometry, which can subsequently be validated using these previously described methods.

In the remainder of the article, the detailed protocol for EMSA and DAPA analysis of genetic variants is provided.

Protocol

1. Preparation of Solutions and Reagents

1. Order custom DNA oligonucleotide probes for use in EMSA and DAPA.
    1. To reduce non-specific protein binding, design short oligos (between 35-45 base pairs (bp) in length)[30], and place the variant of interest directly in the center flanked by its 17 bp endogenous genomic sequence. For EMSA oligos, add a 5' fluorophore. For DAPA oligos, add a 5' biotin tag.
    2. Order both the sense strand and its reverse complement strand. Alternatively, order duplex (pre-annealed) oligos. When naming the oligos, base the nomenclature on an established reference genome.
    Note: "Risk" and "non-risk" designation can be disease and project specific, while "reference" and "non-reference" are more universally relevant.
    3. Upon arrival of the oligos, briefly spin down the contents and resuspend in nuclease-free water to a final concentration of 100 µM. Store resuspended stock at -20 °C. Protect oligos tagged with a fluorophore from light by wrapping with aluminum foil.

Name	Sequence
rs76562819_REF_FOR	GTAATGCCTTAATGAGAGAGAGTTAGTCATCTTCTCACTTC
rs76562819_REF_REV	GAAGTGAGAAGATGACTAACTCTCTCTCATTAAGGCATTAC
rs76562819_NONREF_FOR	GTAATGCCTTAATGAGAGAGGGTTAGTCATCTTCTCACTTC
rs76562819_NONREF_REV	GAAGTGAGAAGATGACTAACCCTCTCTCATTAAGGCATTAC

Table 1: Example EMSA/DAPA oligonucleotide design to test a SNP for differential binding."REF" stands for the reference allele, while "NONREF" stands for the non-reference allele. "FOR" stands for the forward strand, while "REV" indicates its complement. The SNP is seen in red.

2. Prepare cytoplasmic extraction (CE) buffer with a final concentration of 10 mM HEPES (pH 7.9), 10 mM KCl, and 0.1 mM EDTA in deionized water.
3. Prepare nuclear extraction (NE) buffer with a final concentration of 20 mM HEPES (pH 7.9), 0.4 M NaCl, and 1 mM EDTA in deionized water.
4. Prepare annealing buffer with a final concentration of 10 mM Tris (pH 7.5-8.0), 50 mM NaCl, and 1 mM EDTA in deionized water.

2. Preparation of Nuclear Lysate from Cultured Cells

Note: This experimental protocol was optimized using B-lymphoblastoid cell lines, but has been tested in several other unrelated adherent/ suspension cell lines and works universally.

1. Culture B-lymphoblastoid cells in Roswell Park Memorial Institute (RPMI) 1640 with 2 mM L-glutamine, 10% fetal bovine serum, and 1x antibiotic-antimycotic containing 100 units/ml of penicillin, 100 µg/ml of streptomycin, and 250 ng/ml of amphotericin B.
    1. Seed at a range of 200,000-500,000 viable cells/ml and incubate flasks at 37 °C with 5% carbon dioxide in an upright position with vented or loose caps.
    Note: The growth of B-lymphoblastoid cells slows when they reach over 1,000,000 cells/ml. Break up cell blasts by pipetting up and down several times and return cells to 200,000-500,000 cells/ml to maintain a rapid rate of growth.

2. Wash cultured cells twice with 10 ml ice cold phosphate buffered saline (PBS), spin down at 4 °C, 300 x g for 5 min and remove PBS via aspiration.
3. Count cells using a hemocytometer and resuspend pellet as 1 ml ice cold PBS per 10^7 cells.
Note: For example, if lysing 2 x 10^7 cells, resuspend in 2 ml PBS.
4. Aliquot 1 ml to 1.5 ml microcentrifuge tubes so that each tube contains 10^7 cells in PBS. Centrifuge at 3,300 x g for 2 min 4 °C and aspirate off PBS.
5. Prior to use, add 1 mM dithiothreitol (DTT), 1x phosphatase inhibitor, and 1x protease inhibitor to a working stock of CE buffer. Resuspend cell pellet with 400 µl of CE buffer and incubate on ice for 15 minutes.
6. Add 25 µl of 10% Nonidet P-40 and mix by pipetting. Centrifuge at 4 °C, max speed for 3 min. Decant and discard the supernatant.
7. Prior to use, add 1 mM DTT, 1x phosphatase inhibitor, and 1x protease inhibitor to a working stock of NE buffer. Resuspend cell pellet with 30 µl of NE buffer and mix by vortexing.
8. Incubate at 4 °C in a tube rotator or on ice for 10 min. Centrifuge 3,300 x g for 2 min 4 °C.

9. Collect the clear supernatant (nuclear lysate) and aliquot before storing at -80 °C to avoid multiple freeze-thaw cycles that may degrade the protein. Leave a 10 µl aliquot to measure protein concentration using the bichoninic acid assay (BCA)[31].

3. Electrophoretic Mobility Shift Assay (EMSA)

1. Prepare Oligo Working Stock and EMSA Gel.
    1. If oligos were ordered in duplex, thaw the 100 µM stock and dilute 1:2,000 in annealing buffer to achieve a 50 nM working stock.
    2. If oligos were ordered single-stranded, thaw the 100 µM stocks and dilute 1:10 in annealing buffer to achieve 100 nM working stocks. Combine 100 µl of the 100 nM complement strand solution with each other in a microcentrifuge tube.
        1. Place in a heat block at 95 °C for 5 min. Turn off the heat block and allow the oligos to slowly cool down to room temperature for at least one hour prior to use.

    3. Pre-run the EMSA Gel.
        1. Remove the slide from a pre-cast 6% TBE gel and rinse under deionized water several times to remove any buffer from the wells. Prepare 1 L of 0.5x TBE buffer by adding 50 mL of 10x TBE to 950 ml of deionized water.
        2. Assemble the gel electrophoresis apparatus and check for leaks by filling the inner chamber with 0.5x TBE buffer. If no buffer leaks into the outer chamber, fill the outer chamber roughly two thirds of the way.
        3. Pre-run the gel at 100 V for 60 min.
        4. Flush each well with 200 µl of 0.5x TBE buffer.

2. Prepare Binding Buffer Master Mix.
    1. Prepare 10x binding buffer with a final concentration of 100 mM Tris, 500 mM KCl, 10 mM DTT; pH 7.5 in deionized water.
    2. In a microcentrifuge tube, create a master mix consisting of the reagents common to all reactions (10 µl 10x binding buffer, 10 µl DTT/polysorbate, 5 µl Poly d(I-C), and 2.5 µl salmon sperm DNA; **Table 2**). Prepare an additional 10% to account for volume loss due to pipetting.

Reagent	Final Conc.	Rxn #1	Rxn #2	Rxn #3	Rxn #4
Ultrapure Water	to 20 µl vol.	13.5 µl	11.98 µl	13.5µl	11.98 µl
10x Binding Buffer	1x	2 µl	2 µl	2 µl	2 µl
DTT/TW-20	1x	2 µl	2 µl	2 µl	2 µl
Salmon Sperm DNA	500 ng/µl	0.5 µl	0.5 µl	0.5 µl	0.5 µl
1µg/µl Poly d(I-C)	1 µg	1 µl	1 µl	1 µl	1 µl
Nuclear Extract (5.26 ug/µl)	8 µg	-	1.52 µl	-	1.52 µl
NE Buffer		1.52 µl	-	1.52 µl	-
Reference allele oligo	50 fmol	1 µl	1 µl	-	-
Non-Reference allele oligo	50 fmol	-	-	1 µl	1 µl

Table 2: Example EMSA reaction setup. The table illustrates an example EMSA to test the hypothesis that there is genotype-dependent binding of TFs to a specific SNP.

3. Add nuclease-free water to each microcentrifuge tube such that the final volume following addition of all reagents will be 20 µl.
4. Add the appropriate amount (5.5 µl) of master mix to each microcentrifuge tube.
5. Add 8 µg of nuclear lysate to the appropriate microcentrifuge tubes. Include tubes containing the oligo without nuclear extract as negative controls (*e.g.* **Table 2**, Rxn #1 and Rxn #3).
   Note: The optimal amount of lysate per reaction must be determined experimentally by titration. Generally, titrating a range of 2-10 µg of lysate is sufficient.
6. Add 50 fmol of oligo to the appropriate microcentrifuge tubes. Flick to mix and briefly spin the contents to the bottom of the tube. Incubate for 20 min at room temperature.
   Note: If attempting a supershift, incubate the lysate mixture with antibody for 20 min at room temperature prior to the addition of oligos. It is recommended to use 1 µg of a ChIP-grade antibody for best results.
7. Add 2 µl of 10x Orange Loading Dye to each microcentrifuge tube. Pipette up and down to mix.
8. Load the samples into the pre-run 6% TBE gel by pipetting up and down to mix and then expelling each sample into a separate well. Run the gel at 80 V until the orange dye has migrated 2/3 to 3/4 of the way down the gel. This should take approximately 60-75 min.
9. Remove the gel from the plastic cassette by prying it open with a gel knife and place the gel in a container with 0.5% TBE buffer to keep it from drying out.
10. Place the gel on the surface of an infrared and chemiluminescence imaging system, being sure to eliminate any bubbles or contaminants that will disrupt the image.
11. Using the scanning system software, click the "Acquire" tab and then select "Draw New" to draw a box around the area corresponding to where the gel is located on the surface of the scanner.

12. In the "Channels" section of the "Acquire" tab, select the channel corresponding to the wavelength of the fluorophore tag on the oligo. In the "Scanner" section, click "preview" to get a low-quality preview scan. Adjust the scan area by dragging the blue box surrounding the acquired preview image down to the portion of the gel to be imaged.
    Note: For example, if using oligos labeled with a 700 nm fluorophore, make sure the "700 nm" channel is selected before scanning.
13. In the "Scan Controls" section, select the "84 µM" resolution option and the "Medium" quality option. Set the focus to offset to half the thickness of the gel. Note: For example, a 1 mm gel would use a 0.5 mm focus offset.
14. In the "Scanner" section, click "Start" to begin the scan.
    Note: During the scan, the brightness, contrast, and color scheme can often be adjusted manually depending on the manufacturer of the scanning system.
15. After the scan has finished, select the "Image" tab and click "Rotate or Flip" in the "Create" section to correct the orientation. Save the image file by clicking "Export" in the main menu and then select "Single Image View."

4. DNA Affinity Purification Assay (DAPA)

1. Preparation of 5 µM Oligo Working Stock.
    1. If oligos were ordered in duplex, thaw the 100 µM stock and dilute 1:20 in annealing buffer to achieve a 5 µM working stock.
    2. If oligos were ordered single-stranded, thaw the 100 µM stocks and dilute 1:10 in annealing buffer to achieve 10 µM working stocks. Combine 10 µl of the 10 µM complementary strands with each other. Place in a heat block at 95 °C for 5 min. Turn off the heat block and allow the oligos to slowly cool down to room temperature prior to use.

2. Before starting, warm the binding buffer, low stringency wash buffer, high stringency wash buffer, and elution buffer to room temperature.
    Note: A final concentration of 50 ng/mL Poly d(I-C) can be added to the binding buffer, low stringency wash buffer, and high stringency wash buffer to reduce potential non-specific binding of proteins to the oligos.
3. Prepare the binding mixtures for each variant.
    1. Mix 1 volume of nuclear lysate with 2 volumes of binding buffer.
        Note: The required amount of lysate must be determined experimentally due to varying abundance of TFs. Using between 100-250 µg of nuclear lysate per column is sufficient in most cases.
    2. Add 1x phosphatase inhibitor, 1x protease inhibitor, and 1x binding enhancer (optional) and mix by flicking the tube several times.
        Note: 100x binding enhancer consists of 750 mM $MgCl_2$ and 300mM $ZnCl_2$. Add binding enhancer if the binding of the TF to DNA depends on cofactors or reducing agents. If this information is not known, add the binding enhancer.
    3. Add 10 µl of 5 µM biotinylated capture DNA (50 pmol) to each respective binding mixture. Incubate for 20 min at room temperature.
        Note: Incubation time and temperature may vary depending on the TF. The optimal values need to be determined experimentally.

4. Add 100 µl of streptavidin microbeads. Incubate for 10 min at room temperature.
5. For each oligo probe being tested, place a binding column in the magnetic separator. Place a microcentrifuge tube directly under each binding column and apply 100 µl of binding buffer to rinse the column.
6. Pipette the contents of each binding mixture into separate columns and allow the liquid to flow completely through the column into the microcentrifuge tube before proceeding. Make sure to label the columns with the variant oligo that was used in the binding mixture. Label the flow-through samples and replace with new microcentrifuge tubes to collect the low-stringency washes.
7. Apply 100 µl of low-stringency wash buffer to the column; wait until the column reservoir is empty. Repeat wash 4x. Label the low-stringency wash samples and replace with new microcentrifuge tubes to collect the high-stringency washes.
8. Apply 100 µl of high-stringency wash buffer to the column; wait until the column reservoir is empty. Repeat wash 4x. Label the high-stringency wash samples and replace with new microcentrifuge tubes to collect the pre-elution.
9. Add 30 µl of native elution buffer to the column and let stand for 5 min. Label the pre-elution samples and replace with new microcentrifuge tubes to collect the elution.
    Note: This does not elute the bound protein; it washes the remaining high-stringency buffer out of the column and replaces it with elution buffer to maximize the efficiency of the elution.
10. Add an additional 50 µl native elution buffer to elute the bound TFs. For a higher yield but less concentrated eluate, add an additional 50 µl of native elution buffer and collect the flow-through.
    Note: Analyze elution samples through mass spectrometry to determine the identity of the bound TFs[32]. Afterwards, verify the proteomic results through sodium dodecyl sulfite polyacrylamide gel electrophoresis (SDS-PAGE) followed by a Western blot[33]. If mass spectrometry is not available, run a silver stain using standard technique instead of a Western blot to determine the size of the protein(s) showing genotype-dependent binding. Use this information to narrow down the list of predicted TFs from the computational approaches detailed in the introduction.

Representative Results

In this section, representative results of what to expect are provided when performing an EMSA or DAPA, and the variability with regards to the quality of lysate is characterized. For example, it has been suggested that freezing and thawing protein samples multiple times may result in denaturation. In order to explore the reproducibility of EMSA analysis in the context of these "freeze-thaw" cycles, two 35 bp oligos differing at one genetic variant were incubated with a single batch of nuclear lysate that was thawed and re-frozen for the indicated amount of times. **Figure 2** demonstrates that freezing and thawing this particular B-lymphocyte nuclear extract up to 5 times has seemingly no effect on the integrity of the proteins; however, stability of nuclear protein varies by samples and should thus be tested on each individual cell line used. It is also possible to have variation from different batch preparations of nuclear lysates. Whether this variation is due to the stage of the cell cycle prior to lysis, cell passage number, or other factors, it is important to replicate EMSA results using different batches of lysate to ensure real results.

Additionally, it is important to optimize the signal-to-noise ratio for the EMSA. An important variable in this is the oligo concentration. The amount of oligo (5-300 fmol) was titrated to assess how different quantities of oligos affect the signal (**Figure 3**). An increase in the intensity of the band up to 100 fmol of oligo was observed. After this point, the signal plateaus suggesting 100 fmol is the optimal oligo amount for this EMSA reaction.

Lastly, a representative figure from one of our published studies using part of the strategy described in this manuscript is provided (**Figure 4**). In this study[34], we showed that the lupus-associated risk allele of rs6590330 increases binding of STAT1, a transcription factor that participates in both synergistic activation and inhibition of gene expression downstream of the Type 1 IFN receptor complex[35]. In this example, STAT1 was first identified by DAPA followed by proteomic analysis using high performance liquid chromatography coupled to tandem mass spectroscopy (DAPA-MS). A DAPA-Western blot was used to confirm the TF identified from proteomic analysis (STAT1) and confirm that the phosphorylated form of STAT1 was binding to the non-reference (lupus-risk) oligo. This figure illustrates how the various assays in this strategy can be used to identify differential binding of transcription factor to a non-coding variant.

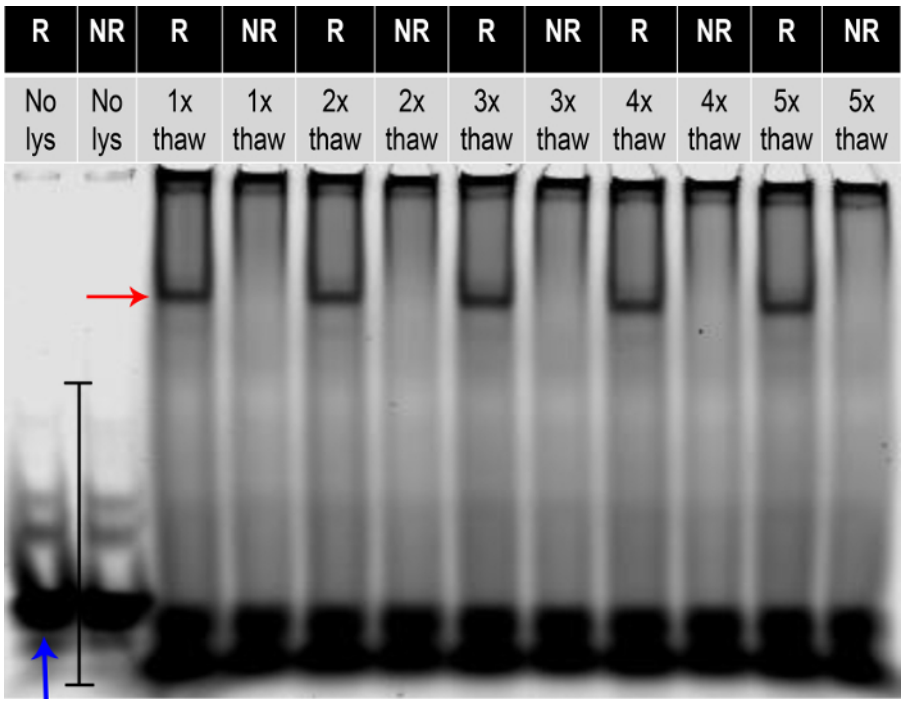

Figure 2: Analysis of reproducibility and the consequences of freeze-thaw cycles on EMSA results. Oligos containing the reference (R) or non-reference (NR) allele of a genetic variant were used to probe the same preparation of lysate after multiple cycles of freezing and thawing. (Lys: Lysate). Each lane contains the free probe at the bottom of the gel (blue arrow). Binding of the TFs to the oligo can be seen as a band in the top half of the gel (red arrow). Bands contained within the bottom half of the gel (see bracket) are non-specific. Please click here to view a larger version of this figure.

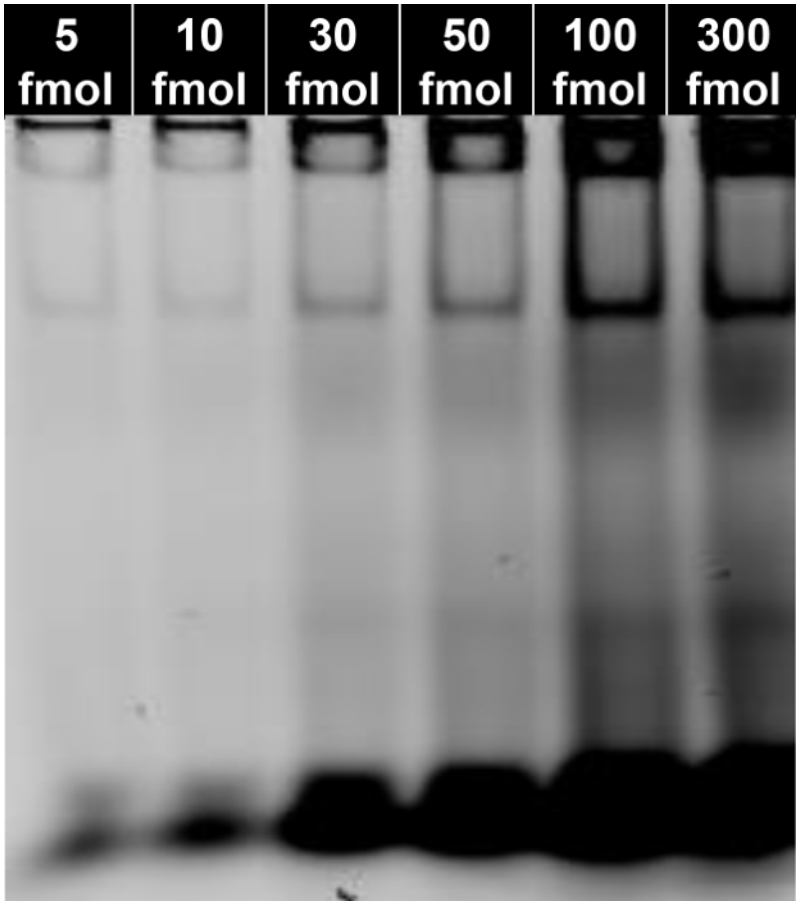

Figure 3: Assessment of the effect of different oligo concentrations. Various concentrations of oligos were used to probe a single preparation of nuclear lysate. Fluorescent signal increases with increased amounts oligo, indicating that the oligo is the limiting reagent. The signal plateaus at the 100-300 fmol lanes, where the amount of protein becomes the limiting reagent. Please click here to view a larger version of this figure.

Figure 4: The lupus-risk allele of rs6590330 increases STAT1 binding, as assessed by DAPA-MS. STAT1 and pSTAT1 exhibit higher binding to oligos containing the rs6590330 risk allele compared to the non-risk allele. Biotin-labeled oligos were incubated with Epstein-Barr virus-transformed B-cell nuclear extract. Proteins bound to the oligo were captured using DAPA. Proteins were then separated by SDS-polyacrylamide gel electrophoresis and detected using anti-pSTAT1 (**A**) or anti-STAT1 (B). M: Marker. NR: oligo containing the non-risk allele of rs6590330; R: oligo containing the risk allele of rs6590330; Mutant: oligo containing a disrupted putative STAT1 binding site downstream of rs6590330; Cell lysate: nuclear extract from B cells. The relative intensities of the bands are indicated below each band. Results are representative of four experiments; while all experiments demonstrated increased STAT1 binding to the probes with the risk allele, in 2/4 experiments no STAT1 or pSTAT1 was detected in the immunoprecipitate from the non-risk oligo, while both were detected in the immunoprecipitate from the risk oligo. This figure has been modified from reference[34]. Please click here to view a larger version of this figure.

Discussion

Although advances in sequencing and genotyping technologies have greatly enhanced our capacity to identify genetic variants associated with disease, our ability to understand the functional mechanisms impacted by these variants is lagging. A major source of the problem is that many disease-associated variants are located in n on-coding regions of the genome, which likely affect harder-to-predict mechanisms controlling gene expression. Here, we present a protocol based on the EMSA and DAPA techniques, valuable molecular tools for identifying genotype-dependent TF binding events that likely contribute to the function of many non-coding variants. Although these two techniques have been used extensively in the past, they have only recently been adapted for genetic variant analysis of TF binding. Beyond TFs, EMSA can also be used to analyze the effect of genetic variants on RNA binding proteins with only minor adjustments to the protocol[36].

The protocol presented herein is simple and easy to perform; nevertheless, certain parts require additional considerations prior to experimentation. First, it is important to generate an initial list of variants for consideration by performing rigorous statistical analyses. An error in this step can derail all downstream analyses, as many variants may bind nuclear lysate differentially without contributing to disease risk. Additionally, the use of functional genomics data performed in pertinent cell types is crucial, as irrelevant cell types may lead to the generation of false positive TF binding predictions. Currently, the most widely-used functional genomic resources include ENCODE[12] and Roadmap Epigenomics[17]. Additional information regarding gene expression levels in cell-types relevant to a disease of interest can be obtained from other resources such as ImmGen[37], BioGPS[38,39], and SNPsea[20]. For example, such resources can be used to filter TF binding predictions to only include those TFs that are expressed in relevant cell types.

It is also important to consider the caveats of *in vitro* experiments such as EMSA and DAPA. In particular, it is necessary to repeat experiments with separate nuclear lysate preparations to reduce false negatives. Moreover, a successful EMSA-supershift can present as a further shift in the EMSA band, in which antibody binds to the TF, or a loss of band, in which antibody blocks the TF's DNA binding domain. In either scenario, including an isotype control antibody and/or an antibody towards a different TF is useful in confirming the specificity of the supershift. Another consideration in performing a supershift is whether to include DTT/polysorbate in the reaction. DTT/polysorbate stabilizes the loading dye allowing for more accurate quantification of unbound DNA; however, it may also reduce disulfide linkages of antibodies resulting in failure of EMSA-supershift. It is recommended to try reactions with and without DTT/polysorbate when attempting to supershift a complex. The optimal amount of Poly d(I-C) and salmon sperm DNA per reaction must be determined experimentally by titration. Generally, titrating a range of 1-6 µg Poly d(I-C) and 50-500 ng salmon sperm is sufficient. One useful positive control for DAPA involves using a consensus sequence for a TF with a well characterized binding motif (obtained from databases such as CIS-BP[23] or Factorbook[40]) and running a Western with an antibody specific to that TF. For both assays, a scrambled oligo can be used to show that any observed binding is specific to the oligos of interest.

Both the EMSA and DAPA techniques have experimental limitations. For example, binding affinity between a TF and DNA is affected in large part by the buffer conditions. Ideally, buffer conditions should mimic the endogenous conditions of the nucleus to allow for optimal binding. For EMSA, improper buffer conditions may result in a weak band or the loss of a band entirely. For DAPA, non-ideal conditions may cause the TF(s) to be eluted during the wash steps. Therefore, each assay is only effective in identifying TF-oligo binding under certain buffer conditions. The most universally useful buffer conditions are presented in the protocol above. A second limitation is that the experimental results from EMSA and DAPA methods provide little information about the mechanisms through which TFs bind to the oligos. TFs could bind to oligos directly or be recruited by other factors. Accordingly, it is important to analyze the oligo sequences computationally to predict how TFs might bind to oligos and verify these predictions experimentally. For example, a specific binding sequence can be mutated or a binding partner can be experimentally depleted. Finally, the amount of extract used needs to be titrated for each experiment to acquire optimal results. Too much lysate may saturate the TF-oligo binding and obscure any differential binding between risk and non-risk alleles. For additional troubleshooting, the reader can refer to several excellent review and method manuscripts[30,41,42].

In addition to the assays described here, there are a variety of *in vivo* assays available to further study the role of variants within living cells (**Figure 1**, bottom). Chromatin immunoprecipitation followed by allele-specific quantitative polymerase chain reactions (ChIP-qPCR) studies can be undertaken to identify if the differential binding observed within the *in vitro* assays is replicated in living cells[34]. For example, if cells heterozygous for a variant of interest are used, qPCR experiments can detect the difference in enrichment between the two alleles in TF immunoprecipitated chromatin. ChIP requires specific ChIP-grade antibodies; however, if ChIP-grade TF antibodies are unavailable, transfecting cells with a flag-tagged TF is an effective alternative[43]. Additionally, the effect of differential binding on target gene expression can be investigated through expression quantitative trait loci (eQTL) analysis in relevant cell types using locally collected expression data from genotyped cells or publically available resources such as those compiled by Genevar[44]. Luciferase reporter assays can also be used to explore the degree to which differential binding of the protein affects gene expression. Such assays can be modified to work regardless of whether the variants are located in a promoter, enhancer, or repressor region[45-47]. Finally, genome-editing technologies, such as CRISPR/Cas9[48], can be used to generate cell lines that differ only at a single variant, which is vital for confirming causation. Such technologies can substantially reduce the experimental variance observed between cell lines derived from genetically diverse subjects, since functional readouts analyzing gene expression or another disease intermediate phenotype can be performed on the edited cell-line, and compared to non-edited cell lines.

The main advantage of the strategy presented is that it enables easy and rapid detection of genotype-dependent TF binding. By prioritizing the key genetic variants, further experiments can be designed to identify their biological effect and demonstrate their causality. Notably, this protocol can be applied to investigate any disease or phenotype associated variant identified from GWAS or fine-mapping. A large, growing trove of genetic data and lists of statistically associated genetic variants is already available. In most cases, the biological mechanisms driving statistical association of these variants is not clear. The strategy outlined in this proposal allows for accurate functional interpretation of the many non-coding disease-associated variants. Such knowledge is vital for the full elucidation of the molecular mechanisms driving any genetic-based disease.

Disclosures

The authors have nothing to disclose.

Acknowledgements

We thank Erin Zoller, Jessica Bene, and Lindsey Hays for input and direction in protocol development. MTW was supported in part by NIH R21 HG008186 and a Trustee Award grant from the Cincinnati Children's Hospital Research Foundation. ZHP was supported in part by T32 GM063483-13.

References

1. Hindorff, L. A. *et al.* Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc Natl Acad Sci U S A.* **106** (23), 9362-9367 (2009).
2. Maurano, M. T. *et al.* Systematic localization of common disease-associated variation in regulatory DNA. *Science.* **337** (6099), 1190-1195 (2012).
3. Ward, L. D., & Kellis, M. Interpreting noncoding genetic variation in complex traits and human disease. *Nat Biotechnol.* **30** (11), 1095-1106 (2012).
4. Paul, D. S., Soranzo, N., & Beck, S. Functional interpretation of non-coding sequence variation: concepts and challenges. *Bioessays.* **36** (2), 191-199 (2014).
5. Zhang, F., & Lupski, J. R. Non-coding genetic variants in human disease. *Hum Mol Genet.* (2015).
6. Lee, T. I., & Young, R. A. Transcriptional regulation and its misregulation in disease. *Cell.* **152** (6), 1237-1251 (2013).
7. Slatkin, M. Linkage disequilibrium--understanding the evolutionary past and mapping the medical future. *Nat Rev Genet.* **9** (6), 477-485 (2008).
8. Bush, W. S., & Moore, J. H. Chapter 11: Genome-wide association studies. *PLoS Comput Biol.* **8** (12), e1002822 (2012).
9. 1000 Genomes Project Consortium, *et al.* An integrated map of genetic variation from 1,092 human genomes. *Nature.* **491** (7422), 56-65 (2012).
10. Chang, C. C. *et al.* Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience.* **4** 7 (2015).
11. Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet.* **81** (3), 559-575 (2007).
12. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature.* **489** (7414), 57-74 (2012).
13. Crawford, G. E. *et al.* Genome-wide mapping of DNase hypersensitive sites using massively parallel signature sequencing (MPSS). *Genome Res.* **16** (1), 123-131 (2006).
14. Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y., & Greenleaf, W. J. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat Methods.* **10** (12), 1213-1218 (2013).
15. Giresi, P. G., Kim, J., McDaniell, R. M., Iyer, V. R., & Lieb, J. D. FAIRE (Formaldehyde-Assisted Isolation of Regulatory Elements) isolates active regulatory elements from human chromatin. *Genome Res.* **17** (6), 877-885 (2007).
16. Kent, W. J. *et al.* The human genome browser at UCSC. *Genome Res.* **12** (6), 996-1006 Article published online before print in May 2002 (2002).
17. Roadmap Epigenomics Consortium, *et al.* Integrative analysis of 111 reference human epigenomes. *Nature.* **518** (7539), 317-330 (2015).
18. Martens, J. H., & Stunnenberg, H. G. BLUEPRINT: mapping human blood cell epigenomes. *Haematologica.* **98** (10), 1487-1489 (2013).
19. Liu, T. *et al.* Cistrome: an integrative platform for transcriptional regulation studies. *Genome Biol.* **12** (8), R83 (2011).
20. Griffon, A. *et al.* Integrative analysis of public ChIP-seq experiments reveals a complex multi-cell regulatory landscape. *Nucleic Acids Res.* **43** (4), e27 (2015).
21. Staudt, L. M. *et al.* A lymphoid-specific protein binding to the octamer motif of immunoglobulin genes. *Nature.* **323** (6089), 640-643 (1986).
22. Singh, H., Sen, R., Baltimore, D., & Sharp, P. A. A nuclear factor that binds to a conserved sequence motif in transcriptional control elements of immunoglobulin genes. *Nature.* **319** (6049), 154-158 (1986).
23. Weirauch, M. T. *et al.* Determination and inference of eukaryotic transcription factor sequence specificity. *Cell.* **158** (6), 1431-1443 (2014).
24. Ward, L. D., & Kellis, M. HaploReg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. *Nucleic Acids Res.* **40** (Database issue), D930-934 (2012).
25. Boyle, A. P. *et al.* Annotation of functional variation in personal genomes using RegulomeDB. *Genome Res.* **22** (9), 1790-1797 (2012).
26. Hume, M. A., Barrera, L. A., Gisselbrecht, S. S., & Bulyk, M. L. UniPROBE, update 2015: new tools and content for the online database of protein-binding microarray data on protein-DNA interactions. *Nucleic Acids Res.* **43** (Database issue), D117-122 (2015).
27. Mathelier, A. *et al.* JASPAR 2014: an extensively expanded and updated open-access database of transcription factor binding profiles. *Nucleic Acids Res.* **42** (Database issue), D142-147 (2014).
28. Smith, M. F., Jr., & Delbary-Gossart, S. Electrophoretic Mobility Shift Assay (EMSA). *Methods Mol Med.* **50** 249-257 (2001).
29. Franza, B. R., Josephs, S. F., Gilman, M. Z., Ryan, W., & Clarkson, B. Characterization of cellular proteins recognizing the HIV enhancer using a microscale DNA-affinity precipitation assay. *Nature.* **330** (6146), 391-395 (1987).
30. LI-COR. *Odyssey Infrared EMSA Kit: Instruction Manual.* <http://www.licor.com/bio/pack_inserts/?pi=829-07910> (2014).
31. Thermo Fisher Scientific, Inc. *Pierce BCA Protein Assay Kit: User Guide.* <https://tools.thermofisher.com/content/sfs/manuals / MAN0011430_Pierce_BCA_Protein_Asy_UG.pdf> (2014).
32. Wijeratne, A. B. *et al.* Phosphopeptide separation using radially aligned titania nanotubes on titanium wire. *ACS Appl Mater Interfaces.* **7** (21), 11155-11164 (2015).
33. Silva, J. M., McMahon, M. The Fastest Western in Town: A Contemporary Twist on the Classic Western Blot Analysis. *J. Vis. Exp.* (84) (2014).
34. Lu, X. *et al.* Lupus Risk Variant Increases pSTAT1 Binding and Decreases ETS1 Expression. *Am J Hum Genet.* **96** (5), 731-739 (2015).

35. Ramana, C. V., Chatterjee-Kishore, M., Nguyen, H., & Stark, G. R. Complex roles of Stat1 in regulating gene expression. *Oncogene.* **19** (21), 2619-2627 (2000).
36. Fillebeen, C., Wilkinson, N., & Pantopoulos, K. Electrophoretic Mobility Shift Assay (EMSA) for the Study of RNA-Protein Interactions: The IRE/IRP Example. *J. Vis. Exp.* (94), e52230 (2014).
37. Heng, T. S., Painter, M. W., & Immunological Genome Project, C. The Immunological Genome Project: networks of gene expression in immune cells. *Nat Immunol.* **9** (10), 1091-1094 (2008).
38. Wu, C. *et al.* BioGPS: an extensible and customizable portal for querying and organizing gene annotation resources. *Genome Biol.* **10** (11), R130 (2009).
39. Wu, C., Macleod, I., & Su, A. I. BioGPS and MyGene.info: organizing online, gene-centric information. *Nucleic Acids Res.* **41** (Database issue), D561-565 (2013).
40. Wang, J. *et al.* Sequence features and chromatin structure around the genomic regions bound by 119 human transcription factors. *Genome Res.* **22** (9), 1798-1812 (2012).
41. Holden, N. S., & Tacon, C. E. Principles and problems of the electrophoretic mobility shift assay. *J Pharmacol Toxicol Methods.* **63** (1), 7-14 (2011).
42. Miltenyi Biotec, Inc. *µMACS FactorFinder Kit: Data sheet.* <http://www.miltenyibiotec.com/en/products-and-services/macsmolecular/protein-research/biotinylated-molecule-isolation/macs-factorfinder-kit.aspx> (2014).
43. Xu, J., Liu, H., Park, J. S., Lan, Y., & Jiang, R. Osr1 acts downstream of and interacts synergistically with Six2 to maintain nephron progenitor cells during kidney organogenesis. *Development.* **141** (7), 1442-1452 (2014).
44. Yang, T.-P. *et al.* Genevar: a database and Java application for the analysis and visualization of SNP-gene associations in eQTL studies. *Bioinformatics.* **26** (19), 2474-2476 (2010).
45. Fort, A. *et al.* A liver enhancer in the fibrinogen gene cluster. *Blood.* **117**(1), 276-282, (2011).
46. Solberg, N., & Krauss, S. Luciferase assay to study the activity of a cloned promoter DNA fragment. *Methods Mol Biol.* **977** 65-78 (2013).
47. Rahman, M. *et al.* A repressor element in the 5'-untranslated region of human Pax5 exon 1A. *Gene.* **263** (1-2), 59-66 (2001).
48. Mali, P. *et al.* RNA-Guided Human Genome Engineering via Cas9. *Science.* **339** (6121), 823-826 (2013).

